METODY I NARZĘDZIA WERYFIKACJI RZETELNOŚCI DANYCH LICZBOWYCH

Tadeusz Grabiński, Marzena Farbaniec,
Marta Woźniak-Zapór, Wacław Zając

METODY I NARZĘDZIA WERYFIKACJI RZETELNOŚCI DANYCH LICZBOWYCH

Kraków 2016

Recenzent:
prof. Jan K. Steczkowski

Projekt typograficzny, okładka:
Joanna Sroka

ISBN 978-83-65208-69-9

Badania dofinansowano ze środków przeznaczonych na działalność statutową
Wydziału Zarządzania i Komunikacji Społecznej (nr projektu WZiKS/DS/9/2016)

Oficyna Wydawnicza AFM
Kraków 2016

Sprzedaż
e-mail: ksiegarnia@kte.pl

Druk:
Eikon Plus

Spis treści

Wstęp

W pracy przedstawiono problematykę rozkładu cyfr znaczących w dużych zbiorach danych liczbowych pochodzących z pomiaru. Empiryczne rozkłady cyfr znaczących opisane są funkcjami matematycznymi ze ściśle sprecyzowanymi parametrami. Odstępstwo rozkładów empirycznych od rozkładów teoretycznych może oznaczać, że analizowany zbiór danych liczbowych zawiera niewiarygodne informacje. Szczegółowa analiza rozbieżności pozwala ponadto wskazać, które dane są najmniej wiarygodne.

Narzędzia analizy rozkładów cyfr znaczących są użyteczne w każdych badaniach empirycznych, gdyż rzetelność danych źródłowych ma fundamentalne znaczenie dla poprawności wniosków wynikających z przeprowadzanych analiz. Prezentowana praca zawiera charakterystykę metod i procedur weryfikacji danych podaną w przystępny sposób i zilustrowaną praktycznymi przykładami.

W rozdziale pierwszym przedstawiono genezę i historię odkryć naukowych związanych z szeroko rozumianymi prawami liczbowymi, m.in. regułę Pareto, ciągi Fibonacciego i Lukasa, prawa Estoupa, Zipfa, Heapsa oraz prawa Newcomba-Benforda.

Szczegółowe omówienie rozkładów cyfr znaczących, testów i mierników służących do oceny stopnia zbieżności rozkładów empirycznych z rozkładami teoretycznymi przytoczone jest w rozdziale drugim. Zamieszczono tu również uzasadnienie teoretyczne praw rządzących rozkładami pierwszych, drugich i kolejnych cyfr znaczących a także rozkładów uogólnionych i alternatywnych.

Rozdział trzeci ma charakter narzędziowy i zawiera opis dostępnych w Internecie, darmowych aplikacji komputerowych wykorzystywanych w procedurach analitycznych związanych z prawami Benforda rozkładu cyfr znaczących. Są to m.in. programy: EZ-R Stats for Excel, Web Computer Assisted Audit Tool (Web CAAT), Digital Analysis Tests and Statistics (DATAS), Benford's Law Utility,

W kolejnym, czwartym rozdziale przedstawiono autorskie narzędzie analizy danych wykorzystujące makroinstrukcje arkusza kalkulacyjnego do

realizacji ciągu obliczeń zgodnie z omówionymi wcześniej procedurami. Aplikacja ta udostępniona jest na stronie www.benford.pl i pozwala przeprowadzać złożone obliczenia w sposób maksymalnie zautomatyzowany.

Ostatni, piąty rozdział zawiera przykład pełnej analizy zbioru danych dotyczących blisko 6 tys. faktur zakupowych apteki X. Analizę przeprowadzono przy pomocy narzędzia omówionego w rozdziale czwartym wykorzystując rozkład pierwszej (F1), drugiej (D2), trzeciej (D3), dwóch pierwszych (F2), trzech pierwszych (F3) oraz ostatniej (L1) cyfry znaczącej.

Reasumując, praca stanowi użyteczne narzędzie praktyczne w zakresie infometrii, zajmującej się oceną i poprawą jakości informacji. Narzędzia wypracowane w ramach tej dyscypliny stosowane są w różnych obszarach dziedzinowych takich jak: naukometria, bibliometria, infobrokering, webometria itd.

Metody omówione w pracy mogą też być wykorzystane w e-learningu, zwłaszcza w procedurach weryfikujących poprawność sprawdzianów wiedzy czy też do oceny wiarygodności opinii studentów pozyskiwanych w trakcie ewaluacji kursów e-learningowych. W dalszych badaniach przewiduje się wdrożenie procedur omówionych w niniejszej monografii do szeroko rozumianej polityki podnoszenia jakości kształcenia w formie e-learningowej.

Tadeusz Grabiński

Rozdział 1
Geneza i historia wybranych praw liczbowych

1.1. Uwagi wstępne

Wpisując do wyszukiwarki Google hasło: *Benford*, *Frank Benford*, *Benford's Law*, *Prawo Benforda* uzyskuje się coraz to większą liczbę odnośników. W tab. 1.1. podano te liczby w latach 2007–2013, przy czym pomiarów dokonywano w miesiącach wakacyjnych. Hasła podawano z reguły w dwóch opcjach: bez ograniczeń językowych oraz tylko w języku polskim.

Tab. 1.1. Liczba wskazań w Google przy hasłach związanych ze słowem Benford w latach 2007–2013

Liczba wskazań Google na hasło	2007	2009	2011	2013
Benford	b.d.	814 000	3 530 000	4 810 000
Benford – język polski	b.d.	160 000	580 000	112 000
%		19,7	16,4	
Frank Benford	1 800	14 600	17 000	10 900
Frank Benford – język polski	13	450	300	830
%	0,7	3,1	1,8	
Benford's Law	39 000	44 000	154 000	312 000
Benford's Law – język polski	b.d.	b.d.	490	3 010
Prawo Benforda	b.d.	b.d.	610	560
Prawo Benforda – język polski	50	690	580	547

Źródło: opracowanie własne.
(*) Hasło: *Benford's Law* w 2005 roku – 10 000 trafień.

Wśród pierwszych znaczeń hasła *Benford* można spotkać strony dotyczące maszyn budowlanych produkowanych przez firmę Terex (Benford to marka pełnoobrotowego miniwozidła drogowego oraz walca wibracyjnego), publikacji pisarza *science fiction* Gregory Benforda, autora podręczników języka angielskiego Michela Benforda, a także witryny zawierające informacje o innych osobach o tym nazwisku, np. Jay Benford, Alec Benford, Mark Benford, itp. Wysoko znajdują się witryny poświęcone synowi Harrisona Forda, właścicielа sieci restauracji w USA – Benowi Fordowi, a także leworęcznemu gwiazdorowi baseballa noszącemu to samo nazwisko. Tym niemniej dwa pierwsze odnośniki (w języku angielskim) odnoszą się do haseł w Wikipedii: *Frank Ben-*

ford oraz *Benford's law*. W języku polskim (w polskiej Wikipedii) na pierwszym miejscu jest odnośnik do hasła: *Frank Benford*.

Wpisując hasło: *Benford w*yszukiwarka Google podpowiada z reguły następujące terminy: *...online, ...law,...gregory centrum galaktyki, ...części, ... dumper, ...terex.*

Witryny dotyczące tematyki prawa Benforda uzyskuje się podając zawężone hasła, np. *Frank Benford,* czy *Benford's Law*. W pierwszym przypadku liczba tych powołań w ostatnich 4 latach na świecie wzrosła dziesięciokrotnie, z 2 tys. w 2007 r. do 17 tys. w 2011 r. Jeszcze większy przyrost trafień miał miejsce w języku polskim – z 13 w 2007 r. do 300 w roku 2011. W relacji do zasobów światowych udział witryn w języku polskim nie jest duży i zawiera się w granicach 2–3%. Ciekawy jest fakt spadku odnośników w języku polskim w roku 2011 w stosunku do roku 2009. Dotyczy to zarówno hasła: *Frank Benford,* jak i hasła: *Prawo Benforda.*

Dla hasła *Benford's Law* wyszukiwarka Google udziela następujących podpowiedzi: *...Excel, ...examples, ...Wiki, ...proof, ...of controversy, ...explanation, ...auditing, ...calculator, ...applications, ...accounting.*

W tab. 1.2. podano pięć pierwszych witryn pojawiających się w odpowiedzi na hasło: *Benford's Law*. Kolejność tych witryn jest prawie identyczna we wszystkich latach. Na pierwszym miejscu jest Wikipedia, następnie portale matematyczne – Wolfram MathWorld, MathPages, Intuitor oraz witryny domowe konsultantów, np. audytora Rexforda Swaina. Zawężając w wyszukiwarce poszukiwania do określonego typu zasobów przy haśle *Benford's Law* pokazuje się (IX 2011) 4700 odnośników do plików w pdf oraz 230 prezentacji w ppt.

Tab. 1.2. Pięć pierwszych witryn pojawiających się w odpowiedzi na hasło: Benford's Law

Adres witryny	2007	2009	2011
www.en.wikipedia.org/wiki/Benford's_law	1	1	1
www.mathworld.wolfram.com/BenfordsLaw.html	2	2	2
www.rexswain.com/benford.html	5	3	3
www.mathpages.com/HOME/kmath302/kmath302.htm	4	4	4
www.intuitor.com/statistics/Benford's%20Law.html	3	5	5

Źródło: opracowanie własne.

W języku polskim podając hasło: *prawo Benforda* na pierwszych pięciu miejscach podawane są witryny polskiej Wikipedii, blogów na witrynie wordpress.com, strona domowa, portal WSIZ w Rzeszowie oraz portal polskich neuroinformatyków.

Tab. 1.3. Pięć pierwszych witryn pojawiających się w odpowiedzi na hasło: prawo Benforda (IX 2011)

Lp.	Adresy witryn
1	www.pl.wikipedia.org/wiki/Rozkład_Benforda
2	www.dataminingalapolonaise.wordpress.com/2010/01/06/prawo-benforda/
3	www.ipipan.waw.pl/~ldebowsk/uslugi/index.html
4	portal.wsiz.rzeszow.pl/plik.aspx?id=2618
5	www.neuroinf.pl/Members/danek/swps/2008/Article.2008-05.../getFile

Źródło: opracowanie własne.

Serwis New Scientist zaliczył prawo Benforda do jednego z pięciu najbardziej znanych odkryć naukowych o nieodpowiedniej nazwie[1]. Na tej liście (poza prawem Benforda) podano następujące odkrycia.

1. Salmonella – bakteria, którą odkrył Theobald Smith, młody pracownik laboratorium kierowanego przez Daniela Salmona.
2. Kometa Halleya, znana już astronomom chińskim w III w., a także J. Keplerowi 75 lat przed tym, jak Halley sformułował hipotezę o okresowości tej komety.
3. Równanie Arrheniusa, opisujące zależność między szybkością reakcji a temperaturą i energią aktywacji, zostało najpierw opisane przez kinetyka holenderskiego van Hoffa, natomiast Arrhenius 5 lat później (powołując się na prace van Hoffa) wyjaśnił istotę tej zależności.
4. Choroba Hansena, czyli trąd. G.A. Hansen wprawdzie pierwszy odkrył bakterie powodujące tę chorobę, ale dopiero jego kolega A. Neissser udowodnił, że bakterie odkryte przez Hansena faktycznie powodują trąd.

Na powyższej liście znajduje się też prawo Benforda *(first digit law)* opisujące rozkład częstości występowania *pierwszych* ***cyfr znaczących (wiodących, leading digit)***[2] w dużych zbiorach liczb, w miarę możliwości wielocyfrowych pochodzących z realnych pomiarów oraz dotyczących empirycznych zjawisk i procesów.

Pierwszy zwrócił uwagę na ten problem amerykański astronom i matematyk Simon Newcomb, który na początku lat 80. XIX wieku zauważył, że strony w tablicach logarytmicznych są bardziej zabrudzone na początku książki niż pod jej koniec (tablice logarytmiczne liczb 7-cyfrowych zawierają ok. 200 stron). Oznaczało to, że z jakichś powodów użytkownicy tablic logarytmicznych częściej korzystali w obliczeniach z liczb mniejszych niż

[1] http://www.newscientist.com/article/dn14461-five-scientific-discoveries-that-got-the-wrong-name.html? DCMP=ILC-hmts&nsref=news10_head_dn14461.

[2] Pierwszą cyfrą znaczącą w liczbie 0,0502 jest $d_1=5$, drugą $d_2=0$, trzecią $d_3=2$.

większych (na początku tablic podawane są logarytmy liczb zaczynających się od 1 000 000 i potem rosną aż do 9 999 999).

Swoje spostrzeżenie Newcomb opublikował w krótkim, dwustronicowym doniesieniu[3], gdzie przytoczył prawdopodobieństwa pojawiania się pierwszej i drugiej cyfry znaczącej, a także stwierdził, że kolejne cyfry znaczące mają rozkład równomierny. Ponadto ustalił, że mantysy logarytmów liczb mają także rozkład równomierny, skąd wynikało, że poszczególne strony tablic antylogarytmów są wykorzystywane z jednakową intensywnością i dlatego w odróżnieniu od tablic logarytmów, mają jednakowo zabrudzone brzegi. Ponieważ użytkownicy tablic logarytmów nie czytają ich tak, jak czyta się np. powieść, lecz traktują je jako narzędzie ułatwiające obliczenia numeryczne (wczesny komputer) i szukają w nich wartości logarytmów dla konkretnych liczb wziętych z rzeczywistych pomiarów, to fakt częstszego korzystania z liczb zaczynających się od mniejszych cyfr nie jest przypadkowy i ma charakter ogólnego prawa.

Formuła, według której można ustalić prawdopodobieństwo (częstość) pojawiania się pierwszych cyfr znaczących d_i ma postać:

$$P(d_i)=\log_{10}(1+1/d_i) \quad \text{dla} \quad d_i=1, 2, \ldots, 9 \tag{1.1}$$

Konkretne wartości tych prawdopodobieństw podano w tabeli 1.4. oraz na rysunku 1.1. Jak się okazuje, wbrew intuicji, która przemawia raczej za rozkładem równomiernym, mamy tu do czynienia z rozkładem szybko malejącym. Co trzecia liczba spotykana w realnych zbiorach danych liczbowych zaczyna się od „1", a tylko jedna na 22 liczby zaczyna się od „9". Inaczej mówiąc, liczb zaczynających się od „1" jest prawie 7 razy więcej niż liczb zaczynających się od „9".

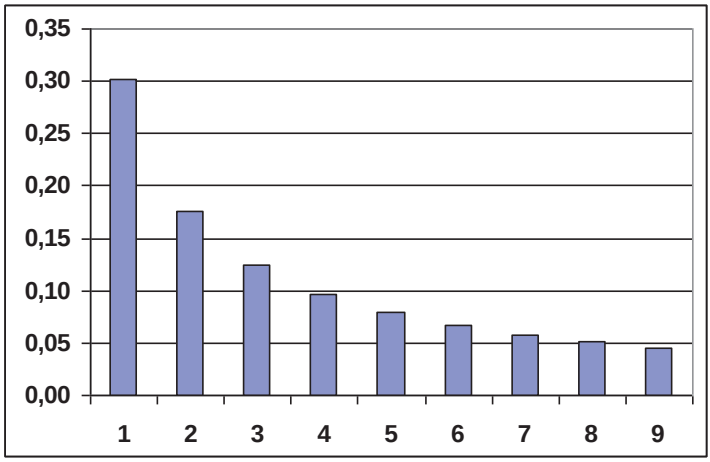
Tab. 1.4. Prawdopodobieństwa i częstości pojawiania się pierwszych cyfr znaczących d_i

d(i)	1	2	3	4	5	6	7	8	9
P(d)	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046
%	30,1	17,6	12,5	9,7	7,9	6,7	5,8	5,1	4,6
100/%	3	6	8	10	13	15	17	20	22

Źródło: opracowanie własne.

[3] S. Newcomb, *Note on the frequency of use different digits in natural numbers,* American Journal of Mathematics, 4/1881, p. 39–40.

Rys. 1.1. Diagram ilustrujący prawdopodobieństwa pojawiania się pierwszych cyfr znaczących d_i



Źródło: opracowanie własne.

Odkrycie Simona Newcomba umykało uwadze przez 60 lat. W 1938 r. fizyk Frank Benford sprawdził doświadczalnie słuszność formuły Newcomba na wielu empirycznych zbiorach liczb, ale nie wyjaśnił dlaczego w badanych zbiorach obserwuje się malejące częstości pojawiania się liczb zaczynających się od coraz to większej cyfry. Dopiero w 1995 r. matematyk Theodore Hill wykazał na czym polega istota tej prawidłowości oraz podał jej własności i uwarunkowania.

Trudno więc powiedzieć, czy prawo o rozkładzie pierwszych cyfr znaczących powinno być nazywane prawem Benforda, czy Newcomba, czy Hilla, czy też prawem Newcomba–Benforda, albo Benforda–Newcomba–Hilla, czy też jeszcze inaczej. Wątpliwości budzi również fakt powszechności działania tego prawa, a więc czy w pełni jest tu uzasadnione używanie terminu *prawo,* czy też należałoby raczej stosować słabsze określenia, np. *prawidłowość, formuła, reguła*.

Jako ciekawostkę można podać, że w literaturze funkcjonuje też prawo Benforda dotyczące stopnia zainteresowania sprawami bulwersującymi, mówiące, że *zaciekawienie jest odwrotnie proporcjonalne do liczby realnie dostępnej informacji*. Autorem tego prawa jest jednak wspomniany wcześniej pisarz *science fiction* Gregory Benford i dotyczy ono zjawisk z zakresu psychologii oraz socjologii.

1.2. Problematyka rozkładu cyfr znaczących w publikacjach naukowych

W 2007 r. pojawiła się w Internecie witryna www.benfordonline.net zawierająca wykaz ważniejszych publikacji z zakresu problematyki dotyczącej prawa Benforda (rys. 1.2). Jej inicjatorem i głównym realizatorem jest Ted Hill, który wykorzystał zasoby bibliograficzne wcześniej zgromadzone przez M. Nigriniego oraz W. Hurlimana.

Wykaz (VIII 2011) zawiera 621 pozycji, które można ujmować w porządku chronologicznym, alfabetycznym (wg tytułów) oraz według nazwisk autorów (por. rys. 1.4–1.5). Każda pozycja zawiera oprócz nazwisk autorów, tytułów, miejsca i roku wydania także informacje dotyczące:

- syntetycznego streszczenia zawartości (autorzy *Bibliografii* proszą na specjalnym formularzu o propozycje w tym zakresie, zarówno co do prac znajdujących się na witrynie, jak i propozycji uwzględnienia prac nowych) – por. rysunek 1.3,
- adres witryny, gdzie dana pozycja jest dostępna w pełnej wersji elektronicznej,
- wykaz prac z *Bibliografii*, które cytują daną pracę – por. rysunek 1.6,
- wykaz prac z *Bibliografii*, które cytuje dana praca – por. rysunek 1.7.

W aneksie A1 przytoczono wykaz prac opublikowanych w latach początkowych, czyli 1881–1970 (37 pozycji) oraz w latach 2009–2010 (66 pozycji). Na podstawie całego zestawienia wyznaczono rozkład liczby prac opublikowanych w poszczególnych latach okresu 1881–1970 (tab. 1.5).

Od opublikowania pierwszej pracy Newcomba minęło 130 lat. Przez połowę tego okresu (do 1945 r.) opublikowano 9 prac, tj. 1,5% dotychczasowego dorobku. Druga praca z zakresu omawianej problematyki pojawiła się w 1912 r. – po ponad 30 latach. W latach 1920–1938, czyli przez prawie 20 lat, nie pojawiła się na świecie ani jedna praca dotycząca prawa Benforda. W latach 1940–1970 opublikowano łączne tylko 30 prac, czyli średnio w ciągu roku wydawano 1 pracę.

W aneksie A2 zebrano zamieszczone na witrynie streszczenia publikacji, które ukazały się przed 1970 rokiem. Na 37 prac z tego okresu dostępnych jest tylko 20 streszczeń. Dla nowszych publikacji liczba streszczeń jest znacznie większa.

Tab. 1.5. Liczba ważniejszych prac na temat prawa Benforda opublikowanych w latach 1881–1970

Rok	L. prac	Odstęp czasowy	Skum. l. prac	Rok	L. prac	Odstęp czasowy	Skum. l. prac
1881	1		1	1952	1	2	14
1912	1	31	2	1956	2	4	16
1916	1	4	3	1957	1	1	17
1917	1	1	4	1961	3	4	20
1920	1	3	5	1963	1	2	21
1938	1	18	6	1964	1	1	22
1939	1	1	7	1965	2	1	24
1944	1	5	8	1966	1	1	25
1945	1	1	9	1967	1	1	26
1946	1	1	10	1968	2	1	28
1948	2	2	12	1969	7	1	35
1950	1	2	13	1970	2	1	37

Źródło: opracowanie własne.

Dopiero po roku 1970 liczba publikacji na temat rozkładu Benforda zaczyna wzrastać, a od roku 1998 następuje gwałtowny ich wzrost. Na rysunkach 1.8–1.10 przytoczono wykresy ilustrujące kształtowanie się liczby publikacji zebranych w omawianej *Bibliografii*:

- w całym okresie 1881–2009 (rys. 1.8),
- w okresie 1970–2009 (rys. 1.9)
- w zagregowanych odcinkach czasu (rys. 1.10).

Z wykresów tych wynika, że rozwój zainteresowania problematyką rozkładu Benforda kształtuje się według funkcji wykładniczej.

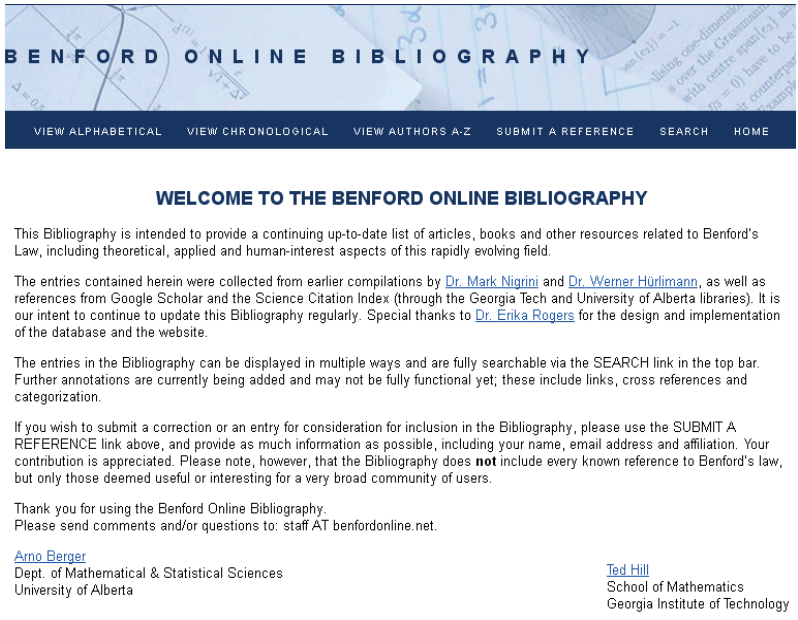
W tabeli 1.6 podano wykaz nazwisk autorów największej liczby prac znajdujących się w wykazie www.benfordonline.net. Zdecydowanymi liderami na tej liście są: M. Nigrini, P. Schatte, T. Hill (po 17–18 prac). Kolejne miejsca zajmują: A. Berger i S. Miller (po 11 prac). Łącznie bibliografia liczy 621 prac 700 autorów. Różnica wynika z faktu występowania prac współautorskich.

Tab. 1.6. Wykaz autorów o największej liczbie opublikowanych prac na temat prawa Benforda

Lp.	Autor	L. prac
1	Nigrini M.J.	18
2	Schatte P.	18
3	Hill T.P.	17
4	Berger A.	11
5	Miller S.J.	11
6	Nagasaka K.	7
7	Turner P.R.	7
8	Bhattacharya S.	6
9	Feldstein A.	6
10	Baird J.C.	5
11	Guan L.	5
12	Lu F.	5
13	Ma B.Q.	5
14	Posch P.N.	5
15	Shao L.	5
16	Shiue J.S.	5
17	Uppuluri V.R.R.	5

Źródło: opracowanie własne.

Rys. 1.2. Witryna www.benfordonline.net



Rys. 1.3. Syntetyczne streszczenie zawartości prac

VIEW COMPLETE REFERENCE

Newcomb, S (1881)

Note on the frequency of use of the different digits in natural numbers

American Journal of Mathematics 4(1), 39-40

ISSN / ISBN: Not available at this time

INTRODUCTION: That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9. The question naturally arises whether the reverse would be true of logarithms. That is, in a table of anti-logarithms, would the last part be more used than the first, or would every part be used equally? The law of frequency in the one case may be deduced from. that in the other. The question we have to consider is, what is the probability that if a natural number be taken at random its first significant digit will be n, its second n', etc

Zentralblatt: JFM 13.0161.01

MathSciNet: MR1505286

For online information, click here.

Bibtex not available at this time.

Reference Type: Journal Article

Subject Area(s): General Interest

Rys. 1.4. Chronologiczny układ prac

BENFORD ONLINE BIBLIOGRAPHY

VIEW ALPHABETICAL | VIEW CHRONOLOGICAL | VIEW AUTHORS A-Z | SUBMIT A REFERENCE | SEARCH | HOME

VIEW CHRONOLOGICAL

(1881) Newcomb, S. Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics 4(1), 39-40.

(1912) Poincaré, H. *Répartition des décimales dans une table numérique.* pp 313-320 in: Calcul des Probabilités, Gauthier-Villars, Paris.

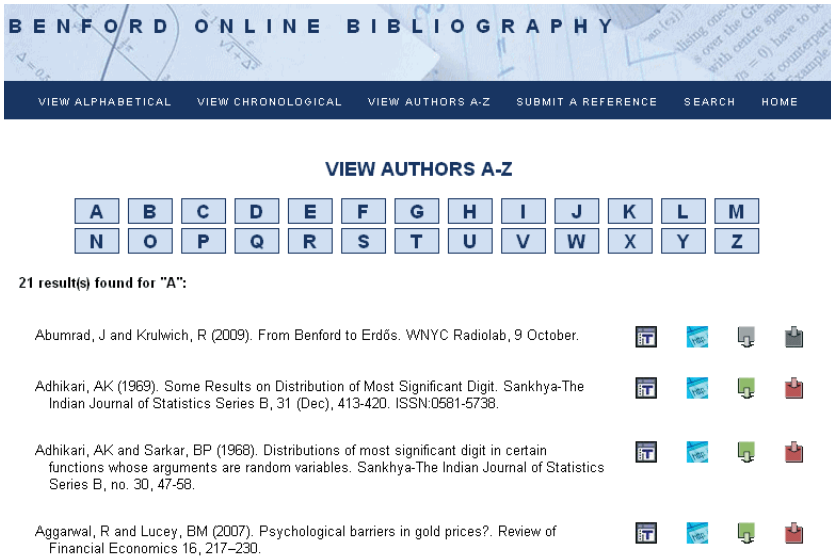
(1916) Weyl, H. Über die Gleichverteilung von Zahlen mod Eins. Mathematische Annalen 77, 313-352.

(1917) Franel, J. A propos des tables de logarithmes. Festschrift der Naturforschenden Gesellschaft in Zürich, Vierteljahrsschrift 62, 286-295.

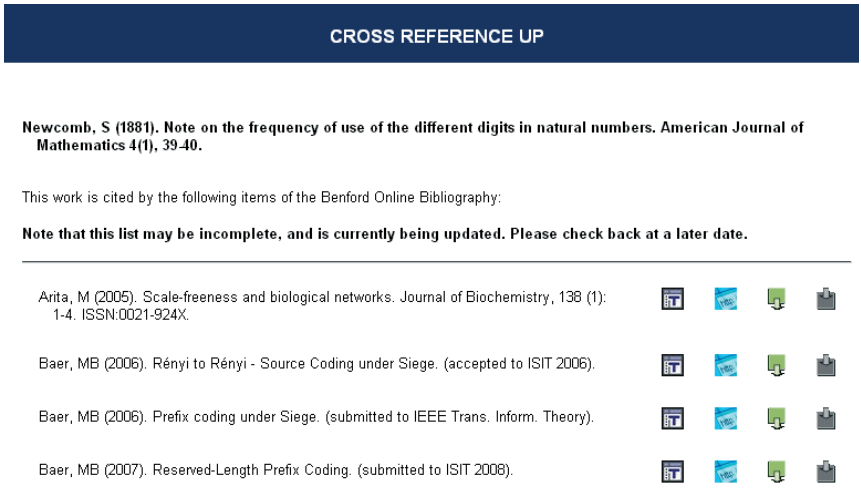
(1920) Boring, EG. The logic of the normal law of error in mental measurement. American Journal of Psychology 31, 1-33. ISSN:0002-9556.

(1938) Benford, F. The law of anomalous numbers. Proceedings of the American Philosophical Society 78, 551-572.

Rys.1.5. Układ prac według nazwisk autorów



Rys. 1.6. Wykaz prac znajdujących się w „Bibliografii", których autorzy powołują się na pracę S. Newcomba (łącznie 201 pozycji)



Rys. 1.7. Wykaz prac znajdujących się w „Bibliografii", na które powołuje się w swojej pracy Z. Szewczak (7 pozycji)

CROSS REFERENCE DOWN

Szewczak, ZS (2010). A limit theorem for random sums modulo 1. Statistics & Probability Letters.

This work cites the following items of the Benford Online Bibliography:

Feller, W (1971). *An Introduction to Probability Theory and Its Applications.* p 63, vol 2, 2nd ed. J. Wiley.

Hürlimann, W (2006). Benford's Law from 1881 to 2006: A Bibliography.

Lévy, P (1939). L'addition des variables aléatoires définies sur une circonférence. Bull. Soc. Math. France 67, 1[99]-41.

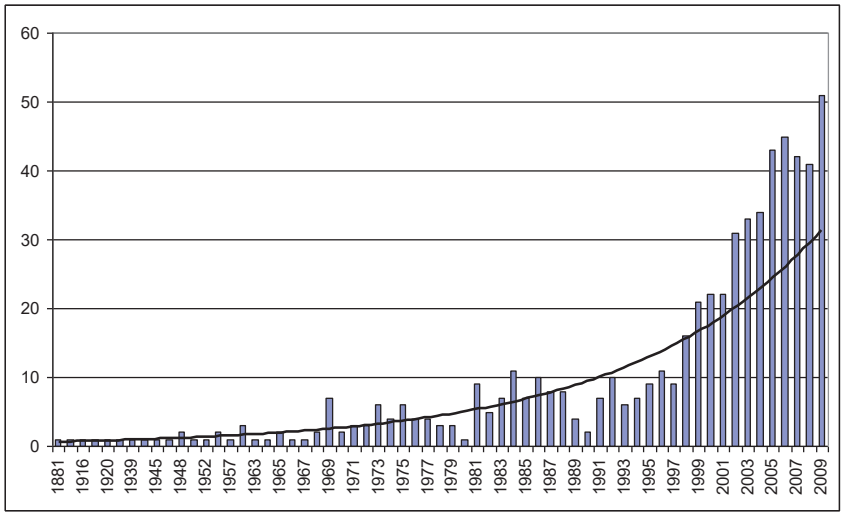
Miller, SJ and Nigrini, MJ (2008). The Modulo 1 Central Limit Theorem and Benford's Law for Products. International Journal of Algebra 2(3), 119 - 130.

Niederreiter, H and Philipp, W (1973). Berry-Esseen bounds and a theorem of Erdős and Turàn on uniform distribution mod 1. Duke Mathematical Journal 40(3), 633-649.

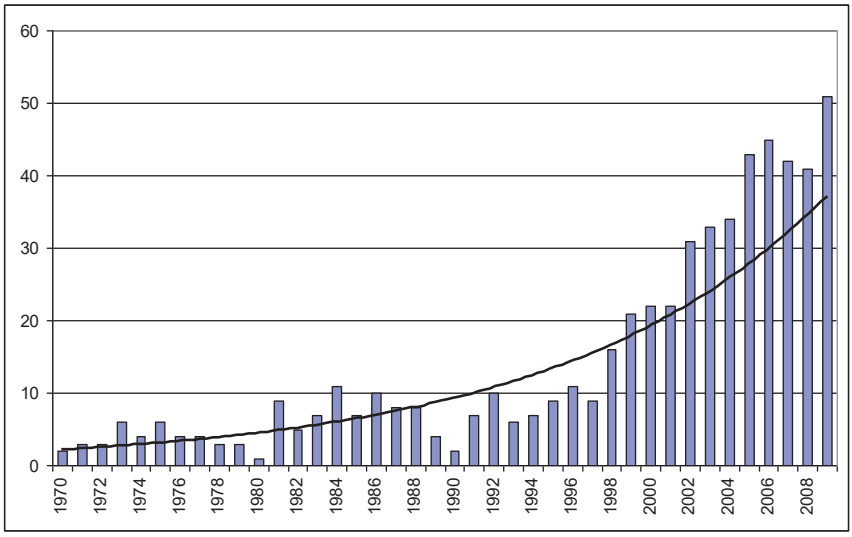
Schatte, P (1984). On the asymptotic uniform distribution of sums reduced mod 1. Math. Nachr. 115, 275-281.

Schatte, P (1988). On mantissa distributions in computing and Benford's law. Journal of Information Processing and Cybernetics EIK 24(9), 443-455.

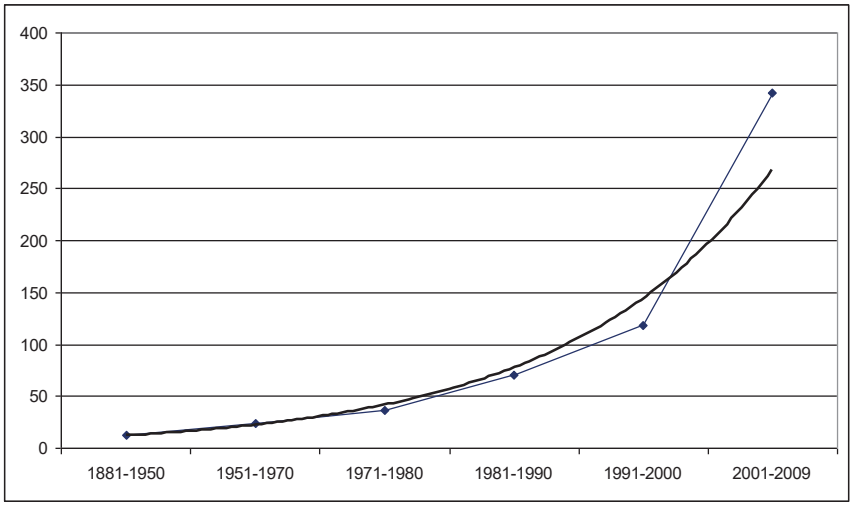
Rys. 1.8. Liczba prac dotyczących rozkładu Benforda opublikowanych w latach 1881–2009



Rys. 1.9. Liczba prac dotyczących rozkładu Benforda opublikowanych w latach 1970–2009



Rys. 1.10. Liczba opublikowanych prac dotyczących rozkładu Benforda w zagregowanych odcinkach czasowych



Warto zwrócić uwagę, że najczęściej cytowaną pracą przez autorów występujących w omawianej *Bibliografii* jest klasyczna praca F. Benforda z 1938 r. *The law of anomalous numbers. Proceedings of the American Phi-*

losophical Society 78, 551–572. Praca ta jest cytowana przez autorów 240 prac spośród wszystkich 621 prac (to jest 39% ogólnej liczby autorów). Na drugim miejscu jest praca S. Newcomba z 1881 r. *Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics 4(1), 39–40,* którą cytowało 201 autorów innych prac (32%).

F. Benford w swojej pracy nie zacytował żadnej z 5 wcześniejszych prac poruszających podobne tematy (ani też nie zacytował żadnej innej publikacji). Co więcej, w następnych latach nie napisał żadnej innej pracy z tego zakresu. Z punktu widzenia historii nauki przypadek ten niewątpliwie zasługuje na uwagę.

Warto nadmienić, że S. Newcomb w swojej pracy używa pojęcia: liczby naturalne, natomiast F. Benford stosuje termin: *anomalous number*. Można by z tego wnioskować, że Benford uważał, iż jego formuła opisuje raczej anomalie niż sytuacje typowe, podczas gdy prawa naukowe powinny z zasady mieć charakter uniwersalny.

Również interesujący jest przypadek napisanej w latach 1945–1946, lecz nieopublikowanej, pracy G.J. Stiglera *The distribution of leading digits in statistical tables*. Autor podał w niej istotną modyfikację prawa Benforda, na którą powołują się autorzy innych prac i która stanowi w wielu sytuacjach alternatywę w stosunku do rozkładu Benforda. G. Stigler w czasie II wojny światowej pracował przy 'projekcie Manhattan' jako matematyk i statystyk. Jako pierwszy (1961) zwrócił uwagę na wartość informacji i stąd uważany jest za inicjatora tej dziedziny ekonomii[4].

1.3. Empiryczne prawa liczbowe w nauce

Poza prawem rozkładu pierwszych cyfr znaczących odkrytym niezależnie przez S. Newcomba (1881) i F. Benforda (1938) w wielu dziedzinach nauki posługujemy się podobnymi regułami. Prawa te na ogół opisują rzeczywistość w dużym przybliżeniu i wynikają z powszechnie znanych prawd wynikających z doświadczenia i noszą nazwę tzw. reguły kciuka (*rule of thumb*)[5].

---

[4] G. Stigler, *The Economics of Information*, Journal of Political Economy, 69/1961.

[5] Nazwa pochodzi z czasów, kiedy kciuk był uproszczonym narzędziem i jednostką miary. Np. zgodnie z regułą browarnika, jeżeli palec włożony do brzeczki nie oparzył się, to można było dodawać do niej drożdże. Reguła kapitana statku stosowana była przy nawigacji wzdłuż wybrzeża i polegała na niezbliżaniu się do linii wybrzeża na odległość kciuka, aby nie wpaść na rafy. Reguła stołu wyrażała się w ułożeniu talerzy od krawędzi stołu na odległość między kciukiem a wskazującym palcem. Według niepotwierdzonych źródeł w prawodawstwie angielskim do końca XIX wieku obowiązywała zasada, że mąż nie mógł bić swojej żony kijem grubszym od swojego kciuka.

Poniżej przytoczono kilka takich reguł głównie z dziedziny ekonomii i informatyki.

Reguły z zakresu ekonomii

- Reguła 72 (niekiedy określana jako reguła 70 lub reguła 69) służy do oceny czasu podwojenia kapitału oprocentowanego na stałym poziomie $r(\%)$[6]. Np. wyjściowy kapitał oprocentowany w skali rocznej na 8% podwoi się po: 72/8=9 latach; przy oprocentowaniu rocznym 6% czas podwojenia kapitału wynosi: 72/6=12 lat itd.

  Czas podwojenia przy stopie r dany jest formułą $T=ln(2)/ln(1+r)$. W tab. 1.7 podano dokładny czas podwojenia T (w latach) przy stopie procentowej r od 1% do 12% (w skali roku) oraz dla różnych podstaw reguły: 72–70–69. W ostatnim wierszu przytoczono średni błąd procentowy modułów różnic pomiędzy faktycznymi okresami podwojenia a czasami wynikającymi z poszczególnych reguł. Jak się okazuje, najlepsze wyniki (średni błąd – 1,5%) uzyskuje się, jeżeli za podstawę reguły przyjmie się liczbę 72. Liczba ta jest bardzo wygodna także dlatego, że dzieli się bez reszty przez 1, 2, 3, 4, 6, 8, 9, 12.

  Podobne reguły można sformułować na potrojenie kapitału, 4-krotne zwiększenie itd. Odpowiednie podstawy i formuły mają postać: dla 3-krotności: $T=114/r$, a dla 4-krotności: $T=144/r$.

Tab. 1.7. Czasy podwojenia kapitału wynikające z reguł 72–70–69

%	T	72	70	69
1	69,7	72,0	70,0	69,0
2	35,0	36,0	35,0	34,5
3	23,4	24,0	23,3	23,0
4	17,7	18,0	17,5	17,3
5	14,2	14,4	14,0	13,8
6	11,9	12,0	11,7	11,5
7	10,2	10,3	10,0	9,9
8	9,0	9,0	8,8	8,6
9	8,0	8,0	7,8	7,7
10	7,3	7,2	7,0	6,9
11	6,6	6,5	6,4	6,3
12	6,1	6,0	5,8	5,8
Błąd średni %		**1,5**	**2,2**	**3,5**

Źródło: opracowanie własne.

[6] Pierwsze wzmianki o tej regule pochodzą z pracy Luki Pacioliego (1445–1514) *Summa de Arithmetica Geometria, Proportioni et Proportionalita,* 1494.

- Reguła Okuna: każdemu wzrostowi bezrobocia o 1% towarzyszy spadek potencjalnego GDP o 2%.
- Reguła „nafciarska": długoterminowa cena ropy naftowej to 3,5-krotność kosztów poszukiwań i wydobycia *(F&D costs).*
- Reguła 2 i 3 sigm: 68% danych znajduje się w przedziale ± 2 odchyleń standardowych od średniej, a 95% danych w przedziale ± 3 sigm.

Reguły z zakresu informatyki

- Reguła Moore'a: moc obliczeniowa komputerów podwaja się co 24 miesiące (przy tym samym koszcie). Dotyczy to liczby tranzystorów w stosunku do powierzchni układu scalonego, mocy obliczeniowej do kosztu, rozmiarów RAM, pojemności dysków twardych, przepustowości sieci itd.
- Reguła Wirtha: oprogramowanie staje się wolniejsze szybciej niż sprzęt staje się szybszy, np. proces rozruchu komputera z nowoczesnym systemem operacyjnym na nowoczesnym PC trwa coraz dłużej.
- Reguła Gatesa: szybkość oprogramowania maleje o połowę co 18 miesięcy.

Inne reguły kciuka

- Reguła Hellina: bliźnięta rodzą się raz na 89 ciąż, trojaczki raz na 89^2 ciąż, natomiast czworaczki raz na 89^3 ciąż.
- Reguła Carnegie College'u: na każdą godzinę spędzoną na zajęciach zorganizowanych student powinien przeznaczyć 2–3 godziny pracy własnej.
- Reguła odległości od pioruna: każdą sekundę od chwili zobaczenia błyskawicy do momentu usłyszenia grzmotu należy pomnożyć przez 300 metrów.

Wymienione powyżej przykładowe reguły nie są bezpośrednio związane z prawami rozkładu cyfr znaczących. Mają jednak z nimi wspólną cechę: mają empiryczny charakter i wynikają z zaobserwowanych prawidłowości statystycznych, zazwyczaj opartych na prawie wielkich liczb.

Poniżej bardziej szczegółowo przedstawiono inne empiryczne prawa liczbowe mające już ścisły związek z prawem Newcomba–Benforda oraz twórców tych praw (por. tab. 1.8).

Leonardo Fibonacci (1175–1250)

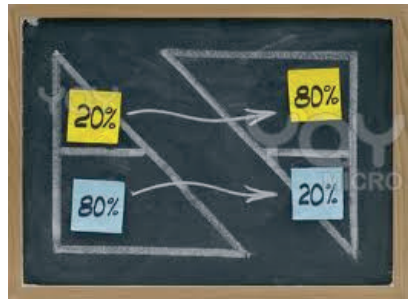
- Włoski matematyk.
- *Liber Abaci* (1202) – opis systemu pozycyjnego, arytmetyka liczb całkowitych, tablica z zapisem liczb rzymskich i indyjskich.
- Ciąg Fibonacciego przytoczony w pracy *Liber Abaci* znany był wcześniej matematykom hinduskim – Gopali (1135), Hemachandrze (1150).

- *Practica geometriae* (1220) – połączenie algebry, geometrii i trygonometrii.
- Sposoby mnożenia liczb tzw. próbą dziesiątkową.
- Rozkład liczb na czynniki pierwsze, cechy podzielności.
- Arytmetyka handlowa oparta na proporcjach.
- Zadania na mieszaninę (ustalenie składników dających stop określonej próby).
- Reguła towarzystwa (podział wielkości proporcjonalnie do części uczestników podziału).
- Reguła poziomów wartości (*figura cata).*

Tab. 1.8. Odkrywcy praw liczbowych

Leonardo Fibbonaci (1175–1250)

Vilfredo Pareto (1848–1923)



George Kingsley Zipf (1902–1950)

Alfred J. Lotka (1880–1949)

Źródło: opracowanie własne.

Ciągi Fibonacciego

Ciągi Fibonacciego mają silny związek, a w sensie rozkładu pierwszej cyfry znaczącej są nawet tożsame z prawem Benforda (por. rozdz. II niniejszej pracy). Są to ciągi liczb naturalnych określonych rekurencyjną formułą:

(1.2) $$\boxed{F_{i+1} = F_i + F_{i-1}}$$

przy czym zakłada się, że $F_1=F_2=1$. Ze wzoru (1.2) wynika, że każdy następny wyraz ciągu Fibonacciego jest sumą dwóch poprzednich. Ciąg ten został podany przez Fibonacciego w 1202 r. jako rozwiązanie zadania o rozmnażaniu królików.[7]

Ciąg (1.2) posiada wiele interesujących własności. Poniżej przykładowo podano kilka z nich[8].

- *n-ty* wyraz ciągu wyrażony wzorem Bineta:

(1.3) $$\boxed{F_n = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^2 - \left(\frac{1-\sqrt{5}}{2}\right)^2\right] = \frac{(1+\sqrt{5})^2-(1-\sqrt{5})^2}{2^n\sqrt{5}}}$$

[7] Każda para królików rodzi co miesiąc nową parę młodych królików. Okres rozrodczy królików trwa 2 miesiące. Na początku mamy jedną parę królików [1]. Po miesiącu rodzi się druga para [1] Po dwóch miesiącach rodzą się 2 nowe pary królików [2]. W kolejnym miesiącu rodzą się 3 pary królików (dwie pary z rodziców, i jedna para z „dziadków"), itd. Ciąg Fibonacciego podaje liczbę nowonarodzonych par królików w kolejnych miesiącach. Suma elementów ciągu określa liczebność populacji królików (przy założeniu, że króliki nie umierają).

[8] Więcej informacji na temat ciągów Fibonacciego znaleźć można m.in. w artykułach zamieszczanych na łamach kwartalnika *Fibonacci Quarterly*, a także w pracach T. Koshy, *Fibonacci and Lucas Numbers with Applications,* Wiley 2001; L.C. Washington, *Benford's Law for Fibonacci and Lucas Number*, FIbonacci Quarterly, vol. 19/1981, p.175–177.

- suma *n* wyrazów ciągu:

(1.4–1.5) $$\sum_{i=1}^{n} F_i = F_{n+2} - 1 \qquad \sum_{i=0}^{n} iF_i = nF_{n+2} - F_{n+3} + 2$$

- wyraz ciągu jako suma kwadratów wyrazów sąsiednich

(1.6–1.7) $$F_{2i} = F_{i+1}^2 - F_{i-1}^2 \qquad F_{2i-1} = F_i^2 + F_{i-1}^2$$

W literaturze można spotkać wiele modyfikacji ciągu Fibonacciego, np.
- Ciąg Fibonacciego, w którym $F_1=0$ oraz $F_2=1$.
- Ciąg Lukasa, w którym $F_1=2$ oraz $F_2=1$.
- Ciąg Tribonacciego, w którym każdy kolejny element powstaje przez zsumowanie trzech poprzedzających go elementów, przy czym $F_1=0$, $F_2=0$, $F_3=1$.
- Ciąg Tetranacciego, w którym każdy kolejny element powstaje przez zsumowanie czterech poprzedzających go elementów, przy czym $F_1=0$, $F_2=0$, $F_3=0$, $F_4=1$.

W tabeli 1.9 oraz na rysunku 1.11 przytoczono początkowe wartości ciągów Fibonacciego. Poza ciągami wymienionymi powyżej, przykładowo podano też ciągi dla $F_1=F_2=2$ oraz $F_1=1$ i $F_2=2$. Jak można zauważyć wszystkie te ciągi mają podobną postać i dają się aproksymować funkcją wykładniczą. Na rysunku 1.11 zamieszczono funkcję wykładniczą dopasowaną do ciągu Fibonacciego $F_1=F_2=1$ ze współczynnikiem determinacji $R^2=0{,}994$.

Ilorazy sąsiednich elementów ciągów Fibonacciego dążą w granicy do tzw. *złotej liczby*. Liczba ta dana jest wzorem:

(1.8) $$\varphi = \frac{F_{i+1}}{F_i} \to \frac{\sqrt{5}+1}{2} = 1{,}61804$$

i wyznacza tzw. *złoty podział* (podział harmoniczny, boska proporcja) odcinka. Jest to podział odcinka na dwie części [a;b] takie, że stosunek długości części dłuższej [a] do krótszej [b] jest taki sam, jak stosunek długości całego odcinka [a+b] do części dłuższej [a]

(1.9) $$\frac{a+b}{a} = \frac{a}{b} = \varphi$$

Złoty podział (liczbę) możemy odnaleźć w wielu przypadkach, np.:

- w proporcjach Wielkiej Piramidy w Gizie (stosunek wysokości ściany bocznej do połowy wymiaru podstawy), katedry w Mediolanie czy Partenonu,
- w przypadku człowieka witruwiańskiego Leonarda da Vinci, czy rzeźby Wenus z Milo złote liczby to stosunek wysokości człowieka do długości dolnej części ciała (od pępka w dół) i stosunek długości dolnej części ciała do górnej (od pępka w górę),
- w ułożeniu łodyg roślin, płatków kwiatów, ziaren w słonecznikach, łusek na szyszce świerkowej itd. (tzw. filotaksja),
- w kształtach muszli ślimaków i głowonogów,
- w muzyce – w konstrukcji skrzypiec Stradivariego, utworach Jana Sebastiana Bacha, V Symfonii Bethovena też zastosowano złoty podział.

Wartości złotej liczby wyznaczonej dla poszczególnych ciągów Fibonacciego zamieszczono w tab. 1.10. Jak można zauważyć, dla każdego z tych ciągów ilorazy sąsiednich wyrazów stabilizują się już przy 12–14 wyrazie na poziomie $\approx$1,61804.

Odwrotność złotej liczby (1,8) wyznacza tzw. *złotą proporcję* wykorzystywaną do ustalania *poziomów Fibonacciego*, m.in. w analizie technicznej do określenia poziomów wsparcia oraz poziomów oporu (punkty zwrotne ruchu cen) na podstawie analizy wykresów cen instrumentów finansowych, indeksów akcji, kontaktów terminowych itp. Poziomy Fibonacciego mogą dotyczyć zarówno pionowej osi cen (miejsca realizacji zysków oraz zleceń obronnych typu *stop loss*), jak i poziomej osi czasu (okresy pomiędzy kolejnymi ekstremami na wykresie).

Zazwyczaj w analizie technicznej wykorzystuje się kilka poziomów Fibonacciego, które wyznacza się jako graniczne wartości ilorazów wyrazów ciągu Fibonacciego, ale nie wyrazów sąsiednich lecz wyrazów przesuniętych względem siebie o 2, 3, (...) pozycje:

(1.10)
$$P_2 = \frac{F_i}{F_{i+2}}; \quad P_3 = \frac{F_i}{F_{i+3}}; \quad P_4 = \frac{F_i}{F_{i+4}}; \quad P_5 = \frac{F_i}{F_{i+5}}$$

W tabeli 1.11 podano dla wyjściowego ciągu Fibonacciego *F1=F2=1* wartości ilorazów P_1, P_2, ... , P_5. Jak można zauważyć ilorazy te stabilizują się już przy 10–11 wyrazie ciągu. Odwrotności tych granicznych wartości można też uzyskać jako kolejne potęgi liczby:

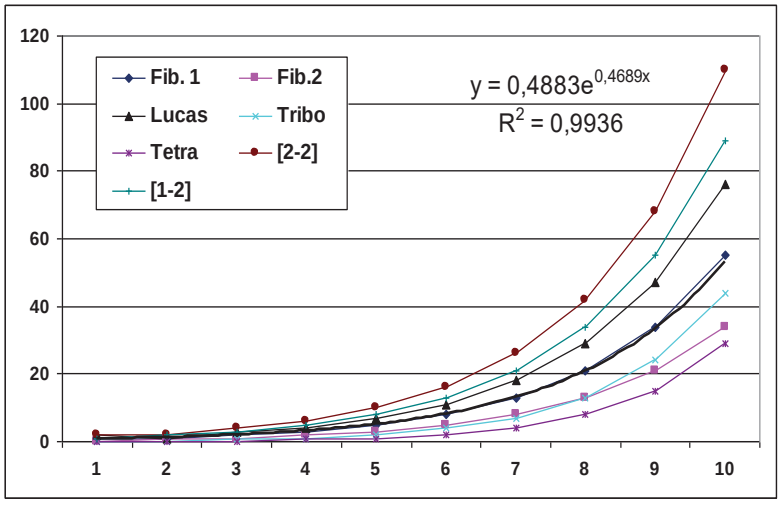
(1.11)
$$P_k = \frac{1}{\varphi^k} \quad (k = \cdots, -3, -2, -1,\ 0,\ 1,\ 2,\ 3, \ldots)$$

Tab. 1.9. Początkowe wyrazy ciągów Fibonacciego

i	Fib. [1;1]	Fib. [0;1]	Lucas	Tribo	Tetra	[2–2]	[1–2]
1	1	0	2	0	0	2	1
2	1	1	1	0	0	2	2
3	2	1	3	1	0	4	3
4	3	2	4	1	1	6	5
5	5	3	7	2	1	10	8
6	8	5	11	4	2	16	13
7	13	8	18	7	4	26	21
8	21	13	29	13	8	42	34
9	34	21	47	24	15	68	55
10	55	34	76	44	29	110	89
11	89	55	123	81	56	178	144
12	144	89	199	149	108	288	233
13	233	144	322	274	208	466	377
14	377	233	521	504	401	754	610
15	610	377	843	927	773	1220	987
16	987	610	1364	1705	1490	1974	1597
17	1597	987	2207	3136	2872	3194	2584
18	2584	1597	3571	5768	5536	5168	4181
19	4181	2584	5778	10609	10671	8362	6765
20	6765	4181	9349	19513	20569	13530	10946

Źródło: opracowanie własne

Rys. 1.11. Wykresy dziesięciu początkowych wyrazów ciągów Fibonacciego



Tab. 1.10. Ilorazy sąsiednich wyrazów Fi+1/Fi ciągów FIbonacciego

i	Fib. [1;1]	Fib. [0;1]	Lucas	Tribo	Tetra	[2−2]	[1-2]
1	1,00000		0,50000			1,00000	2,00000
2	2,00000	1,00000	3,00000			2,00000	1,50000
3	1,50000	2,00000	1,33333	1,00000		1,50000	1,66667
4	1,66667	1,50000	1,75000	2,00000	1,00000	1,66667	1,60000
5	1,60000	1,66667	1,57143	2,00000	2,00000	1,60000	1,62500
6	1,62500	1,60000	1,63636	1,75000	2,00000	1,62500	1,61538
7	1,61538	1,62500	1,61111	1,85714	2,00000	1,61538	1,61905
8	1,61905	1,61538	1,62069	1,84615	1,87500	1,61905	1,61765
9	1,61765	1,61905	1,61702	1,83333	1,93333	1,61765	1,61818
10	1,61818	1,61765	1,61842	1,84091	1,93103	1,61818	1,61798
11	1,61798	1,61818	1,61789	1,83951	1,92857	1,61798	1,61806
12	1,61806	1,61798	1,61809	1,83893	1,92593	1,61806	1,61803
13	1,61803	1,61806	1,61801	1,83942	1,92788	1,61803	1,61804
14	1,61804	1,61803	1,61804	1,83929	1,92768	1,61804	1,61803
15	1,61803	1,61804	1,61803	1,83927	1,92755	1,61803	1,61803
16	1,61803	1,61803	1,61804	1,83930	1,92752	1,61803	1,61803
17	1,61803	1,61803	1,61803	1,83929	1,92758	1,61803	1,61803
18	1,61803	1,61803	1,61803	1,83929	1,92757	1,61803	1,61803
19	1,61803	1,61803	1,61803	1,83929	1,92756	1,61803	1,61803

Źródło: opracowanie własne.

Tab. 1.11. Kolejne poziomy Fibonacciego wyznaczone na podstawie ciągu F1=F2=1

i	Fib. [1;1]	1	2	3	4	5
1	1	1,00000	0,50000	0,33333	0,20000	0,12500
2	1	0,50000	0,33333	0,20000	0,12500	0,07692
3	2	0,66667	0,40000	0,25000	0,15385	0,09524
4	3	0,60000	0,37500	0,23077	0,14286	0,08824
5	5	0,62500	0,38462	0,23810	0,14706	0,09091
6	8	0,61538	0,38095	0,23529	0,14545	0,08989
7	13	0,61905	0,38235	0,23636	0,14607	0,09028
8	21	0,61765	0,38182	0,23596	0,14583	0,09013
9	34	0,61818	0,38202	0,23611	0,14592	0,09019
10	55	0,61798	0,38194	0,23605	0,14589	0,09016
11	89	0,61806	0,38197	0,23607	0,14590	0,09017
12	144	0,61803	0,38196	0,23607	0,14590	0,09017
13	233	0,61804	0,38197	0,23607	0,14590	0,09017
14	377	0,61803	0,38197	0,23607	0,14590	0,09017
15	610	0,61803	0,38197	0,23607	0,14590	0,09017
16	987	0,61803	0,38197	0,23607	0,14590	
17	1597	0,61803	0,38197	0,23607		
18	2584	0,61803	0,38197			

19	4181	0,61803				
	6765	0,61803	0,38197	0,23607	0,14590	0,09017
odwr.		1,61803	2,61803	4,23607	6,85410	11,09017

Źródło: opracowanie własne.

Dokładne wartości poziomów Fibonacciego (w %) dla wybranych wartości parametru *k* z przedziału [-3;3] zebrano w tab. 1.12.

Tab. 1.12. Poziomy Fibonacciego jako potęgi liczby φ

Wykł. pot. k	φ = 1,61803	Poziomy Fib. (%)
-3	0,23607	23,6
-2	0,38197	38,2
-1	0,61803	61,8
-0,5	0,78615	78,6
0	1,00000	100,0
0,5	1,27202	127,2
1	1,61803	161,8
2	2,61803	261,8
3	4,23607	423,6

Źródło: opracowanie własne.

Vilfredo Federico Damasco Pareto (1848–1923)

- Włoski ekonomista i socjolog, współtwórca tzw. lozańskiej szkoły w ekonomii.
- W zakresie ekonomii zajmował się teorią ogólnej równowagi ekonomicznej i podziału dobrobytu oraz zastosowaniami metod matematycznych w ekonomii.
- Twórca pojęcia 'optymalność Pareta': nie można powiększyć dobrobytu jednostki bez zmniejszenia dobrobytu innej jednostki.
- W socjologii twórca teorii krążenia elit jako grupy ludzi mających w danej dziedzinie najwyższe osiągnięcia oraz teorii rezyduów (trwałych dyspozycji psychicznych) i derywacji (zmiennych elementów działań człowieka).

Reguła Pareto (zasada 80/20)

Pareto na podstawie analizy dochodów ludności we Włoszech (1897) stwierdził, że 80% majątku jest w posiadaniu 20% mieszkańców. W trakcie innych badań nad koncentracją okazało się, że w praktyce bardzo często większość (80%) problemów powoduje mała część (20%) możliwych przyczyn. Na przykład 20% klientów generuje 80% zysków, 20% tekstu pozwa-

la zrozumieć 80% zawartych w nim treści, 20% kierowców powoduje 80% wypadków, 20% powierzchni regionu zamieszkuje 80% ludności, 20% życia przynosi 80% szczęścia. Należy podkreślić, że liczby 20–80 są tylko przybliżeniem. Istota prawa polega na tym, że nie jest tak, że aby uzyskać 100% efektów trzeba koniecznie ponieść 100% nakładów.

Reguła Pareto stanowi uproszczoną metodę diagnozowania zjawisk i planowania przedsięwzięć, ułatwia organizację czasu pracy w grupie, pozwala ustalać priorytety działań. Jest podstawą metody ABC jako narzędzia zarządzania jakością w TQM. Regułę Pareto wykorzystuje się w badaniach marketingowych i metodach portfelowych w celu analizy potencjału strategicznego przedsiębiorstw.

W bibliotekoznawstwie reguła Pareto jest wykorzystywana w postaci ***prawa Bradforda***. Zgodnie z tym prawem (zwanym prawem rozproszenia), w każdej dziedzinie nauki istnieje ograniczony, nieliczny zestaw najważniejszych czasopism, w których drukowana jest znacząca liczba (ok. 1/3) wszystkich wartościowych prac z danej dziedziny.

Jeżeli czasopisma uporządkuje się według malejącej liczby artykułów na dany temat, to można wyróżnić grupę czasopism podstawowych oraz kilka grup czasopism z taką samą liczbą artykułów, co grupa podstawowa. Liczba czasopism w kolejnych grupach rośnie wtedy geometrycznie [1, n, n^2 ,...]. Wystarczy zanalizować rejestr wypożyczeń, aby zracjonalizować decyzje o prenumeracie i koszcie zakupu czasopism z danych dziedzin. Prawo Bradforda ma związek z listą *Impact factor* prowadzoną przez Instytut Filadelfijski.

George Kingsley Zipf (1902–1950)

- Amerykański lingwista i filolog.
- Praca doktorska (1929) *Relative Frequency as a Determinant of Phonetic Change*, Harvard University.
- W 1932 r. sformułował twierdzenie dotyczące częstości występowania słów, nie powołując się na znane od 6 lat zbliżone prawo Lotki, nie dokonując analiz statystycznych ani nie opisując formułowanych zależności przy pomocy odpowiednich wzorów matematycznych.
- Obecnie uchodzi za twórcę ilościowej lingwistyki (*zipfian linguistic*) jako składowej informetrii.

Prawa Estoupa, Zipfa (1932), Heapsa

Zajmując się analizą częstości występowania słów w różnych językach G. Zipf doszedł do wniosku, że ***większość słów używana jest rzadko, natomiast liczba słów często wykorzystywanych w tekstach jest stosunko-***

wo nieduża. Jeżeli uporządkować słowa według częstości ich występowania w tekstach to częstość *c* wystąpienia r-tego słowa jest proporcjonalna do $1/r^a$ (gdzie *a* to wykładnik potęgowy bliski jedności). Pierwsze (najczęściej występujące w badanym tekście) słowo występuje więc 2x częściej niż drugie słowo, trzy razy częściej niż trzecie słowo, itd. Warto nadmienić, że nie jest potwierdzona formuła Zipfa dla par słów, oraz dla fraz, tzw. N-gramów.

Podobna zasada jak w prawie Zipfa wyrażona jest w postaci ***formuły Estoupa–Zipfa***, zgodnie z którą ***iloczyn pozycji danego słowa na liście ich częstości występowania (r) przez częstość (c) jest stały r*c=const i zależy od długości analizowanego tekstu***.

W tab. 1.13 podaje się przykłady analizy tekstów zawierających:

- ponad 40 mln wyrazów zaczerpniętych z Wall Street Journal (WSJ) z lat 1987–1989,
- ponad 37 mln wyrazów zaczerpniętych z artykułów opublikowanych przez Associated Press (AP) w roku 1989.

W tabeli 1.13 uwzględniono tylko 10 słów najczęściej występujących w analizowanych tekstach. W przypadku zbioru (korpusu) WSJ analizie poddano zarówno częstości występowania pojedynczych słów jak i tzw. bigramów i trigramów. W przypadku bazy AP dostępne były tylko informacje o pojedynczych słowach.

Zgodność list 10 najczęściej wykorzystywanych słów jest duża. W obydwóch zbiorach na obydwóch listach znajduje się po 7 słów na tych samych pozycjach: *the, of, to, a, and, in* oraz *for*. Kolejne słowo *that* jest także na obydwóch listach, ale na różnych pozycjach. Listy różnią się tylko dwoma elementami – w zbiorze WSJ występują słowa *that* oraz *is* natomiast w zbiorze AP – *said* oraz *was.*

Łącznie 10 najczęstszych słów tworzy 19% (w zbiorze WSJ) oraz 23% (w zbiorze AP) ogólnej zawartości analizowanych tekstów.

Liczba fraz dwuwyrazowych i trzywyrazowych jest wyraźnie mniejsza. Liczba bigramów stanowi tylko 13% liczby pojedynczych wyrazów, natomiast liczba trigramów – tyko 2,6%. Dla uzyskania porównywalności pomiędzy tymi trzema zbiorami, częstości bigramów i trigramów przeliczono zakładając, że ich ogólna liczba równa jest liczbie pojedynczych wyrazów w całym zbiorze (N=40 mln).

Dla tak przeliczonych częstości w ujęciu procentowym wyznaczono iloczyny pozycji słów *(r)* przez częstości ich pojawiania się (*c)*. Jak można zauważyć, są one zbliżone do stałej wartości *const=0,1* – średnia wartość iloczynów dla zbioru WSJ wynosi 0,074 a dla zbioru AP – 0,088. W ostatnich trzech kolumnach tabeli 1.13 podano odpowiednie wartości iloczynów *r*c* z dokładnością do trzech, dwóch i jednego miejsca po przecinku. Porównując stopień zgodności iloczy-

nów *r*c* dla różnych zbiorów obserwuje się, że najmniejsze ich zróżnicowanie (mierzone rozstępem, czyli różnicą pomiędzy maksymalnymi oraz minimalnymi wartościami iloczynów) ma miejsce dla pojedynczych słów (0,058), nieco większe jest dla bigramów (0,068) i największe dla trigramów (0,076).

Drugi przykład dotyczy tekstu w języku polskim. Analizie poddano utwór Michała Bułhakowa *Mistrz i Małgorzata*. Obliczenia wykonano przy pomocy programu *Hermetic Word Frequency Counter* dostępnego na stronie www.hermetic Firma Hermetic Systems specjalizuje się w programach lingwistycznych, kryptograficznych, matematycznych, aplikacjach internetowych (kalendarze, konwertery).

W analizowanym tekście znajduje się ogółem 73 175 wyrazów, w tym 18 132 wyrazów unikalnych. W tabeli 1.14 przytoczono listę 40 najczęstszych słów wraz z ich częstością (liczba oraz wskaźnik procentowy udziału w stosunku do ogólnej liczby słów).

Natomiast w tabeli 1.15 zebrano wybrane parametry opisujące własności analizowanego zbioru. Porównując wyniki analizy zbioru w języku polskim ze zbiorami w języku angielskim daje zauważyć się duże podobieństwo. Dla przykładu 10 najczęstszych słów zarówno w zbiorze w języku polskim, jak i w zbiorze WSJ tworzy 19,2% całości tekstu. Iloczyny *r*c* w zbiorze w języku polskim kształtują się średnio na poziomie 0,078, natomiast w zbiorze WSJ na podobnym poziomie 0,074.

W zbiorze w języku polskim 40 najczęstszych wyrazów pozwala stworzyć 30% całości tekstu. Te 40 wyrazów to zaledwie 0,2% ogólnej liczby unikalnych wyrazów występujących w całym tekście. Na rysunkach 1.12–1.15 przedstawiono wykresy ilustrujące prawo Zipfa na przykładzie zbioru w języku polskim.

Pierwsze dwa rysunki przedstawiają zależność w układzie liniowym pomiędzy częstością wyrazów *c* a ich pozycją (rys. 1.12) dla pierwszych 100 najczęstszych wyrazów, a na rysunku 1.13 – dla pierwszego 1000 wyrazów (spośród ponad 18 000). Obydwa wykresy mają postać hiperboli, przy czym uwzględnienie na wykresie dużej liczby wyrazów powoduje tłumienie przebiegu funkcji dla początkowych najczęstszych wyrazów.

Wykresy 1.14 i 1.15 przedstawiają tę samą zależność (częstość występowania a pozycja słów), ale nie w układzie liniowym, lecz w logarytmicznym. Podobnie, jak poprzednio, uwzględnienie mniejszej liczby słów pozwala dokładniej przeanalizować kształtowanie się wykresu dla słów z „najwyższej półki". Wykresy w układzie podwójnie logarytmicznym przyjmują postać funkcji prostoliniowej, z zawirowaniami na jej krańcach (to znaczy dla wyrazów znajdujących się na początkowych i końcowych pozycjach).

Tab. 1.13. Przykłady ilustrujące prawo Zipfa

WSJ 1-GRAM	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)
THE	1	2 057 968	5,145	0,051	0,05	0,1
OF	2	973 650	2,434	0,049	0,05	0,0
TO	3	940 525	2,351	0,071	0,07	0,1
A	4	853 342	2,133	0,085	0,09	0,1
AND	5	825 489	2,064	0,103	0,10	0,1
IN	6	711 462	1,779	0,107	0,11	0,1
THAT	7	368 012	0,920	0,064	0,06	0,1
FOR	8	362 771	0,907	0,073	0,07	0,1
ONE	9	298 646	0,747	0,067	0,07	0,1
IS	10	281 190	0,703	0,070	0,07	0,1
	S1	7 673 055	19,2	0,107	=max	0,074
	N	40 000 000		0,049	=min	0,058

WSJ BIGRAM	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)
OF THE	1	217 427	4,140	0,041	0,04	0,0
IN THE	2	173 797	3,309	0,066	0,07	0,1
MILLION DOLLARS	3	110 291	2,100	0,063	0,06	0,1
U. S.	4	89 184	1,698	0,068	0,07	0,1
NINETEEN EIGHTY	5	83 799	1,596	0,080	0,08	0,1
FOR THE	6	76 187	1,451	0,087	0,09	0,1
TO THE	7	72 312	1,377	0,096	0,10	0,1
ON THE	8	65 565	1,248	0,100	0,10	0,1
ONE HUNDRED	9	63 838	1,216	0,109	0,11	0,1
THAT THE	10	55 014	1,048	0,105	0,10	0,1
	S2	1 007 414		0,109	max	
U2=S2/S1*100	13,1	5 251 697	U2*N/100	0,041	min	0,068

Źródło: opracowanie własne na podstawie prac: D.B. Paul, J.M. Baker *The Design for the Wall Street Journal – based CSR Corpus*, Proc. ICSLP, 1992, p. 899–902 oraz Le Quan Ha, E.I. Sicilia-Garcia, Ji Ming, F.J. Smith, *Extension of Zipf's Law to Words and Phrases,* The Association for Computational Linguistics, A Digital Archive of Research Papers, www.aclweb.org/anthology/C/C02/C02-1117.pdf.

	Associated.Press	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)
T	THE	1	2 420 778	6,487	0,065	0,06	0,1
T	OF	2	1 045 733	2,802	0,056	0,06	0,1
T	TO	3	968 882	2,596	0,078	0,08	0,1
T	A	4	892 429	2,392	0,096	0,1	0,1
T	AND	5	865 644	2,32	0,116	0,12	0,1
T	IN	6	847 825	2,272	0,136	0,14	0,1
	SAID	7	504 593	1,352	0,095	0,09	0,1
T	FOR	8	363 865	0,975	0,078	0,08	0,1
X	THAT	9	347 072	0,93	0,084	0,08	0,1
	WAS	10	293 027	0,785	0,079	0,08	0,1
			8 549 848	22,9	0,088	=średnia	

WSJ TRIGRAM	Poz (r)	Częst. c	c (%)	rc (3)	rc (2)	rc (1)
THE U. S.	1	42 030	4,081	0,041	0,04	0,0
IN NINETEEN EIGHTY	2	27 260	2,647	0,053	0,05	0,1
CENTS A SHARE	3	24 165	2,346	0,070	0,07	0,1
NINETEEN EIGHTY SIX	4	18 233	1,770	0,071	0,07	0,1
NINETEEN EIGHTY SEVEN	5	16 786	1,630	0,081	0,08	0,1
FIVE MILLION DOLLARS	6	15 316	1,487	0,089	0,09	0,1
MILLION DOLLARS OR	7	14 943	1,451	0,102	0,10	0,1
MILLION DOLLARS IN	8	14 517	1,410	0,113	0,11	0,1
IN NEW YORK	9	12 327	1,197	0,108	0,11	0,1
A YEAR EARLIER	10	11 981	1,163	0,116	0,12	0,1
	S3	197 558		0,116	max	
U3=S3/S1*100	2,6	1 029 879	U3*N/100	0,041	min	0,076

Tab. 1.14. Analiza Zipfa utworu Michała Bułhakowa „Mistrz i Małgorzata"

wyraz	rangar	częstc	c (%)	rc/100	wyraz	rangar	częstc	c (%)	rc/100
się	1	2224	3,040	0,030	go	21	258	0,353	0,074
i	2	2217	3,030	0,061	pan	22	255	0,349	0,077
w	3	1983	2,710	0,081	było	23	252	0,344	0,079
na	4	1649	2,254	0,090	jego	24	223	0,305	0,073
nie	5	1507	2,060	0,103	mu	25	218	0,298	0,074
z	6	1265	1,729	0,104	od	26	204	0,279	0,072
że	7	905	1,237	0,087	tylko	27	204	0,279	0,075
to	8	863	1,180	0,094	jeszcze	28	201	0,275	0,077
do	9	828	1,132	0,102	pod	29	198	0,271	0,078
a	10	586	0,801	0,080	jej	30	194	0,265	0,080
o	11	461	0,630	0,069	przez	31	192	0,262	0,081
ale	12	412	0,563	0,068	tego	32	186	0,254	0,081
jak	13	404	0,552	0,072	ze	33	184	0,251	0,083
co	14	370	0,506	0,071	był	34	168	0,230	0,078
po	15	370	0,506	0,076	kiedy	35	164	0,224	0,078
już	16	354	0,484	0,077	mi	36	158	0,216	0,078
za	17	320	0,437	0,074	sobie	37	156	0,213	0,079
tak	18	304	0,416	0,075	ma	38	154	0,210	0,080
tym	19	267	0,365	0,069	mnie	39	153	0,209	0,082
jest	20	266	0,364	0,073	powiedział	40	153	0,209	0,084

Źródło: opracowanie własne.

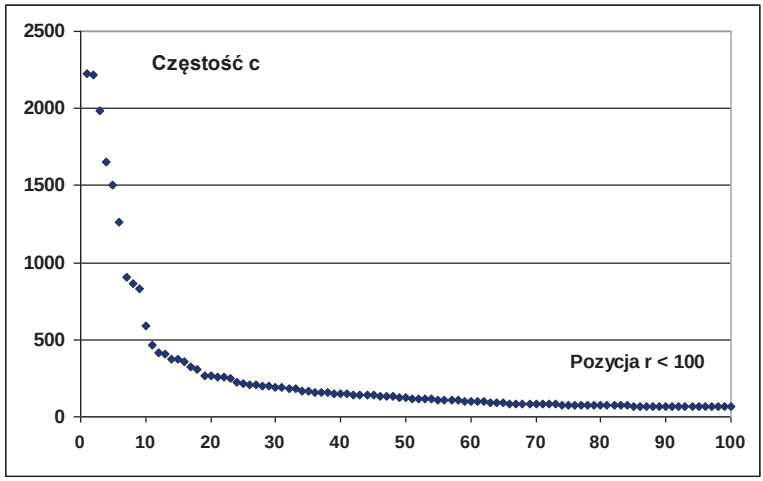
Tab. 1.15. Wyniki analizy Zipfa utworu Michała Bułhakowa „Mistrz i Małgorzata"

L. wyraz.	Częst. c	c (%)	c (% kum)	słowa (%)	rc
10	14 027	19,17	19,2	0,055	0,083
20	17 555	4,82	24,0	0,110	0,078
30	19 762	3,02	27,0	0,165	0,077
40	21 430	2,28	29,3	0,221	0,078
N	73 163			18 132	

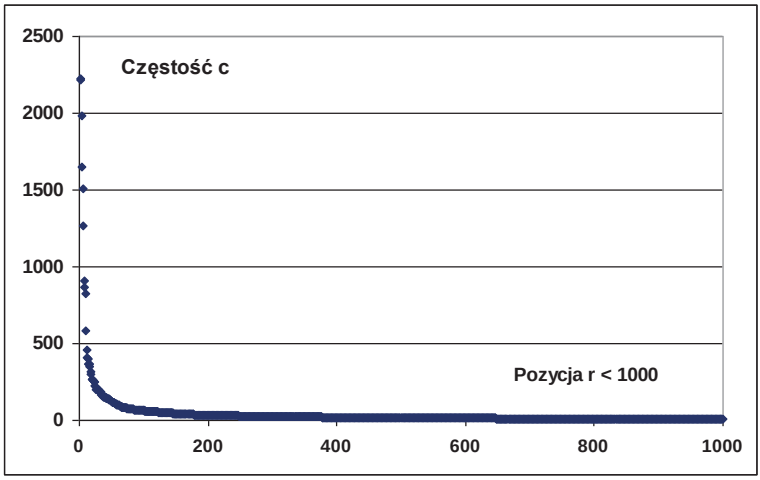
Źródło: opracowanie własne.

Kolejna reguła podobna do prawa Zipfa to ***prawo Heapsa*** opisujące zależność[9] między ***rozmiarem tekstu N a liczbą użytych w nim różnych wyrazów V***. Relację ta wyraża wzór $V=K*N^a$, w którym parametry K oraz a ustala się empirycznie. Dla języka angielskiego zazwyczaj K ≈ [30;100] natomiast a ≈ [0,4;0,6]. Dla innych języków parametry te przyjmują inne wartości. Reguła Heapsa może być wykorzystana np. w projektowaniu tekstowych baz danych do ustalenia rozmiarów indeksu jako funkcji rozmiaru bazy danych.

Rys. 1.12. Wykres Zipfa typu częstość–pozycja dla 100 najczęstszych wyrazów

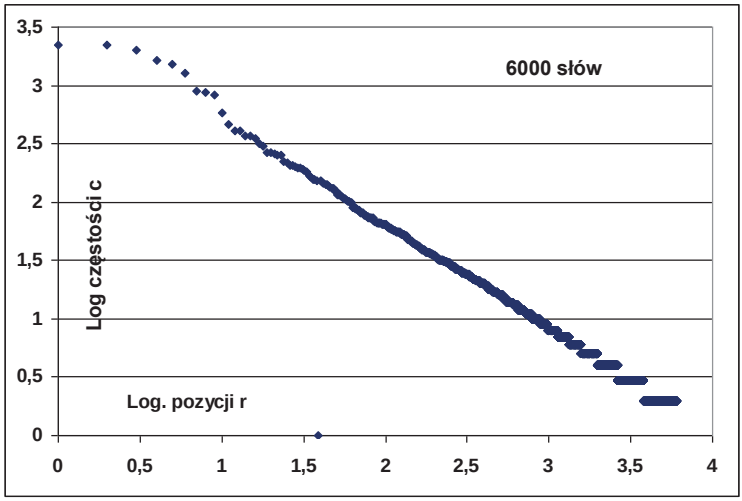


Rys. 1.13. Wykres Zipfa typu częstość–pozycja dla 1000 najczęstszych wyrazów

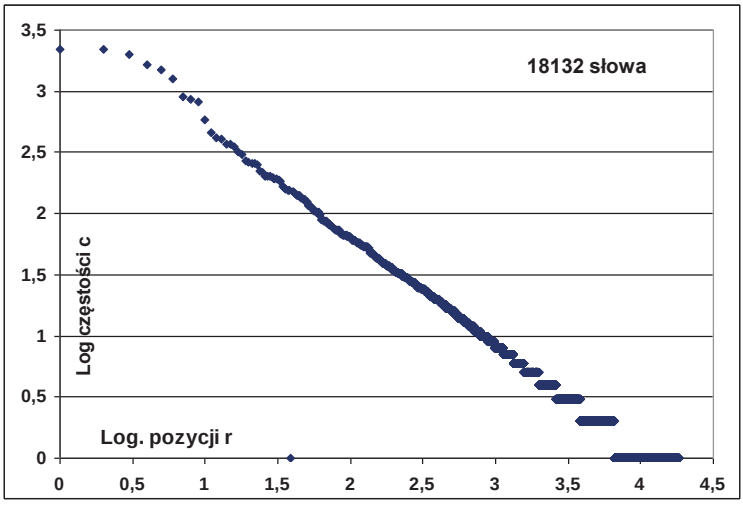


[9] Por. D.C. van Leijenhorst, T.P. van der Weide, *A formal derivation of Heaps' Law,* Information Science, 170/2005, p. 263–272.

Rys. 1.14. Logarytmiczny wykres Zipfa typu częstość–pozycja dla 6000 najczęstszych wyrazów



Rys. 1.15. Logarytmiczny wykres Zipfa typu częstość-pozycja dla wszystkich 18 000 wyrazów



Prawo Zipfa dotyczy także ludności i powierzchni miast, przychodów i rozmiarów korporacji gospodarczych, rozkładu dochodów osób, rozkładu liczby trzęsień ziemi od najsłabszych do najsilniejszych. Równanie Zipfa pozwala opisać popularność witryn internetowych, dzięki czemu można bardziej wydajnie zaprojektować tabele adresowe serwerów[10].

[10] Inne przykłady zastosowania prawa Zipfa znaleźć można m.in. w pracach A.A. Adamic, B.A. Huberman *Zipf's Law and the Internet*, Glottometrics, 3/2002, p. 143–150, Y.M. Ioannides, H.G. Overman, *Zipf's law for cities: an empirical examination*. Regional Science and Urban Economics 2002; X. Gabaix, *Zipf's Law and the growth of cities*, American Economic Review, 89/1999, p. 129–132; M. Aida, N. Takahashi, T. Abe, *A proposal of dual Zipfian model forde-*

W pracy z 1941 r. Zipf stwierdził[11], że krzywa rozkładów dochodów dla Indonezji wyraźnie odbiega od reguły rozkładu potęgowego, stąd wysnuł tezę, że w tym kraju będą duże napięcia społeczne. Rewolucja w Indonezji rzeczywiście zaczęła się 4 lata później, w 1945 roku.

Prawo Zipfa zakłada, że istnieje ścisły związek pomiędzy rangami obiektów (miasto, firma, osoba) w ich uporządkowaniu ze względu na analizowaną cechę a wielkością tej cechy (rank-size rule). Odwrotnością prawa Zipfa jest prawo proporcjonalnego wzrostu Gibrata (1904–1980), zgodne z którym wielkość firmy i stopień jej wzrostu są niezależne od siebie[12].

Alfred James Lotka (1880–1949)

- Amerykański matematyk, chemik, statystyk, biolog, demograf.
- Prezydent Population Association of Americ (1938–1939), American Statistical Association (1942).
- Statystyk w agencji ubezpieczeniowej w Nowym Jorku (1924–1947).
- Współtwórca modelu drapieżnik–ofiara (model Lotki–Volterry). Vito Volterra (1926) podał równanie opisujące populację ryb odławianych w Morzu Adriatyckim, natomiast A. Lotka (1910–1920) równanie opisujące oscylację stężeń substancji w reakcji chemicznej. Model Lotki–Volterry pozwala analizować układy dynamiczne w ekosystemach, w demografii a także w gospodarce.
- Sformułował równanie łączące strukturę wieku ludności, płodność i umieralność, co w demografii dało początek koncepcji ludności ustabilizowanej.

***Prawo Lotki* (1926)**

Prawo to nazywane ***prawem produktywności pracowników naukowych***, opisuje zależność między liczbą publikacji (X) a liczbą autorów (Y) mających daną liczbę publikacji (w zadanym okresie czasu i w ustalonym obszarze merytorycznym)[13]. Odpowiedni wzór ma postać $Y=const/X^{a,}$ gdzie *const* oraz *a* to stałe. Parametr $a \approx 2,$ natomiast *const* przyjmuje wartość zależną od analizowanej dziedziny i w przybliżeniu równy jest liczbie pra-

scribing HTTP access trends and its application to address cache design, IEICE Transactions on Communications, 81(7)/1998, p. 1475–1485.

[11] G.K. Zipf *National Unity and Disunity. The Nation as a Bio-Social Organism*, Principia Press, Bloomingotn Indiana, Princeton Press, 1941.

[12] Więcej informacji o G.K. Zipfie i jego osiągnięcich zob. *To honor G.K. Zipf,* Glottometrics, 3/2002 (m.in. prace: R. Rousseau *George Kingsley Zipf: life, ideas, his law and informetrics,* p. 11–18; A. Altman, *Zipfian linquistics,* p. 19–26).

[13] A.J. Lotka, *The frequency distribution of scientific productivity*, Journal of the Washington Academy of Science, 16/1926, p. 317–323.

cowników, którzy napisali tylko jeden artykuł w badanym obszarze tematycznym.

Inaczej mówiąc, ***liczba autorów (Y), z których każdy napisał pewną liczbę prac, jest odwrotnie proporcjonalna do kwadratu liczby tych prac (X)***. Większość autorów (60%) pisze jedną publikację, a bardzo mały odsetek autorów tworzy dużą część publikacji. Dla przykładu jeżeli przyjmiemy za *const*=200 to liczba autorów, którzy napisali od 1 do 12 artykułów kształtuje się następująco:

X - liczba artykułów	1	2	3	4	5	6	7	8	9	10	11	12
Y - liczba autorów z liczbą artykułów 1,2,...	200	50	22	13	8	6	4	3	2	2	2	1

Podobne zależności mają miejsce na przykład w lotnictwie wojskowym, dla odzwierciedlenia liczby pilotów wojskowych w zależności od liczby zestrzelonych przez nich samolotów.

Prawa Bradforda, Zipfa, Heapsa, Lotki i ich mutacje są podstawą dyscypliny zwanej ***informetrią* (*infometrics*)** i będącą częścią informatologii (informologii) jako nauki zajmującej się teoretycznymi i praktycznymi aspektami wiedzy o informacji. Informetria kładzie nacisk na wykorzystanie metod ilościowych i statystycznych do badania praw rządzących informacją.

Innymi działami informatologii[14] są:

- bibliometria analizująca zjawiska i procesy, w których biorą udział dokumenty (wydzielenie najczęściej cytowanych czasopism, relacje pomiędzy czasopismami krajowymi i zagranicznymi, ocena efektywności działalności bibliotecznej),
- naukometria, zajmująca się ilościową charakterystyką struktury nauki, określeniem dynamiki i kierunków jej rozwoju,
- webometria, badająca dynamikę zmian w środowisku www,
- cybermetria, poszerzająca zakres badań webometrii o zasoby elektroniczne,
- infobrokering, tworzy zasady i metody efektywnego pozyskiwania adekwatnych informacji, zwłaszcza w postaci elektronicznej.

[14] Więcej informacji znaleźć można w pracach: M. Dembowska, *Nauka o informacji naukowej (informatologia). Organizacja i problematyka badań w Polsce*, Instytut Informacji Naukowej, Technicznej, Ekonomicznej IINTE, Warszawa 1991; J. Ratajewski, *Wybrane problemy metodologiczne informologii nauki (informacji naukowej),* Prace Naukowe UŚ, Katowice 1994; B. Sordylowa, *Informacja naukowa w Polsce. Problemy teoretyczne, źródła, organizacja*, Ossolineum, Wrocław 1987; E. Ścibor, *Informacja naukowa w Polsce: tradycja i współczesność*, Olsztyn 1998; Z. Żmigrodzki (red.) i in., *Informacja naukowa: rozwój, metody, organizacja*, Wyd. SBP, Warszawa 2006.

1.4. Odkrywcy prawa Benforda

Poniżej przytoczono krótkie charakterystyki osób (por. tab. 1.16), które szczególnie przyczyniły się do powstania i popularności prawa Benforda. Należy do nich zaliczyć przede wszystkim nieżyjących już twórców prawa – S. Newcomba oraz F. Benforda. Jak się wydaje, do tej listy można dołączyć także T. Hilla oraz M. Nigriniego, którzy od wielu lat prowadzą najbardziej intensywne badania nad prawami rozkładu cyfr znaczących.

Simon Newcomb (1835–1909)

- Studiowal frenologię, naukę o powiązaniu kształtu czaszek z osobowością.
- W wieku 21 lat zatrudniony jako „computer" (osoba ds. obliczeń) w Nautical Almanac Office, Cambridge, Massachusetts.
- Później dyrektor Nautical Almanac Office.
- Profesor matematyki i astronomii w John Hopkins University.
- Przyczynił się do pomiaru szybkości światła, ustalenia dokładnej orbity Księżyca.
- Stworzył system stałych astronomicznych.
- Założyciel i prezes: American Astronomical Society, American Mathematical Society, American Association for the Advancement of Science, Philosophical Society of Washington.
- Najbardziej znany amerykański astronom. Otrzymał najwyższe nagrody naukowe w zakresie astronomii w USA, Wielkiej Brytanii, Holandii i Niemczech.
- W Kanadzie przyznawana jest przez Royal Astronomical Society nagroda jego imienia – Simon Newcomb Award.
- Praca na temat rozkładu pierwszych cyfr znaczących ze wzorem

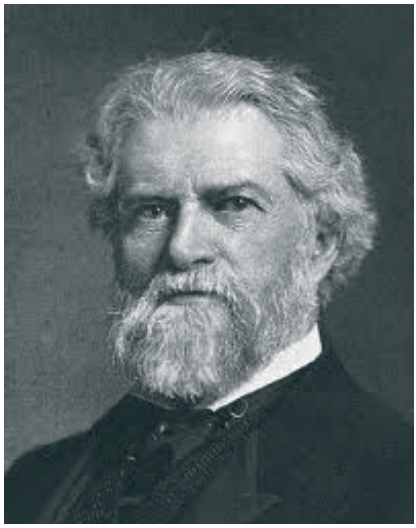
$$p(d)=\log(1+1/d)$$

  liczyła 2 strony i nie została przez 60 lat przez nikogo zauważona.
- W 1885 r. przyczynił się do powstania prawa Irvinga Fishera (1911) w zakresie ilościowej teorii pieniądza *(Quantity Theory of Money)* wyrażonej wzorem $MV=PT$, gdzie M – liczebność pieniądza w obiegu, V – szybkość obiegu pieniądza (transakcji), P – poziom cen, T – liczba transakcji. Z podanego wzoru wynika, że jeżeli relacja T do V jest stała, to wzrost pieniądza w obiegu powoduje wzrost cen (inflację).

Frank Benfrod (1883–1948)

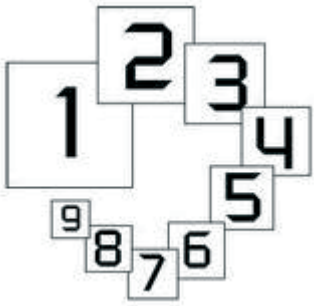
- Amerykański inżynier, fizyk, elektrotechnik.
- Po ukończeniu Uniwersytetu w Michagan w 1910 r., pracował w firmie General Electric, najpierw przez 18 lat w Illuminating Engineering Lab, a potem przez 20 lat w Research Lab.
- W 1937 r. skonstruował instrument pomiaru załamywania światła.
- Ekspert pomiarów optycznych, autor 109 dokumentów z zakresu optyki i matematyki.
- Opracował 20 patentów na przyrządy optyczne.
- W 1938 r. opublikował jedyną w swoim dorobku pracę na temat rozkładu cyfr znaczących. W swojej pracy nie wykorzystał informacji wynikających z wcześniejszej pracy S. Newcomba z 1881 r. Publikacja ta nie została zauważona przez 30 lat.

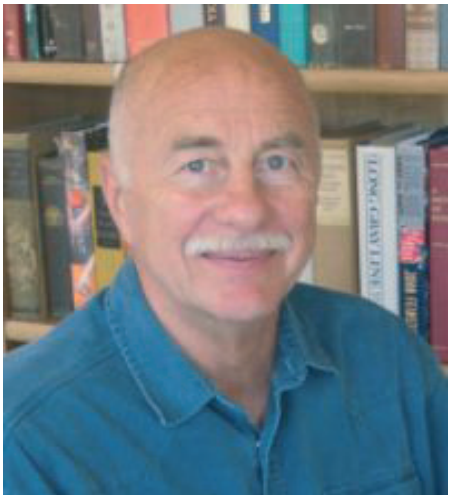
Tab. 1.16. Odkrywcy prawa Benforda



Simon Newcomb (1835–1909)

Frank Benford (1883–1948)





Theodore Hill (1943- ...)

Mark Nigrini (1957 - ...)

Źródło: opracowanie własne.

Theodore Preston Hill (1943– ...)

- Matematyk zatrudniony w George Institute of Technology, specjalizujący się w teorii prawdopodobieństwa, a w szczególności w prawie Benforda.
- Absolwent Akademii West Point (1966), rocznika o największym procencie ofiar wojny w Wietnamie, uczestnik tej wojny.
- Autor prac dotyczących wyboru najlepszego obiektu (problem łowcy posagu, problem sekretarki) ze skończonego zbioru istotnie różnych propozycji, prezentowanych w losowej kolejności (reguła zatrzymywania, *optimal stopping theory*).
- Autor prac związanych z problemem sprawiedliwego podziału *(fair division problem)* oraz strategią *Cut-And-Choose.*
- Twórca serwisu internetowego benfordonline.net, zawierającego bibliografię prac związanych z prawem Benforda, a także dwóch innych portali motherfunctor.org oraz earlyamericanmathbooks.org
- Pełna dokumentacja dorobku naukowego znajduje się na witrynie tphill.net.

Mark Nigrini (1957– ...)

- Profesor w St. Michael's College of New Jersey, specjalizacja: rachunkowość finansowa, rachunkowość zarządcza, audyt, zastosowania matematyki, systemy informatyczne, wykrywanie oszustw finansowych.
- Członek American Accounting Assotiation, Institute of Internal Auditors.
- W 1992 r. obronił pracę doktorską (University of Cincinnati) na temat *The Detecion of Income Tax Invasion Through an Analysis of Digital Distributions.*

- Analiza zwrotów podatków Billa Clintona w latach 1977–1992.
- Pierwszy z naukowców, który próbuje zarobić pieniądze na znajomości prawa Benforda – płatne seminaria naukowe (>150 dol.), książki w cenie 32 centy za stronę.
- Zebrał wiele przypadków opisujących oszustwa finansowo-księgowe.

Wyczerpujące informacje o M. Nigrinim i jego prac znaleźć można w witrynie nigrini.com.

W tabeli 1.17 przedstawiono kalendarium ważniejszych wydarzeń, jakie miały miejsce w związku z prawem Benfora przed rokiem 2000. Wskazano tu na rolę odkrywców tego prawa, a także na osiągnięcia kilka innych badaczy, którzy przyczynili się w początkowym okresie do rozwoju, popularyzacji i zastosowań prawa pierwszych cyfr znaczących.

Tab. 1.17. Kalendarium ważniejszych osiągnięć w zakresie prawa Benforda przed 2000 rokiem

Rok	Nazwisko	Wyszczególnienie	Praca
1881	S. Newcomb	Pierwsza praca na temat rozkładu cyfr znaczących. Niezauważona przez 60 lat.	Note on the frequency of use of the different digits in natural numbers, American Journal of Mathematics 4(1)/1881, p. 39–40.
1938	F. Benford	Fundamentalna praca dająca początek teorii i analizy rozkładu cyfr znaczących.	The law of anomalous numbers, Proceedings of the American Philosophical Society 78/1938, p. 551–572.
1944	S.A. Goudsmith W.H. Furry	Teza, że prawo Benforda wynika ze sposobu zapisu liczb.	Significant figure of numbers in statistical tables, Nature, 154/1944, p. 800–801.
1945	G.J. Stigler	Modyfikacja formuły wyznaczania pierwszej cyfry znaczącej oparta na innych założeniach niż formuła Benforda.	The distribution of leading digits in statistical tables (praca nieopublikowana).
1948	L.V. Furlan	Prawo Benforda oddaje harmoniczną istotę rzeczywistości (skala logarytmiczna).	Das Harmoniesgestez der Statistik,Eine Untersuchung uber die metrische Interdependenz der sozialeen Erscheinungen, Bael, Switzerland, Verlag fur Recht und Geselschaft, XIII/1948.
1961	R.S. Pinkham	Dowód na niezmienniczość skali rozkładu Benforda względem dowolnej operacji arytmetycznej (mnożenia, dzielenia, potęgowania).	On the Distribution of First Significant Digits. Annals of Mathematical Statistics 32(4)/1961, p. 1223–1230.
1969	R.A. Raimi	Popularyzacja problematyki prawa Benforda.	The Peculiar Distribution of First Digits. Scientific American 221(6)/1969, 109–119, On Distribution of First Significant Figures. American Mathematical Monthly 76(4)/1969, p. 342–348.

Rok	Nazwisko	Wyszczególnienie	Praca
1988	C. Carslaw	Pierwsze zastosowania prawa Benforda do analizy danych finansowych w poszukiwaniu błędów.	Anomalies in Income Numbers: Evidence of Goal Oriented Behavior,The Accounting Review 63(2)/1988, p. 321–327.
1993	M.J. Nigrini	Pierwsze próby wykorzystania prawa Benforda w audycie księgowym.	Can Benford's law be used in forensic accounting?. The Balance Sheet, VI/1993, 7–8. Using digital frequencies to detect fraud. Fraud Magazine,The White Paper Index 8(2)/1994, 3–6; A taxpayer compliance application of Benford's law. Journal of the American Taxation Association 18(1)/1996, p. 72–91.
1995	T.P. Hill	Dowód na niezmienniczość podstawy systemu liczbowego, którą to własność rozkład Benforda posiada jako jedyny	Base-Invariance Implies Benford's Law. Proceedings of the American Mathematical Society 123(3)/1995, 887–895, The Significant-Digit Phenomenon. American Mathematical Monthly 102(4)/1995, p. 322–327.
1996	T.P. Hill	Prawo Benforda opisuje "rozkład rozkładów" – losowo dobrane próby z losowo dobranych rozkładów.	A Statistical Derivation of the Significant-Digit Law. Statistical Science 10(4)/1996, 354–363 The first digital phenomenon, American Scientist, 86/1998, p. 358–363.
1996	E. Leo	Pierwsze zastosowania prawa Benforda do analizy danych giełdowych.	On the Peculiar Distribution of the US Stock Indexes' Digits. American Statistician 50(4)/1996, p. 311–313.
1997	M.J. Nigrini	Powstaje *digital analysis* – system procedur do analizy własności rozkładów cyfr i ich kombinacji w celu poszukiwania nietypowych wartości.	The use of Benford's Law as an aid in analytical procedures. Auditing – A Journal of Practice & Theory 16(2)/1997, 52-67 (M.J. Nigrini, L.J. Mittermaier) Numerology for Accountants. Journal of Accountancy, November 1998, p. 15. Adding value with digital analysis. The Internal Auditor 56(1)/1999, p. 21–23.

Rozdział 2
Istota prawa Benforda

2.1. Podstawowe informacje

Frank Benford, fizyk zatrudniony w laboratorium General Electrics, zainspirowany pracą Newcomba również zwrócił uwagę na nierównomierne zabrudzenie tablic logarytmicznych i podjął kilkuletnie badania mające na celu sprawdzenie, czy rzeczywiście mamy do czynienia z sytuacją, w której częściej na pierwszym miejscu liczb występują raczej cyfry mniejsze niż większe. W swojej pracy[1] opublikował wyniki analizy ponad 20 tys. liczb charakteryzujących 20 różnych zjawisk (m.in. powierzchnia rzek, liczba mieszkańców w jednostkach administracyjnych, dane adresowe osób z *Amercian Men of Science*, wyniki rozgrywek w baseballu, wskaźniki śmiertelności, liczby znajdujące się w artykułach zamieszczonych w *Readers Digest,* itp.). W większości przypadków F. Bedford uzyskał podobne wyniki, które doprowadziły go do sformułowania prawa opisującego częstość występowania cyfr od 1 do 9 na początku liczb.

Zgodnie z tym prawem prawdopodobieństwo $P(d_i)$ z jakim pojawia się ***pierwsza*** cyfra znacząca d_i w liczbach ***wielocyfrowych*** wziętych z dużego zbioru liczb dana jest wzorem:

(2.1)

$$P(d_i) = \frac{\log(d_{i+1}) - \log(d_i)}{\log(10) - \log(1)} = \log\left(\frac{d_i + 1}{d_i}\right) = \log\left(1 + \frac{1}{d_i}\right) = \log(d_{i+1}) - \log(d_i) \quad (d_i = 1,2,\ldots,9)$$

W tabeli 2.1. ukazano wyniki analiz uzyskane przez F. Benforda. Zbiory danych uporządkowane są według malejących wartości statystyki chi-kwadrat oraz odpowiadającemu tej statystyce prawdopodobieństwu *p* zdarzenia polegającego na odrzuceniu prawdziwej hipotezy o zgodności danego rozkładu empirycznego z rozkładem Benforda.

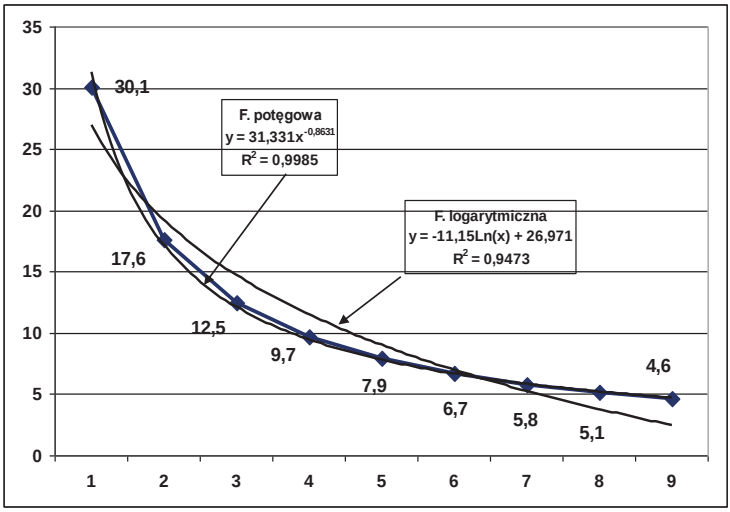
Jak można zauważyć (por. rys. 2.1) w rozkładzie pierwszych cyfr znaczących podanym przez Benforda cyfra 1 występuje z częstością 30%, następne cyfry pojawiają się coraz rzadziej, aż do cyfry 9, dla której częstość wynosi tylko 4,6%. Rozkład Benforda z dużą dokładnością (R^2=0,999) można aproksymować funkcją potęgową:

[1] F. Benford, *The Law of Anomalous Numbers*, Proceedings of the Amercian Philosophical Society, 78/1938, p. 551–572.

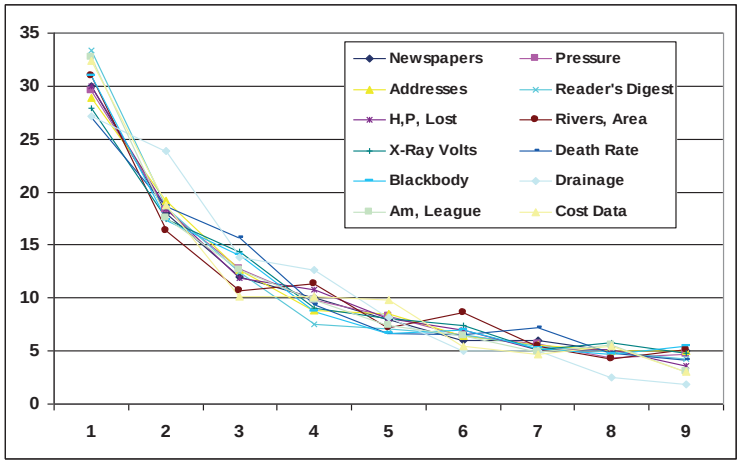
(2.2)
$$\begin{aligned} P(d_i) &= 0{,}31331 * d_i^{-0{,}8631} \quad (d_i = 1,2,...,9) \\ 100 * P(d_i) &= 31{,}331 * d_i^{-0{,}8631} \quad (d_i = 1,2,...,9) \end{aligned}$$

Inne funkcje, np. logarytmiczna nie oddają tak dobrze przebiegu krzywej Benforda.

Rys. 2.1. Prawo Benforda – rozkład częstości pierwszych cyfr znaczących



Rys. 2.2. Rozkłady częstości pierwszych cyfr znaczących w 12 zbiorach FB najbardziej zgodnych z prawem Benforda

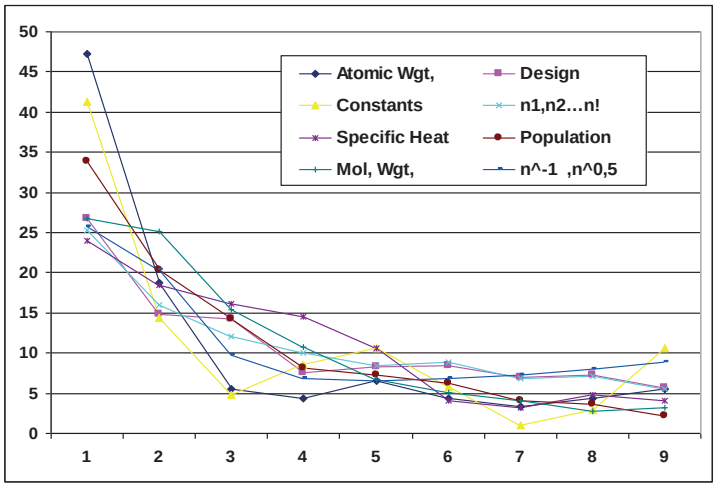


Tab. 2.1. Rozkłady pierwszych cyfr znaczących w zbiorach analizowanych przez Franka Benforda uporządkowane wg wartości statystyki chi-kwadrat

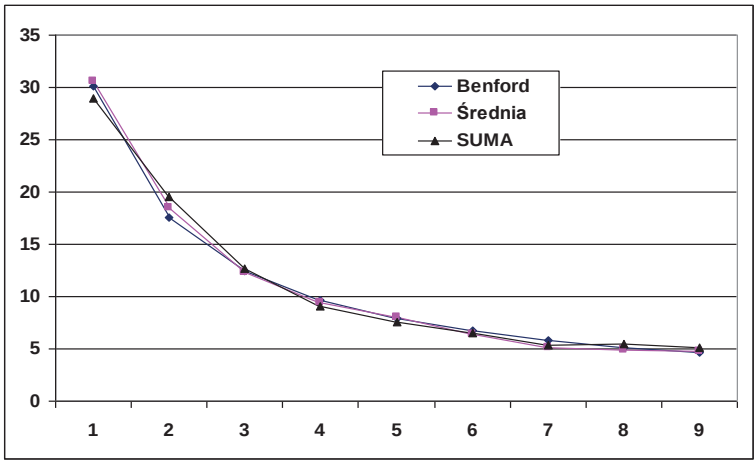
Nr	Symbol	Nazwa	1	2	3	4	5	6	7	8	9	n	chi kw.	p
1	D	Newspapers	30,0	18,0	12,0	10,0	8,0	6,0	6,0	5,0	5,0	100	0,2	1,000
2	F	Pressure	29,6	18,3	12,8	9,8	8,3	6,4	5,7	4,4	4,7	703	1,3	0,996
3	R	Addresses	28,9	19,2	12,6	8,8	8,5	6,4	5,6	5,0	5,0	342	1,3	0,996
4	M	Reader's Digest	33,4	18,5	12,4	7,5	7,1	6,5	5,5	4,9	4,2	308	3,2	0,921
5	G	H, P, Lost	30,0	18,4	11,9	10,8	8,1	7,0	5,1	5,1	3,6	690	3,5	0,899
6	A	Rivers, Area	31,0	16,4	10,7	11,3	7,2	8,6	5,5	4,2	5,1	335	5,0	0,758
7	O	X-Ray Volts	27,9	17,5	14,4	9,0	8,1	7,4	5,1	5,8	4,8	707	5,4	0,714
8	T	Death Rate	27,0	18,6	15,7	9,4	6,7	6,5	7,2	4,8	4,1	418	7,6	0,473
9	Q	Blackbody	31,0	17,3	14,1	8,7	6,6	7,0	5,2	4,7	5,4	1165	9,5	0,302
10	**I**	**Drainage**	**27,1**	**23,9**	**13,8**	**12,6**	**8,2**	**5,0**	**5,0**	**2,5**	**1,9**	**159**	**11,1**	0,196
11	P	Am, League	32,7	17,6	12,6	9,8	7,4	6,4	4,9	5,6	3,0	1458	14,6	0,067
12	N	Cost Data	32,4	18,8	10,1	10,1	9,8	5,5	4,7	5,5	3,1	741	15,6	0,048
13	J	Atomic Wgt,	47,2	18,7	5,5	4,4	6,6	4,4	3,3	4,4	5,5	91	17,2	0,028
14	L	Design	26,8	14,8	14,3	7,5	8,3	8,4	7,0	7,3	5,6	560	19,2	0,014
15	C	Constants	41,3	14,4	4,8	8,6	10,6	5,8	1,0	2,9	10,6	104	24,4	0,002
16	S	n1,n2...n!	25,3	16,0	12,0	10,0	8,5	8,8	6,8	7,1	5,5	900	25,0	0,002
17	E	Specific Heat	24,0	18,4	16,2	14,6	10,6	4,1	3,2	4,8	4,1	1389	111,2	0,0000000
18	B	Population	33,9	20,4	14,2	8,1	7,2	6,2	4,1	3,7	2,2	3259	118,6	0,0000000
19	H	Mol, Wgt,	26,7	25,2	15,4	10,8	6,7	5,1	4,1	2,8	3,2	1800	125,8	0,0000000
20	K	n-1, n1/2	25,7	20,3	9,7	6,8	6,6	6,8	7,2	8,0	8,9	5000	440,8	0,0000000
		Benford	**30,1**	**17,6**	**12,5**	**9,7**	**7,9**	**6,7**	**5,8**	**5,1**	**4,6**			
		Średnia	30,6	18,5	12,3	9,4	8	6,4	5,1	4,9	4,8	1011	1,8	0,987
		SUMA	28,9	19,5	12,7	9,1	7,6	6,5	5,4	5,5	5,1	20229	85,4	0,0000000

Źródło: opracowanie własne.

Rys. 2.3. Rozkłady częstości pierwszych cyfr znaczących w 8 zbiorach FB najmniej zgodnych z prawem Benforda



Rys. 2.4. Rozkłady częstości pierwszych cyfr znaczących wg prawa Benforda oraz dla sumy i dla wartości średnich z 20 zbiorów FB



W tabeli 2.2 zestawiono wartości funkcji potęgowej (2.2) z częstościami rozkładu Benforda. Różnice pomiędzy tymi rozkładami nie są duże – średnia z modułów różnic częstości wynosi 0,3%, natomiast średnia z modułów różnic częstości względnych (w stosunku do częstości w prawie Benforda) – 2%.

Tab. 2.2. Funkcja potęgowa $P_i=33,331*d_i\hat{}-0,8631$ aproksymująca rozkład Benforda

d	F. potęg.	F. Benf.	P–B	(P–B)/B %
1	31,3	30,1	1,2	4,1
2	17,2	17,6	-0,4	-2,2
3	12,1	12,5	-0,4	-2,8
4	9,5	9,7	-0,2	-2,3
5	7,8	7,9	-0,1	-1,4
6	6,7	6,7	0,0	-0,3
7	5,8	5,8	0,0	0,7
8	5,2	5,1	0,1	1,8
9	4,7	4,6	0,1	2,8
	100,3	100,0	**0,3**	2,0

Źródło: opracowanie własne.

Na rysunku 2.2 podano wykresy częstości pierwszych cyfr znaczących dla 12 zbiorów danych analizowanych przez F. Benforda, w których zaobserwowano największą zgodność rozkładu empirycznego z teoretycznym. Prawdopodobieństwo popełnienia błędu odrzucenia hipotezy o zgodności porównywanych rozkładów kształtuje się w tych przypadkach na poziomie nie większym niż 0,05.

Na rysunku 2.3 przytoczono wykresy dla pozostałych 8 zbiorów danych, w których zgodność z rozkładem Benforda była stosunkowo mniejsza. W przypadku dwóch zbiorów: *(J) masy atomowej* oraz *(L) Dane z projektów* zgodność z prawem Benforda obserwuje się przy ostrzejszym poziomie istotności [>0,01]. Również w przypadku następnych dwóch zbiorów: *C – Stałe* oraz *S – potęgi liczb naturalnych*, obniżając poziom istotności do 0,001, można by przyjąć hipotezę o zgodności rozkładów. Jedynie w 4 ostatnich zbiorach (na 20 analizowanych) trzeba zdecydowanie odrzucić hipotezę o zgodności rozkładów cyfr znaczących z rozkładem Benforda.

Na rysunku 2.4 przedstawiono rozkłady częstości dla zbioru sumarycznego (dla n=20229) powstałego jako mieszanka wszystkich 20 zbiorów danych *(SUMA)* oraz rozkład częstości ustalonych jako średnia z częstości dla poszczególnych zbiorów danych *(Średnia)*. Sądząc z wykresu, obydwa te rozkłady charakteryzują się dużą wzajemną zgodnością, jak również zgodnością z prawem Benforda. Jednak rzut oka na statystki chi-kwadrat wskazuje, że wniosek o zgodności porównywanych rozkładów uzasadniony jest tylko dla rozkładu uśrednionego *(Średnia)* natomiast w przypadku rozkładu sumarycznego *(SUMA)* hipotezę o zgodności z rozkładem Benforda należy odrzucić. Związane to jest z liczebnością zbioru danych *n.* Liczebności te są

bardzo zróżnicowane – liczebność sumaryczna przekracza 20 000, natomiast liczebność średnia jest 20x mniejsza.

Z podanego przykładu wynika, że wzrokowa pobieżna ocena podobieństwa wykresów może być zawodna we wnioskowaniu o zgodności rozkładów. Trzeba zwrócić uwagę na wskazania mierników i testów statystycznych służących do obiektywnego ustalania, czy porównywane rozkłady są zbieżne, czy też różnice pomiędzy nimi są zbyt duże, aby je uznać za równorzędne.

Prawo Benforda powiązane jest ściśle z ciągiem Fibonacciego oraz mniej znanym ciągiem Lukasa. Ciągi te określone są rekurencyjnym wzorem:

(2.3) $$\boxed{F(i+1) = F(i) + F(i-1)}$$

- ciąg Fibonacciego F(0)=0 F(1)=1 {1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, ...}
- ciąg Lukasa F(0)=2 F(1)=1 {1, 3, 4, 7, 11, 18, 29, 47, 76, 123, 199, ...}

Analiza rozkładu 1475 liczb tych ciągów doprowadza do wniosku, że w obu przypadkach rozkłady pierwszych cyfr znaczących są całkowicie zgodne z rozkładem Benforda (por. tab. 2.3). Na marginesie warto dodać, że z uwagi na przekroczenie maksymalnego rzędu dokładności w Excelu, w obydwóch ciągach nie udało się uzyskać więcej liczb niż 1475. Ostatnie (możliwie największe) wartości w tych ciągach wynoszą odpowiednio:

– dla ciągu Fibonacciego 4,99225460547777000E+307

– dla ciągu Lukasa 1,11630206588347000E+308

Również inne ciągi, w których kolejny element jest sumą dwóch poprzednich elementów o dowolnych wartościach, są zgodne z rozkładem Benforda. W tabeli 2.3 podano rozkłady pierwszych cyfr dla dwóch innych ciągów, w których F(0)=F(1)=2 oraz F(0)=3 i F(1)=1. Różnice w rozkładach wynikają z zaokrągleń i nie przekraczają 1.

Tab. 2.3. Rozkłady pierwszych cyfr znaczących w ciągach Fibonacciego i Lukasa wraz z rozkładem Benforda dla n=1475

d	Benford	F0=0 F1=1	F0=2 F1=1	F0=F1=2	F0=3 F1=2
1	444	444	444	444	443
2	260	260	260	261	259
3	184	184	184	183	185

4	143	143	143	143	142
5	117	117	117	117	118
6	99	98	98	99	98
7	86	85	85	85	86
8	75	77	77	75	75
9	67	67	67	68	68
	1475	**1475**	**1475**	**1475**	**1474**
		Fibbonacci	**Lukas**		

Źródło: opracowanie własne.

2.2. Testy zgodności empirycznych rozkładów cyfr z rozkładami wynikającymi z prawa Benforda

W podręcznikach statystyki znanych jest wiele metod oceny zgodności rozkładów. Są to:

- test chi-kwadrat,
- testy Kołmogorowa–Smirnowa,
- testy istotności oparte na statystyce z,
- mierniki zgodności.

Metody te wykorzystywane są do odpowiedzi na pytanie, czy empiryczne rozkłady następujących cyfr:

- F1 – pierwszej cyfry znaczącej,
- F2 – dwóch pierwszych cyfr znaczących,
- F3 – trzech pierwszych cyfr znaczących,
- D2 – dokładnie drugiej cyfry znaczącej,
- D3 – dokładnie trzeciej cyfry znaczącej,
- L1 – ostatniej cyfry

są zgodne z rozkładami wynikającymi z prawa Benforda.

Poniżej podano wzory i definicje poszczególnych parametrów. We wzorach przyjęto następującą konwencję oznaczeń.

- n – ogólna liczba obserwacji w analizowanym zbiorze
- k – liczba kombinacji cyfr (w teście F1 – $k=9$, w testach D2, D3, L1 – $k=10$, w teście F2 $k=90$ natomiast w teście F3 – $k=900$)
- i – subskrypt oznaczający numer kolejny kombinacji cyfr ($i=1,2,...,k$)
- n_i oraz $\hat{n}_i$ ($i=1,2,...,k$) – liczebności empiryczne i teoretyczne pojawienia się i-tej cyfry (lub i-tej kombinacji cyfr)
- c_i oraz $\hat{c}_i$ ($i=1,2,...,k$) – częstości empiryczne i teoretyczne dane wzorami:

(2.4) $$c_i = \frac{n_i}{n}100 \qquad \hat{c}_i = \frac{\hat{n}_i}{n}100$$

- p_i oraz $\hat{p}_i$ ($i=1,2,...,k$) – prawdopodobieństwa empiryczne i teoretyczne dane wzorami:

(2.5) $$p_i = \frac{n_i}{n} \qquad \hat{p}_i = \frac{\hat{n}_i}{n}$$

- f_i oraz $\hat{f}_i$ $(i=1,2,...,k)$ - wartości dystrybuanty (kumulanty) empirycznego i teoretycznego rozkładu częstości cyfr dane wzorami:

(2.6) $$f_i = \sum_{l=1}^{i} p_l \qquad \hat{f}_i = \sum_{l=1}^{i} \hat{p}_l$$

Do oceny zgodności rozkładów najczęściej stosowany jest test chi-kwadrat, w którym wyznacza się wartość parametru:

(2.7) $$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i} = n\sum_{i=1}^{k} \frac{(p_i - \hat{p}_i)^2}{\hat{p}_i} = \frac{n}{100}\sum_{i=1}^{k} \frac{(c_i - \hat{c}_i)^2}{\hat{c}_i}$$

Statystykę (2.7) porównuje się z wartością krytyczną testu χ^2 dla założonego poziomu istotności α oraz *k-1* stopni swobody[2]. W tabeli 2.4, podano wartości krytyczne testu χ^2 dla wybranych poziomów istotności α = 0,1; 0,05, 0,01 i 0,001 oraz dla stopni swobody właściwych dla testów F1–F3, D2, D3 i L1[3].

[2] Niektórzy autorzy twierdzą, że test chi-kwadrat jest zbyt rygorystyczny, gdyż jego wskazania w dużym stopniu zależą od liczby elementów w analizowanym zbiorze *n*. W analizach związanych z rozkładami Benforda zwykle mamy do czynienia z bardzo dużymi zbiorami i w takich przypadkach nawet niewielkie odchylenia liczebności teoretycznych i empirycznych mogą sugerować istotną niezgodność porównywanych rozkładów. Por. E. Ley, *On the Peculiar Distribution of the U.S. Stock Indexes' Digits*, The American Statistician, 50/1996, p. 311–313; D.E.A.Giles, *Benford's Law and Naturally Occurring Prices in Certain e-Bay Auctions*, Applied Economics Letters, 14/2007, p.157–161.

[3] Wartości te można uzyskać stosując w Excelu funkcję =ROZKŁAD.CHI.ODWR (α; ss), gdzie ss to liczba stopni swobody (ss=k-1)

Tab. 2.4. Wybrane wartości krytyczne testu χ^2_α

Test	K	a =0,1	a =0,05	a =0,01	a =0,001
F1	9	13,4	15,5	20,1	26,1
D2, D3, L1	10	14,7	16,9	21,7	27,9
F2	90	106,5	112,0	122,9	136,0
F3	900	953,8	969,9	1000,6	1035,8

Źródło: opracowanie własne.

Im wyższa jest wartość empiryczna statystyki χ^2 tym bardziej różnią się porównywane rozkłady. Jeżeli $\chi^2 \geqslant \chi^2_\alpha$ to z prawdopodobieństwem $1-\alpha$ można twierdzić, że rozkład empiryczny ***nie jest zgodny*** z regułami prawa Benforda.

W analizach korzysta się także z parametru wskazującego jaki poziom istotności α odpowiada ustalonej wartości empirycznej testu χ^2. W tabeli 2.5 podano wartości tych prawdopodobieństw dla statystyk[4] z przedziału w którym zawierają się typowe wartości poziomów istotności od 0,001 do 0,10. Warto tu zwrócić uwagę, na fakt, że przy 900 stopniach swobody (test F3) oraz statystyce χ^2=890 funkcja w Excelu zwraca błąd „LICZBA!"

Tab. 2.5. Wartości prawdopodobieństw w rozkładzie χ^2 *dla zadanej wartości testu (chi emp) oraz przy liczbie stopni swobody właściwej dla testów Benforda*

Test	F1	D2,D3,L1		F2		F3
chi emp	9	10	**chi emp**	90	**chi emp**	900
4	0,857	0,911	**80**	0,742	**860**	0,821
6	0,647	0,740	**85**	0,600	**875**	0,711
8	0,433	0,534	**90**	0,450	**890**	#LICZBA!
10	0,265	0,350	**95**	0,312	**905**	0,438
12	0,151	0,213	**100**	0,200	**920**	0,306
14	0,082	0,122	**105**	0,118	**935**	0,197
16	0,042	0,067	**110**	0,065	**950**	0,116
18	0,021	0,035	**115**	0,033	**965**	0,062
20	0,010	0,018	**120**	0,016	**980**	0,031
22	0,005	0,009	**125**	0,007	**995**	0,014
24	0,002	0,004	**130**	0,003	**1010**	0,006
26	0,001	0,002	**135**	0,001	**1025**	0,002
28	0,000	0,001	**140**	0,000	**1040**	0,001
30	0,000	0,000	**145**	0,000	**1055**	0,000

Źródło: opracowanie własne.

[4] W tym przypadku korzysta się z funkcji w Excelu =ROZKŁAD.CHI (stat; ss) gdzie stat=χ^2.

Tab. 2.6. Wartości prawdopodobieństw w rozkładzie χ^2 dla zadanej wartości testu chi-kwadrat [870;900;1] oraz liczbie stopni swobody z przedziału [870;920;10]

chi emp / ss	870	880	890	900	910	920
870	0,484	#LICZBA!	0,669	0,750	0,819	0,875
871	0,475	#LICZBA!	0,661	0,743	0,813	0,870
872	0,465	#LICZBA!	0,652	0,735	0,806	0,864
873	0,455	#LICZBA!	0,643	0,727	0,800	0,859
874	0,446	#LICZBA!	0,634	0,719	0,793	0,854
875	0,437	#LICZBA!	#LICZBA!	0,711	0,786	0,848
876	0,427	#LICZBA!	#LICZBA!	0,702	0,779	0,842
877	0,418	#LICZBA!	#LICZBA!	0,694	0,772	0,836
878	0,409	#LICZBA!	#LICZBA!	0,686	0,764	0,830
879	0,399	0,494	#LICZBA!	0,677	0,757	0,824
880	0,390	0,484	#LICZBA!	0,668	0,749	0,818
881	0,381	0,475	#LICZBA!	0,660	0,741	0,811
882	0,372	0,465	#LICZBA!	0,651	0,734	0,805
883	0,363	0,456	#LICZBA!	0,642	0,726	0,798
884	0,354	0,446	#LICZBA!	#LICZBA!	0,718	0,791
885	0,346	0,437	#LICZBA!	#LICZBA!	0,710	0,784
886	0,337	0,428	#LICZBA!	#LICZBA!	0,701	0,777
887	0,328	0,418	#LICZBA!	#LICZBA!	0,693	0,770
888	0,320	0,409	#LICZBA!	#LICZBA!	0,685	0,763
889	0,311	0,400	0,494	#LICZBA!	0,676	0,756
890	0,303	0,391	0,484	#LICZBA!	0,668	0,748
891	0,295	0,382	0,475	#LICZBA!	0,659	0,740
892	0,287	0,373	0,465	#LICZBA!	0,650	0,733
893	0,279	0,364	0,456	#LICZBA!	#LICZBA!	0,725
894	0,271	0,355	0,447	#LICZBA!	#LICZBA!	0,717
895	0,263	0,346	0,437	#LICZBA!	#LICZBA!	0,709
896	0,256	0,338	0,428	#LICZBA!	#LICZBA!	0,700
897	0,248	0,329	0,419	#LICZBA!	#LICZBA!	0,692
898	0,241	0,321	0,410	#LICZBA!	#LICZBA!	0,684
899	0,233	0,312	0,400	0,494	#LICZBA!	0,675

Źródło: opracowanie własne.

W celu wyjaśnienia tego błędu przeanalizowano ciągi formuł =ROZKŁAD.CHI (chi; ss) dla wartości statystyki χ^2 z przedziału [870;900] ze skokiem co 1 oraz dla liczby stopni swobody z przedziału [870;920] ze skokiem co 10 (tab. 2.6). Jak się okazuje, błąd zlokalizowany jest na odcinkach o długości 15 jednostek i zaczyna się dla wartości χ^2 o 15 mniejszej niż zadana liczba stopni swobody.

Do ustalenia, czy omawiany błąd ma charakter lokalny, czy globalny, wyznaczono wartości formuły =ROZKŁAD.CHI (chi; ss) dla parametrów w szerokim

przedziale zmienności od 100 do 1800. W tabeli 2.7 przytoczono fragment tych obliczeń, począwszy od miejsca, gdzie po raz pierwszy stwierdzono obecność tego błędu. Jak można zauważyć, błąd w omawianej formule pojawia się przy parametrach funkcji ROZKŁAD.CHI na poziomie chi=ss=800. Początkowo błąd ten „trwa" krótko, ale stopniowo okres jego obecności systematycznie się wydłuża – co 100 o 12 jednostek. W tabeli 2.8 podano początkowe i końcowe wartości statystyki χ^2, przy których dla danej liczby stopni swobody zaczyna się wyświetlać błąd: LICZBA! W ostatniej kolumnie tej tabeli znajdują się informacje o długości odcinka liczbowego, na którym błąd ten jest obecny.

Tab. 2.7. Lokalizacja długości odcinków na których wyświetlany jest błąd LICZBA! w funkcji ROZKŁAD.CHI

Początek	Koniec/ss	Długość
797	800	3
885	900	15
973	1000	27
1060	1100	40
1148	1200	52
1236	1300	64
1323	1400	77
1411	1500	89
1499	1600	101
1586	1700	114
1674	1800	126

Źródło: opracowanie własne.

Kolejny problem, jaki pojawia się przy wyznaczaniu wartości mierników dopasowania, związany jest z kwestią zaokrągleń. Poniżej przedstawiono wyniki obliczeń wartości testu χ^2 na podstawie danych analizowanych w klasycznej pracy Benforda[5]. Rezultaty ujęte są w formie dwóch tabel.

W tabeli 2.9 wzięto pod uwagę empiryczny rozkład pierwszych cyfr znaczących wynikający z sumy wszystkich 20 zbiorów analizowanych przez Benforda. Tabela składa się z 6 modułów (A–F), w których przytoczono wyznaczane statystyki χ^2 w przypadku, gdy punktem wyjścia były rozkłady procentowe wynikające z rozkładu Benforda (B%) oraz empiryczne rozkłady cyfr znaczących (E%). Na podstawie tych udziałów procentowych wyznaczano teoretyczne $[B_i]$ oraz empiryczne $[E_i]$ liczebności *absolutne* n(i) rozkładów (łączna liczba obserwacji wynosiła tu 20 229) oraz poszczególne elementy składowe statystyki $\chi^2 - (B\text{-}E)^2/B$.

[5] F. Benford, *The Law of Anomalous Numbers*, Proceedings of the American Philosfical Society, 78/1938, p. 551–572.

Tab. 2.8. Lokalizacja „usterki" funkcji =ROZKŁAD.CHI w Excelu

Chi/ss	790	800	810	820	830	840	850	860	870	880	890	900	910	920	930	940	950	960	970
794	0,44	0,54	0,64	0,73	0,80	0,86	0,91	0,94	0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
796	0,42	0,52	0,62	0,71	0,79	0,85	0,90	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
798	0,40	#LICZBA!	0,60	0,69	0,77	0,84	0,89	0,93	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
800	0,38	0,48	0,58	0,68	0,76	0,83	0,88	0,93	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
802	0,37	0,46	0,56	0,66	0,74	0,82	0,87	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
804	0,35	0,44	0,54	0,64	0,73	0,80	0,86	0,91	0,94	0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
806	0,33	0,42	#LICZBA!	0,62	0,71	0,79	0,85	0,90	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
808	0,31	0,40	#LICZBA!	0,60	0,69	0,77	0,84	0,89	0,93	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00	1,00
810	0,29	0,39	0,48	0,58	0,68	0,76	0,83	0,88	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
812	0,28	0,37	0,46	0,56	0,66	0,74	0,81	0,87	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00	1,00
814	0,26	0,35	0,44	#LICZBA!	0,64	0,73	0,80	0,86	0,91	0,94	0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00
816	0,25	0,33	0,42	#LICZBA!	0,62	0,71	0,79	0,85	0,90	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00	1,00
818	0,23	0,31	0,41	#LICZBA!	0,60	0,69	0,77	0,84	0,89	0,93	0,96	0,97	0,99	0,99	1,00	1,00	1,00	1,00	1,00
820	0,22	0,30	0,39	0,48	0,58	0,67	0,76	0,83	0,88	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00	1,00
822	0,20	0,28	0,37	0,46	0,56	0,66	0,74	0,81	0,87	0,92	0,95	0,97	0,98	0,99	0,99	1,00	1,00	1,00	1,00
824	0,19	0,26	0,35	0,44	#LICZBA!	0,64	0,72	0,80	0,86	0,91	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00
826	0,18	0,25	0,33	0,43	#LICZBA!	0,62	0,71	0,79	0,85	0,90	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00	1,00
828	0,16	0,23	0,31	0,41	#LICZBA!	0,60	0,69	0,77	0,84	0,89	0,93	0,96	0,97	0,99	0,99	1,00	1,00	1,00	1,00
830	0,15	0,22	0,30	0,39	0,48	0,58	0,67	0,76	0,82	0,88	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00	1,00
832	0,14	0,20	0,28	0,37	0,46	#LICZBA!	0,66	0,74	0,81	0,87	0,91	0,95	0,97	0,98	0,99	0,99	1,00	1,00	1,00
834	0,13	0,19	0,26	0,35	0,44	#LICZBA!	0,64	0,72	0,80	0,86	0,91	0,94	0,96	0,98	0,99	0,99	1,00	1,00	1,00
836	0,12	0,18	0,25	0,33	0,43	#LICZBA!	0,62	0,71	0,78	0,85	0,90	0,93	0,96	0,98	0,99	0,99	1,00	1,00	1,00
838	0,11	0,16	0,23	0,31	0,41	#LICZBA!	0,60	0,69	0,77	0,84	0,89	0,93	0,95	0,97	0,98	0,99	1,00	1,00	1,00
840	0,10	0,15	0,22	0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,88	0,92	0,95	0,97	0,98	0,99	1,00	1,00	1,00
842	0,09	0,14	0,20	0,28	0,37	0,46	#LICZBA!	0,65	0,74	0,81	0,87	0,91	0,94	0,97	0,98	0,99	0,99	1,00	1,00
844	0,09	0,13	0,19	0,27	0,35	0,45	#LICZBA!	0,64	0,72	0,80	0,86	0,90	0,94	0,96	0,98	0,99	0,99	1,00	1,00
846	0,08	0,12	0,18	0,25	0,33	0,43	#LICZBA!	0,62	0,71	0,78	0,85	0,90	0,93	0,96	0,98	0,99	0,99	1,00	1,00
848	0,07	0,11	0,17	0,23	0,32	0,41	#LICZBA!	0,60	0,69	0,77	0,83	0,89	0,93	0,95	0,97	0,98	0,99	1,00	1,00
850	0,06	0,10	0,15	0,22	0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,88	0,92	0,95	0,97	0,98	0,99	0,99	1,00

852	0,06	0,09	0,14	0,21	0,28	0,37	0,46	#LICZBA!	0,65	0,74	0,81	0,87	0,91	0,94	0,97	0,98	0,99	0,99	1,00
854	0,05	0,09	0,13	0,19	0,27	0,35	0,45	#LICZBA!	0,64	0,72	0,80	0,86	0,90	0,94	0,96	0,98	0,99	0,99	1,00
856	0,05	0,08	0,12	0,18	0,25	0,33	0,43	#LICZBA!	0,62	0,70	0,78	0,84	0,89	0,93	0,96	0,97	0,99	0,99	1,00
858	0,04	0,07	0,11	0,17	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,69	0,77	0,83	0,89	0,93	0,95	0,97	0,98	0,99	1,00
860	0,04	0,07	0,10	0,16	0,22	0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,88	0,92	0,95	0,97	0,98	0,99	0,99
862	0,04	0,06	0,10	0,14	0,21	0,28	0,37	0,46	#LICZBA!	0,65	0,74	0,81	0,87	0,91	0,94	0,97	0,98	0,99	0,99
864	0,03	0,05	0,09	0,13	0,19	0,27	0,35	0,45	#LICZBA!	0,63	0,72	0,79	0,85	0,90	0,94	0,96	0,98	0,99	0,99
866	0,03	0,05	0,08	0,12	0,18	0,25	0,34	0,43	#LICZBA!	0,62	0,70	0,78	0,84	0,89	0,93	0,96	0,97	0,99	0,99
868	0,03	0,05	0,07	0,11	0,17	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,69	0,77	0,83	0,88	0,92	0,95	0,97	0,98	0,99
870	0,02	0,04	0,07	0,11	0,16	0,22	0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,87	0,92	0,95	0,97	0,98	0,99
872	0,02	0,04	0,06	0,10	0,15	0,21	0,28	0,37	0,46	#LICZBA!	0,65	0,73	0,81	0,86	0,91	0,94	0,96	0,98	0,99
874	0,02	0,03	0,06	0,09	0,14	0,20	0,27	0,35	0,45	#LICZBA!	0,63	0,72	0,79	0,85	0,90	0,94	0,96	0,98	0,99
876	0,02	0,03	0,05	0,08	0,13	0,18	0,25	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,78	0,84	0,89	0,93	0,96	0,97	0,98
878	0,01	0,03	0,05	0,07	0,12	0,17	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,69	0,76	0,83	0,88	0,92	0,95	0,97	0,98
880	0,01	0,02	0,04	0,07	0,11	0,16	0,22	0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,87	0,92	0,95	0,97	0,98
882	0,01	0,02	0,04	0,06	0,10	0,15	0,21	0,29	0,37	0,47	#LICZBA!	0,65	0,73	0,80	0,86	0,91	0,94	0,96	0,98
884	0,01	0,02	0,03	0,06	0,09	0,14	0,20	0,27	0,35	0,45	#LICZBA!	#LICZBA!	0,72	0,79	0,85	0,90	0,93	0,96	0,98
886	0,01	0,02	0,03	0,05	0,08	0,13	0,18	0,25	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,78	0,84	0,89	0,93	0,96	0,97
888	0,01	0,02	0,03	0,05	0,08	0,12	0,17	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,83	0,88	0,92	0,95	0,97
890	0,01	0,01	0,02	0,04	0,07	0,11	0,16	0,23	0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,87	0,91	0,95	0,97
892	0,01	0,01	0,02	0,04	0,06	0,10	0,15	0,21	0,29	0,37	0,47	#LICZBA!	0,65	0,73	0,80	0,86	0,91	0,94	0,96
894	0,01	0,01	0,02	0,03	0,06	0,09	0,14	0,20	0,27	0,36	0,45	#LICZBA!	#LICZBA!	0,72	0,79	0,85	0,90	0,93	0,96
896	0,00	0,01	0,02	0,03	0,05	0,08	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,78	0,84	0,89	0,93	0,95
898	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,17	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,83	0,88	0,92	0,95
900	0,00	0,01	0,01	0,03	0,04	0,07	0,11	0,16	0,23	0,30	0,39	0,48	#LICZBA!	0,67	0,75	0,82	0,87	0,91	0,94
902	0,00	0,01	0,01	0,02	0,04	0,06	0,10	0,15	0,21	0,29	0,37	0,47	#LICZBA!	#LICZBA!	0,73	0,80	0,86	0,91	0,94
904	0,00	0,01	0,01	0,02	0,04	0,06	0,09	0,14	0,20	0,27	0,36	0,45	#LICZBA!	#LICZBA!	0,72	0,79	0,85	0,90	0,93
906	0,00	0,00	0,01	0,02	0,03	0,05	0,09	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77	0,84	0,89	0,93
908	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,17	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,83	0,88	0,92
910	0,00	0,00	0,01	0,01	0,03	0,04	0,07	0,11	0,16	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,75	0,81	0,87	0,91

912	0,00	0,00	0,01	0,01	0,02	0,04	0,07	0,10	0,15	0,21	0,29	0,37	0,47	#LICZBA!	#LICZBA!	0,73	0,80	0,86	0,90
914	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,09	0,14	0,20	0,27	0,36	0,45	#LICZBA!	#LICZBA!	0,71	0,79	0,85	0,90
916	0,00	0,00	0,01	0,01	0,02	0,03	0,05	0,09	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77	0,84	0,89
918	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,18	0,24	0,32	0,41	#LICZBA!	#LICZBA!	0,68	0,76	0,82	0,88
920	0,00	0,00	0,00	0,01	0,01	0,03	0,05	0,07	0,11	0,16	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,74	0,81	0,87
922	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,07	0,10	0,15	0,22	0,29	0,37	0,47	#LICZBA!	#LICZBA!	0,73	0,80	0,86
924	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,10	0,14	0,20	0,27	0,36	0,45	#LICZBA!	#LICZBA!	0,71	0,79	0,85
926	0,00	0,00	0,00	0,01	0,01	0,02	0,03	0,06	0,09	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77	0,84
928	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,18	0,24	0,32	0,41	#LICZBA!	#LICZBA!	#LICZBA!	0,76	0,82
930	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,07	0,11	0,17	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,74	0,81
932	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,07	0,10	0,15	0,22	0,29	0,38	0,47	#LICZBA!	#LICZBA!	0,73	0,80
934	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,10	0,14	0,20	0,28	0,36	0,45	#LICZBA!	#LICZBA!	0,71	0,79
936	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,03	0,06	0,09	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	0,70	0,77
938	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,18	0,25	0,32	0,41	#LICZBA!	#LICZBA!	#LICZBA!	0,76
940	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,11	0,17	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!	0,74
942	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,03	0,04	0,07	0,11	0,16	0,22	0,29	0,38	0,47	#LICZBA!	#LICZBA!	0,73
944	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,10	0,14	0,20	0,28	0,36	0,45	#LICZBA!	#LICZBA!	0,71
946	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,09	0,13	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!	#LICZBA!
948	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,03	0,05	0,08	0,12	0,18	0,25	0,33	0,41	#LICZBA!	#LICZBA!	#LICZBA!
950	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,17	0,23	0,31	0,39	0,48	#LICZBA!	#LICZBA!
952	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,03	0,04	0,07	0,11	0,16	0,22	0,29	0,38	0,47	#LICZBA!	#LICZBA!
954	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,10	0,15	0,21	0,28	0,36	0,45	#LICZBA!	#LICZBA!
956	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,09	0,14	0,19	0,26	0,34	0,43	#LICZBA!	#LICZBA!
958	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,03	0,05	0,08	0,13	0,18	0,25	0,33	0,41	#LICZBA!	#LICZBA!
960	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,02	0,03	0,05	0,08	0,12	0,17	0,23	0,31	0,39	0,48	#LICZBA!
962	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,03	0,04	0,07	0,11	0,16	0,22	0,29	0,38	0,47	#LICZBA!
964	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,07	0,10	0,15	0,21	0,28	0,36	0,45	#LICZBA!
966	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,04	0,06	0,09	0,14	0,19	0,26	0,34	0,43	#LICZBA!
968	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,01	0,02	0,03	0,05	0,09	0,13	0,18	0,25	0,33	0,41	#LICZBA!

Źródło: opracowanie własne.

W kolejnych modułach tabeli 2.9 przytoczono rozkłady procentowe w postaci:

A. wyjściowej (zgodnej z danymi przytoczonymi w pracy Benforda),
B. zaokrąglonej do najbliższej liczby całkowitej przy pomocy funkcji =ZAOKR(x;0),
C. zaokrąglonej do najbliższej liczby całkowitej (jak w module B) oraz 2 korektach mających na celu doprowadzenie sumy liczby obserwacji do 100%,
D. zaokrąglonej do najbliższej liczby całkowitej (jak w module B) oraz 2 innych korektach mających na celu doprowadzenie sumy liczby obserwacji do 100%,
E. zaokrąglonej *w dół* do najbliższej liczby całkowitej przy pomocy funkcji =ZAOKR.DO.CAŁK(x),
F. zaokrąglonej do najbliższej liczby całkowitej (jak w module E) oraz 9 korektach mających na celu doprowadzenie sumy liczby obserwacji do 100%.

Jak wynika z podanego przykładu zaokrąglanie elementów rozkładu powoduje wyraźne zmiany w statystyce χ^2. Poprawna, wyjściowa wartość tej statystyki wynosi tu 85,2, natomiast wszystkie pozostałe statystyki są większe (nawet trzykrotnie) i zawierają się w granicach od 117,7 do 279,3.

W kolejnym eksperymencie założono, że punktem wyjścia są *identyczne* rozkłady procentowe, jak w poprzednim przykładzie, ale z tą różnicą, że były one wyznaczane na podstawie nie 20 229 lecz tylko 5000 obserwacji. Wyniki analizy (w analogicznym układzie, jak poprzednio) zawiera tabela 2.10. Jak można zauważyć, wartości statystyki χ^2 w każdym przypadku zmniejszyły się, przy czym podobnie, jak poprzednio, najmniejsze wartości uzyskano dla danych wyjściowych (χ^2 = 2,3), a dla wszystkich pozostałych zbiorów danych statystyki χ^2 kształtowały się na wyższym poziomie, w przedziale od 29,1 do 69,0.

Przy okazji warto zauważyć, że wartości statystyki χ^2 zależą od ogólnej liczby obserwacji. W omawianym przykładzie statystyki te w tabeli 2.9 są czterokrotnie wyższe niż w tabeli 2.10, gdyż liczba obserwacji w tabeli 2.9 (20 229) jest czterokrotnie większa niż w tabeli 2.10 (5000).

Natomiast wartość krytyczna testu χ^2 nie zależy od liczby obserwacji n, lecz od poziomu istotności α oraz liczby przedziałów *k* porównywanych rozkładów, czyli liczby stopni swobody ss=k−1. Oznacza to, że im wyższa jest liczba obserwacji, tym większy musi być stopień podobieństwa rozkładów, aby można było je uznać za identyczne przy danym poziomie istotności.

Wzór pozwalający ustalić liczbę obserwacji, dla której można byłoby uznać, że porównywane rozkłady nie różnią się, przy wyjściowej wartości statystyki χ^2 oraz zadanym poziomie istotności α ma postać:

(2.8)
$$n' = \frac{n\chi_{\alpha}^2}{\chi^2}$$

gdzie n to rzeczywista liczba obserwacji, natomiast χ^2_{α} to wartość krytyczna testu chi-kwadrat ustalona przy ss=k–1 stopniach swobody i poziomie istotności α.

Jeżeli w omawianym przykładzie:
- n=20229,
- wartość empiryczna testu chi-kwadrat χ^2=85,2,
- liczba stopni swobody k-1=9-1=8,
- poziom istotności α=0,05,
- wartość krytyczna testu chi-kwadrat χ^2_{α}=15,5,

to liczba obserwacji przy której można byłoby uznać, że porównywane rozkłady nie różnią się od siebie na zadanym poziomie istotności α wynosi:

$$n' = \frac{20229 * 15{,}5}{85{,}2} = 3680$$

Zależność pomiędzy liczbą obserwacji a wartością statystyki χ^2 może być niekiedy powodem błędnych wniosków. Poniżej podano przykład, w którym wyznaczono rozkład pierwszych cyfr znaczących dla liczb będących silnią kolejnych liczb naturalnych od 1 do 170.

Dla n=170 wartość 170! wynosi 7,257415615308E+306 i jest to granica dokładności, jaką można uzyskać w Excelu. Na podstawie zbioru wartości tych 170 silni wyznaczono rozkłady pierwszych cyfr znaczących i obliczono wartość testu chi-kwadrat, przy czym analizę wykonano w 4 wariantach – dla zbiorów będących wielokrotnością (3-, 4-, 10- oraz 20-krotność) zbioru źródłowego.

Jak wynika z tabeli 2.11 duże zbiory danych (3400- i 1700-elementowe) w świetle testu chi-kwadrat należy uznać za niezgodne z rozkładem Benforda. Natomiast w przypadku małych zbiorów (340- i 510-elementowych) można przyjąć hipotezę o zgodności rozkładów pierwszych cyfr znaczących wartości n! z rozkładem Benforda – w pierwszym przypadku na poziomie istotności 0,003, natomiast w drugim – na poziomie istotności 0,047.

Tab. 2.9. Efekt zaokrągleń przy wyznaczaniu statystyki χ^2 na podstawie danych Benforda, dla n=20229

Źródło: opracowanie własne.

(A) Lp.	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,1	28,9	6090	5846	9,7
2	17,6	19,5	3562	3945	41,1
3	12,5	12,7	2527	2569	0,7
4	9,7	9,1	1960	1841	7,3
5	7,9	7,5	1602	1517	4,5
6	6,7	6,4	1354	1295	2,6
7	5,8	5,4	1173	1092	5,6
8	5,1	5,5	1035	1113	5,9
9	4,6	5,0	926	1011	8,0
Suma	100,0	100,0	20229	20229	85,2

=ZAOKR(x;0)

(B) Lp.	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2	18,0	20,0	3641	4046	45,0
3	12,0	13,0	2427	2630	16,9
4	10,0	9,0	2023	1821	20,2
5	8,0	8,0	1618	1618	
6	7,0	6,0	1416	1214	28,9
7	6,0	5,0	1214	1011	33,7
8	5,0	6,0	1011	1214	40,5
9	5,0	5,0	1011	1011	
Suma	101,0	101,0	20431	20431	191,9

=ZAOKR.DO.CAŁK(x)

(E) Lp.	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	28,0	6069	5664	27,0
2	17,0	19,0	3439	3844	47,6
3	12,0	12,0	2427	2427	
4	9,0	9,0	1821	1821	
5	7,0	7,0	1416	1416	
6	6,0	6,0	1214	1214	
7	5,0	5,0	1011	1011	
8	5,0	5,0	1011	1011	
9	4,0	5,0	809	1011	50,6
Suma	95,0	96,0	19218	19420	125,1

(C) Lp.	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2	18,0	20,0	3641	4046	45,0
3	12,0	13,0	2427	2630	16,9
4	10,0	9,0	2023	1821	20,2
5	8,0	8,0	1618	1618	
6	7,0	6,0	1416	1214	28,9
7	5,0	5,0	1011	1011	
8	5,0		1011	1011	
9	5,0	5,0	1011	1011	
Suma	100,0	100,0	20229	20229	117,7

=ZAOKR(x;0) + 2 korekty

(D) Lp.	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2		20,0	3439	4046	107,1
3	12,0	13,0	2427	2630	16,9
4	10,0	9,0	2023	1821	20,2
5	8,0	7,0	1618	1416	25,3
6	7,0	6,0	1416	1214	28,9
7	6,0	5,0	1214	1011	33,7
8.	5,0	6,0	1011	1214	40,5
9	5,0	5,0	1011	1011	
Suma	100,0	100,0	20229	20229	279,3

=ZAOKR.DO.CAŁK(x) + 9 korekt

(F) Lp.	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	6069	5866	6,7
2	18,0	20,0	3641	4046	45,0
3	13,0	13,0	2630	2630	
4	10,0	9,0	2023	1821	20,2
5	8,0		1618	1618	
6	7,0	6,0	1416	1214	28,9
7	5,0	5,0	1011	1011	
8	5,0	5,0	1011	1011	
9	4,0	5,0	809	1011	50,6
Suma	100,0	100,0	20229	20229	151,4

Tab. 2.10. Efekt zaokrągleń przy wyznaczaniu statystyki χ^2 na podstawie danych Benforda dla n=5000

(A) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,1	28,9	1505	1445	2,4
2	17,6	19,5	880	975	10,2
3	12,5	12,7	625	635	0,2
4	9,7	9,1	485	455	1,8
5	7,9	7,5	396	375	1,1
6	6,7	6,4	335	320	0,6
7	5,8	5,4	290	270	1,4
8	5,1	5,5	256	275	1,4
9	4,6	5,0	229	250	2,0
Suma	100,0	100,0	5000	5000	21,1

=ZAOKR(x;0)

(C) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	18,0	20,0	900	1000	11,1
3	12,0	13,0	600	650	4,2
4	10,0	9,0	500	450	5,0
5	8,0	8,0	400	400	0,0
6	7,0	6,0	350	300	7,1
7	5,0	5,0	250	250	0,0
8	5,0	5,0	250	250	0,0
9	5,0	5,0	250	250	0,0
Suma	100,0	100,0	5000	5000	29,1

=ZAOKR(x;0) + 2 korekty

(B) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	18,0	20,0	900	1000	11,1
3	12,0	13,0	600	650	4,2
4	10,0	9,0	500	450	5,0
5	8,0	8,0	400	400	0,0
6	7,0	6,0	350	300	7,1
7	6,0	5,0	300	250	8,3
8	5,0	6,0	250	300	10,0
9	5,0	5,0	250	250	0,0
Suma	101,0	101,0	5050	5050	47,4

=ZAOKR.DO.CAŁK(x)

(D) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	17,0	20,0	850	1000	26,5
3	12,0	13,0	600	650	4,2
4	10,0	9,0	500	450	5,0
5	8,0	7,0	400	350	6,3
6	7,0	6,0	350	300	7,1
7	6,0	5,0	300	250	8,3
8	5,0	6,0	250	300	10,0
9	5,0	5,0	250	250	0,0
Suma	100,0	100,0	5000	5000	69,0

=ZAOKR.DO.CAŁK(x) + 9 korekt

(E) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	28,0	1500	1400	6,7
2	17,0	19,0	850	950	11,8
3	12,0	12,0	600	600	0,0
4	9,0	9,0	450	450	0,0
5	7,0	7,0	350	350	0,0
6	6,0	6,0	300	300	0,0
7	5,0	5,0	250	250	0,0
8	5,0	5,0	250	250	0,0
9	4,0	5,0	200	250	12,5
Suma	95,0	96,0	4750	4800	30,9

(F) Lp	B%	E%	B n(i)	E n(i)	(B-E)^2/B
1	30,0	29,0	1500	1450	1,7
2	18,0	20,0	900	1000	11,1
3	13,0	13,0	650	650	0,0
4	10,0	9,0	500	450	5,0
5	8,0	8,0	400	400	0,0
6	7,0	6,0	350	300	7,1
7	5,0	5,0	250	250	0,0
8	5,0	5,0	250	250	0,0
9	4,0	5,0	200	250	12,5
Suma	100,0	100,0	5000	5000	37,4

Źródło: opracowanie własne.

Tab. 2.11. Test χ^2 dla wartości n! w zależności od wielkości zbioru danych

d	E	B	(E-B)^2/B	E	B	(E-B)^2/B
1	1080	1024	3,1	540	512	1,6
2	580	599	0,6	290	299	0,3
3	440	425	0,5	220	212	0,3
4	240	329	24,3	120	165	12,2
5	240	269	3,2	120	135	1,6
6	200	228	3,4	100	114	1,7
7	120	197	30,2	60	99	15,1
8	280	174	64,7	140	87	32,4
9	220	156	26,7	110	78	13,3
SUMA	3400	3400	156,7	1700	1700	78,3
	x20	x20	0,000000	x10	x10	0,000000

d	E	B	(E-B)^2/B	E	B	(E-B)^2/B
1	162	154	0,5	108	102	0,3
2	87	90	0,1	58	60	0,1
3	66	64	0,1	44	42	0,1
4	36	49	3,6	24	33	2,4
5	36	40	0,5	24	27	0,3
6	30	34	0,5	20	23	0,3
7	18	30	4,5	12	20	3,0
8	42	26	9,7	28	17	6,5
9	33	23	4,00	22	16	2,7
SUMA	510	510	23,5	340	340	15,7
	x3	x3	0,003	x2	x2	0,047

Źródło: opracowanie własne.

Poza testem chi-kwadrat w analizach często korzysta się z testu *z*, w którym dla każdej wartości występującej w rozkładzie wyznacza się statystyki:

(2.9)
$$z_i = \frac{p_i - \hat{p}_i}{\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}} \qquad (i = 1,...,k)$$

Jeżeli wartość z_i co do modułu jest większa od wartości krytycznej tego testu, to należy odrzucić hipotezę o zgodności rozkładu empirycznego z rozkładem Benforda dla *i*-tej cyfry (lub *i*-tej kombinacji cyfr). Wartość krytyczna tego testu dla α=0,05 wynosi 1,96, natomiast dla α=0,01 odpowiednia wartość krytyczna wynosi 2,58.

Podobnie, jak test chi-kwadrat, także i test *z* jest zależny od rozmiarów analizowanego zbioru danych. Im większy jest parametr *n,* tym trudniej jest uzyskać zgodność porównywanych udziałów.

Innym testem służącym do oceny zgodności rozkładów jest test Kołmogorowa–Smirnowa. Jest to test niezależny od wielkości zbioru *n*. W jego wersji oryginalnej wyznacza się statystykę:

(2.10) $$KS1 = D\sqrt{\frac{n^2}{2n}} \qquad D = \max_i \left| f_i - \hat{f}_i \right| \qquad (i = 1,...,k)$$

W literaturze proponuje się także inne wersje tego testu. W jednej z nich wyznacza się statystykę:

(2.11) $$KS2 = D\sqrt{n} \qquad D = \max_i \left| f_i - \hat{f}_i \right| \qquad (i = 1,...,k)$$

Inna modyfikacja testu Kołmogorowa–Smirnowa zaproponowana została przez Kuipera[6]. Uwzględnia ona fakt *cykliczności* analizowanych rozkładów (*circular distrubution*). Chodzi o to, że różnica pomiędzy liczbami 99,99 a 100,01 jest minimalna, podczas gdy w sensie analizy rozkładu pierwszych cyfr znaczących odpowiadające im liczby 9 i 1 znajdują się na przeciwnych biegunach skali. W związku z powyższym proponuje się wykorzystywać statystykę:

(2.12) $$KS3 = V_N * [\sqrt{N} + 0{,}155 + 0{,}24N^{-1/2}] \qquad gdzie \qquad (i = 1,..., k)$$
$$V_N = D_N^+ + D_N^- \quad D_N^+ = \sup_i \left[f_i - \hat{f}_i \right] \quad D_N^- = \sup_i \left[\hat{f}_i - f_i \right] \quad N = \frac{n^2}{2n}$$

Wartości krytyczne testów KS1–KS2 wynoszą 1,36 dla α=0,05 oraz 1,63 dla α=0,01. W przypadku zmodyfikowanego testu Kołomogorowa–Smirnowa KS3 wartości krytyczne wynoszą odpowiednio: 1,747 dla α=0,05 oraz 2,001 dla α=0,01.

[6] N.H. Kuiper, *Alternative proof of a theorem of Birnhaum and Pyke*, Annals of Mathematical Statistics 30/1959, p. 251–252.

2.3. Mierniki zgodności empirycznych rozkładów cyfr z rozkładami wynikającymi z prawa Benforda

Alternatywnym sposobem pomiaru zgodności dwóch rozkładów są miary ich podobieństwa. Można tu wymienić następujące mierniki.

(2.13) $$M_1 = \frac{100}{k}\sum_{i=1}^{k}\left|\frac{c_i - \hat{c}_i}{\hat{c}_i}\right|$$

(2.14) $$M_2 = \frac{1}{k}\sqrt{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}$$

(2.15) $$M_3 = \sqrt{\frac{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}{k}}$$

(2.16) $$M_4 = \frac{100\sqrt{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}}{\sqrt{\sum_{i=1}^{k}\hat{c}_i^2}}$$

(2.17) $$M_5 = \frac{100\sum_{i=1}^{k}|n_i - \hat{n}_i|}{n}$$

Mierniki te są niezależne od wielkości zbioru *n* i przyjmują tym mniejsze wartości, im bardziej zgodne ze sobą są porównywane rozkłady częstości. Generalnie wskazują one na przeciętną wielkość różnic pomiędzy częstościami rzeczywistymi a częstościami teoretycznymi w danym teście[7].

Mierniki M2 oraz M3 wskazują na przeciętną różnicę pomiędzy częstościami empirycznymi a teoretycznymi porównywanych rozkładów. W literaturze preferuje się miernik M3, jakkolwiek bardziej naturalny wydaje się miernik M2, gdyż jego interpretacja jest bardziej zbliżona do przeciętnej wielkości różnicy pomiędzy empirycznymi i teoretycznymi częstościami. Miernik M3 jest większy od miernika M2 (w przypadku testów F1, D2, D3,

[7] Można sformułować analogiczne formuły, w których zamiast częstości c_i występują liczebności n_i- z identyczną interpretacją i wartościami.

L1 – trzykrotnie większy) i szacuje z dużym nadmiarem wielkość różnic pomiędzy porównywanymi rozkładami.

Mierniki M1 oraz M4 wskazują, jaka jest przeciętna różnica między częstościami empirycznymi a teoretycznymi w relacji do częstości teoretycznych rozkładu. Miernik M5 pokazuje, jaką część wszystkich obserwacji trzeba by zamienić miejscami, aby rozkłady empiryczne pokryły się z rozkładami teoretycznymi. Wielkości tych trzech mierników (M1, M4, M5) wyrażone są w procentach.

Kolejnym miernikiem zgodności może być współczynnik korelacji linowej *r* pomiędzy empirycznymi i teoretycznymi liczebnościami rozkładu:

(2.18)
$$r = \frac{\sum_{i=1}^{k}\left(n_i - \overline{n}\right)\left(\hat{n}_i - \overline{\hat{n}}\right)}{\sqrt{\sum_{i=1}^{k}\left(n_i - \overline{n}\right)^2 \sum_{i=1}^{k}\left(\hat{n}_i - \overline{\hat{n}}\right)^2}}$$

W powyższym wzorze $\overline{n}$ oraz $\overline{\hat{n}}$ to średnie arytmetyczne z empirycznych i teoretycznych liczebności. Identyczne wartości współczynników korelacji uzyskuje się, jeżeli we wzorze 2.18 przyjmie się nie liczebności rozkładów, ale ich częstości lub prawdopodobieństwa. Współczynnika *r* nie można wyznaczyć dla testu L1, a także D3, gdyż częstości teoretyczne w tych przypadkach są identyczne i nie mają żadnej zmienności.

W tabeli 2.12 podano rozkłady pierwszych cyfr znaczących w zbiorach analizowanych przez F. Benforda wraz z wyznaczonymi na ich podstawie miernikami M1–M5, statystykami χ^2, statystykami testu Kolmogorowa–Smirnowa (KS1, KS2, KS3) oraz parametrami wynikającymi z wartości testu **z.** W tym ostatnim przypadku jest to średnia z modułów statystyk **z** (ostatnia kolumna tabeli 2.12) oraz liczba cyfr, dla których statystyki **z** wskazują na istotną rozbieżność pomiędzy porównywanymi częstościami rozkładów. Liczba ta może przyjmować wartości od 0 do 9 i jest ustalana w czterech wariantach różniących się poziomem istotności, przy którym należy uznać, że porównywane częstości są istotnie różne. Przyjęto następujące progi wartości krytycznych testu **z**:

- z>1,64, dla poziomu istotności α=0,1
- z>1,96, dla poziomu istotności α=0,05
- z>2,58, dla poziomu istotności α=0,01
- z>3,29, dla poziomu istotności α=0,001

Ponadto wyznaczono współczynniki korelacji pomiędzy empirycznymi a teoretycznymi składowymi rozkładu pierwszych cyfrowych wraz z prawdopodobieństwem, przy którym należy odrzucić hipotezę o braku istotnego skorelowania. Ze względu na ujemne skorelowania z pozostałymi miernikami tych dwóch ostatnich parametrów zostały one zastąpione dopełnieniami do jedności ich modułów: *1–mod(r) 1–p(r)*. Dzięki tej operacji przy interpretacji wszystkich mierników dopasowania i testów można przyjąć zasadę, że im mniejsze są wartości tych parametrów, tym lepiej dany rozkład empiryczny jest dopasowany do rozkładu Benforda.

Ta zasada nie dotyczy liczebności zbioru danych (pierwsza kolumna tabeli 2.12). Parametr *n* uwzględniono dla sprawdzenia, czy istnieje związek pomiędzy wielkością zbioru danych a miernikami charakteryzującymi stopień dopasowania rozkładów.

Zbiory uporządkowane są według rosnących wartości testu χ^2, od zbiorów najbardziej do najmniej zgodnych z rozkładem Benforda. Jak można zauważyć tylko część (12 na 22) zbiorów ma rozkłady cyfr znaczących zgodne z prawem Benforda w sensie statystyki chi-kwadrat – są to zbiory, dla których wartość empiryczna statystyki chi-kwadrat jest mniejsza od 15,5, przy założeniu 5% poziomu istotności. 5 zbiorów wymienionych na końcu tabeli 2.12 (jest wśród nich zbiorowość sumaryczna) ewidentnie mają rozkłady niezgodne z prawem Benforda. W przypadku pozostałych 5 zbiorów wymienionych w środku tabeli 2.12 można przyjąć założenie o zgodności rozkładu pierwszych cyfr znaczących z rozkładem Benforda, pod warunkiem obniżenia poziomu istotności do $\alpha = 0{,}01$ (3 zbiory) lub $\alpha = 0{,}001$ (2 zbiory).

W tabeli 2.13 podano rangi zbiorów danych odnoszące się do wartości mierników ich dopasowania do rozkładu Benforda (1 – zbiór najlepiej dopasowany, 22 – zbiór najgorzej dopasowany). Zbiory danych uporządkowano według średniej rangi ze względu na wszystkie analizowane mierniki (ostatnia kolumna w tab. 2.13). Stworzenie takiej syntetycznej miary dopasowania zbiorów ze względu na wszystkie mierniki zawarte w poprzedniej tabeli (tab. 2.12) było niemożliwe z uwagi na różne znaczenie i różny zakres wartości poszczególnych mierników.

Porównując uporządkowanie zbiorów danych z tabeli 2.12 (wg statystyki χ^2) z ich uporządkowaniem w tabeli 2.13 (ze względu na wszystkie mierniki) można stwierdzić nieznaczne różnice. Sprowadzają się one do przesunięcia kilku zbiorów danych o jedną (rzadziej o dwie) pozycje w górę lub w dół. Tak więc można przyjąć, że test χ^2 jest miarodajnym miernikiem poprawności wnioskowania o podobieństwie rozkładów cyfr znaczących.

W tabeli 2.14 przytoczono współczynniki korelacji pomiędzy poszczególnymi miernikami dopasowania. W ostatnich wierszach tej tabeli podano liczbę dodatnich oraz ujemnych współczynników korelacji oraz średnią z modułów tych współczynników dla każdego miernika dopasowania. Jak można zauważyć, za wyjątkiem kilku współczynników korelacji pomiędzy liczebnością zbiorów *n* a miernikami M1–M5 i współczynnikiem *1–r*, wszystkie pozostałe współczynniki korelacji są dodatnie. Nawet te ujemne współczynniki korelacji są niewielkie i statystycznie nieistotne (dla α=0,05 oraz przy n=20 wartość krytyczna modułu współczynnika korelacji wynosi 0,47). Oznacza to, że wszystkie mierniki dopasowania mają identyczne zasady interpretacji – im większe przyjmują wartości, tym dany rozkład jest mniej podobny do rozkładu Benforda.

Warto zwrócić uwagę na dużą zgodność wskazań mierników w ramach ich grup definicyjnych. I tak, mierniki M2–M3–M4 dają identyczne oceny stopnia zgodności (przy różnych wartościach mierników). Wynika stąd, że w analizach wystarczy skorzystać z dowolnego miernika z tej grupy. Podobne rezultaty jak mierniki M2–M4 (r=0,98), daje miernik M5, a nieco bardziej odmienne (r=0,92) – miernik M1.

Również identyczne wnioski uzyskuje się w przypadku statystyk Kołmogorowa–Smirnowa KS1–KS2. Zgodność wskazań trzeciego testu KS3 z dwoma poprzednimi jest też wysoka (r=0,95).

Miary oparte na statystyce ***z*** cechuje duży, jakkolwiek mniejszy niż w poprzednich przypadkach, stopień zgodności wskazań. Interesujący jest współczynnik korelacji pomiędzy testem χ^2 a poziomem istotności odpowiadającym tej statystyce. Jest on zadziwiająco niski (r=0,45). Można nawet powiedzieć, że te dwie miary są ze sobą nieskorelowane.

Biorąc pod uwagę sumaryczny stopień skorelowania danego miernika z pozostałymi (ostatni wiersz w tab. 2.14), najbardziej diagnostyczne w sensie zgodności wskazań są statystyki Kołmogorowa–Smirnowa, zwłaszcza statystyka KS3, a w następnej kolejności mierniki oparte na statystyce ***z***.

W tabeli 2.15 przedstawiono wyniki podziału współczynników korelacji na cztery kategorie ze względu na ich poziom. Kryterium podziału stanowiły wartości krytyczne wynikające z testu Studenta dla różnych poziomów istotności[8]:

- Kategoria 3 – poziom wysoki r>0,8, α1= α2=0,00001,
- Kategoria 2 – poziom średni (0,5<r<0,8), α1=0,009, α2=0,019,
- Kategoria 1 – poziom niski (0,35<r<0,5),α1=0,056, α2=0,112,
- Kategoria 0 – brak korelacji r<0,35.

[8] Podano tu wartości krytyczne testu Studenta zarówno dla jednostronnego α1 jak i dwustronnego α2 obszaru krytycznego. Z uwagi na dodatniość wszystkich współczynników korelacji bardziej właściwy wydaje się jednostronny obszar krytyczny. Wartości rozkładu Studenta wyznaczane są przy pomocy funkcji Excela ROZKŁAD.T (t; ss; 1) gdzie $t=[r/(1-r^2)^{1/2}]*ss^{1/2}$ oraz ss=n-2, gdzie n to liczba obserwacji (w tym przypadku n=22).

W ostatniej kolumnie tabeli podano średnie wartości rang przypisanych poszczególnym kategoriom. Stanowią one syntetyczną miarę zgodności danego miernika z pozostałymi. Kolumny i wiersze macierzy współczynników uporządkowane zostały wg malejących wartości tej średniej. Jak się okazuje, wśród mierników dających wskazania najbardziej podobne do wskazań innych należą dwa mierniki oparte na statystyce *z* (*z>1,64* oraz *z>1,96*) i statystyka Kołmogorowa–Smirnowa KS3. Z drugiej strony znajdują się mierniki o najmniejszej zgodności z pozostałymi: *–r(p), M2, M3, M4,* a także statystyka χ^2.

Podobną analizę można przeprowadzić wykorzystując nie średnie wartości rang. lecz rangi danego typu, np. tylko kategorii „3". W ostatnich wierszach tabeli 2.15 podano liczbę mierników zaliczonych do poszczególnych kategorii. Najwięcej kategorii „3" jest przypisanych do pięciu mierników: *KS1–KS2–KS3, z sred* oraz *z>2,58*. Jest to więc nieco inny zestaw mierników, niż wynikało to z ich uporządkowania według średnich wartości rang.

Tabela 2.16 ma analogiczną konstrukcję jak tabela 2.15 z tym, że zawiera informacje uzyskane nie na podstawie macierzy współczynników korelacji liniowej, lecz macierzy współczynników korelacji rang Spearmana, obrazujących stopień zgodności pomiędzy poszczególnymi miernikami opisujących podobieństwo rozkładów cyfr znaczących w zbiorach Benforda z prawem Benforda.

Przyjęto tu nieco wyższe wartości graniczne określające poszczególne kategorie współczynników. Zamiast [0,35–0,5–0,8] są to wartości [0,5–0,75–0,9]. Wynikało to z wyższego poziomu wartości współczynników korelacji Spearmana niż współczynników korelacji liniowej Pearsona. Pomimo podwyższonych wartości granicznych, średni poziom zgodności wszystkich mierników zawartych w macierzy (prawy dolny narożnik tabel) w przypadku współczynników Spearmana i tak jest wyższy (1,81) niż dla współczynników Pearsona (1,67).

Analiza parametrów zawartych w tabeli 2.16 tylko częściowo potwierdza poprzednio uzyskane rezultaty. Do „zgodnych" mierników nadal zalicza się statystyki KS1–KS2–KS3, ale także statystykę χ^2, która poprzednio była zaliczona do grupy mierników dających odmienne wskazania niż pozostałe. Wśród parametrów o niskiej zgodności, analiza współczynników Spearmana nie wykazuje mierników M2–M4, które to mierniki wynikały z analizy współczynników Pearsona.

Reasumując, problem oceny diagnostyczności testów i mierników charakteryzujących podobieństwo rozkładów nie jest prosty i wymaga dużej rozwagi oraz dalszych badań. Badania te powinny mieć charakter symulacyjny i opierać się na danych generowanych z kontrolowanym stopniem podobieństwa rozkładów.

Macierz mierników podobieństwa rozkładów (tab. 2.12), którą w ogólnej postaci można zapisać:

(2.19) $$[X_{ij}] \quad (i = 1,2 \ldots, p; \quad j = 1,2, \ldots q)$$

gdzie *p* to liczba analizowanych zbiorów danych, zaś *q* – liczba mierników podobieństwa (w przypadku analizy zbiorów Benforda p=20, q=18), może stanowić punkt wyjścia analizy taksonometrycznej[9]. Jej celem jest klasyfikacja zbioru obiektów (dane Benforda lub mierniki podobieństwa) na bardziej jednorodne podzbiory. Chodzi o taki podział, aby elementy danego podzbioru były jak najbardziej podobne do siebie z punktu widzenia opisujących je charakterystyk, a jednocześnie jak najmniej podobne do obiektów tworzących inne podzbiory.

Klasyfikacja pozwala wprowadzić porządek do analizowanych zjawisk, tym samym ułatwia i sprzyja wyciąganiu poprawnych wniosków. Dla przykładu, jeżeli przeprowadzimy klasyfikację 20 mierników podobieństwa i okaże się, że dają się one podzielić np. na 3 grupy (jednorodne podzbiory), to w analizie można nie uwzględniać wszystkich 20 mierników, a tylko 3, będące reprezentantami grup, do których te mierniki należą.

Klasyfikacji mogą podlegać obiekty znajdujące się w wierszach macierzy danych, które opisane są charakterystykami (cechami) ujętymi w kolumnach macierzy 2.18, względnie cechy traktowane jako punkty w wielowymiarowej przestrzeni obiektów (tzw. zadanie dualne). W literaturze znanych jest wiele metod klasyfikacji. W niniejszej pracy wykorzystano najstarszą metodę taksonometryczną opracowaną przez polskiego antropologa Jana Czekanowskiego. Program obliczeniowy dostępny jest w witrynie P. Jaskulskiego Archeo-Data poświęconej problematyce wykorzystania komputerów w antropologii i archeologii[10].

Rezultaty analizy taksonometrycznej przedstawiono w postaci tzw. diagramów Czekanowskiego (rys. 2.5–2.8 – klasyfikacja zbiorów Benforda w przestrzeni miar zgodności rozkładów oraz rys. 2.9–2.12 – klasyfikacja miar zgodności rozkładów w przestrzeni zbiorów Benforda). W diagramach Czekanowskiego symbolami graficznymi o coraz to większym stopniu zaczernienia przedstawia się poziomy miar podobieństwa klasyfikowanych obiektów (zbiorów Benforda, miar zgodności rozkładów). Im bardziej zaczerniony jest element diagramu, tym bardziej podobne są do siebie obiekty przyporządkowane temu elementowi.

---

[9] Przegląd metod taksonometrii znaleźć można m.in. w pracach: T. Grabiński, *Metody taksonometrii*, Akademia Ekonomiczna w Krakowie, Kraków 1991; T. Grabiński, *Analiza taksonometryczna krajów Europy w ujęciu regionów*, Akademia Ekonomiczna w Krakowie, Kraków 2003.

[10] http://eskimo73.republika.pl/maczek.html oraz eskimo73.republika.pl/download/manual_30_pl.pdf.

Tab. 2.12. Miary dopasowania dla 20 zbiorów analizowanych przez F. Benforda

Symb.	Nazwa	n	1-r	r(p)	M5	M1	M2	M3	M4	chi	1-chi(p)	KS1	KS2	KS3	z>1,64	z>1,96	z>2,58	z>3,29	z sred
D	Czasopisma	100	0,001	0,000	2,8	4,0	0,12	0,37	2,7	0,16	0,000	0,04	0,05	0,06					0,11
F	Ciśnienie	703	0,001	0,000	3,2	4,1	0,14	0,42	3,1	1,27	0,004	0,18	0,26	0,28					0,33
R	Adresy	342	0,005	0,000	5,4	5,5	0,26	0,78	5,7	1,30	0,004	0,16	0,22	0,22					0,35
	Średnia	*1011*	*0,001*	*0,000*	*3,2*	*3,9*	*0,15*	*0,44*	*3,3*	*1,75*	*0,012*	*0,31*	*0,44*	*0,34*					*0,38*
M	Dane Reader's Digest	308	0,004	0,000	8,4	7,8	0,46	1,39	10,3	3,23	0,081	0,52	0,73	0,53					0,49
G	Wiatr	690	0,003	0,000	4,8	6,8	0,22	0,65	4,8	3,46	0,098	0,31	0,44	0,34					0,51
A	Dorzecza rzek	335	0,012	0,000	9,9	12,5	0,41	1,22	9,0	4,96	0,238	0,27	0,39	0,39					0,70
O	Promienie X	707	0,007	0,000	7,4	8,2	0,36	1,08	8,0	5,43	0,289	0,43	0,61	0,44					0,70
T	Śmiertelność	418	0,022	0,000	11,2	11,5	0,55	1,66	12,2	7,55	0,522	0,45	0,63	0,61	1	1			0,80
Q	Prom. ciała czarnego	1165	0,006	0,000	7,3	9,5	0,30	0,91	6,7	9,52	0,700	0,53	0,75	0,73	2				0,97
I	Drenaż	159	0,058	0,000	21,6	26,5	0,97	2,91	21,5	11,14	0,806	0,69	0,98	0,98	1	1			1,02
P	Amer. liga baseball	1458	0,002	0,000	6,6	9,0	0,36	1,09	8,0	14,60	0,932	0,76	1,07	0,76	2	2	1		0,98
N	Koszty	741	0,011	0,000	12,3	15,3	0,51	1,53	11,3	15,60	0,952	0,67	0,95	0,68	3	1			1,27
J	Masa atomowa	91	0,048	0,000	38,2	33,5	2,18	6,54	48,2	17,25	0,972	1,23	1,74	1,33	3	2	1	1	1,18
L	Dane z projektów	560	0,021	0,000	16,6	20,0	0,68	2,03	15,0	19,21	0,986	1,09	1,54	1,10	4	1			1,46
C	Stałe	104	0,095	0,001	39,8	48,1	1,82	5,46	40,3	24,44	0,998	0,81	1,14	1,27	4	4	1		1,50
S	n^1, n^2, ... n!	900	0,003	0,000	13,8	16,4	0,67	2,02	14,9	24,99	0,998	1,46	2,07	1,48	3	3	2		1,52
	Suma	20229	0,005	0,000	5,7	6,2	0,27	0,82	6,0	84,10	1,000	1,19	1,69	2,03	7	7	4	2	2,87
E	Ciepło	1389	0,093	0,001	24,2	26,6	1,09	3,28	24,2	111,21	1,000	1,61	2,27	3,21	6	6	6	6	3,24
B	Populacja	3259	0,004	0,000	16,6	20,4	0,69	2,08	15,3	118,63	1,000	3,35	4,73	3,36	7	7	7	5	3,54
H	Masa cząsteczkowa	1800	0,068	0,000	23,2	25,9	1,07	3,22	23,7	125,76	1,000	2,46	3,48	3,50	8	7	7	3	3,54
K	n^(−1), n^(0,5)	5000	0,066	0,000	22,8	30,6	0,95	2,86	21,1	440,76	1,000	4,36	6,16	4,37	8	8	8	8	6,28
α	Wart. teoret. chi kw.	0,05	**15,507**	0,01	**20,09**	0,001	**26,124**												

Źródło: opracowanie własne na podstawie pracy: F. Benford, The Law of Anomalous Numbers, Proceedings of the Amercian Philosophical Society, 78/1938, p. 551–572.

Tab. 2.13. Rangi zbiorów Benforda według mierników zgodności rozkładów cyfr znaczących

Lp.	Symb	Nazwa	n	1-r	r(p)	M5	M1	M2	M3	M4	chi	1-ch(p)	KS1	KS2	KS3	z>1,64	z>1,96	z>2,58	z>3,29	z sred	R
1	D	Czasopisma	2	2	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1,2
2	F	Ciśnienie	11	3	3	3	3	2	2	2	2	2	3	3	3	1	1	1	1	2	2,7
3		Średnia	15	1	1	2	1	3	3	3	4	4	5	5	5	1	1	1	1	4	3,3
4	R	Adresy	7	9	9	5	4	5	5	5	3	3	2	2	2	1	1	1	1	3	3,8
5	G	Wiatr	10	6	6	4	6	4	4	4	6	6	6	6	4	1	1	1	1	6	4,6
6	M	Reader's Digest	5	7	7	10	7	11	11	11	5	5	9	9	8	1	1	1	1	5	6,3
7	A	Dorzecza rzek	6	14	14	11	12	10	10	10	7	7	4	4	6	1	1	1	1	7	7,0
8	O	Promienie X	12	12	12	9	8	8	8	8	8	8	7	7	7	1	1	1	1	8	7,0
9	Q	Prom. ciała czarnego	16	11	11	8	10	7	7	7	10	10	10	10	11	11	1	1	1	10	8,4
10	T	Śmiertelność	8	16	16	12	11	13	13	13	9	9	8	8	9	9	10	1	1	9	9,7
11	P	Amer. liga baseball	18	4	4	7	9	9	9	9	12	12	13	13	12	11	14	14	1	11	10,1
12	N	Koszty	13	13	13	13	13	12	12	12	13	13	11	11	10	13	10	1	1	14	11,0
13	I	Drenaż	4	18	18	17	18	18	18	18	11	11	12	12	13	9	10	1	1	12	12,3
14	L	Dane z projektów	9	15	15	16	15	15	15	15	15	15	15	15	14	16	10	1	1	15	12,9
15	S	n^1, n^2, ... n!	14	5	5	14	14	14	14	14	17	17	18	18	17	13	16	17	1	17	13,6
16		SUMA	22	10	10	6	5	6	6	6	18	18	16	16	18	19	19	18	18	18	13,8
17	J	Masa atomowa	1	17	17	21	21	22	22	22	14	14	17	17	16	13	14	14	17	13	16,2
18	C	Stałe	3	22	22	22	22	21	21	21	16	16	14	14	15	16	17	14	1	16	16,3
19	B	Populacja	20	8	8	15	16	16	16	16	20	19	21	21	20	19	19	20	20	21	17,5
20	E	Ciepło	17	21	21	20	19	20	20	20	19	19	19	19	19	18	18	19	21	19	19,3
21	H	Masa cząsteczkowa	19	20	20	19	17	19	19	19	21	19	20	20	21	21	19	20	19	20	19,6
22	K	n^(-1), n^(0,5)	21	19	19	18	20	17	17	17	22	19	22	22	22	21	22	22	22	22	20,2

Źródło: opracowanie własne.

Tab. 2.14. Współczynniki korelacji liniowej pomiędzy miernikami zgodności

	n	1-r	r(p)	M5	M1	M2	M3	M4	chi	1-chi(p)	KS1	KS2	KS3	z>1,64	z>1,96	z>2,58	z>3,29	z sred
n	1,00	-0,08	-0,07	-0,14	-0,13	-0,14	-0,14	-0,14	0,32	0,30	0,27	0,27	0,35	0,51	0,55	0,40	0,31	0,43
1-r	-0,08	1,00	0,89	0,85	0,87	0,79	0,79	0,79	0,44	0,53	0,42	0,42	0,58	0,53	0,54	0,48	0,50	0,53
r(p)	-0,07	0,89	1,00	0,70	0,74	0,62	0,62	0,62	0,28	0,40	0,25	0,25	0,45	0,43	0,47	0,39	0,42	0,38
M5	-0,14	0,85	0,70	1,00	0,98	0,98	0,98	0,98	0,31	0,66	0,45	0,45	0,51	0,52	0,48	0,37	0,36	0,43
M1	-0,13	0,87	0,74	0,98	1,00	0,92	0,92	0,92	0,40	0,69	0,51	0,51	0,56	0,56	0,53	0,42	0,41	0,51
M2	-0,14	0,79	0,62	0,98	0,92	1,00	1,00	1,00	0,23	0,61	0,38	0,38	0,44	0,44	0,41	0,30	0,29	0,35
M3	-0,14	0,79	0,62	0,98	0,92	1,00	1,00	1,00	0,23	0,61	0,38	0,38	0,44	0,44	0,41	0,30	0,29	0,35
M4	-0,14	0,79	0,62	0,98	0,92	1,00	1,00	1,00	0,23	0,61	0,38	0,38	0,44	0,44	0,41	0,30	0,29	0,35
chi	0,32	0,44	0,28	0,31	0,40	0,23	0,23	0,23	1,00	0,43	0,89	0,89	0,85	0,73	0,76	0,82	0,89	0,93
1-chi(p)	0,30	0,53	0,40	0,66	0,69	0,61	0,61	0,61	0,43	1,00	0,63	0,63	0,68	0,81	0,71	0,57	0,46	0,65
KS1	0,27	0,42	0,25	0,45	0,51	0,38	0,38	0,38	0,89	0,63	1,00	1,00	0,95	0,86	0,86	0,91	0,88	0,95
KS2	0,27	0,42	0,25	0,45	0,51	0,38	0,38	0,38	0,89	0,63	1,00	1,00	0,95	0,86	0,86	0,91	0,88	0,95
KS3	0,35	0,58	0,45	0,51	0,56	0,44	0,44	0,44	0,85	0,68	0,95	0,95	1,00	0,93	0,95	0,97	0,92	0,97
z>1,64	0,51	0,53	0,43	0,52	0,56	0,44	0,44	0,44	0,73	0,81	0,86	0,86	0,93	1,00	0,96	0,90	0,79	0,91
z>1,96	0,55	0,54	0,47	0,48	0,53	0,41	0,41	0,41	0,76	0,71	0,86	0,86	0,95	0,96	1,00	0,95	0,84	0,92
z>2,58	0,40	0,48	0,39	0,37	0,42	0,30	0,30	0,30	0,82	0,57	0,91	0,91	0,97	0,90	0,95	1,00	0,93	0,94
z>3,29	0,31	0,50	0,42	0,36	0,41	0,29	0,29	0,29	0,89	0,46	0,88	0,88	0,92	0,79	0,84	0,93	1,00	0,93
z sred	0,43	0,53	0,38	0,43	0,51	0,35	0,35	0,35	0,93	0,65	0,95	0,95	0,97	0,91	0,92	0,94	0,93	1,00
>0	**10**	**16**	**16**	**16**	**16**	**16**	**16**	**16**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**	**17**
<0	**7**	**1**	**1**	**1**	**1**	**1**	**1**	**1**										
srednia	**0,31**	**0,61**	**0,50**	**0,62**	**0,64**	**0,57**	**0,57**	**0,57**	**0,59**	**0,61**	0,67	0,67	0,72	0,70	0,70	0,66	0,63	**0,69**

Źródło: opracowanie własne.

Tab. 2.15. Kategorie współczynników korelacji liniowej r między miernikami zgodności rozkładów cyfr znaczących ze względu na ich poziom [0<0,35]; [1 <0,35;0,5]; [2 <0,5;0,8]; [3 >0,8]

	z>1,64	KS3	z>1,96	M1	KS1	KS2	z sred	1-r	z>2,58	1-chi(p)	M5	z>3,29	chi	M2	M3	M4	r(p)	n	Średnia
z>1,64	3	3	3	2	3	3	3	2	3	3	2	2	2	1	1	1	1	2	2,18
KS3	3	3	3	2	3	3	3	2	3	2	2	3	3	1	1	1	1		2,12
z>1,96	3	3	3	2	3	3	3	2	3	2	1	3	2	1	1	1	1	2	2,12
M1	2	2	2	3	2	2	2	3	1	2	3	1	1	3	3	3	2		2,00
KS1	3	3	3	2	3	3	3	1	3	2	1	3	3	1	1	1			1,94
KS2	3	3	3	2	3	3	3	1	3	2	1	3	3	1	1	1			1,94
z sred	3	3	3	2	3	3	3	2	3	2	1	3	3				1	1	1,94
1-r	2	2	2	3	1	1	2	3	1	2	3	2	1	2	2	2	3		1,82
z>2,58	3	3	3	1	3	3	3	1	3	2	1	3	3				1	1	1,82
1-chi(p)	3	2	2	2	2	2	2	2	2	3	2	1	1	2	2	2	1		1,76
M5	2	2	1	3	1	1	1	3	1	2	3	1		3	3	3	2		1,71
z>3,29	2	3	3	1	3	3	3	2	3	1	1	3	3				1		1,71
chi	2	3	2	1	3	3	3	1	3	1		3	3						1,47
M2	1	1	1	3	1	1		2		2	3			3	3	3	2		1,35
M3	1	1	1	3	1	1		2		2	3			3	3	3	2		1,35
M4	1	1	1	3	1	1		2		2	3			3	3	3	2		1,35
r(p)	1	1	1	2			1	3	1	1	2	1		2	2	2	3		1,18
n	2		2				1		1									3	0,35
L.3>0,8	**7**	**8**	**7**	**5**	**8**	**8**	**8**	**3**	**8**	**1**	**5**	**7**	**6**	**4**	**4**	**4**	**1**		**91**
L.2<0,8	**6**	**4**	**5**	**8**	**2**	**2**	**3**	**9**	**1**	**12**	**4**	**2**	**2**	**3**	**3**	**3**	**5**	**2**	**73**
L.1<0,5	**4**	**4**	**5**	**3**	**5**	**5**	**3**	**4**	**6**	**4**	**6**	**4**	**3**	**5**	**5**	**5**	**8**	**2**	**78**
L.0<0,35		**1**		**1**	**2**	**2**	**3**	**1**	**2**		**2**	**4**	**6**	**5**	**5**	**5**	**3**	**13**	**52**
Średnia	2,18	2,12	2,12	2,00	1,94	1,94	1,94	1,82	1,82	1,76	1,71	1,71	1,47	1,35	1,35	1,35	1,18	0,35	1,67

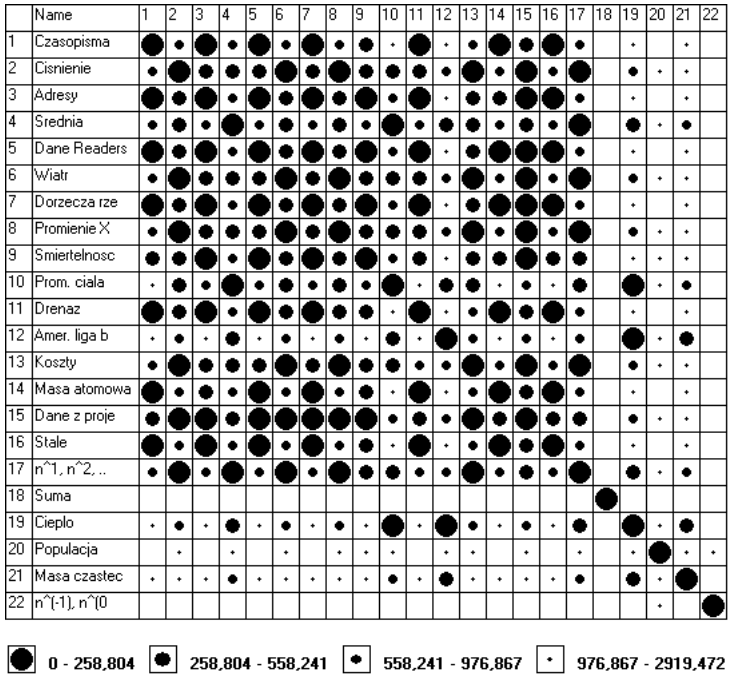
Źródło: opracowanie własne.

Tab. 2.16. Kategorie współczynników korelacji rang Spearmana między miernikami zgodności rozkładów cyfr znaczących ze względu na ich poziom [0 <0,5]; [1 <0,5;0,75]; [2 <0,75;0,9]; [3 >0,9]

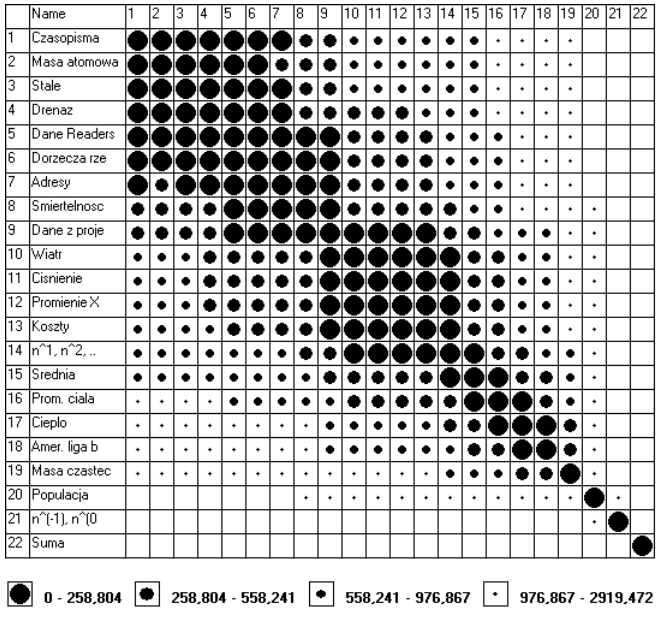
	KS1	KS2	KS3	chi	1-chi(p)	z sred	R	z>1,64	z>1,96	M2	M3	M4	M5	M1	z>2,58	1-r	r(p)	z>3,29	n	Średnia
KS1	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2	1	2,28
KS2	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2	1	2,28
KS3	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2	1	2,28
chi	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	1	1	2,22
1-chi(p)	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	1	1	2,22
z sred	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	1	1	2,22
R	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	1	1	2		2,22
z>1,64	3	3	3	3	3	3	3	3	3	1	1	1	1	1	2	1	1	1	1	1,94
z>1,96	3	3	3	3	3	3	3	3	3	1	1	1	1	1	3	1	1	1		1,94
M2	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2	1		1,89
M3	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2	1		1,89
M4	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2	1		1,89
M5	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2			1,83
M1	2	2	2	2	2	2	2	1	1	3	3	3	3	3	1	2	2			1,83
z>2,58	2	2	2	2	2	2	2	2	3	1	1	1	1	1	3			2	1	1,50
1-r	1	1	1	1	1	1	1	1	1	2	2	2	2	2		3	3			1,22
r(p)	1	1	1	1	1	1	1	1	1	2	2	2	2	2		3	3			1,22
z>3,29	2	2	2	1	1	1	2	1	1	1	1	1			2			3	1	1,06
n	1	1	1	1	1	1		1							1			1	3	0,50
L.3>0,9	**8**	**8**	**8**	**8**	**8**	**8**	**8**	**8**	**9**	**4**	**4**	**4**	**4**	**4**	**1**	**1**	**1**			93
L.2<0,9	**7**	**7**	**7**	**6**	**6**	**6**	**7**	**1**		**9**	**9**	**9**	**9**	**9**	**9**	**5**	**5**	**5**		113
L.1<0,75	**3**	**3**	**3**	**4**	**4**	**4**	**2**	**9**	**8**	**4**	**4**	**4**	**3**	**3**	**6**	**9**	**9**	**9**	**9**	97
L.0<0,5	1	1	1	1	1	1	2	1	2	2	2	2	3	3	3	**4**	**4**	**5**	**10**	46
Średnia	2,28	2,28	2,28	2,22	2,22	2,22	2,22	1,94	1,94	1,89	1,89	1,89	1,83	1,83	1,50	1,22	1,22	1,06	0,50	1,81

Źródło: opracowanie własne.

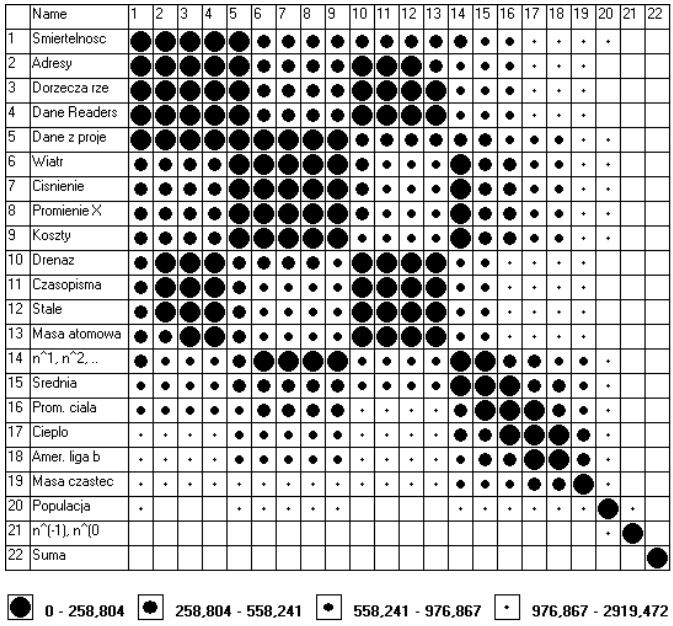
Rys. 2.5. Nieuporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących



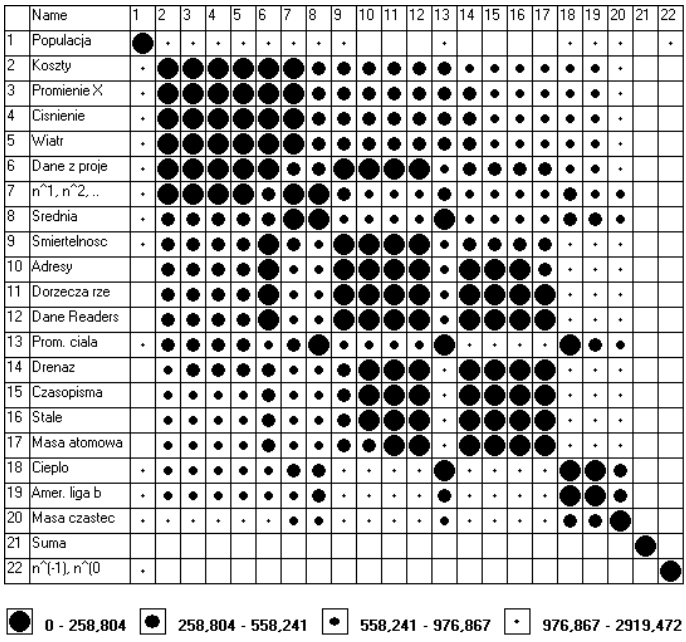
Rys. 2.6. Uporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących (metoda A)



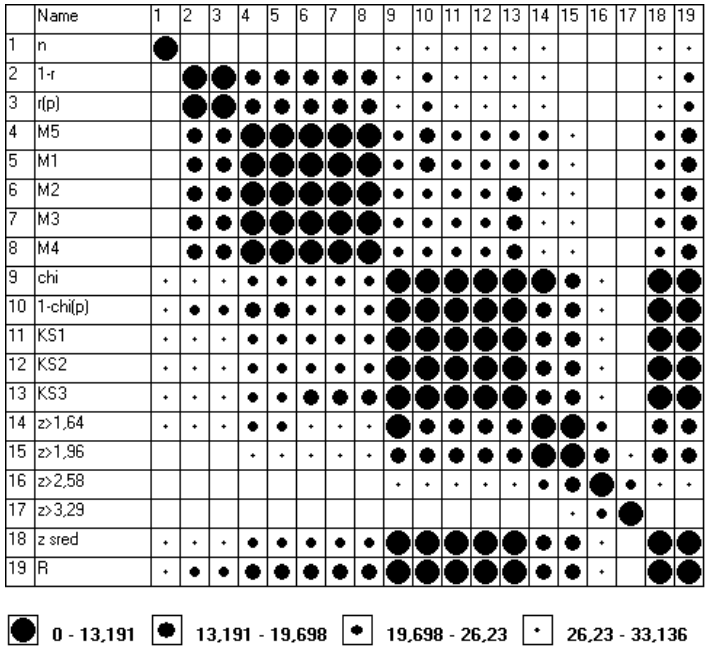
Rys. 2.7. Uporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących (metoda B)



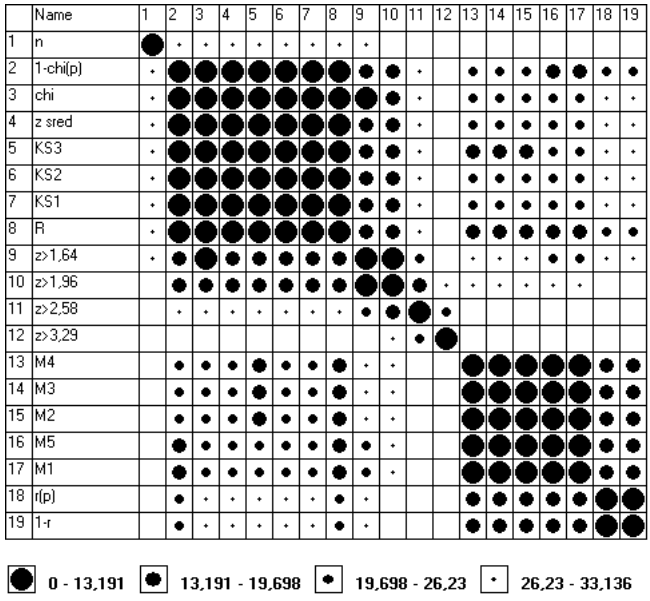
Rys. 2.8. Uporządkowany diagram Czekanowskiego zbiorów Benforda w przestrzeni mierników zgodności rozkładów cyfr znaczących (metoda C)



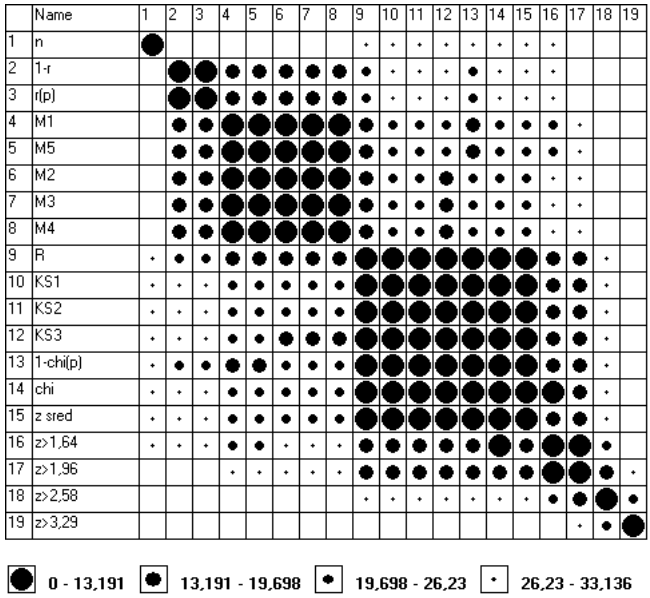
Rys. 2.9. Nieuporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda



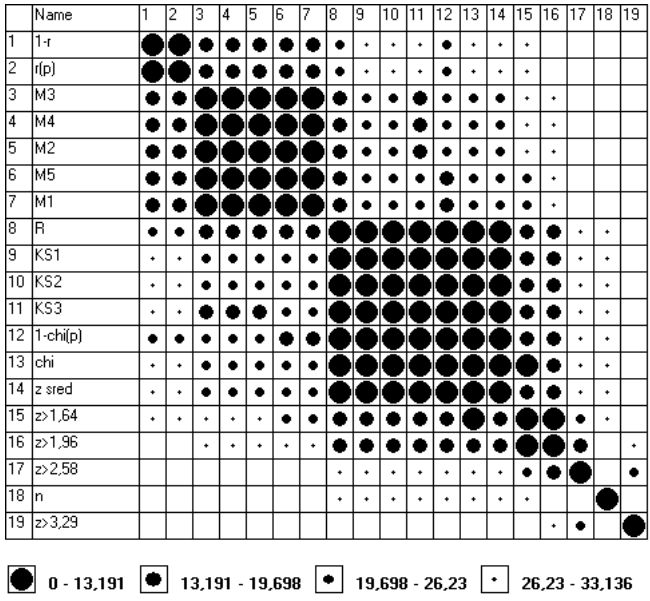
Rys. 2.10. Uporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda (metoda A)



Rys. 2.11. Uporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda (metoda B)



Rys. 2.12. Uporządkowany diagram Czekanowskiego mierników zgodności rozkładów cyfr znaczących w przestrzeni zbiorów Benforda (metoda C)



Istota metody Czekanowskiego[11] polega na takim uporządkowaniu zbioru klasyfikowanych obiektów, aby wzdłuż głównej przekątnej diagramu znalazły się elementy jak najbardziej zaczernione (podobne) oraz im dalej od głównej przekątnej, tym bardziej te elementy powinny być mniej zaczernione.

Na rysunku 2.5 przedstawiono nieuporządkowany diagram Czekanowskiego dla zbiorów Benforda, w którym przyjęto 5 klas podobieństwa. Trudno w tym diagramie zauważyć jakikolwiek porządek. Układ i konfiguracja elementów diagramu są chaotyczne.

Rysunki 2.6–2.8 zawierają uporządkowane przy pomocy różnych algorytmów (A–B–C) diagramy Czekanowskiego wygenerowane w programie MaCzek. Analiza tych diagramów prowadzi do wniosku, że wśród 20 zbiorów Benforda można wyróżnić 4 podgrupy zbiorów o podobnych wartościach mierników zgodności oraz trzy zbiory jednoelementowe (populacja, potęga liczb naturalnych, zbiór sumaryczny), dla których miary zgodności kształtują się inaczej niż dla pozostałych zbiorów Benforda.

Podobną analizę wykonano dla zbioru miar zgodności rozkładów (rys. 2.9–2.12). Z uporządkowanych diagramów Czekanowskiego wynika, że w rozpatrywanym zbiorze miar zgodności można wyróżnić następujące bardziej jednorodne podzbiory:

- mierniki M1–M5,
- współczynniki *1–r* oraz *r(p),*
- statystyki *KS1, KS2, KS3, z sred* oraz statystyki *chi* i *1–ch(p),*
- statystyki *z>1,64* oraz *z>1,96,*
- trzy zbiory jednoelementowe, zawierające mierniki *z>2,58; z>3,29* oraz *n.*

Ponadto daje się zauważyć podobieństwo podzbiorów a) i b) oraz c) i d). Wyniki te ująć można w postaci tabeli (2.17). Z przeprowadzonej analizy można wyciągnąć wniosek, że w praktyce należy przy ocenie zgodności rozkładów cyfr znaczących uwzględnić nie 18, ale co najwyżej 6 mierników – po jednym mierniku z każdej z grup od (a) do (d) oraz mierniki tworzące grupy (e) oraz (f).

Alternatywnym rozwiązaniem jest uwzględnienie 3 mierników. Jeden powinien reprezentować podzbiór {a;b}, drugi podzbiór {c; d} natomiast trzeci miernik – podzbiór {e;f}.

[11] Por. np. T. Grabiński *Propozycje w zakresie porządkowania diagramu Jana Czekanowskiego*, [w:] pr. zbior. *Studia z zakresu metod ilościowych w ekonomii, demografii i socjologii*, Prace Komisji Socjologicznej PAN, O. Kraków, nr 40/1977, Wrocław–Warszawa–Kraków–Gdańsk.

Tab. 2.17. Wyniki klasyfikacji miar zgodności rozkładów

(a)	M2	M3	M4	M5	**M1**		(b)	**1-r**	r(p)
(c)	KS1	KS2	KS3	chi	1-chi(p)	z sred	(d)	**z>1,64**	z>1,96
(e)	z>2,58	(f)	z>3,29		n				

Źródło: opracowanie własne.

Odrębną kwestią jest wybór reprezentanta poszczególnych podzbiorów. Można tu posłużyć się kategoryzacją miar zgodności (por. tab. 2.15–2.16). Biorąc pod uwagę te parametry, w wariancie oszczędnym „najlepszymi" miarami zgodności rozkładów (w sensie reprezentatywności całego zbioru miar) mogą być: miernik *M1*, statystyka *KS3* oraz *z>2,58*. W wariancie poszerzonym można by dodatkowo uwzględnić miary *1–r* oraz *z>1,64.*

Warto zauważyć, że wśród wskazanych miar nie ma statystyki χ^2. Biorąc jednak pod uwagę jej popularność w praktyce warto też uwzględniać i ten parametr.

2.4. Interpretacja prawa Benforda

W literaturze[12] można spotkać przykłady pozwalające zrozumieć istotę prawa Benforda. Niech będzie dana dowolna początkowa wielkość (kwota lokaty, wielkość produkcji, liczba mieszkańców), która przyrasta z okresu na okres o określoną wielkość (np. 1%) przez dłuższy czas (np. 240 miesięcy, to jest 20 lat). W tabeli 2.18 podano rozkłady pierwszych cyfr znaczących uzyskanych wielkości przy założeniu, że tempo wzrostu wynosi od 1% do 5%. Obliczenia wykonano w trzech wersjach, zakładając, że w okresie wyjściowym początkowa wielkość wynosiła 1, 2 i 3 jednostki. Jak można zauważyć, zgodność rozkładów z prawem Benforda jest wysoka (por. ostatnie dwie kolumny tabeli 2.18, gdzie zamieszczono różnice pomiędzy częstościami empirycznymi a rozkładem Benforda oraz statystyki chi-kwadrat i mierniki M1–M3). Obserwuje się tylko względną nadwyżkę częstości empirycznych w stosunku do częstości teoretycznych dla cyfry odpowiadającej założonej wielkości wyjściowej. Nadwyżka ta rekompensowana jest w pozostałych przedziałach odpowiadających innym cyfrom znaczącym.

[12] M. Nigrini, *I've got your number: How a mathematical phenomenon can help CPAs uncover fraud and Rother irregulaties*, AICPA Journal of Accountancy Online Journal, 5/1999, www.aicpa.org/pubs/jofa/may1999/ nigrini.htm

Zwiększenie lub zmniejszenie liczby obserwacji, zmiana wielkości początkowej, zmiana tempa przyrostu nie spowodują w tym przykładzie żadnych istotnych zmian. Można stąd wysnuć wniosek, że jeżeli w zbiorze danych empirycznych mamy do czynienia z pomiarami wielkości o różnej „długości życia", to najczęściej występują w nim wartości zaczynające się od małych, a nie od dużych cyfr.

Intuicyjnie prawo Benforda można zilustrować wychodząc od początkowych wielkości na poziomie jednostkowym (1, 10, 100, 1000 itd.). Aby przejść do wielkości zaczynających się od 2, 20, 200, ..., trzeba podwoić wielkość wyjściową (wzrost o 100%). Aby przejść od 2, 20, 200,... do 3, 30, 300,..., trzeba zwiększyć podstawę tylko o 50%, czyli dwukrotnie mniej niż w poprzednim przypadku. Idąc dalej, aby przejść od 8, 80, 800,... do 9, 90, 900,... wystarczy, aby wielkość początkowa zwiększyła się o 12,5%, aby przejść od 9, 90, 900,... do 10,100,1000 trzeba zwiększyć liczby tylko o 11,1%. Potem te stopy wzrostu cyklicznie powtarzają się: 100%, 50%, 33%, 25%, 20% itd. Ten mechanizm pokazany jest w tabeli 2.19 oraz na rysunku 2.13.

Tab. 2.18. Procenty przyrostu niezbędne do zmiany pierwszej cyfry znaczącej w kolejnych przedziałach liczbowych

Zmiana od–do itd.	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9	9–10
	10–20	20–30	30–40	40–50	50–60	60–**70**	**70**–80	80–90	90–100
	ooo	ooo	ooo	ooo	ooo	ooo	ooo	ooo	ooo
%	100,0	50,0	33,3	25,0	20,0	16,7	14,3	12,5	11,1

Źródło: opracowanie własne.

Tab. 2.19. Rozkłady pierwszych cyfr znaczących w szeregach zaczynających się od 1,2 i 3 z tempem wzrostu wielkości od 1% do 5% w ciągu 240 okresów

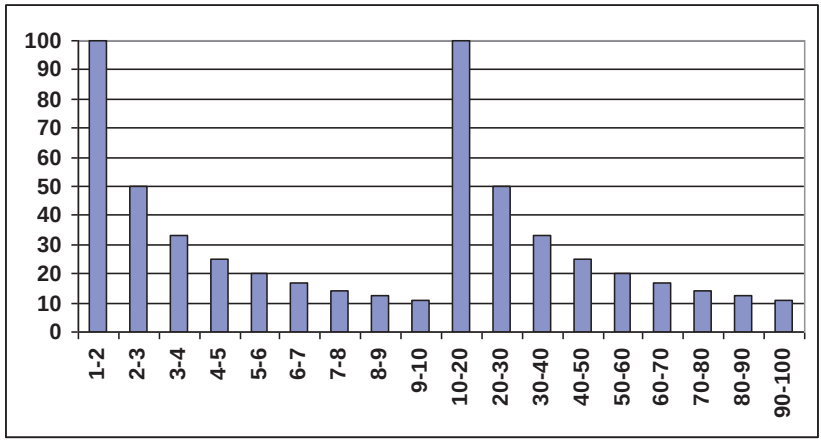
do = 1	1%	2%	3%	4%	5%	Średnia	Benford	P(d)	B(d)	P(d)–B(d)	Chi=
1	78	78	78	76	75	77,0	72,2	32,1	30,1	1,98	0,49
2	41	40	41	42	42	41,2	42,3	17,2	17,6	-0,44	
3	29	30	28	29	29	29,0	30,0	12,1	12,5	-0,41	M1=
4	22	22	24	23	23	22,8	23,3	9,5	9,7	-0,19	0,45
5	19	18	18	18	19	18,4	19,0	7,7	7,9	-0,25	
6	15	16	15	16	16	15,6	16,1	6,5	6,7	-0,19	M2=
7	13	14	15	14	14	14,0	13,9	5,8	5,8	0,03	0,24
8	12	11	12	12	12	11,8	12,3	4,9	5,1	-0,20	
9	11	11	9	10	10	10,2	11,0	4,3	4,6	-0,33	M3=
SUMA	240	240	240	240	240	240	240	1	100	0,00	0,71

do = 2	1%	2%	3%	4%	5%	Średnia	Benford	P(d)	B(d)	P(d)-B(d)	Chi=
1	**70**	**70**	69	**70**	71	**70,0**	72,2	29,2	30,1	-0,94	0,83
2	49	48	48	47	46	47,6	42,3	19,8	17,6	2,22	
3	29	30	30	29	29	29,4	30,0	12,3	12,5	-0,24	M1=
4	23	22	21	23	23	22,4	23,3	9,3	9,7	-0,36	0,49
5	18	18	20	19	19	18,8	19,0	7,8	7,9	-0,08	
6	15	16	16	15	16	15,6	16,1	6,5	6,7	-0,19	M2=
7	14	14	12	14	13	13,4	13,9	5,6	5,8	-0,22	0,27
8	12	11	12	12	13	12,0	12,3	5,0	5,1	-0,12	
9	10	11	12	11	10	10,8	11,0	4,5	4,6	-0,08	M3=
SUMA	240	240	240	240	240	240	240	1	100	0,00	0,82

do = 3	1%	2%	3%	4%	5%	Średnia	Benford	P(d)	B(d)	P(d)-B(d)	Chi=
1	**70**	**70**	71	71	71	**70**,6	72,2	29,4	30,1	-0,69	1,23
2	41	41	40	41	41	40,8	42,3	17,0	17,6	-0,61	
3	37	36	36	35	34	35,6	30,0	14,8	12,5	2,34	M1=
4	23	23	24	23	23	23,2	23,3	9,7	9,7	-0,02	0,52
5	18	19	18	18	18	18,2	19,0	7,6	7,9	-0,33	
6	16	15	15	16	16	15,6	16,1	6,5	6,7	-0,19	M2=
7	13	13	14	14	14	13,6	13,9	5,7	5,8	-0,13	0,28
8	12	12	12	12	12	12,0	12,3	5,0	5,1	-0,12	
9	10	11	10	10	11	10,4	11,0	4,3	4,6	-0,24	M3=
SUMA	240	240	240	240	240	240	240	1	100	0,00	0,85

Źródło: opracowanie własne.

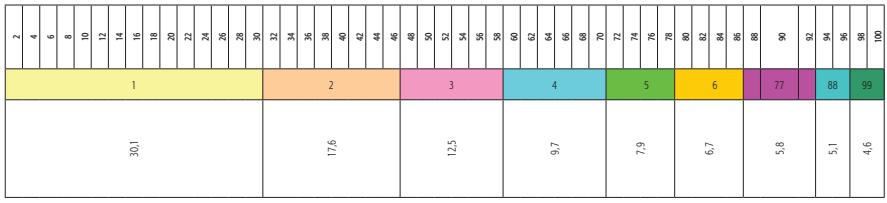
Rys. 2.13. Procenty przyrostu niezbędne do zmiany pierwszej cyfry znaczącej w kolejnych przedziałach liczbowych



Generalnie mamy więcej małych i średnich przedsiębiorstw niż dużych koncernów, więcej małych miast niż dużych, więcej rzeczek i strumyków niż dużych rzek. Zaczynamy zwykle od jednostek i staramy się *zwiększyć* daną wielkość. Wychodząc od początkowych, jednostkowych wielkości nie zawsze udaje się dotrzeć do kolejnych rzędów wielkości. Zwykle istnieje jakaś granica wzrostu. Żaden z podmiotów nie potrafi zwiększać danego parametru w nieskończoność, wiele z nich zatrzymuje się „po drodze" na pewnym poziomie jego wielkości. Niewiele podmiotów przechodzi od wyjściowej wartości skali cyfr (1, 10, 100,...) do jej końca (9 ,99, 999,...), na ogół podmioty te docierają do wcześniejszych szczebli, najwięcej z nich (bo jest to najłatwiejsze) osiąga drugi szczebel skali (2, 20, 200,...).

Inna interpretacja prawa Benforda wynika z wzoru definicyjnego (2.1), z którego można wnosić, że długość dystansu pomiędzy sąsiednimi, pierwszymi cyframi znaczącymi dzielona przez odcinek jednostkowy daje kolejne elementy rozkładu Benforda (por. rys. 2.14). Liczba 1 zajmuje tu 30,1% ogólnej powierzchni prostokąta (długości skali), liczba 2 – 17,6% itd. Jeżeli liczba zaczyna się od 1 to pojawienie się liczby 2 wymaga podwojenia (wzrost o 100%) pierwotnej wartości. Jeżeli liczba zaczyna się od 9 to pojawienie się liczby 1 jest bardziej prawdopodobne, gdyż wymaga zmiany pierwotnej wartości tylko o 11%.

Rys. 2.14. Skala logarytmiczna



2.5. Próby wyjaśnienia prawa Benforda

Przez 50 lat od badań Franka Benforda praktycznie nikt ich nie kontynuował. Pojawiały się tylko próby wyjaśniania istoty prawa Benforda. M.in. w 1944 r. S.A.Goudsmith i W.H.Furry[13] sformułowali tezę, że powodem tej prawidłowości jest sposób zapisu liczb. W 1948 r. L.V. Furlan[14] stwierdził, że prawo Benforda oddaje harmoniczną w swojej istocie naturę otaczają-

[13] S.A. Goudsmith i W.H.Furry, *Significant figure of numbers in statistical tables,* Nature, 154/1944, p. 800-801.

[14] L.V.Furlan, *Das Harmoniesgestez der Statistik,Eine Untersuchung uber die metrische Interdependenz der sozialeen Erscheinungen*, Bael, Switzerland, Verlag fur Recht und Geselschaft, xiii/1948.

cej nas rzeczywistości. Dla przykładu, skala wrażliwości naszych zmysłów (słuchu, wzroku), a także wiele zjawisk naturalnych, np. fal sejsmicznych w czasie trzęsień ziemi, przebiega raczej w skalach logarytmicznych, a nie linowych.

Dopiero w latach 90. udało się uzyskać bardziej przekonujące interpretacje prawa Benforda. T. Hill[15] dowodzi, że jeżeli weźmiemy losowo dobrane próby z losowo dobranych rozkładów („*random samples from random distributions*"), wtedy łączna próba, składająca się z wszystkich prób pochodzących z różnych populacji, będzie podlegała prawu Benforda. Prawo to opisuje więc postać tzw. „rozkładu rozkładów" („*Distribution of Distributions*", „*Second Generation Distributions*"). Np. jeżeli weźmiemy dane dotyczące powierzchni rzek, wyników meczów sportowych, numerów losów na loterii, obrotów na giełdzie oraz liczby cytowań prac na kolejnych stronach artykułów, to pomimo, iż żaden ze zbiorów danych pochodzących z tych populacji może nie podlegać prawu Benforda, to ich łączna kombinacja daje zbiór liczb, który prawo Benforda ma dużą szansę spełnić. Jeżeli weźmiemy pod uwagę wielkość obrotów, to są one iloczynem ilości sprzedanych towarów lub usług oraz ceny. Wolumen sprzedaży oraz cena mogą mieć różne rozkłady, ale ich złożenie (obroty w wyrażeniu wartościowym) daje rozkład łączny, prawdopodobnie spełniający założenia prawa Benforda.

Obecnie do opisu częstości występowania pierwszych cyfr znaczących można spotkać następujące próby wyjaśnienia przyczyn działania formuły Benforda.

Pierwsza interpretacja wynika z udowodnionego przez T. Hilla twierdzenia zgodnie z którym dane, które są iloczynem wielu liczb, podlegają rozkładowi Benforda. W istocie rzeczy wiele informacji ma charakter iloczynowy, np. wartość transakcji to iloczyn ceny jednostkowej przez wolumen sprzedaży. Twierdzenie to nawiązuje do centralnego twierdzenia granicznego mówiącego nie o iloczynie, ale o sumie zmiennych losowych.

Druga interpretacja związana jest z twierdzeniem zgodnie z którym każda „mieszanka" liczb pobranych losowo z różnych zbiorowości podlega rozkładowi Benforda. Tak więc, jeżeli wylosujemy z różnych tabel liczby z rocznika statystycznego, to ich pierwsze cyfry znaczące będą podlegały rozkładowi Benforda, nawet jeżeli liczby z poszczególnych tabel temu rozkładowi nie będą podlegały. Jest to więc prawo opisujące postać analityczną „rozkładu rozkładów".

---

[15] T.P. Hill, *A statistical derivation of the significant digit law,* Statistical Science, 10/1996, p. 354–363; T.P.Hill, *The first digital phenomenon,* American Scientist, 86/1998, p. 58–363, www.mccombs.utexas.edu/faculty/jonathan.koehler/docs/sta309h/Benford_1998.pdf; www.math.gatech.edu/ ~ hill/publications/cv.dir/1st-dig.pdf.

Trzecia interpretacja wyraża się w przekonaniu, iż istnieje ogólne prawo przyrody dotyczące pierwszych cyfr znaczących oddające harmoniczną naturę rzeczywistości (wrażliwości zmysłów, a także wiele zjawisk przyrody, przebiega raczej w skalach logarytmicznych niż linowych).

2.6. Własności rozkładów cyfr znaczących

Prawo Benforda ma dwie istotne własności, a mianowicie: (1) niezmienniczość skali oraz (2) niezmienniczość podstawy.

Niezmienniczość skali oznacza, że jeżeli pomnożymy (podzielimy, podniesiemy do dowolnej potęgi) dane o rozkładzie Benforda przez niezerową stałą, to otrzymamy rozkład, który nadal będzie podlegał prawu Benforda[16]. Nie ma więc znaczenia, czy wielkości wyrażone są w dolarach, czy w euro, czy w kilometrach, czy w milach, czy w kilogramach, czy w funtach. Odwrotności liczb spełniających prawo Benforda również to prawo spełniają, np. wielkość obrotów przypadających na 1 akcję w obrocie oraz liczbę akcji przypadających na 1 dolara obrotu.

Niezmienniczość podstawy oznacza, że prawo Benforda stosuje się nie tylko dla liczb zapisanych w systemie o podstawie 10, ale również w przypadku innych podstaw systemów liczbowych. T. Hill udowodnił ponadto, że jest to ***jedyny*** rozkład, który posiada tę własność.

Prawo Benforda funkcjonuje najlepiej dla danych mających następujące własności:

- ***dostatecznie duża zmienność*** – im bardziej zróżnicowane są dane, tym lepiej, np. długości plików MP3 nie podlegają prawu Benforda, w odróżnieniu od długości wszystkich plików zapisanych na dysku komputera,
- ***brak ustalonego maksimum*** lub dopuszczalnych granic zmienności,
- ***duża próba*** – im więcej danych, tym lepiej,
- ***dodatnia asymetria*** rozkładu (średnia arytmetyczna jest większa od mediany), inaczej mówiąc, powinno być więcej małych jednostek niż dużych,
- ***losowy dobór wielu populacji***, z których pochodzą dane,
- ***dane wynikają ze zliczania lub pomiaru***,
- ***dane na poziomie transakcyjnym***, np. notowania giełdowe, oświadczenia o bieżących wydatkach, faktury sprzedaży,
- ***dane pochodzące z działań matematycznych***, np. wartość transakcji jako iloczyn ceny jednostkowej i liczby sprzedanych jednostek.

[16] Por. R. Pinkham, *On the distribution of first significant digits*, Annals of Mathematical Statistics, 32/1961, p. 1223–1230.

Prawo Benforda **nie funkcjonuje** w takich sytuacjach jak:

- ***istnienie maksymalnych i minimalnych*** progów wartości, np. oceny z egzaminu (od 2 do 5), ilorazy inteligencji (od 50 do 200), wiek absolwentów szkoły średniej (od 14 do 18), wyniki biegu na 400 m na zawodach lekkoatletycznych,
- ***wyspecyfikowane dopuszczalne limity wartości***, np. granica tygodniowych wydatków biurowych bez akceptacji dyrekcji na poziomie 100 zł,
- ***ograniczenia formalnoprawne***, np. konieczność uiszczenia podatków od transakcji kupna–sprzedaży powyżej 1000 zł,
- ***obecność liczb „psychologicznych"***, np. ceny na poziomie „wszystko poniżej $1,99"[17],
- ***dane identyfikacyjne,*** np. numery rejestracyjne pojazdów, dowodów osobistych, telefonów, kody pocztowe, kody kreskowe, numery rachunków bankowych, numery rejestrów PESEL, NIP, REGON, ISBN, ISSN, ISMN, IACS (ewentualne liczby w tych informacjach należy raczej traktować jako symbole i znaki graficzne),
- ***te same dane powtarzane wielokrotnie***,
- ***dane pochodzące z generatorów liczb losowych,*** np. wyniki loterii.

2.7. Uogólnienia rozkładu Benforda

Prawo Benforda można uogólnić i wykorzystywać w bardziej złożonej analizie, w ramach której stosuje się następujące testy:

F1 – pierwszej cyfry znaczącej,
F2 – dwóch pierwszych cyfr znaczących,
F3 – trzech pierwszych cyfr znaczących,
L2 – dwóch ostatnich cyfr znaczących,
L1 – ostatniej cyfry znaczącej,
D2 – dokładnie drugiej cyfry znaczącej,
D3 – dokładnie trzeciej cyfry znaczącej.

Wzór 2.1 określa częstość dla ***pierwszej*** cyfry znaczącej (**test F1**). Można go uogólnić na kolejne cyfry występujące w liczbach wielocyfrowych.

Dla ***drugiej*** w kolejności cyfry **(test D2*)*** wzór ten przyjmuje postać:

$$(2.20)\quad \boxed{P(D_2 = d_2) = \sum_{k=1}^{9} \log[1 + \frac{1}{10k + d_2}] \qquad (d_2 = 0,1,2,...,9)}$$

[17] Terminem $1,99 określa się psychologiczne przekonanie klientów, że cena 1,99 $ jest znacznie niższa od ceny 2,00 $ i zachęci ich do zakupu towaru. Por. C. Carlslaw, *Anomalies in income numbers: Evidence of goal oriented behavior*, The Accounting Review, 63/1988, p.321-327.

Na przykład prawdopodobieństwo, że drugą cyfrą znaczącą będzie 3 wynosi:

$$\log(1+1/13)+\log(1+1/23)+\log(1+1/33)+,\ldots,+\log(1+1/93)=0{,}17$$

Z powyższego wzoru można też uzyskać częstość wystąpienia dowolnej dwuelementowej kombinacji cyfr na ***dwóch pierwszych*** miejscach znaczących **(test F2)**.

$$P(D_1 = d_1; D_2 = d_2) = \log\left[1+\frac{1}{10d_1 + d_2}\right] \tag{2.21}$$

Dla przykładu częstość wystąpienia na dwóch pierwszych miejscach kombinacji ***23*** wynosi:

$$P(D_1=2;\ D_2=3)=\log[1+1/23]=\log(1{,}043478)=0{,}0184$$

Natomiast dla dowolnej n-elementowej kombinacji cyfr odpowiedni wzór ma postać:

$$P(D_1 = d_1; D_2 = d_2;\ldots, D_n = d_n) = \log\left[1+1/\sum_{i=1}^{n} d_i 10^{n-1}\right] \tag{2.22}$$

Na przykład dla kombinacji trzyelementowej **(test F3)** wzór ten przyjmuje postać:

$$P(D_1 = d_1; D_2 = d_2; D_3 = d_3) = \log\left[1+\frac{1}{100d_1 + 10d_2 + d_3}\right] \tag{2.23}$$

Jeżeli mamy trzy cyfry w następującej kolejności 3, 5 i 6, to ich szacowana częstość występowania powinna wynieść.

$$P(D_1=3;D_2=5;D_3=6)=\log(1+1/356)=0{,}001218.$$

Pojawianie się określonych cyfr na poszczególnych miejscach liczb nie jest od siebie niezależne. Wiedza o tym, jaka konkretnie cyfra występuje przed lub po innej cyfrze wpływa na prawdopodobieństwo jej pojawienia

się. Wzór na częstość warunkową pojawienia się na drugim miejscu cyfry d_2 pod warunkiem, że na pierwszym miejscu pojawiła się cyfra d_1 ma postać:

(2.24)
$$P(D_2 = d_2 | D_1 = d_1) = \frac{\log[(10 * d_1 + d_2 + 1)/(10 * d_1 + d_2)]}{\log[(d_1 + 1)/d_1]}$$

Na przykład $P(D_2=2 \mid D_1=1)= \log(13/12)/\log(2/1)=0{,}115$.

Bezwarunkowe prawdopodobieństwo pojawienia się na drugim miejscu cyfry 2 jest mniejsze i wynosi 0,109.

W tabeli 2.20 podano prawdopodobieństwa bezwarunkowe wystąpienia poszczególnych cyfr od 0 do 9 na kolejnych miejscach liczb wielocyfrowych od pierwszego do piątego.

Tab. 2.20. Prawdopodobieństwa pojawienia się cyfr od 0 do 9 na kolejnych miejscach od I do V liczb wielocyfrowych

d	I	II	III	IV	V
0		0,11968	0,10178	0,1002	0,1000
1	0,30103	0,11389	0,10138	0,1001	0,1000
2	0,17609	0,10882	0,10097	0,1001	0,1000
3	0,12494	0,10432	0,10057	0,1001	0,1000
4	0,09691	0,10031	0,10018	0,1000	0,1000
5	0,07928	0,09668	0,09979	0,1000	0,1000
6	0,06695	0,09337	0,09940	0,0999	0,1000
7	0,05799	0,09035	0,09902	0,0999	0,1000
8	0,05115	0,08757	0,09864	0,0999	0,1000
9	0,04576	0,08499	0,09827	0,0998	0,1000

Źródło: opracowanie własne.

Prawo Benforda działa niezależnie od ***podstawy systemu liczenia***. Dla dowolnej podstawy systemu liczenia *B* uogólniony wzór Benforda ma postać:

(2.25)
$$P(d) = \frac{1}{\ln(B)} \sum_{k=1}^{B-1} \ln\left[1 + \frac{1}{kB + d}\right] = \frac{1}{\ln(B)} \ln \prod_{k=1}^{B-1} \left[1 + \frac{1}{kB + d}\right] \quad n \in N \quad d = (1,...,9)$$

natomiast dla pierwszej cyfry znaczącej w systemie o podstawie *B* wzór ten upraszcza się:

(2.26)
$$P(D_1 = d_1) = \frac{\ln(1 + 1/d)}{\ln(B)} \quad d_1 = (1,2,...,B-1)$$

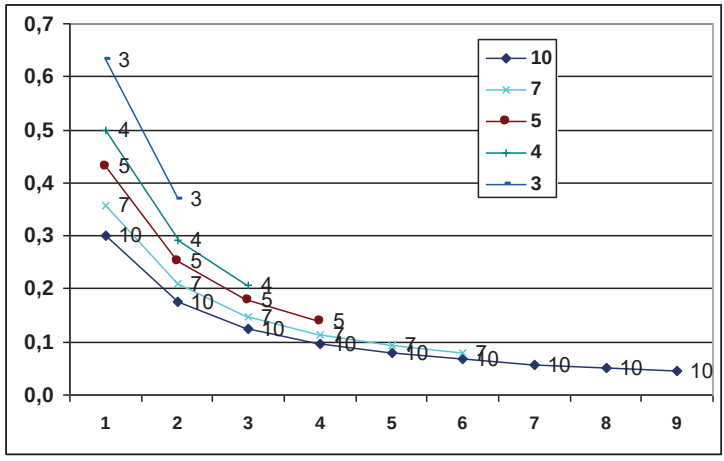
W tabeli 2.21 i na rysunku 2.15 przedstawiono częstości pojawiania się pierwszych cyfr znaczących przy różnych podstawach systemu liczbowego. Kształt rozkładu jest w każdym przypadku identyczny, przy różnej ilości możliwych liczb. Warto zauważyć, że w przypadku dwójkowego systemu liczbowego w ***każdej*** (za wyjątkiem zera) liczbie pierwszą cyfrą znaczącą musi być jedynka. Można stąd wnosić, że system binarny jest najbardziej odporny na błędy, gdyż w innych systemach pierwsze cyfry znaczące mogą być różne.

Tab. 2.21. Częstości pojawiania się pierwszych cyfr znaczących przy różnych podstawach systemu liczbowego od B=0 do B=2

d	Podstawa systemu liczbowego B								
	10	9	8	7	6	5	4	3	2
1	0,301	0,315	0,333	0,356	0,387	0,431	0,500	0,631	1,000
2	0,176	0,185	0,195	0,208	0,226	0,252	0,292	0,369	
3	0,125	0,131	0,138	0,148	0,161	0,179	0,208		
4	0,097	0,102	0,107	0,115	0,125	0,139			
5	0,079	0,083	0,088	0,094	0,102				
6	0,067	0,070	0,074	0,079					
7	0,058	0,061	0,064						
8	0,051	0,054							
9	0,046								

Źródło: opracowanie własne.

Rys. 2.15. Rozkład częstości pierwszych cyfr znaczących przy różnych podstawach systemu liczbowego dla B=3 ,4, 5,...,10



Ostatnim uogólnieniem prawa Benforda jest przypadek odnoszący się do ***liczb „ograniczonych"***, mających tylko *r* cyfr (r=1, 2, 3,...). Związane to

jest ze stosowanym w praktyce zaokrąglaniem liczb lub ich obcinaniem[18]. W przypadku pierwszej cyfry znaczącej odpowiedni wzór ma postać:

(2.27) $$P(D_1 = d_1; r) = \frac{1}{N}\left[\ln\frac{10(2*10^{r-1}-1)}{10^{r-1}-1} + \frac{8}{10^r}\right]$$

natomiast dla pozostałych cyfr znaczących:

(2.28) $$P(D_i = d_i; r) = \frac{1}{N}\left[\ln\frac{(d_i+1)10^{r-1}-1}{d_i 10^{r-1}-1} + \frac{1}{10^r}\right]$$

W tabeli 2.22 podano prawdopodobieństwa pojawienia się cyfr od 1 do 9 dla *r=1*. Jak łatwo zauważyć są one większe od odpowiadających im wielkości dla wyjściowego rozkładu Benforda (dowolne *r)*.

Tab. 2.22. Rozkłady prawdopodobieństw wystąpienia pierwszych cyfr w liczbach jednocyfrowych (r=1) oraz w liczbach wielocyfrowych (Benford)

Cyfra	1	2	3	4	5	6	7	8	9
r=1	0,393	0,257	0,133	0,081	0,053	0,036	0,024	0,015	0,008
Benford	0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Źródło: opracowanie własne.

2.8. Rozkłady alternatywne

W 1945 roku G.J. Stigler opublikował pracę, w której przedstawił inną koncepcję rozkładu pierwszych cyfr znaczących niż F. Benford[19]. W późniejszym okresie pojawiły się kolejne propozycje alternatywnych rozkładów, które zdaniem ich twórców mają podobne własności, jak rozkład Benforda i mogą być wykorzystywane do tych samych zadań.

W rozkładzie Stiglera punktem wyjścia było założenie, że w zbiorach danych (G. Stigler miał tu na myśli dane statystyczne) o ***największych*** wartościach każda z cyfr od 1 do 9 będąca pierwszą cyfrą znaczącą może pojawić się z identycznym prawdopodobieństwem wynoszącym 1/9. Wszystkie następne (mniejsze) liczby też mają równomierny rozkład pierwszych cyfr zna-

[18] Problem zaokrąglania pierwszych liczb w górę w zakresie zarobków opisał. J.K. Thomas, *Unusual patterns in reported earnings*, The Accounting Review, 64/1989, p. 773–787.

[19] Por. G.J. Stigler *The distribution of leading digits in statistical tables,* Chicago, 1945, za: J.Lee, W.K. Tam Cho, G.G. Judge *Stigler's approach to recovering the distribution of first significant digits in natural data sets*, Statistical and Probability Letters, 80/2010, p. 82–88.

czących. Jeżeli podobne założenia przyjmiemy dla kolejnych największych liczb w zbiorze, aż do ostatniego jego elementu, to uśrednione prawdopodobieństwa pojawiania się pierwszych cyfr znaczących dane są wzorem:

(2.29) $$P(D_i = d_i) = \frac{d_i \ln(d_i) - d_{i+1} \ln(d_{i+1}) + m}{9}$$

gdzie

(2.30) $$m = \frac{\sum_{i=1}^{9} i^2 \ln(d_i) - d_i d_{i+1} \ln(d_{i+1})}{9 - \sum_{i=1}^{9} d_i} = 1 + \frac{10}{9} \ln(10) = 3{,}558428$$

W tabeli 2.23 podano wartości rozkładu Stiglera oraz porównano je z odpowiednimi wartościami rozkładu Benforda (rys. 2.16). Do obydwóch rozkładów dopasowano funkcje logarytmiczne (rys. 2.17) oraz potęgowe (rys. 2.18). Rozkład Stiglera jest lepiej aproksymowany przez funkcję logarytmiczną (R^2=0,998) niż potęgową, natomiast w przypadku rozkładu Benforda sytuacja jest odwrotna – bardziej odpowiednia jest tu funkcja potęgowa (R^2=0,999) niż logarytmiczna.

Oba rozkłady mają zbliżoną postać krzywej malejącej, przy czym w przypadku rozkładu Benforda jedynek jest o 6% więcej niż w rozkładzie Stiglera (30%, a nie 24%). Odwrotna sytuacja ma miejsce dla trójek, czwórek oraz piątek, dla których rozkład Stiglera przewiduje większą częstość pojawiania (o 2%). Relatywne znaczenie tych rozbieżności jest podobne i wynosi ok. 20% w stosunku do częstości w rozkładzie Benforda. Ponadto, jeszcze większe znaczenie (25%) w ujęciu względnym ma różnica 1,2% dla dziewiątek.

W tabeli 2.23 przytoczono także wartości testu chi-kwadrat oraz testu ***z*** przy założeniu różnych rozmiarów zbioru danych, dla n={480, 550, 700, 800, 1500 oraz 2000}. Okazuje się, że obydwa rozkłady nie różnią się od siebie pod warunkiem, że liczebność zbioru danych nie przekracza

- 480 dla poziomu istotności 0,10,
- 550 dla poziomu istotności 0,05,
- 700 dla poziomu istotności 0,01.

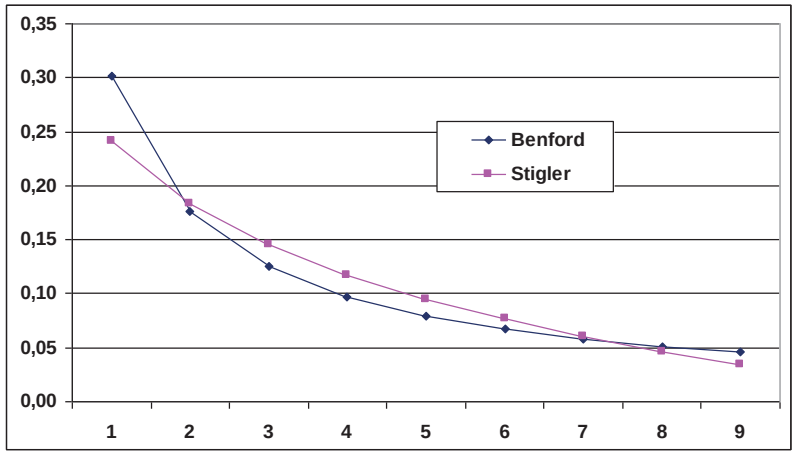
Dla większych zbiorów danych są to rozkłady istotnie różne od siebie.

Tab. 2.23. Analiza zgodności rozkładu Benforda z rozkładem Stiglera

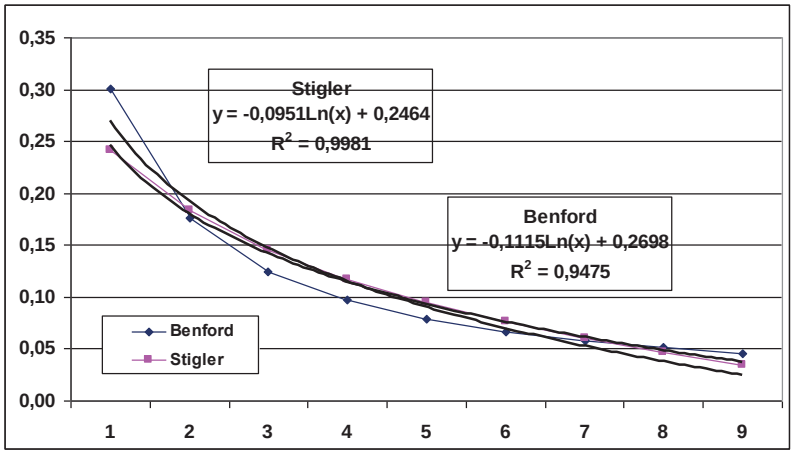
d	Benford – B	Stigler – S	B–S	S/B %	Statystyki *z*					
1	30,1	24,1	6,0	19,8	2,85	3,05	3,44	4,01	5,04	5,82
2	17,6	18,3	-0,7	-4,0	-0,41	-0,44	-0,49	-0,58	-0,72	-0,84
3	12,5	14,5	-2,1	-16,4	-1,36	-1,46	-1,64	-1,91	-2,40	-2,77
4	9,7	11,7	-2,0	-21,1	-1,52	-1,62	-1,83	-2,13	-2,68	-3,09
5	7,9	9,5	-1,6	-20,0	-1,28	-1,37	-1,55	-1,81	-2,27	-2,62
6	6,7	7,6	-0,9	-14,1	-0,83	-0,89	-1,00	-1,17	-1,47	-1,69
7	5,8	6,0	-0,2	-4,3	-0,23	-0,25	-0,28	-0,33	-0,41	-0,47
8	5,1	4,7	0,5	9,0	0,46	0,49	0,55	0,64	0,81	0,93
9	4,6	3,4	1,2	25,3	1,21	1,30	1,47	1,71	2,15	2,48
	100,0	100,0	0,0	**n**	**480**	**550**	**700**	**950**	**1500**	**2000**
				chi	13,3	15,3	19,4	26,4	41,6	55,5
				p	0,102	0,054	0,013	0,001	0,000002	0,000000

Źródło: opracowanie własne.

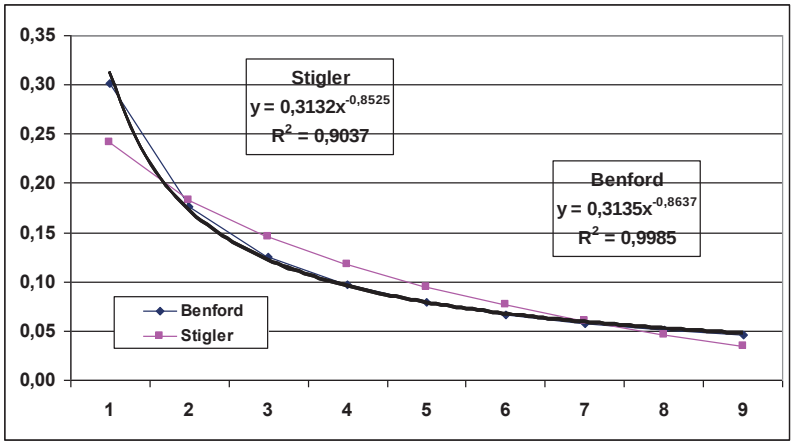
Rys. 2.16. Rozkład pierwszych cyfr znaczących według Benforda i Stiglera



Rys. 2.17. Funkcje logarytmiczne aproksymujące rozkłady pierwszych cyfr znaczących wg Benforda i Stiglera



Rys. 2.18. Funkcje potęgowe aproksymujące rozkłady pierwszych cyfr znaczących wg Benforda i Stiglera



Podobne wnioski dotyczące związku pomiędzy liczebnością zbiorów danych a zgodnością rozkładu Benforda z rozkładem Stiglera wynikają z analizy statystyk ***z***. Sytuacja jest tu zróżnicowana w zależności od tego, z jaką cyfrą mamy do czynienia. Dla jedynek, a następnie trójek, czwórek, piątek, jak również dziewiątek brak jest zgodności już dla małych zbiorów danych, podczas gdy dla pozostałych cyfr: {2; 6; 7; 8} obserwuje się zgodność częstości nawet dla zbiorów liczących więcej niż 2 000 elementów.

W literaturze można spotkać jeszcze szereg innych propozycji rozkładów konkurencyjnych w stosunku do rozkładu Newcomba–Benforda. Jednym z nich jest tzw. uogólniony rozkład Benforda[20] *(generalized Benford's Law GBL)*. Prawdopodobieństwa pojawienia się pierwszych cyfr znaczących dane są tu wzorem:

$$(2.31) \quad P(d_i\} = \frac{1}{10^{1-\alpha}-1}[(i+1)^{1-\alpha} - i^{1-\alpha}]$$

gdzie α może przyjąć dowolną wartość. Jeżeli α=0 to otrzymuje się rozkład równomierny, jeżeli α=1 to uzyskuje się rozkład Benforda. W tabeli 2.24 podano wartości rozkładów wynikających z wzoru (2.31) dla α za przedziału od –1 do 6. Na rysunkach 2.19–2.22 przedstawiono przebieg tych wartości w różnych konfiguracjach:

- rys. 2.19 dla parametru α z przedziału od 0 do 2,
- rys. 2.20 dla parametru α z przedziału od -1 do 1,
- rys. 2.21 dla parametru α z przedziału od -1 do 0,
- rys. 2.22 dla parametru α z przedziału od 2 do 6.

Na rysunku 2.19 rozkład Benforda zajmuje pozycję środkową, na rysunku 2.20 przedstawiono rozkłady umiejscowione „poniżej" rozkładu Benforda (w kierunku rozkładu równomiernego), na rysunku 2.21 – rozkłady położone „poniżej" rozkładu równomiernego, dla których kolejne wartości nie maleją, lecz wzrastają, i wreszcie na rysunku 2.22 – rozkłady dla wyższych wartości parametru α>2. W tym ostatnim przypadku rozkłady szybko dążą do postaci zdegenerowanych, w których widoczne są tylko 2–3 pierwsze cyfry znaczące.

Analiza przytoczonych wykresów pozwala sformułować wniosek, że praktyczne znaczenie mogą mieć rozkłady zbliżone do rozkładu Benforda, czyli te, dla których parametr α zawiera się w przybliżeniu w przedziale od 0,6 do 2,0. Taki wniosek nasuwa alternatywną metodę analizy rozkładu pierwszych cyfr znaczących.

W dotychczasowych analizach przyjmowano założenie, że „poprawny" zbiór danych ma rozkład zgodny z prawem Benforda. Jeżeli stwierdzano niezgodność pomiędzy rozkładem empirycznym a rozkładem Benforda to formułowano wniosek, że prawdopodobna przyczyna leży w systematycznych błędach pomiarowych danych, celowym ich zniekształcaniu, itp.

---

[20] L. Pietronero, E. Tossati, V. Tossati, A. Vespignani, *Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf*, Physica A, 293/2001, p. 297–304; M.J. Nigrini, S.J. Miller, *Benford's law applied to hydrological data – results and relevance to other geophysical data,* Math. Geol. 39/2007, p. 469–490; B. Luque, L. Lacasa, *The first-digit frequencies of prime numbers and Riemann zeta zeros,* Proceedings of the Royal Society A, 465/2009, p. 2197–2216.

Tymczasem rzeczywiste pomiary mogą nieco różnić się od rozkładu Benforda i być lepiej odzwierciedlane przez rozkłady zbliżone do niego, np. przez rozkład uogólniony GBL z parametrem α niewiele odbiegającym od jedności: α=1,1 lub α=0,9. W takiej sytuacji wydaje się, że wniosek o celowym zafałszowaniu danych jest zbyt daleko idący.

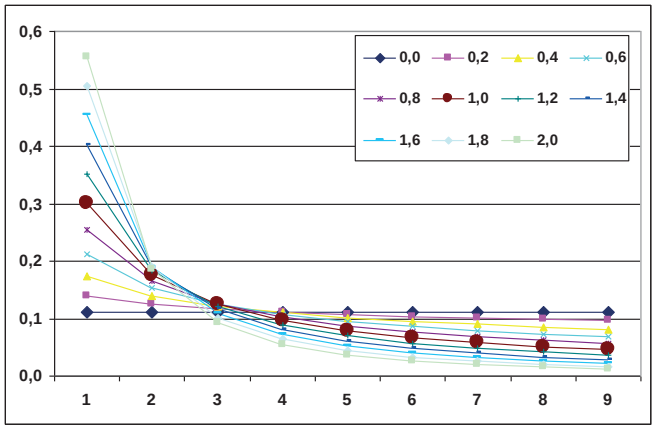
Tab. 2.24. Wartości uogólnionego rozkładu Benforda (GBL) dla parametru α z przedziału od –1 do 6

d	–1,0	–0,8	–0,6	–0,4	–0,2	0,0	0,2	0,4
1	0,030	0,040	0,052	0,068	0,087	0,111	0,140	0,173
2	0,051	0,060	0,071	0,084	0,097	0,111	0,126	0,140
3	0,071	0,079	0,087	0,096	0,104	0,111	0,117	0,122
4	0,091	0,097	0,102	0,106	0,109	0,111	0,112	0,110
5	0,111	0,113	0,115	0,115	0,114	0,111	0,107	0,102
6	0,131	0,130	0,127	0,123	0,117	0,111	0,104	0,095
7	0,152	0,145	0,138	0,130	0,121	0,111	0,101	0,090
8	0,172	0,161	0,149	0,137	0,124	0,111	0,098	0,086
9	0,192	0,176	0,159	0,143	0,127	0,111	0,096	0,082
d	**0,6**	**0,8**	**1,0**	**1,2**	**1,4**	**1,6**	**1,8**	**2,0**
1	0,211	0,254	**0,301**	0,351	0,402	0,454	0,506	0,556
2	0,154	0,166	**0,176**	0,184	0,189	0,190	0,189	0,185
3	0,125	0,126	**0,125**	0,122	0,116	0,110	0,101	0,093
4	0,108	0,103	**0,097**	0,090	0,081	0,073	0,064	0,056
5	0,095	0,088	**0,079**	0,070	0,061	0,053	0,045	0,037
6	0,086	0,077	**0,067**	0,057	0,049	0,040	0,033	0,026
7	0,079	0,068	**0,058**	0,048	0,040	0,032	0,025	0,020
8	0,073	0,062	**0,051**	0,042	0,033	0,026	0,020	0,015
9	0,069	0,057	**0,046**	0,036	0,028	0,022	0,017	0,012
d	**2,5**	**3,0**	**3,5**	**4,0**	**4,5**	**5,0**	**5,5**	**6,0**
1	0,668	0,758	0,826	0,876	0,912	0,938	0,956	0,969
2	0,166	0,140	0,113	0,088	0,067	0,050	0,037	0,027
3	0,070	0,049	0,033	0,021	0,014	0,008	0,005	0,003
4	0,037	0,023	0,013	0,008	0,004	0,002	0,001	0,001
5	0,022	0,012	0,007	0,003	0,002	0,001	0,000	0,000
6	0,015	0,007	0,004	0,002	0,001	0,000	0,000	0,000
7	0,010	0,005	0,002	0,001	0,000	0,000	0,000	0,000
8	0,007	0,003	0,001	0,001	0,000	0,000	0,000	0,000
9	0,006	0,002	0,001	0,000	0,000	0,000	0,000	0,000

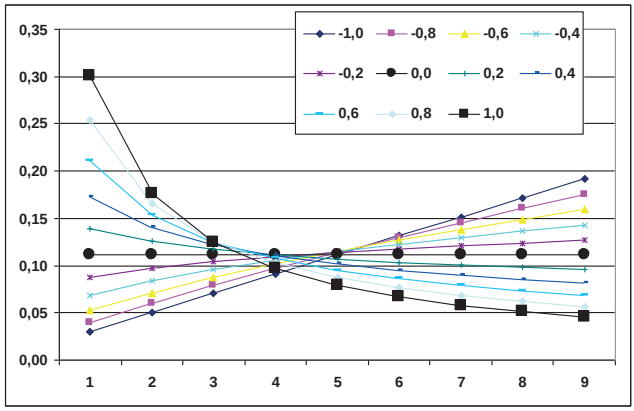
Biorąc to pod uwagę, można zaproponować następującą analizę poprawności danych, bazującą na rozkładzie pierwszych cyfr znaczących.

1. Na podstawie analizowanego zbioru danych tworzy się empiryczny rozkład pierwszych cyfr znaczących.
2. Wykorzystując algorytmy optymalizacyjne (np. Solver Excela) szuka się takiego parametru α, aby wybrany miernik zgodności (np. test χ^2) pomiędzy rozkładem (2.30) a empirycznym rozkładem pierwszych cyfr wskazywał na jak największą zgodność tych rozkładów.
3. Jeżeli uda się znaleźć taki parametr α, dla którego test zgodności rozkładów da zadawalające wyniki (np. zgodność przy zadanym z góry poziomie istotności), to można uznać, że badany zbiór nie wykazuje nieprawidłowości.
4. Zmiana w stosunku do standardowej analizy polega więc na poszerzeniu możliwości uznania poprawności zbioru danych w kontekście zgodności charakteryzującego go rozkładowi pierwszych cyfr znaczących z szerszą skalą rozkładów „wzorcowych".

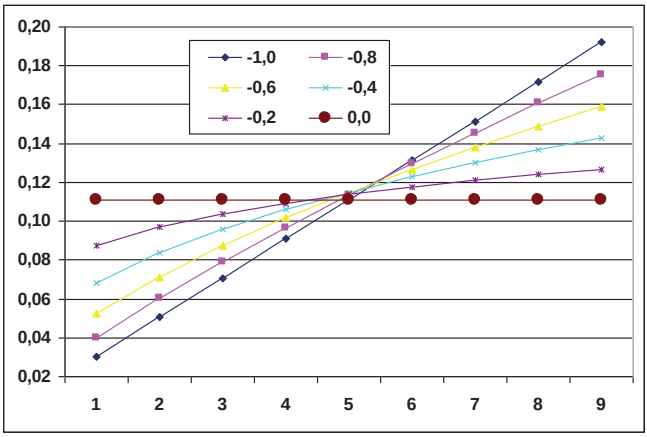
Rys. 2.19. Uogólnione rozkłady Benforda (GBL) dla parametru α od 0 do 2



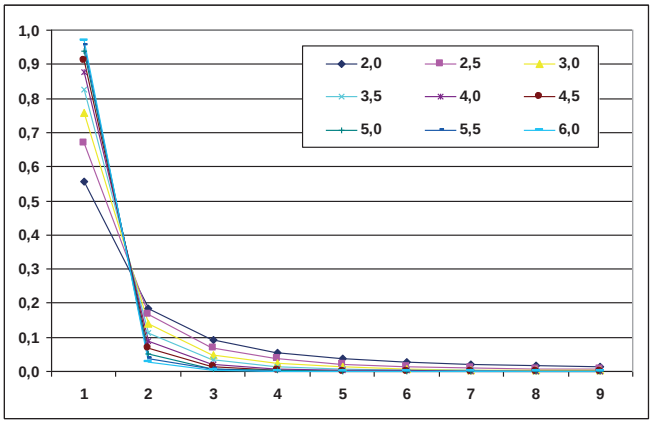
Rys. 2.20. Uogólnione rozkłady Benforda (GBL) dla parametru α od −1 do 1



Rys. 2.21. Uogólnione rozkłady Benforda (GBL) dla parametru α od –1 do 0



Rys. 2.22. Uogólnione rozkłady Benforda (GBL) dla parametru α od 2 do 6



Analiza optymalnych wartości parametru α dla wielu empirycznych zbiorów danych pozwoli sformuować hipotezę, czy typową sytuacją jest:

- „świat wg Benforda, przy $\alpha=1$”,
- „świat zbliżony do Benforda, przy $\alpha\approx1$”,
- „świat z tendencją do unitaryzmu, przy $\alpha<1$” – rozkłady wskazujące na większą równomierność rozkładu pierwszych cyfr znaczących,
- „świat z tendencją do uniformizmu, przy $\alpha>1$” – rozkłady z wyraźną dominacją początkowych pierwszych cyfr znaczących.

Może się okazać, że są różne „światy” w zależności od obszaru analizy (dane finansowe, giełdowe, makroekonomiczne, techniczne, przyrodnicze). Każdy z przedstawionych powyżej rozkładów można uznać za „normalny”.

Dopiero wtedy, gdy nie uzyska się żadnego „regularnego" rozkładu klasy GBL zgodnego z faktycznym rozkładem pierwszych cyfr znaczących, to można doszukiwać się w zbiorze danych symptomów nieprawidłowości.

Kolejnym rozkładem alternatywnym w stosunku do rozkładu Benforda jest dwustronna funkcja potęgowa TSPB (*two-sided power Benford distribution*)[21]. Zgodnie z tą propozycją rozkład pierwszej cyfry znaczącej dany jest poprzez funkcję gęstości:

(2.32)

$$g_i = [\log(1+i)^c - \log(i)^c - (1-\log(1+i))^c + (1-\log(i))^c] * 0{,}5 \quad (i = 1{,}2, \ldots, 9)$$

Prawdopodobieństwo pojawienia się pierwszych cyfr znaczących wyznacza się ze wzoru:

(2.33) $$p_i = g_i - g_{i-1}; \quad (i = 2, \ldots, 9) \quad p_1 = 1 + g_1$$

Jeżeli parametr *c=1* to rozkład dany wzorem (2.31) jest rozkładem Benforda. Dla innych wartości parametru *c* otrzymuje się inne wartości rozkładu, jakkolwiek układ tych wartości jest zbliżony (por. tab. 2.25 oraz rys. 2.23).

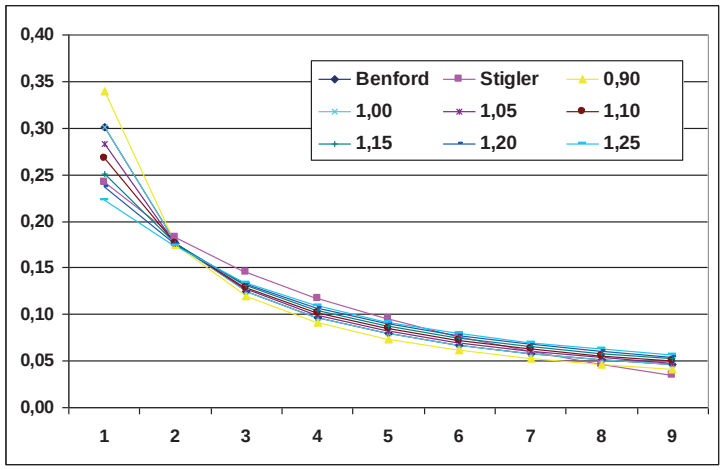
Tab. 2.25. Rozkłady Benforda, Stiglera oraz TSPB

			TSPB dla różnych wartości parametru c						
d	**Benford**	**Stigler**	**0,90**	**1,00**	**1,05**	**1,10**	**1,15**	**1,20**	**1,25**
1	0,301	0,241	0,339	0,301	0,283	0,267	0,251	0,237	0,223
2	0,176	0,183	0,174	0,176	0,176	0,176	0,176	0,175	0,174
3	0,125	0,145	0,120	0,125	0,127	0,129	0,131	0,132	0,134
4	0,097	0,117	0,091	0,097	0,100	0,102	0,104	0,107	0,109
5	0,079	0,095	0,073	0,079	0,082	0,084	0,087	0,089	0,092
6	0,067	0,076	0,062	0,067	0,070	0,072	0,075	0,077	0,079
7	0,058	0,060	0,053	0,058	0,060	0,063	0,065	0,068	0,070
8	0,051	0,047	0,046	0,051	0,054	0,056	0,058	0,060	0,063
9	0,046	0,034	0,041	0,046	0,048	0,050	0,052	0,055	0,057

Źródło: opracowanie własne.

[21] Por. W. Hurlimann, *A generalized Benford Law and its applications,* Advances and Applications in Statistics, vol. 3, 3/2003, p. 217–228; W. Hurlimann, *Generalizing Benford's Law Using Power Law: Application to Integer Sequences,* International Journal of Mathematics and Mathematical Sciences, vol. 2009.

Rys. 2.23. Rozkłady pierwszej cyfry znaczącej według Benforda, Stiglera oraz TSPB



W pracy[22] zaproponowano jeszcze inną funkcję alternatywną, tzw. dwuparametryczną funkcję gęstości Pareto–Benforda (PB) daną wzorem:

$$g_i = \frac{\alpha\beta}{\alpha+\beta i}\,\frac{1}{\ln 10}\left(\frac{\ln(i)}{\ln 10}\right)^{\beta-1} \qquad 1 < i \leq 10 \tag{2.34}$$

$$g_i = \frac{\alpha\beta}{\alpha+\beta i}\,\frac{1}{\ln 10}\left(\frac{\ln(i)}{\ln 10}\right)^{-\alpha-1} \qquad i > 10 \tag{2.35}$$

Jeżeli parametr $\beta=1$ oraz parametr α dąży do nieskończoności to rozkład Pareto–Benforda przechodzi w rozkład Benforda. Zwykle parametr β przyjmuje wartości z przedziału od 1 do 3 natomiast parametr α wartości większe – rzędu kilkunastu lub kilkudziesięciu[23].

Zaproponowaną przy okazji omawiania rozkładów GBL metodę analizy można oczywiście poszerzyć uwzględniając w niej dodatkowo zarówno funkcje TSPB jak i funkcje BP.

[22] W.J. Reed, *The Pareto, Zipf and other power laws,* Economic Letters, vol.74/2001, 15–19.

[23] Kolejne propozycje rozkładów zbliżonych do rozkładu Benforda znaleźć można m.in. w pracach: A.V. Kantorovich, S.J. Miller, *Benford's law, values of L-functions and the 3x+1 problem,* Acta Arithmetica, vol. 120, 3/2005, 269-297; K. Schurger, *Extensions of Black-Scholes processes and Benford's law,* Stochastic Processes and Their Applications, vol.118, 7/2008, p. 1219–1243.

Rozdział 3
Narzędzia wspomagające analizę rozkładów częstości cyfr

3.1.Przegląd programów obliczeniowych

Programy obliczeniowe stworzone przez wyspecjalizowane zespoły statystyków oraz programistów zyskują coraz większą popularność pośród osób zainteresowanych obliczeniami i analizą posiadanych danych. Aktualnie na rynku, w zależności od dziedziny i branży, można znaleźć duży wybór oprogramowania, które będzie spełniać oczekiwane wymagania.

Rozdział ten został poświęcony programom, które w nowatorski sposób pozwalają przeprowadzić analizę rozkładu cyfr znaczących i mogą pomóc w dostarczeniu informacji o analizowanych zbiorach.

Na początku można zadać pytanie, dlaczego audytorzy i kontrolerzy powinni korzystać z takiego oprogramowania? Otóż, dzięki niemu szybko można sprawdzić zależności statystyczne i dokonać zaawansowanych analiz. Dodatkowo, ich działanie polega zwykle na przypisaniu oczekiwanej częstotliwości do każdej liczby z badanej populacji, a następnie wyodrębnieniu takich, które wykraczają poza teoretyczne wartości. Ułatwia to zrozumienie wyników i upraszcza sposób prezentacji danych.

Najczęściej wykorzystuje się do tego rodzaju analiz arkusza kalkulacyjnego Microsoft Excel. Używany przy codziennych obliczeniach statystycznych pozwala oszczędzić zarówno czas, jak i pieniądze. Niestety, jest to program ogólny i aby wykorzystać jego możliwości oraz uzyskać wyniki analiz w sposób zadowalający, niezbędna jest wiedza programistyczna. Problemem jest również wyświetlanie danych oraz wpływ ustawień systemowych komputera na stworzone funkcje oraz często pojawiający się problem z datami czy liczbami zmiennoprzecinkowymi. Ogrom funkcji, które należy utworzyć oraz czas, jaki trzeba na to poświęcić, skłania do zainteresowania się programami, które zostały stworzone tylko i wyłącznie do obliczeń statystycznych.

Jednym z pionierów na rynku jest firma Statsoft, której programy z rodziny Statistica wykorzystywane są na całym świecie przez naukowców, uczelnie, firmy pragnące przeanalizować dane. Programy te zyskały popularność dzięki możliwościom zastosowania w wielu dziedzinach (prognozowanie i zarządzanie, badania rynku i badania marketingowe, przemysł, sterowanie i zarządzanie jakością sześć sigma, technika i badania innowacyjne, analizy

ekonomiczne i społeczne, bankowość i finanse, ubezpieczenia, medycyna, farmacja, badania naukowe i wiele innych)[1].

Naturalne stało się, iż zaczęły pojawiać się na rynku programy do analiz zbiorów danych przy użyciu prawa Benforda. Mowa tu o programie EZ-R Stats for Excel (for Windows), Web CAAT, DATAS2009 czy Benford's Law Utility. W poniższych podrozdziałach przedstawiono ich funkcjonalność, zalety oraz wady.

3.2. Program EZ-R Stats for Excel

EZ-R Stats for Excel jest darmowym dodatkiem, który został przygotowany z myślą o użytkownikach programu Excel 2003. Może być używany w każdym celu, zarówno komercyjnym, jak i prywatnym.

Nazwa programu pochodzi od angielskiego stwierdzenia „Easier Stats", co w wolnym tłumaczeniu może być rozumiane jako „łatwa statystyka".

Program ułatwia i wspiera analizę danych, dostarczając licznych narzędzi i wykresów, wykorzystując obiekty wewnątrz MS Excel.

Jego głównymi zaletami są:

- sterowanie przez menu z poziomu programu Excel,
- kompatybilność z innymi produktami Microsoft Office,
- posiadanie skomplikowanych funkcji analitycznych,
- dostępność na licencji *freeware*.

Wedle producenta EZ-R Stats (EZSW) usprawnia i upraszcza proces analizy danych, a tym samym pozwala analitykom wykorzystać swój czas bardziej efektywnie.

Jest to możliwe dzięki[2]:

- zmniejszeniu czasu potrzebnego na opracowanie specyfikacji kodu komputerowego potrzebnego do analizowania danych,
- zmniejszeniu czasu potrzebnego na naukę języka programowania i czasu potrzebnego na zaprogramowanie potrzebnych funkcjonalności,
- obniżeniu poziomu doświadczenia wymaganego do opracowania skutecznych analitycznych specyfikacji kodu,
- wykorzystaniu standardowych formatów plików, aby uprościć przekaz danych pomiędzy systemami.

[1] Źródło: statsoft.pl.

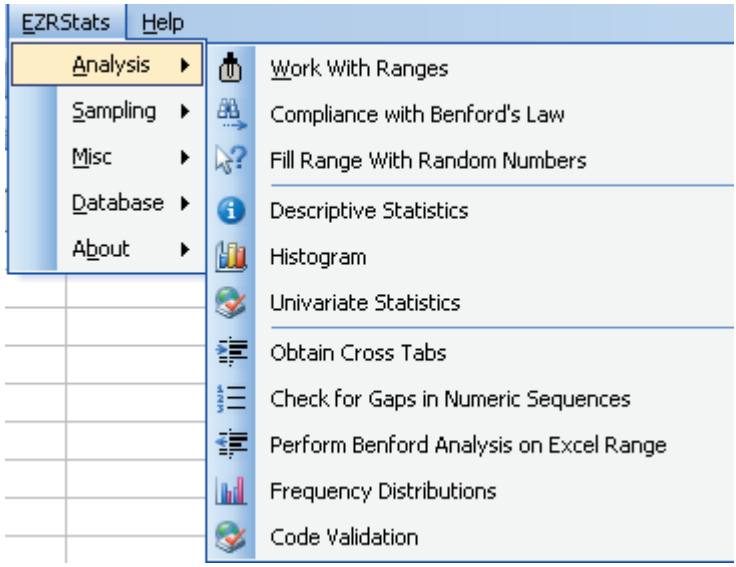
[2] Źródło: http://www2.statistics.com/resources/software/commercial/e/Ezrstat.php.

Zarys funkcji i użyteczności

Instalacja programu przebiega bardzo sprawnie i intuicyjnie, a instalator przedstawia wszystkie niezbędne kroki do poprawnego funkcjonowania dodatku. Niepodważalną zaletą programu jest *Menu*, które zostaje dodane do Excela. Czytelne i jednoznaczne nazwy od razu sugerują możliwości wykorzystania danej funkcjonalności w analizie.

Osoby przyzwyczajone do widoku ikon funkcyjnych, zwrócą uwagę na pojawienie się niektórych z nich tuż obok nazw.

Rys. 3.1. Menu programu EZ-R for Excel



Liczba opracowanych funkcji wymaga zapoznania się z dokumentacją, która jasno i czytelnie opisuje każdą z nich. Uwagę zwraca również fakt, że zawiera ona przykłady wraz ze zdjęciami, dzięki czemu użytkownik szybciej rozumie możliwości programu.

Analizę rozpocząć można od funkcjonalności o nazwie *Work With Ranges*, która dostarcza wielu ciekawych funkcji do przygotowywania danych. Niektóre z możliwości to:

- usuwanie i dodawanie kolorów w komórce o odpowiedniej wartości,
- usuwanie komentarzy i pustych znaków,
- sumowanie, usuwanie, sortowanie pogrubionych wartości oraz wskazanie duplikatów liczb.

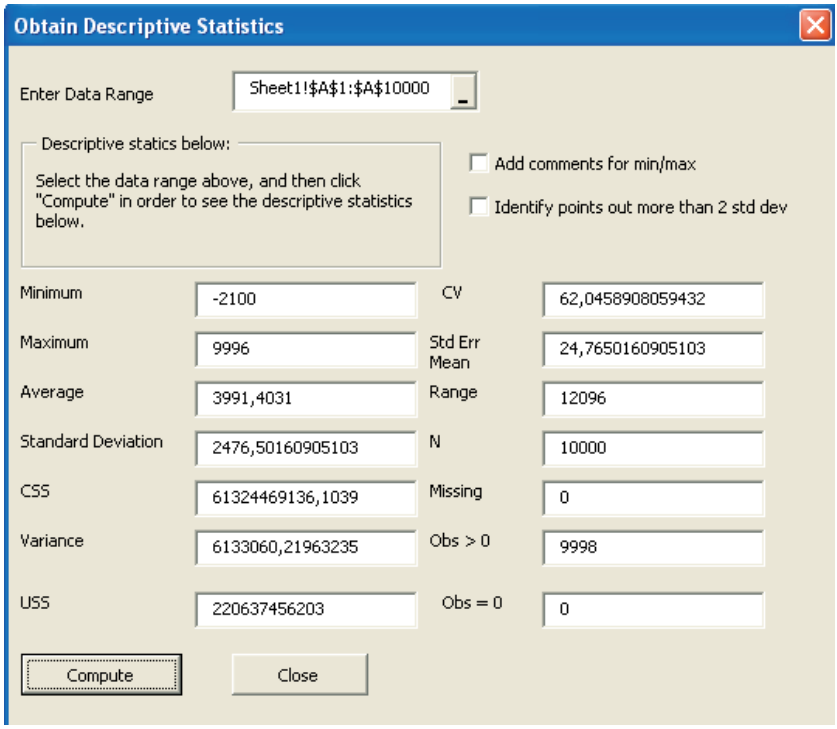
Pozornie wydaje się, że są to zabiegi kosmetyczne, jednak w momencie pracy nad zbiorem danych zaczyna się doceniać ich użyteczność, a prezentowane dane stają się bardziej czytelne.

Kolejną zaletą jest możliwość podłączenia się do bazy danych i analizowania danych bez wcześniejszego eksportu. Dzięki takiemu zabiegowi zapewniona jest pełna kontrola nad danymi umożliwiając szybkie pobieranie, sortowanie, analizowanie, sumowanie i raportowanie wyników. Pozwala na łączenie danych z różnych plików, eliminując potrzebę dwukrotnego wprowadzania tych samych informacji.

Uwagę zwracają również zróżnicowane obliczenia statystyczne względem zakresu danych, jaki został wzięty pod uwagę. W kilka sekund użytownik może uzyskać podstawowe informacje o wielkościach mierników statystycznych takich, jak:

- średnia arytmetyczna,
- odchylenie standardowe,
- wariancja,
- minimalna/maksymalna wartość,
- liczności dodatnich i ujemnych liczb.

Rys. 3.2. Wynik obliczeń podstawowych mierników statystycznych



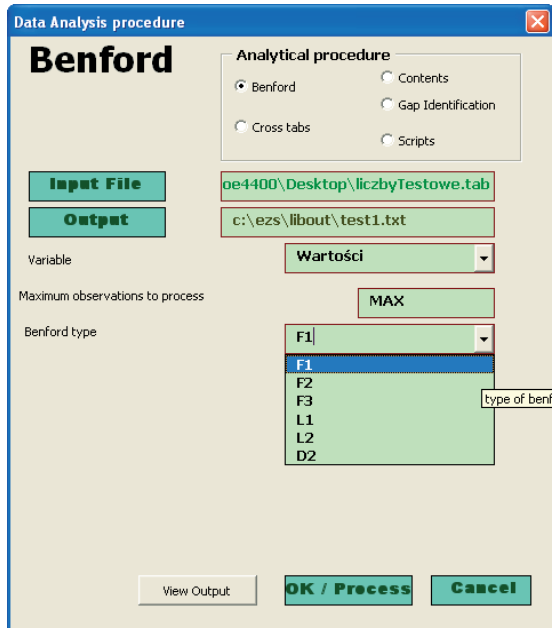
Funkcje i użyteczność programu w zastosowaniu do prawa Benforda

Do analizy prawa Benforda EZ-R Stats używa dwóch funkcjonalności. Pierwsza pozwala nam obliczać rozkład, pobierając dane z pliku o rozszerzeniu *.tab, a druga analizuje liczby znajdujące się w komórkach, które zdefiniuje użytkownik.

Przed analizą, za pomocą pierwszej z nich mamy możliwość wybrania, czy chcemy analizować cały plik, czy tylko wartości do wskazanego n-tego rekordu w pliku. Program sam daje możliwość wybrania rozkładu do przeanalizowania:

- F1 – analiza pierwszej cyfry znaczącej,
- F2 – analiza dwóch pierwszych cyfr w liczbie,
- F3 – analiza pierwszych trzech cyfr w liczbie,
- D2 – analiza tylko drugiej cyfry w liczbie,
- L1 – analiza ostatniej cyfry w liczbie,
- L2 – analiza dwóch ostatnich cyfr w liczbie.

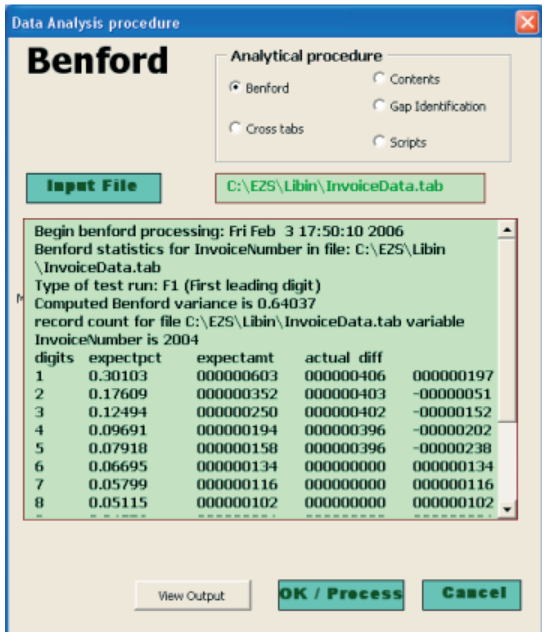
Rys. 3.3. Ustalanie warunków analizy



Po analizie pliku wejściowego, raport końcowy jest zapisywany do pliku z informacją zwrotną. Raport ten pokazuje zarówno oczekiwane, jak i obserwowane wartości oraz szacuje prawdopodobieństwo wskazujące, czy odchylenie od oczekiwanej wartości jest dziełem przypadku, czy też nie.

Prawdopodobieństwo liczone jest za pomocą testu chi-kwadrat i znajduje się w przedziale od 0,00004 (mało prawdopodobne) do 0,99999 (bardzo prawdopodobne).

Rys. 3.4. Wyniki analizy



Pomimo tego, że program zawiera możliwość analizy zbioru danych za pomocą rozkładu Benforda, przedstawione wyniki i obliczone mierniki mają charakter podstawowy. Brak spójnych, eleganckich i wyczerpujących raportów skłania do poszukiwań innych, doskonalszych programów. Minusem EZ-R for Excel jest również brak wartości procentowych dla rozkładów oraz mało precyzyjne wykresy.

Rodzi się konkluzja, że nazwę programu „Łatwa statystyka", można rozumieć dosłownie, ponieważ wykorzystanie podstawowych mierników nie przyda się w zastosowaniu go w szerszej analizie.

Często zdarza się, że po wstępnej analizie zbioru danych trzeba porzucić pracę i dopiero po jakimś czasie do niej powrócić, aby uzyskać wyniki. Program nie daje możliwości jednolitego wglądu, a przede wszystkim nie opisuje, co udało się zrobić do tej pory.

Sam fakt zapisywania części analizy może nie jest tak potrzebny (zawsze możemy przeanalizować dane jeszcze raz), jak dokładny opis danych na wykresach, dostępność większej liczby współczynników, czy czytelność uzy-

skanych wyników. Co więcej, często możemy spotkać się z komunikatem o błędach, który to komunikat nie zawiera istotnych informacji z punktu widzenia użytkownika, ale to uniemożliwia wywołanie niektórych funkcji.

Program może być przede wszystkim używany do realizacji typowych funkcji analitycznych, tj.: statystyk ludności, rozkładów częstotliwości, histogramów, analizy Benforda, identyfikacji luk, czy wyboru losowych próbek tych samych wzorców. Należy też wziąć pod uwagę fakt, że prezentowane wyniki nie są zbyt czytelne, a program przestał być rozwijany.

3.3. Web CAAT (Computer Assisted Audit Tool)

Produktem tej samej firmy, co program EZ-R Stats for Excel jest program Web CAAT, który został rozszerzony o szereg nowych funkcji – wedle producenta zawiera on ponad 100 procedur potrzebnych do analizy danych, które mają pomóc w procesie audytu. Nazwa jest akronimem „Computer Assisted Audit Tool", co oznacza *komputerowe narzędzie wspomagające audyt*.

Innowacyjność programu polega na tym, że jest on oparty na przeglądarce internetowej. W zależności od tego, czy chcemy go używać online, w środowisku intranetowym, czy w wersji stacjonarnej – program daje taką możliwość bez konieczności instalacji dodatkowych komponentów. Co więcej zawiera on wbudowaną bazę danych MySQL, która wspomaga utrzymywanie, przechowywanie, modyfikację i proces audytu danych.

Jego podstawowe zalety, to:

- aplikacja dostępna z poziomu przeglądarki internetowej,
- wbudowane funkcje analityczne,
- darmowa licencja,
- oparcie na bazie MySQL, dzięki czemu wyniki analiz dostępne są dla każdego użytkownika mającego dostęp do tabeli,
- szybkość wykonywania obliczeń.

Funkcje w procesie audytu

Proces instalacji jest bardzo prosty, wystarczy ściągnąć spakowaną wersję ze strony producenta[3], rozpakować i korzystać z programu. Po otwarciu, należy zalogować się do bazy, którą system automatycznie dostarcza wraz z pakietem instalacyjnym.

Główną zaletą programu jest wspomniane wykorzystanie bazy danych, przez co obliczenia trwają bardzo krótko, a możliwości modyfikacji danych są znacznie prostsze. Oczywiście, trzeba mieć podstawową znajomość baz

[3] www.ezrstats.com.

danych, aby zrozumieć, co dzieje się w programie, jak działa, jakie aktualnie wykonuje instrukcje i jaki jest ich rezultat.

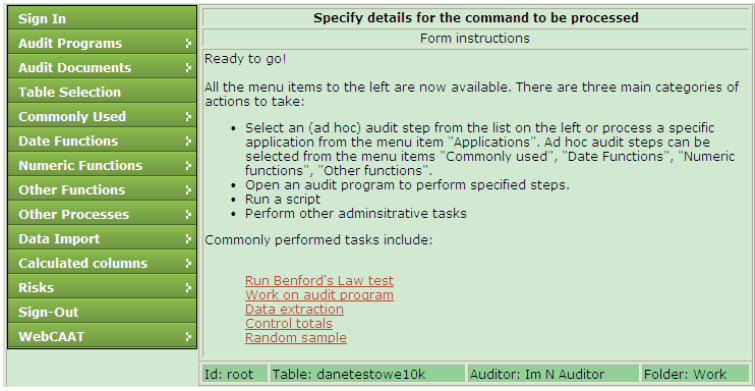
Zakres możliwości programu jest bardzo szeroki i mogą one zostać wykorzystane między innymi do poszukiwania nieprawidłowości w:

- inwestycjach kupna/sprzedaży,
- sprawdzaniu rejestrów min. zamówień,
- historii cen i sprzedaży,
- jednostkowych kosztach zapasów,
- informacjach o przelewach,
- wartościach polisy ubezpieczeniowej,
- aktywach i pasywach konta,
- raportach wydatków,
- inwentarzu.

Praca z programem rozpoczyna się od zalogowania do systemu, a następnie wybrania tabeli z danymi, które będą analizowane. Jeśli dane nie zostały jeszcze załadowane, mamy możliwość importu do nowo utworzonej tabeli bądź do tabeli już istniejącej.

Poprzez zakładkę *Table Selection* wybieramy załadowany zbiór danych, po czym otrzymujemy komunikat informujący o możliwości rozpoczęcia pracy.

Rys. 3.5. Ekran wskazujący na gotowość programu Web CAAT do pracy



Niewątpliwą zaletą tego programu, oprócz przechowywania danych, jest możliwość zebrania i utrzymywania na serwerze dokumentacji, raportów, czy wyników obliczeń aplikacji pochodnych. Dzięki takiemu rozwiązaniu użytkownik przechowuje dane w jednym miejscu, a jeżeli korzysta z aplikacji poprzez internet, ma możliwość wglądu w tworzone raporty i wyniki z każdego miejsca i o każdej porze.

Program, podobnie jak jego poprzednik, oblicza między innymi takie mierniki statystyczne jak:

- średnia arytmetyczna,
- odchylenie standardowe,
- wariancja,
- minimalna, maksymalna wartość,
- liczności dodatnich i ujemnych liczb,
- generowanie histogramów..

Rys. 3.6. Wyniki podstawowych mierników statystycznych w programie Web CAAT

Totals	39,915,956,491.00
Count	100,001
Minimum	0.00
Maximum	999,991.00
Average	399,155.57
Standard Deviation	247,657.43
Debit totals	39,915,956,491.00
Debit Count	100,000
Credit totals	0.00
Credit Count	0
Count of Zeros	1
Net totals	39,915,956,491.00
Counts	100,001

Na powyższym rysunku widać raport końcowy po obliczeniu podstawowych mierników. Od razu można zauważyć podobieństwo do prezentacji wyników z poprzedniej wersji programu, gdzie dane również wyświetlane są za pomocą prostego i tabelarycznego układu. Niestety, również tutaj nie ma możliwości dalszej pracy nad wynikami i aby móc chociażby skopiować dane, użytkownik musi wykonać eksport do pliku.

Wykorzystanie w analizie prawa Benforda

W przypadku analizy danych w oparciu o prawo Benforda, program również nie wypada najlepiej. Choć proces analizy przebiega bardzo szybko i wyniki otrzymujemy w ciągu paru sekund (dla zbioru 100 000 liczb analiza trwała niecałe 2 sekundy), to jednak sposób ich przedstawienia oraz brak możliwości pracy nad wynikami, nie pogłębia wiedzy na temat analizy danych za pomocą prawa Benforda.

Podobnie jak w poprzednim programie, analizie mogą podlegać poniższe rozkłady:

- F1 – analiza pierwszej cyfry znaczącej,
- F2 – analiza dwóch pierwszych cyfr w liczbie,
- F3 – analiza pierwszych trzech cyfr w liczbie,
- D2 – analiza tylko drugiej cyfry w liczbie,
- L1 – analiza ostatniej cyfry w liczbie,
- L2 – analiza dwóch ostatnich cyfr w liczbie.

Dodatkowo, przed rozpoczęciem analizy możliwe jest zawężenie sprawdzanych danych np. do rekordów pochodzących z jednego miasta, regionu, czy takich, które zostały utworzone przez daną jednostkę rozliczeniową, dzięki czemu badany zbiór może być modelowany wedle potrzeb.

Rys. 3.7. Wyniki analizy rozkładu Benforda w programie Web CAAT

Digit	Theoretical	Observed
1	30,103	30,100
2	17,609	17,600
3	12,494	12,500
4	9,691	9,700
5	7,918	7,900
6	6,695	6,700
7	5,799	5,800
8	5,115	5,100
9	4,576	4,600
Chisq		0.115302041677
DF		8
Critical value at 95%		2.17973074725

Powyższy rysunek przedstawia wyniki obliczeń dla sprawdzanego zbioru. Oprócz miernika chi-kwadrat i wartości rozkładów, nie dostajemy pełniejszych informacji.

Program nie jest w stanie przedstawić wyników w formie wizualnej, czyli za pomocą wykresów. Z informacji zawartych na stronie producenta wynika, że ma on pomagać między innymi w kontroli wewnętrznej, czy w audycie. Prezentacja wyników w postaci zwięzłego raportu wraz wykresami powinna być standardem w każdym programie, który został stworzony dla takich potrzeb.

Firma EZ-R Stats LLC skupiła się przede wszystkim na funkcjach audytorskich, które mogą być pomocne dla osób zajmujących się audytem i kontrolą. Biorąc pod uwagę fakt szybkości realizacji wbudowanych funkcji oraz ich różnorodność, program rzeczywiście stanowi narzędzie wspomagające tego rodzaju operacje. Niemniej jednak prezentacja danych nie jest rozwiązana w sposób zadowalający, a sam interfejs i wyświetlanie informacji pomocniczych nie zachęca do dłuższej pracy.

Niewątpliwie wykorzystanie bazy danych i wbudowanie dużej ilości funkcji było bardzo dobrym pomysłem. Osoby mobilne, które często znajdują się poza stałym miejscem pracy, ucieszą się z możliwości korzystania z programu online jak również z jego szybkości.

Brak raportów i prezentacji w formie graficznej oraz uboga funkcjonalność wspomagająca analizę rozkładu Benforda mogą stanowić barierę przy wyborze programu Web CAAT jako kompleksowego narzędzia analitycznego.

3.4. DATAS 2009 (Digital Analysis Tests and Statistics)

Program DATAS 2009 został stworzony przez Marka Nigriniego, jednego z największych ekspertów zajmujących się prawem Benforda. Od kilkunastu lat zajmuje się on analizą zbiorów danych oraz wpływu ich przekształceń i zachowań na uzyskiwane wyniki w oparciu o rozbudowane analizy. Informacje na temat jego działalności i opublikowanych prac można znaleźć na jego stronie internetowej.

Nazwa programu oznacza cyfrowe testy analizy i statystyki, a sam program został wdrożony w 2009 roku. Zawiera on 3 odrębne arkusze kalkulacyjne z wbudowanymi poleceniami makro do obliczeń statystycznych, a wyniki prezentowane są w formie tabelarycznej i graficznej.

Jego największą zaletą jest fakt, że został on zaprojektowany do badania właściwości zbiorów tylko za pomocą rozkładu Benforda, dzięki czemu program jest ciekawym narzędziem do analizy danych poprzez wgląd na częstotliwość występowania cyfr.

Program nie jest udostępniany na licencji *freeware*, a jego koszt to wydatek rzędu 39 dolarów.

Arkusz BenfordsLawFirstSecondFirstTwo

W zależności od liczności badanych liczb program krok po kroku oblicza następujące mierniki:

- ogólny profil danych,
- rozkład F1,
- rozkład F2,
- rozkład D2,
- wykresy dla poszczególnych rozkładów,
- różnice między wartościami teoretycznymi, a badanymi,
- granice pokazujące limit odchyleń dla poszczególnych rozkładów na poziomie istotności 0,05.

Arkusz zawiera 6 zakładek: *Profile, Tables, Bounds, FirstDigits, SecondDigits, FirstTwoDigits*.

Profil danych w pierwszej zakładce oblicza sumę oraz wartość dla badanych liczb oraz określa przedziały ich zakresu. Wiąże się to z przekazaniem podstawowych informacji na temat badanego zbioru. Autor określił tę funkcję jako „lepsze poznanie swoich liczb".

Oprócz wyświetlania danych w grupach, profil danych służy również celom kontroli i zasadności badanego zbioru. Określa liczebność wystąpień w przedziałach:

- liczby większe lub równe 10;
- liczby z przedziału od 0,01 do 9,99,
- liczby równe 0,
- liczby z przedziału od –0,01 do –9,99,
- liczby mniejsze lub równe –10,

Dodatkowo system sprawdza liczebność wystąpień w przedziałach:
- od 0,01 do 50, oraz
- liczb większych od 100 000.

Rys. 3.8. Określony profil danych w programie DATAS 2009

DATA PROFILE

Details		Count	% of Total Count	$
Numbers	10.00 and over	10,000	100.00	$394,546.36
Numbers	0.01 to 9.99	0	0.00	0.00
Numbers	equal to zero	0	0.00	0.00
Numbers	-0.01 to -9.99	0	0.00	0.00
Numbers	-10.00 and under	0	0.00	0.00
		10,000	100.00	$394,546.36
Low-value numbers				
Numbers	0.01 to 50.00	6,990	69.90	$176,911.00
High-value numbers				
Numbers	100,000 and higher	0	0.00	$0.00

Ostatnie dwie pozycje sprawdzane są pod względem użyteczności dla kont płatniczych, gdzie zachodzi wiele operacji na kwoty w tych przedziałach, uczulając audytorów na wartości oraz liczebność niskich i wysokich transakcji.

Sumowanie badanego zbioru ma za zadanie porównanie badanych liczb np. z dokumentacją finansową, a zestawienie liczb w sekcji *Low/High-value numbers* ma pokazać odsetek liczb o niskiej wartości, często spotykanych w przypadku kart płatniczych i transakcji zawieranych przez pracowników na rzecz firmy.

Ustalanie liczby zerowych transakcji wiąże się natomiast z określeniem roszczeń gwarancyjnych, które często przetwarzane są jako normalne zakupy. Dodatkowo makro sprawdza liczbę wystąpień liczb ujemnych w badanym zbiorze, w którym takie dane nie powinny się znajdować.

Zakładka *Tables* przedstawia mierniki dla analizowanych rozkładów oraz porównuje uzyskane wyniki z wartościami teoretycznymi.

Rys. 3.9. Główny raport analizy w programie DATAS 2009

S	TwoDig	First2	First	Second	Digit	Count	Actual	Benford's Law	Difference	Signif	FT Digit	Count	Actual	Benford's Law	Difference	Signif	
98.99	98.99	98	9	8	1	3010	0.301	0.301	0.000	0.007	10	0	0.000	0.041	-0.041	20.754	10000
98.91	98.91	98	9	8	2	1760	0.176	0.176	0.000	0.011	11	359	0.036	0.038	-0.002	0.965	
98.9	98.9	98	9	8	3	1250	0.125	0.125	0.000	0.003	12	350	0.035	0.035	0.000	0.104	10001
98.9	98.9	98	9	8	4	970	0.097	0.097	0.000	0.014	13	396	0.040	0.032	0.007	4.176	
98.87	98.87	98	9	8	5	790	0.079	0.079	0.000	0.048	14	398	0.040	0.030	0.010	5.743	9999
98.83	98.83	98	9	8	6	670	0.067	0.067	0.000	0.000	15	377	0.038	0.028	0.010	5.828	
98.78	98.78	98	9	8	7	580	0.058	0.058	0.000	0.004	16	382	0.038	0.026	0.012	7.382	9998
98.76	98.76	98	9	8	8	510	0.051	0.051	0.000	0.045	17	346	0.035	0.025	0.010	6.254	
98.7	98.7	98	9	8	9	460	0.046	0.046	0.000	0.091	18	402	0.040	0.023	0.017	11.009	10000
98.69	98.69	98	9	8			Actual	Benford's Law			19	0	0.000	0.022	-0.022	15.062	
98.68	98.68	98	9	8	0	0	0.000	0.120	-0.120	36.856	20	0	0.000	0.021	-0.021	14.679	
98.67	98.67	98	9	8	1	1208	0.121	0.114	0.007	2.159	21	205	0.021	0.020	0.000	0.178	
98.67	98.67	98	9	8	2	1228	0.123	0.109	0.014	4.473	22	209	0.021	0.019	0.002	1.119	
98.63	98.63	98	9	8	3	1302	0.130	0.104	0.026	8.447	23	238	0.024	0.018	0.005	3.913	
98.62	98.62	98	9	8	4	1256	0.126	0.100	0.025	8.402	24	231	0.023	0.018	0.005	4.031	
98.59	98.59	98	9	8	5	1286	0.129	0.097	0.032	10.784	25	228	0.023	0.017	0.006	4.421	
98.57	98.57	98	9	8	6	1246	0.125	0.093	0.031	10.717	26	217	0.022	0.016	0.005	4.143	
98.57	98.57	98	9	8	7	1203	0.120	0.090	0.030	10.430	27	214	0.021	0.016	0.006	4.460	
98.56	98.56	98	9	8	8	1271	0.127	0.088	0.040	13.967	28	218	0.022	0.015	0.007	5.314	
98.56	98.56	98	9	8	9	0	0.000	0.085	-0.085	30.461	29	0	0.000	0.015	-0.015	12.181	

Ujęcie wartości w formie tabelarycznej oraz umieszczenie podstawowych mierników na jednej stronie pozwala szybko sprawdzić zależności pomiędzy badanymi liczbami.

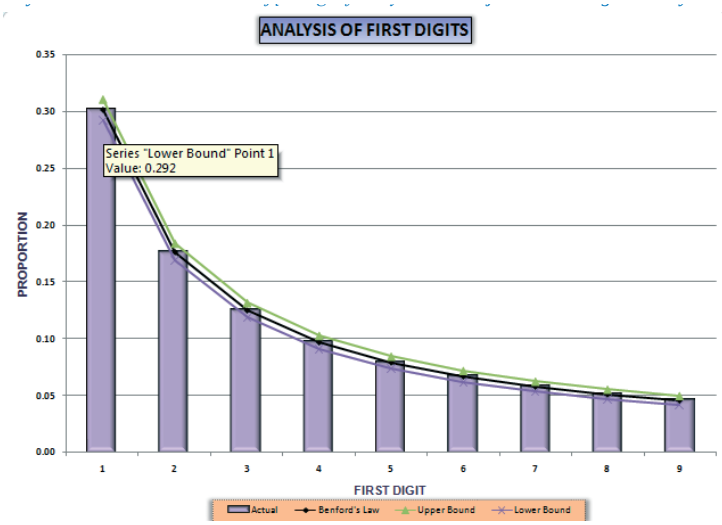
Zakładka *Bounds* ma za zadanie wskazanie przedziału granicznego, w którym audytor, analizując swoje dane, powinien się poruszać. Dodatkowo przedstawiany jest wynik dla testu ***z***.

Rys. 3.10. Analiza przedziałów granicznych w programie DATAS 2009

Upper Bound	Lower Bound	Upper Bound	Lower Bound			Z-stat	First	First-Two
0.310	0.292	0.045	0.037			1.9600	0.0000	0.0414
0.184	0.169	0.042	0.034				0.0001	0.0019
0.131	0.118	0.038	0.031	**Mean Absolute Deviation**			0.0001	0.0002
0.103	0.091	0.036	0.029		**Your data**		0.0001	0.0074
0.085	0.074	0.033	0.027	First	0.0001		0.0002	0.0098
0.072	0.062	0.031	0.025				0.0001	0.0097
0.063	0.053	0.030	0.023	Second	0.0409		0.0000	0.0119
0.056	0.047	0.028	0.022				0.0002	0.0098
0.050	0.042	0.026	0.020	First-Two	0.0046		0.0002	0.0167
Upper Bound	Lower Bound	0.025	0.019				Second	0.0223
0.126	0.113	0.024	0.018				0.1197	0.0212
0.120	0.108	0.023	0.017				0.0069	0.0003
0.115	0.103	0.022	0.017				0.0140	0.0016
0.110	0.098	0.021	0.016				0.0259	0.0053
0.106	0.094	0.020	0.015				0.0253	0.0054
0.103	0.091	0.020	0.014				0.0319	0.0058
0.099	0.088	0.019	0.014				0.0312	0.0053
0.096	0.085	0.018	0.013				0.0300	0.0056
0.093	0.082	0.018	0.013				0.0395	0.0066
0.091	0.079	0.017	0.012				0.0850	0.0147

Kolejne zakładki stanowią interpretację graficzną wartości obliczonych w powyższych tabelach.

Rys. 3.11. Rozkład F1 w ujęciu graficznym wraz z jednostkami granicznymi



Powyższy rysunek przedstawia jeden z trzech wykresów, które są tworzone podczas sprawdzania rozkładu Benforda. Dodatkowe wykresy prezentują rozkład F2 oraz D2.

Arkusz NumberFrequencies

Arkusz ten w prosty i przystępny sposób oblicza liczebność wystąpień poszczególnych liczb w zbiorze. Test ten stanowi interesujące źródło informacji w przypadku, gdy na wykresie pojawiają się duże skoki. Dzięki niemu jesteśmy w stanie określić, jakie liczby powodują anomalie, co z kolei określi konkretne wartości faktur, które w pierwszej kolejności powinny zostać poddane dogłębniejszej analizie.

Dzięki zastosowanej metodzie, analityk sprawdzający zbiór będzie przede wszystkim zainteresowany liczbami, które:

- występują zbyt często w badanym zbiorze,
- powodują znaczne odchylenia od wartości teoretycznych,
- posiadają zaokrąglone wartości dziesiętne, np. 1,99; 1,49; 0,45,
- najprawdopodobniej zostały zaokrąglone na rzecz większych transakcji czy darowizn,
- liczby, które nie pasują do pozostałych, np. badane liczby są wartościami całkowitymi, a w zbiorze pojawiają się pojedyncze przypadki wystąpień liczb dziesiętnych,
- liczby, które pojawiły się szacunkowo częściej niż pozostałe.

Rys. 3.12. Liczebność wystąpień liczb z analizowanego zbioru

Number	Frequency
14.1	15
16.4	14
14.4	13
13.4	13
12.5	13
18.45	12
17.9	12
16	12
14.6	12
18.1	11
17.2	11
16.1	11
15.5	11
18.9	10
18.8	10
18.4	10
18.2	10
16.9	10
16.5	10
16.3	10
16.2	10
15.8	10
15.1	10
12.4	10
11.97	10
26.5	9

Zaleta tego testu polega na tym, że można go używać dla każdej waluty na świecie. Może być wykorzystywany również do sprawdzenia między innymi:

- zapasów,
- odczytów temperatury,
- roszczeń zdrowotnych,
- zwrotów biletów lotniczych,
- ilości sprzedawanego alkoholu na pokładzie,
- odczytów liczników energii elektrycznej.

Test ten został użyty przez audytorów z linii lotniczych dla danych zawierających zebrane przez stałych klientów punkty milowe. Po przeanalizowaniu danych, audytorzy doszli do wniosku, że najczęściej występującą liczbą jest 500 mil na pasażera. To jednak nie dziwi, ponieważ ta wartość przypisywana jest każdemu zarejestrowanemu pasażerowi jako minimalna wartość za każdy zakupiony przez niego lot.

Dużą częstością wystąpień była również liczba 817 mil, która zostawała przyznawana pasażerom podróżującym na jednej z głównych linii przewoźnika. Jak się okazało, większa częstość występowania tej liczby wynikała z dużej liczby lotów na tej trasie.

Kolejnym przykładem wykorzystania tego testu jest przypadek firmy z Tennessee, która użyła go do poszukiwania fikcyjnych pracowników. Audytor użył tego testu do sprawdzenia listy płac w poszukiwaniu powielonych numerów rachunków bankowych. Więcej niż jeden rachunek, na który mają wpłynąć wynagrodzenia, może być wskaźnikiem oszustwa.

Audytor odnalazł 2 przypadki, w których na liście kont istniały takie same rachunki. W pierwszym przypadku było to małżeństwo, a w drugim młodzi pracownicy, którzy wspólnie wynajmowali mieszkanie i korzystali z jednego konta. Sytuację tę pracownicy wyjaśnili niemożnością założenia rachunku bankowego dla jednego z mieszkańców[4].

Arkusz BenfordLawSecondOrderTests

Funkcjonalność i prezentacja danych tego arkusza bazuje na tym samym wzorcu, co pierwszy z opisywanych. Różnica polega na tym, że badane dane sprawdzane są za pomocą zmodyfikowanego testu D2.

Nowy test diagnozuje relacje i wzorce znalezione w danych transakcyjnych i opiera się na różnicy w cyfrach pomiędzy wartościami posortowanymi od najmniejszej do największej. M. Nigrini wykorzystał te badania do:

- analizy kwot zobowiązań,
- analizy księgi wpływów,
- analizy rocznych kosztów i przychodów.

Dzięki przeprowadzonym analizom zauważone zostały anomalie w pobieranych kwotach, zaokrąglanie danych, czy wykorzystanie danych wygenerowanych statystycznie, zamiast rzeczywistych danych transakcyjnych. Powyższe przykłady, powołując się na słowa autora, nie zostałyby wykryte, gdyby nie wprowadzenie testu porządkowego D2.

Definicja tego prawa opisana jest w sposób następujący:

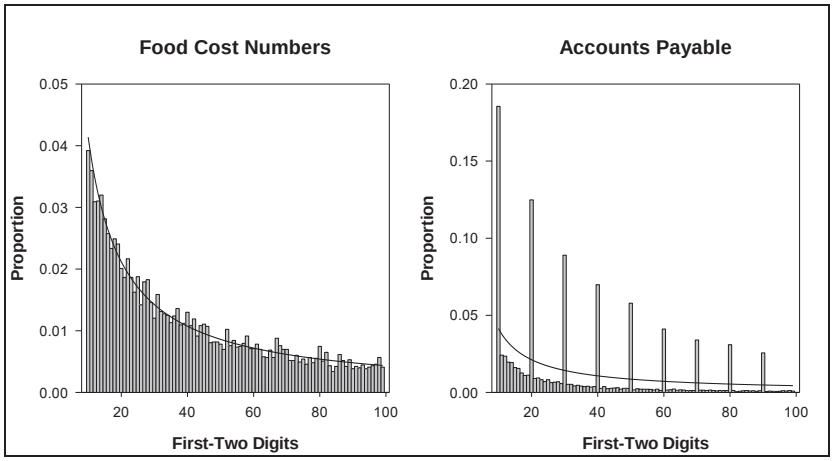
Niech x1, ..., xN będzie zestawem danych zawierającym obserwacje z danego zbioru, i niech y1, ..., yN będzie posortowanym zbiorem xi w porządku rosnącym. Następnie, dla wielu naturalnych zbiorów danych, dużej wartości dla N, cyfry różnic pomiędzy obserwacjami (yi+1 – yi) powinny być bliskie rozkładowi Benforda. W przypadku wystąpień dużych odchyleń dane, dla których pojawiły się nieprawidłowości, powinny zostać zbadane bardziej szczegółowo[5].

[4] M. Nigrini, "Program_Details_2009.doc", DATAS 2009.

[5] M. Nigrini, S. Miller, W. College, *Data diagnostics using second order tests of Benford`s Law,* A Journal of Practice and Theory, 2009.

Powołując się na słowa autora, wyniki badania tego testu powinny zmierzać do wyników przedstawionych na wykresach poniżej.

Rys. 3.13. Wykres rozkładu dla posortowanego D2[6]



3.5. Benford's Law Utility

Program ten został napisany przez Johna Morrowa w 2007 r. i udostępniony wraz z pracą zatytuowaną *Detecting Problems in Survey Data using Benford's Law*.

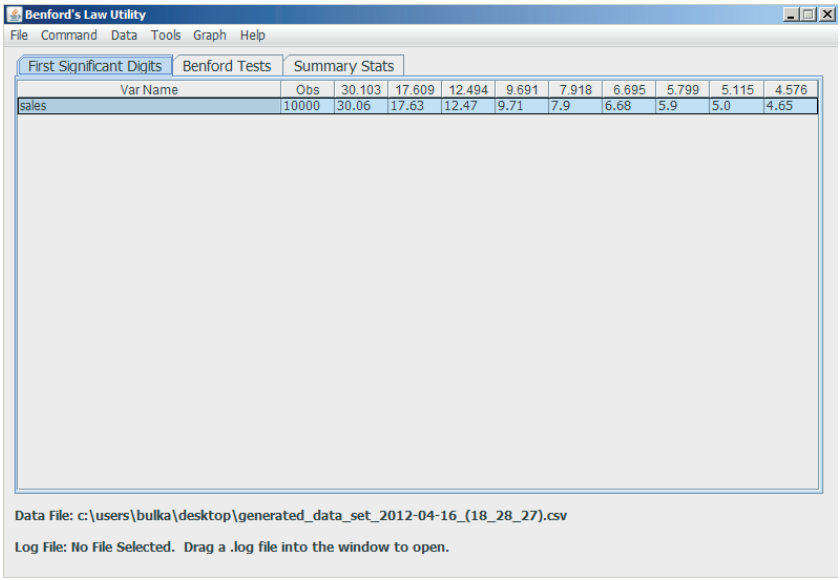
Program, tak samo, jak DATAS 2009, został przygotowany jako narzędzie specjalistyczne badające właściwości zbioru pod względem zgodności z rozkładem Benforda. W porównaniu do programu DATAS 2009, jego funkcjonalność, czas pracy oraz użyteczność wydaje się być zdecydowanie większa.

Poniżej zamieszcza sie listę mierników, których wartości są prezentowane dla rozkładu F1:

- wartość współczynnika korelacji,
- statystyka testu chi-kwadrat,
- średnia wartość dla rozkładu F1,
- test Vn,
- test D,
- wartość miernika dopasowania,
- odchylenie standardowe,
- kurtoza.

[6] M. Nigrini, S. Miller, W. College, *Data diagnostics using second order tests of Benford`s Law*, A Journal of Practice and Theory, 2009.

Rys. 3.14. Okno główne programu Benford's Law Utility

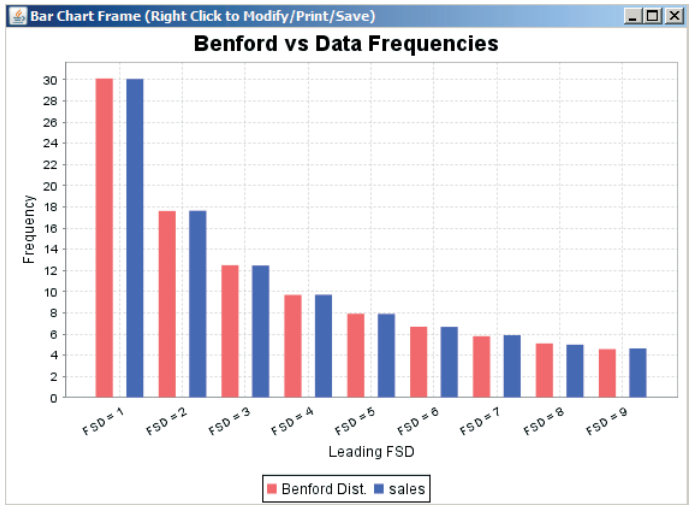


Pod względem liczby i jakości mierników, program jest bardziej precyzyjny i użyteczny od innych, a jego prosty i czytelny interfejs pozwala na szybkie wykonanie obliczeń. Możliwość przeglądania wykresów i odczytanie wartości dla poszczególnych cyfr jest nie tylko bardzo proste, ale też bardzo użyteczne.

Gdyby nie brak jednej funkcjonalności, można by powiedzieć, że program jest wręcz perfekcyjny i idealnie nadaje się jako narzędzie do analizy danych dla rozkładu Benforda. Mowa tu o możliwości wyboru, jakiego testu/testów mają dotyczyć obliczenia. Aktualnie program wykonuje obliczenia tylko dla testu F1, co przy specyficznych danych może nie być wystarczające.

Brak możliwości zastosowania większej liczby testów znacząco wpływa na użyteczność programu. Inne narzędzia analityczne umożliwiają badanie wartości nawet do sześciu pierwszych znaków (rozkład F6). Ma to ogromne znaczenie np. przy wysokich kwotach transakcji bankowych, podatków, ubezpieczeń itp. Nie każdy rodzaj danych można ograniczyć do testu F1. Niektóre specyficzne zbiory danych wymagają analizy n-tej cyfry znaczącej. Opisywany program nie stanowi więc poważnego narzędzia analitycznego. Wszystkie zaprezentowane obliczenia można wykonać w arkuszu kalkulacyjnym.

Rys. 3.15. Wykres dla rozkładu F1 w programie Benford's Law Utility



3.6. ACL (Audit Command Language)

18.10.2010 roku Ministerstwo Finansów ogłosiło przetarg na „Dostawę 427 pakietów oprogramowania do kontroli podatków prowadzących księgowość w formie elektronicznej".

Dokument SIWZ[7] zawierał szereg wymogów, które spełniać miał zakupiony system, a jednym z nich była funkcjonalność badająca dane metodą Benforda:

„Użycie Prawa Benforda (bazującego na analizie sekwencji cyfr) do wykrywania możliwych błędów, potencjalnych nadużyć lub innych nieprawidłowości".

Był to w Polsce pierwszy tak duży przetarg na dostawę systemu audytowego wykorzystującego rozkład Benforda. Zwycięzcą okazała się firma SKG S.A. z Bielska Białej ze swym produktem – ACL (Audit Command Language).

ACL stanowi gotowe rozwiązanie dla audytorów i jest systemem tzw. „pudełkowym" – dostępnym od ręki, niewymagającym specjalnego procesu wdrożenia. Przy zakupie systemu klient otrzymuje dożywotnią licencję. Sama aplikacja składa się z trzech głównych komponentów. Są nimi:

- Komponent importu danych – umożliwia import danych z wielu formatów: MS Excel, CSV, tekstowy, PDF, podglądy wydruków oraz bazy danych. Moduł ten jest zaopatrzony w prosty interfejs umożliwiający

[7] Specyfikacja istotnych warunków zamówienia.

osobom bez zaawansowanej wiedzy technicznej łatwą obsługę za pomocą przyjaznych okien. Sam system jest bezpieczny i nie ingeruje w dane. Nie ma możliwości zmiany informacji w bazie danych za pomocą ACL. Po zakończeniu importu użytkownik może zweryfikować jakość danych: wyszukać luki, zdublowane lub naruszone dane.

- Komponent analityczny – dostępnych jest kilkadziesiąt poleceń analitycznych (od prostych grupowań, po zaawansowane analizy). Komponent ten oferuje własny język programowania. Wybrane polecenia można uruchamiać w dowolnym czasie za pomocą skryptu i wykonywać je w formie zadań. Analiza rozkładu Benforda stanowi jedno z poleceń analitycznych. Istnieje możliwość łączenia poleceń – obudowywania jednych drugimi oraz budowania relacji między tabelami. Każde wykonane polecenie analityczne zapisywane jest w logu. Można odtworzyć poszczególny krok wykonywanej analizy danych.
- Komponent raportów – możliwość generowania wykresów, raportów. Każdy raport można wysłać pocztą elektroniczną na wybrany adres e-mail.

System pomocy ACL opisuje dokładnie sposób działania analizy Benforda:

Polecenie zlicza liczbę wystąpień każdej cyfry wiodącej lub kombinacji cyfr w zestawie danych i porównuje rzeczywistą liczność do oczekiwanej liczności. Oczekiwana liczność obliczana jest z zastosowaniem formuły Benforda. Poniższa lista zawiera dodatkowe informacje o analizie Benforda.

– Określić można do sześciu analizowanych cyfr wiodących. Dla analiz powyżej czterech cyfr wiodących wyniki należy zapisać do pliku zamiast wyświetlania ich na ekranie, czy wysyłania na drukarkę.

– Zależnie od przetwarzanej liczby rekordów analiza pięciu i więcej cyfr wiodących może zająć kilka minut. Niezależnie od zadanej liczby cyfr do analizy, wykonywanie polecenia można przerwać w dowolnym momencie, naciskając przycisk Esc.

– Efektywna analiza Benforda wymaga dużych zestawów danych. W wynikach ACL wyświetla ostrzeżenie, jeżeli zestaw danych jest zbyt mały dla określonej liczby cyfr wiodących.

– Nienormalne dane widoczne są lepiej podczas odrębnej analizy wartości dodatnich i ujemnych. Można zastosować filtr do rozdzielenia ich przed rozpoczęciem analizy.

– Rekordy o wartości zerowej są pomijane, a ich liczba jest raportowana. Wiodące zera, formatowanie pola numerycznego takie, jak przecinki, czy znak dolara, inne oznaczenia nienumeryczne i rekordy niespełniające kryteriów testu, również są raportowane. Jeżeli wynikowa liczba cyfr jest mniejsza niż określona, ACL dodaje zero po prawej stronie wyniku.

Wyniki analizy Benforda obejmują:

– Cyfry wiodące – Wyświetla liczbę sprawdzonych cyfr wiodących. Przykładowo, jeżeli zadano jedną cyfrę wiodącą, wyświetlane są cyfry od 1 do 9. Jeżeli zadano dwie cyfry wiodące, wyświetlane są liczby od 10 do 99.

– Liczba rzeczywista – Prezentuje wykrytą w populacji liczność, dla każdej cyfry wiodącej lub kombinacji cyfr wiodących.

– Liczba oczekiwana – Prezentuje oczekiwaną w populacji liczność dla każdej cyfry wiodącej lub kombinacji cyfr wiodących obliczoną w oparciu o formułę Benforda.

– Test z – Wyświetla wartość statystyki z dla każdej kombinacji cyfr. Statystyka z określa, jak bardzo uzyskany rezultat różni się od oczekiwanego i wyrażona jest odchyleniem standardowym. Przykładowo statystyka z o wartości 0,500 to połowa odchylenia standardowego.

– Dolna granica (opcjonalnie) – Wyświetla cyfry wiodące, dla których częstość jest znacznie niższa niż oczekiwana.

– Górna granica (opcjonalnie) – Wyświetla cyfry wiodące, dla których częstość jest znacznie wyższa niż oczekiwana[8].

ACL umożliwia łączenie się ze wszystkimi rodzajami baz danych (m.in. Oracle DB2, SQL Server) bezpośrednio lub za pomocą ODBC[9].

Obecnie system ACL w sektorze publicznym jest stosowany nie tylko w Ministerstwie Finansów. Innymi klientami są np. Ministerstwo Sprawiedliwości, Agencja Restrukturyzacji i Modernizacji Rolnictwa, Służba Celna, Najwyższa Izba Kontroli.

Poniżej zaprezentowano proces analizy za pomocą rozkładu Benforda.

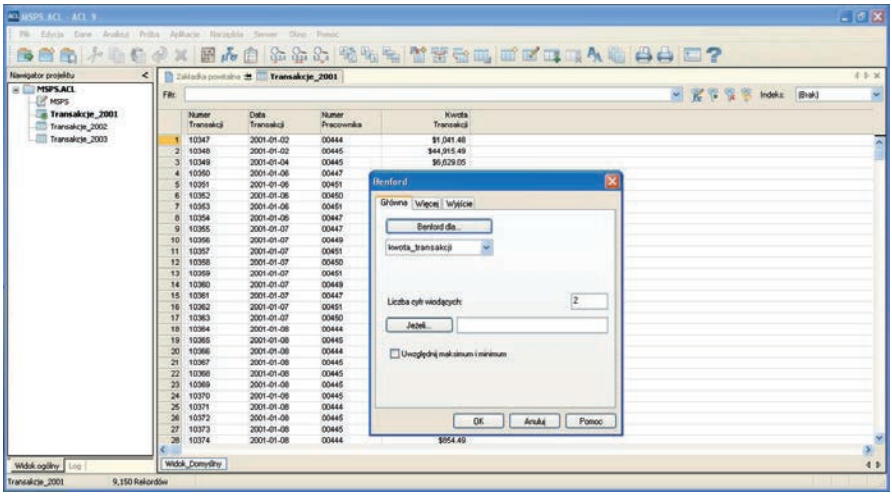
Grafika (rys. 3.16) prezentuje okno aplikacji ACL po zaimportowaniu tabeli zawierającej 9150 rekordów przedstawiających kwoty transakcji pracowników. Uruchomione zostaje polecenie Benford(), gdzie argumentem są wspomniane kwoty zakupów.

Wyniki zostają zaprezentowane w postaci czterokolumnowej tabeli zawierającej liczności wystąpienia danej kombinacji cyfr, wartość oczekiwaną wyliczoną na podstawie prawa Benforda oraz wartość testu z.

[8] Źródło: http://www.skg.pl/acl.

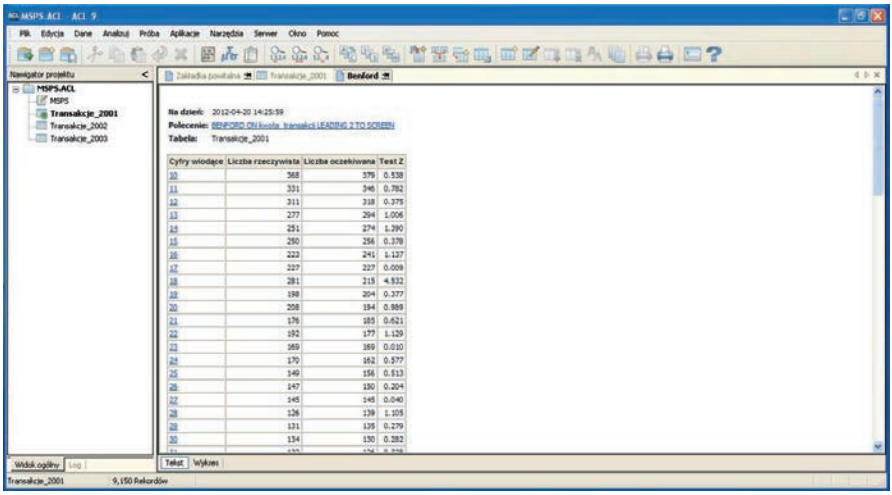
[9] Open DataBase Connectivity.

Rys. 3.16. Okno aplikacji ACL – uruchomienie analizy



Źródło: http://www.skg.pl

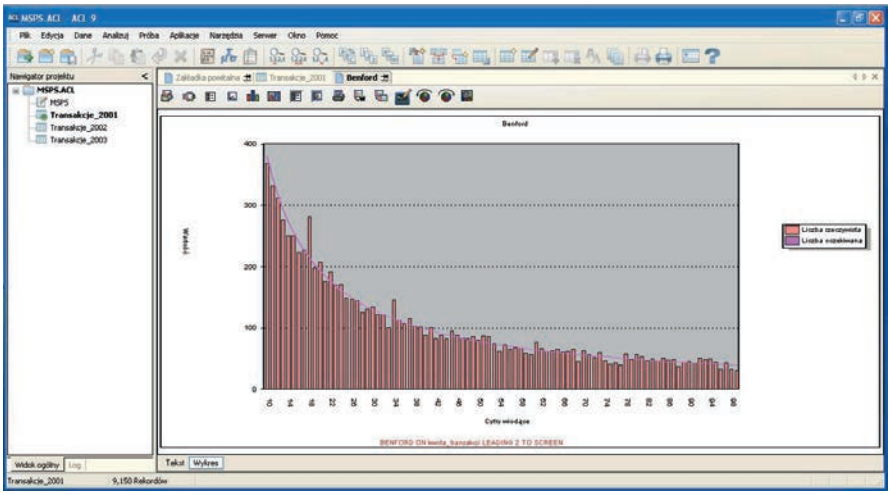
Rys. 3.17. Okno aplikacji ACL – wyniki analizy



Źródło: http://www.skg.pl

Poniżej (rys. 3.18) znajduje się ilustracja wyników analizy. Pionowe kolumny na wykresie wyrażają empiryczną wartość wystąpień kombinacji cyfr, natomiast linia ciągła stanowi jej teoretyczną wartość.

Rys. 3.18. Okno aplikacji ACL – wykres wyników analizy



Źródło: http://www.skg.pl

3.7. Uwagi dotyczące zastosowań aplikacji analitycznych

Obecnie dostępnych jest szereg aplikacji umożliwiających wykonanie analizy danych za pomocą badania rozkładów częstości cyfr znaczących. Większość z nich nie jest jednak aktualizowana. Są to przede wszystkim dodatki do arkusza kalkulacyjnego, makra lub proste programy – kalkulatory. Żadna z przedstawionych aplikacji nie umożliwia głębszej oceny wyników rozbudowanymi miernikami.

Analityk w przypadku wykrycia nieprawidłowości powinien mieć możliwość dotarcia do źródeł powodujących zakłócenia.

Niezbędną właściwością dobrego narzędzia powinna być spójna prezentacja danych z możliwością raportowania, drukowania, zapisywania czy modyfikowania uzyskanych wyników. Brak takiej funkcjonalności powoduje, że prezentowane dane są nieczytelne i nie ma możliwości edytowania oraz łączenia ich z wynikami pochodzącymi z innych analiz.

Grupa badawcza skupiona wokół Uniwersytetu Ekonomicznego w Krakowie pracuje nad wersją oprogramowania *stand-alone* oraz wersją online, która będzie stanowić centrum informacji na temat analizowanego zbioru. Celem jest uzyskanie, w zależności od potrzeb i posiadanych danych, jasnych i czytelnych wyników zawierających opisy poszczególnych mierników i sposób ich interpretacji. Zalążkiem takiego oprogramowania jest niżej opisane makro w MS Excel.

Rozdział 4
Propozycja narzędzia analitycznego

4.1. Przygotowanie danych

Dane do analizy powinny być przygotowane w odrębnym arkuszu MS Excel oraz spełniać wymogi wynikające z założeń i warunków działania prawa Benforda. W szczególności należy tu zwrócić uwagę na:

1. dużą liczbę obserwacji, rzędu co najmniej 1000 i więcej,
2. szeroki zakres wartości, rzędu co najmniej 4 cyfr znaczących, tj. maksymalna wartość powinna przekraczać wielkość 10 000,
3. wysoka zmienność obserwacji mierzona np. współczynnikiem zmienności, rzędu co najmniej 70%.

Są to dolne ograniczenia, sformułowane na podstawie doświadczeń autorów. Należy jednak podkreślić, że nie ma w tym zakresie precyzyjnych badań, na podstawie których można byłoby sformułować bardziej jednoznaczne wskazówki dotyczące pożądanych własności zbiorów danych dla potrzeb analizy z wykorzystaniem prawa Benforda.

Warto wspomnieć o zasadach postępowania w przypadku pojawienia się wartości nietypowych lub budzących wątpliwości.

1. W przypadku braku danych, kiedy w materiałach źródłowych są informacje typu [b.i.], [.]. [-], [*] wszystkie je należy pomijać.
2. Wartości zerowe także należy najczęściej wyeliminować z analiz.
3. Należy zdecydować o rzędzie dokładności danych. Jeżeli np. część danych jest wyrażona z dokładnością do pełnych złotych, a część z dokładnością do groszy, to warto zaokrąglić wszystkie dane do pełnych złotych. Jeżeli wszystkie dane wyrażone są z dokładnością do grosza, to warto je przemnożyć przez 100, aby pozbyć się ułamka dziesiętnego.
4. W przypadku wartości ujemnych (np. wynik finansowy), rozwiązaniem jest ustalenie największej ujemnej wartości oraz zwiększenie każdej obserwacji o tę wielkość przemnożoną przez arbitralnie ustaloną stałą, np. 1,1 (aby uniknąć wartości zerowych).
5. Analizę można prowadzić tylko na danych w skali ilorazowej. Pomiary na skalach słabszych z reguły nie nadają się do analiz przy użyciu prawa Benforda.

Ciekawym kierunkiem badań może być ***analiza warstwowa***. Jej istota polega na dzieleniu zbioru danych na warstwy. Liczba warstw może wynosić od 2 do 5. Przy dwóch warstwach dzieli się zbiór na dwa równe podzbiory (według mediany) i przeprowadza się analizę odrębnie dla obserwacji mniejszych od mediany, i odrębnie dla obserwacji większych od mediany.

Przy trzech warstwach dzieli się zbiór na trzy równe podzbiory przy pomocy odpowiednich kwantyli, itd. Obserwacja stopnia zgodności z rozkładem Benforda w poszczególnych warstwach pozwala na wyciągnięcie z analizy bardziej uzasadnionych wniosków.

Odmianą analizy warstwowej jest analiza danych przekrojowo-czasowych. Zazwyczaj te dane ujęte są w postaci macierzy, w których kolejne kolumny zawierają informacje z różnych okresów czasu, a kolejne wiersze – obserwacje dla różnych obiektów (miast, firm, miejsc pomiarowych itp.). W takich sytuacjach także warto stworzyć odrębne warstwy analizy nie tylko wg obiektów, ale także wg okresów czasu, np. odrębna analiza dla danych z początkowych okresów i odrębna dla danych z końcowych okresów.

Istotą analizy rozkładu Benforda jest potwierdzanie wiarygodności danych i poszukiwanie ewentualnych nieprawidłowości w zbiorach informacji liczbowych. Te negatywne zjawiska mogą wystąpić w określonym miejscu i czasie, stąd analiza dla pełnego zakresu danych niekiedy może nie wskazywać nieprawidłowości, gdyż dane poprawne zostaną przemieszane z danymi niepoprawnymi.

Dysponowanie uniwersalnym narzędziem analitycznym pozwala też na analizy symulacyjne. Można np. wygenerować sztuczne zbiory danych liczbowych, które w 100% będą spełniały kryteria rozkładu Benforda. Następnie wprowadzić do tego generatora funkcję zakłócającą strukturę danych w coraz to większym stopniu i obserwować, jak zachowują się poszczególne miary zgodności rozkładów i odpowiadające im wykresy.

4.2. Sposób uruchamiania

Po uruchomieniu makra i odblokowaniu zabezpieczeń, użytkownik znajduje się w arkuszu BF_INPUT, do którego wprowadza się dane wejściowe podlegające analizie.

W makro przewidziano następujące opcje analizy:

- Struktura ujęcia danych:
  - w kolumnie,
  - w postaci macierzy.
- Zakres analizy:
  - tylko testy F1, D2, D3 i L1,
  - dodatkowo test F2 i/lub test F3.
- Zakres analizowanych danych
  - cały dostępny zbiór,
  - bez zadanej części (w %) najmniejszych i/lub największych obserwacji.

Dane wprowadza się (kopiując je z innego arkusza źródłowego) począwszy od komórki a10. W przypadku, gdy dane ujęte są w postaci macierzy, to w komórkach g2 i g3 należy podać także adres ostatniego elementu macierzy danych (prawy dolny róg). Struktura macierzowa przydatna jest w sytuacjach, gdy analizie podlegają dane przekrojowo-czasowe.

Zaleca się stosowanie wszystkich testów (wraz z testami F2 i F3), jakkolwiek czas wykonania makro jest wtedy wyraźnie dłuższy. Dobrym rozwiązaniem w takim przypadku jest praca „wsadowa" polegająca na uruchomieniu kilku analiz jednocześnie i zaplanowanie dłuższej przerwy czasowej związanej z oczekiwaniem na wyniki obliczeń.

Ostatnia możliwa opcja analizy dotyczy zakresu przetwarzanych danych. Standardem jest analiza całego dostępnego zboru. Niekiedy jednak przydatna jest eliminacja najmniejszych lub największych obserwacji. W makro ustala się w tym celu dwa parametry (komórki i3 oraz i4) określające procentowo liczbę eliminowanych obserwacji. Można eliminować: tylko najmniejsze, tylko największe lub zarówno najmniejsze, jak i największe obserwacje na różnych poziomach eliminacji.

Makro uruchamiane jest przyciskami *Analyze!* Wyniki analizy podawane są w sąsiednim arkuszu BF_OUTPUT_H.

Ponadto, w arkuszu BF_DATA znajduje się czerwony przycisk *Clear data*, który umożliwia czyszczenie wprowadzonych danych z arkusza i umożliwia wprowadzenie nowego zbioru danych do analizy. Identyczny przycisk znajduje się w arkuszu BF_OUTPUT_H. Powoduje on eliminację istniejących w arkuszu wyników analizy.

W arkuszu BF_DATA można wprowadzić także informacje opisujące analizowany zbiór danych:

- nazwa cechy – komórka B5,
- symbol cechy – komórka A6,
- źródło danych – komórka C6.

Informacje te kopiowane są do tabeli z rezultatami analiz i w przypadku obszernych badań ułatwiają identyfikację wyników.

Przykład ekranu z parametrami określającymi zakres analizy przedstawiono na rysunku 4.1.

4.3. Zakres analizy

W ramach makro uwzględniono sześć typowych testów, w których analizowane są rozkłady częstości pojawiania się następujących kombinacji cyfr:

1. pierwsza znacząca cyfra– test F1,
2. dwie pierwsze znaczące cyfry – test F2,

3. trzy pierwsze znaczące cyfry – test F3,
4. dokładnie druga cyfra – test D2,
5. dokładnie trzecia cyfra – test D3,
6. ostatnia cyfra – test L1.

Obliczenia dla testów F2 i F3 wykonywane są opcjonalnie, ale wskazane jest uwzględnianie w analizach wszystkich 6 testów.

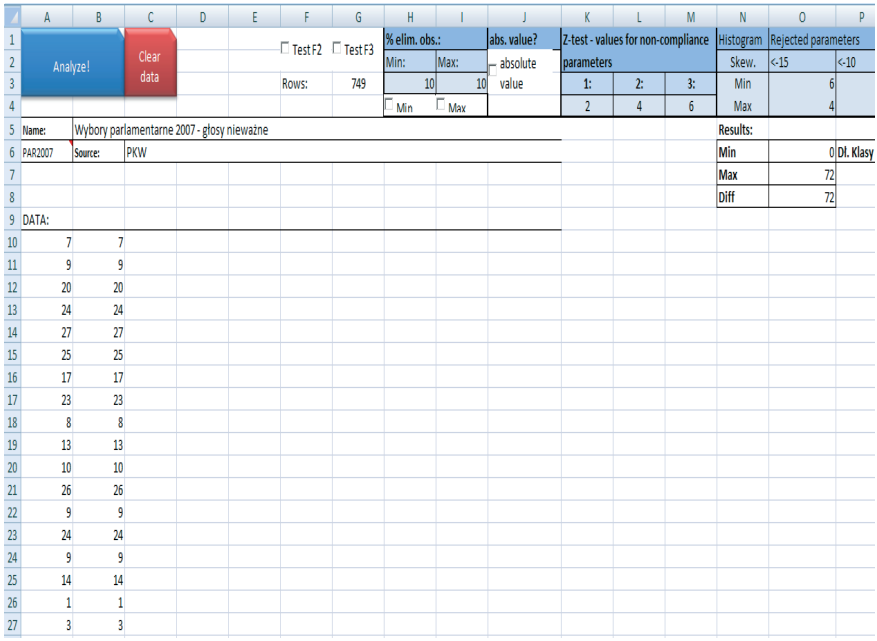
W ramach analizy sporzadzanych jest 6 wykresów – każdy z nich odnosi się do innego testu. Wykresy mają postać histogramów, w których obok siebie podawane są słupki ilustrujące częstości empiryczne (kolor niebieski, symbol EMP) oraz częstości teoretyczne, wynikające z prawa Benforda (kolor czerwony, symbol BF). Do wartości empirycznych dopasowywane są automatycznie funkcje trendu – potęgowe dla testów F1, F2, F3 oraz liniowe dla testów D2, D3, L1. Na osi pionowej Y wykresów podawane są częstości (%) natomiast na osi poziomej X – kombinacje cyfr, dla których wyznaczane są częstości ich występowania.

Druga część analizy dostarcza parametrów statystycznych opisujących analizowany zbiór danych (wartości skrajne, średnia arytmetyczna, współczynniki zmienności, skośności i kurtozy).

Trzecia część analizy obejmuje miary zgodności rozkładów empirycznych z rozkładami teoretycznymi wynikającymi z prawa Benforda (mierniki M1-M5), współczynniki korelacji między liczebnościami empirycznymi i teoretycznymi oraz statystyki testu zgodności rozkładów chi-kwadrat, testu ***z*** oraz testów Kołmogorowa–Smirnowa (w trzech wersjach KS1–KS3). W tabeli wynikowej podawane są także wartości krytyczne testów zgodności dla dwóch poziomów istotności α=0,05 oraz α=0,01. Porównanie wartości krytycznych ze statystykami testów pozwala podjąć decyzje o zgodności lub braku zgodności analizowanych rozkładów. W przypadku testu ***z*** podawane są liczby przedziałów, w których na danym poziomie istotności należy odrzucić hipotezę o zgodności częstości występowania danej kombinacji cyfr. Liczbę tę należy porównywać z ogólną liczbą przedziałów k.

W tabeli z parametrami, poza wymienionymi powyżej statystykami (dla każdego testu odrębnie), podawane są także informacje o nazwie badanej cechy, jej symbolu i źródle danych oraz informacja o tym, czy analizowany był cały dostępny zbiór danych, czy też zbiór ograniczony o x% najmniejszych i/lub x% największych obserwacji.

Rys. 4.1. Ekran sterujący z parametrami wyznaczającymi zakres analizy



	A	B
10	7	7
11	9	9
12	20	20
13	24	24
14	27	27
15	25	25
16	17	17
17	23	23
18	8	8
19	13	13
20	10	10
21	26	26
22	9	9
23	24	24
24	9	9
25	14	14
26	1	1
27	3	3

4.4. Sumaryczna tabela wynikowa

W wyniku uruchomienia makro w arkuszu BF_OUTPUT_H pojawiają się wyniki obliczeń w postaci tabeli sumarycznej zawierającej parametry statystyczne, tabeli z wartościami testu ***z***, wykresy rozkładów częstości oraz tabel roboczych (dla każdego testu odrębnie).

Tabela sumaryczna (por. tab. 4.1) ma wymiary pozwalające na jej wklejanie do tekstów w układzie poziomym i składa się z trzech części.

W pierwszej części znajdują się następujące informacje.

1. nazwa analizowanej cechy i jej symbol,
2. źródło danych liczbowych,
3. Wartości krytyczne testów Kołomogorowa–Smirnowa KS1–KS3 dla dwóch poziomów istotności α=0,05 oraz α=0,01,
4. Parametry określające procent najmniejszych – min[-]% i/lub procent największych – max[-]% obserwacji wyeliminowanych z całego dostępnego zbioru danych. Parametry 0–0 oznaczają, że analizie podlega cały zbiór danych.

Informacje wymienione w poz. 1, 2 kopiowane są z arkusza BF_DATA. Jeżeli użytkownik pominie te informacje, to w tabeli pojawią się puste komórki.

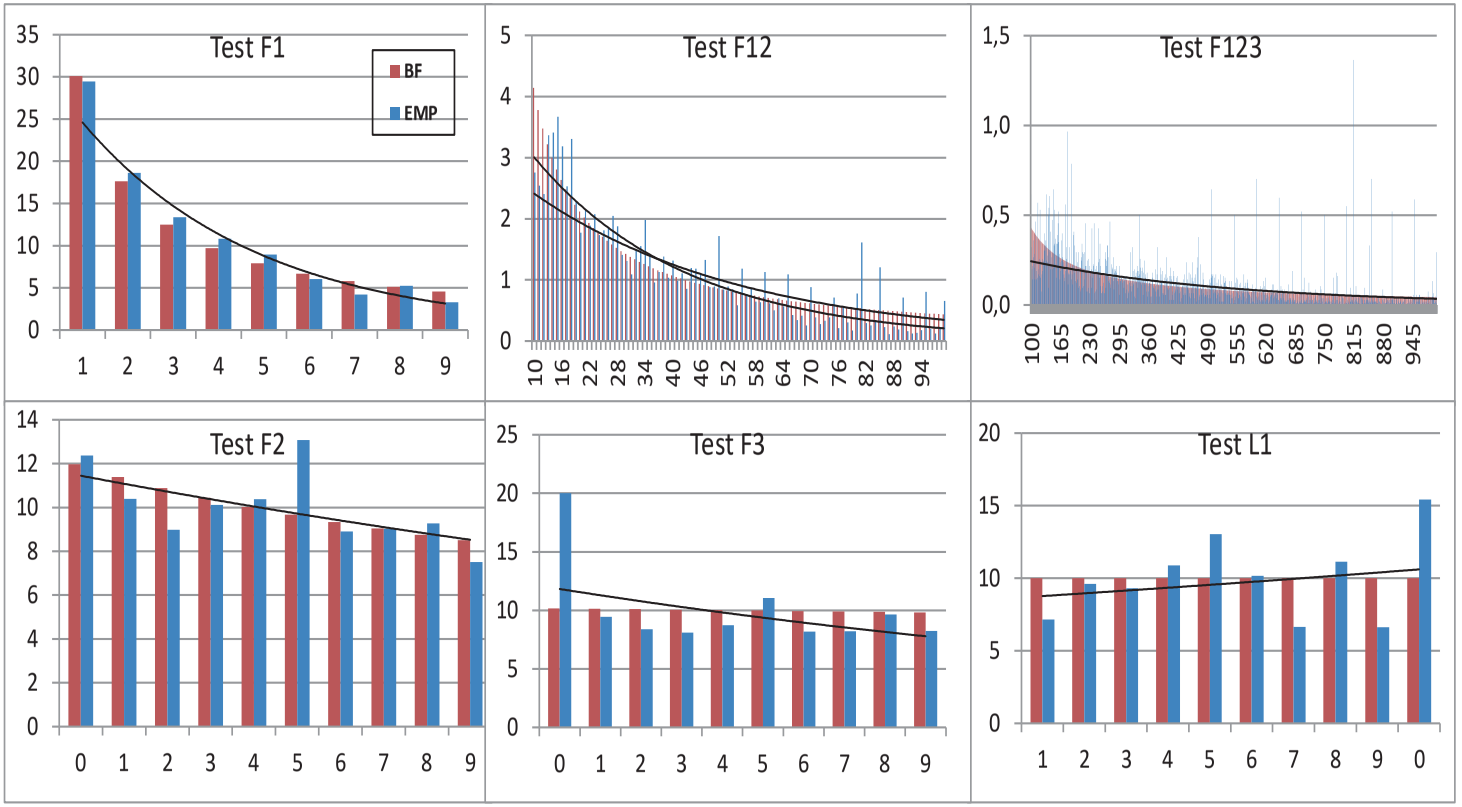
Druga część tabeli zawiera dla każdego z sześciu uwzględnionych testów (F1, F2, F3, D2, D3, L1) następujące parametry.

1. liczba obserwacji *n* (dla poszczególnych testów liczba ta może być różna),
2. wartości mierników dopasowania M1–M5,
3. wartości współczynników korelacji liniowej pomiędzy liczebnościami empirycznymi i teoretycznymi – r,
4. wartości statystyki testu chi-kwadrat – chi,
5. wartości krytyczne testu chi-kwadrat na poziomach istotności α=0,05 (chi – 0,05) oraz α=0,01 (chi – 0,01),
6. ogólna liczba przedziałów klasowych *k* w badanym rozkładzie,
7. liczba przedziałów klasowych dla których test z wskazał na poziomie istotności α=0,05 (z – 0,05) oraz α=0,01 (z – 0,01) na ***istotne rozbieżności*** pomiędzy teoretyczną a empiryczną częstością występowania danej kombinacji cyfr. Jeżeli ta liczba jest niewielka w porównaniu z parametrem *k*, to ewentualne rozbieżności pomiędzy prawem Benforda a rozkładem empirycznym ograniczają się tylko do wskazanych przez test z kombinacji cyfr.
8. statystyki empiryczne testów zgodności Kołmogorowa–Smirnowa KS1––KS3, które należy porównywać ze znajdującymi się powyżej wartościami krytycznymi.
9. parametry statystyczne opisujące własności rozkładu analizowanej cechy; są to:
   – wartość maksymalna,
   – wartość minimalna,
   – średnia arytmetyczna,
   – współczynnik zmienności zdefiniowany jako iloraz odchylenia standardowego przez średnią arytmetyczną x 100,
   – współczynnik skośności (asymetrii)
   – współczynnik kurtozy (spłaszczenia).

Trzecia część tabeli składa się z 6 wykresów ilustrujących zgodność rozkładów empirycznych z rozkładami wynikającymi z prawa Benforda. W pierwszym rzędzie podawane są wykresy dla testów F1, F2, F3, gdzie właściwą funkcją oddającą istotę prawa Benforda jest funkcja potęgowa. W drugim rzędzie natomiast przytoczone są wykresy dla testów D2, D3 i L1, gdzie częstości powinny układać się według funkcji prostoliniowej. W przypadku funkcji D2 powinna to być funkcja lekko opadająca, w przypadku testu D3 – prawie równoległa do osi poziomej, a w przypadku testu L1 – dokładnie równoległa do osi poziomej.

Tab 4.1. Przykład sumarycznej tabeli wynikowej

Name	**Wartość netto sprzedaży w okresie I-X 2008 wg faktur sprzedażowych**													α=0,05	**1,36**	1,36	1,75	–	min [-] %
FK1	**Source:**	Firma usługowa zajmująca się transportem. Dane z programu księgowego Ramzes.												α=0,01	1,63	1,63	2,00	–	max [-] %
	n	**M1**	**M2**	**M3**	**M4**	**M5**	**r**	**p (r)**	**chi**	**chi-0,05**	**p(chi)**	**k**	**z - 0,05**	**z - 0,01**	**KS1**	**KS2**	**KS3**	**stat.**	**param.**
F1	10554	11,9	0,34	1,01	7,4	881,2	0,992	0,000	132,6	15,5	0,000	9	7	7	2,46	3,48	2,94	**max**	290 000
F12	10554	34,0	0,04	0,40	28,8	2748,6	0,889	0,000	1379,7	112	0,000	90	57	46	2,80	3,96	5,48	**min**	45,0
F123	10552	64,0	0,00	0,11	76,8	5619,6	0,588	0,000	12353,3	969,9	0,000	900	270	109	2,91	4,12	5,59	**avg:**	2 860
F2	10554	9,4	0,42	1,34	13,4	983,4	0,552	0,049	192,2	16,9	0,000	10	4	4	2,06	2,91	2,77	**v-factor**	427,0
F3	10552	21,7	1,07	3,40	34,0	2302,0	-	-	1199,5	16,9	0,000	10	9	8	7,15	10,10	7,15	**kurt.**	213,6
L1	10554	21,3	0,85	2,69	26,9	2248,0	-	-	762,1	16,9	0,000	10	8	6	3,95	5,58	4,06	**skew.**	13,0


Test F1
BF
EMP
35
30
25
20
15
10
5
0
1 2 3 4 5 6 7 8 9
Test F12
5
4
3
2
1
0
10 16 22 28 34 40 46 52 58 64 70 76 82 88 94
Test F123
1,5
1,0
0,5
0,0
100 165 230 295 360 425 490 555 620 685 750 815 880 945
Test F2
14
12
10
8
6
4
2
0
0 1 2 3 4 5 6 7 8 9
Test F3
25
20
15
10
5
0
0 1 2 3 4 5 6 7 8 9
Test L1
20
15
10
5
0
1 2 3 4 5 6 7 8 9 0

4.5. Parametry wynikowe

W tabeli sumarycznej przytoczono mierniki podobieństwa rozkładów empirycznych z rozkładami teoretycznymi wynikającymi z rozkładu Benforda oraz wartości testów zgodności i parametrów opisowych rozkładu analizowanej zmiennej. Poniżej podano wzory i definicje poszczególnych parametrów. We wzorach przyjęto następującą konwencję oznaczeń:

- n – ogólna liczba obserwacji w analizowanym zbiorze,
- k – liczba kombinacji cyfr (w teście F1 – $k=9$, w testach D2, D3, L1 – $k=10$, w teście F2 – $k=90$, a w teście F3 – $k=900$)
- n_i oraz $\hat{n}_i$ ($i=1,2,...,k$) – liczebności empiryczne i teoretyczne pojawienia się na określonym miejscu i-tej cyfry (lub i–tej kombinacji cyfr),

c_i oraz $\hat{c}_i$ ($i=1,2,...,k$) – częstości empiryczne i teoretyczne dane wzorami:

(4.1) $$c_i = \frac{n_i}{n}100 \qquad \hat{c}_i = \frac{\hat{n}_i}{n}100$$

p_i oraz $\hat{p}_i$ ($i=1,2,...,k$) – prawdopodobieństwa empiryczne i teoretyczne dane wzorami:

(4.2) $$p_i = \frac{n_i}{n} \qquad \hat{p}_i = \frac{\hat{n}_i}{n}$$

f_i oraz $\hat{f}_i$ ($i=1,2,...,k$) – wartości dystrybuanty empirycznego i teoretycznego rozkładu częstości cyfr znaczących dane wzorami:

(4.3) $$f_i = \sum_{l=1}^{i} p_l \qquad \hat{f}_i = \sum_{l=1}^{i} \hat{p}_l$$

Mierniki podobieństwa rozkładów empirycznych i teoretycznych dane są wzorami:

(4.4) $$M_1 = \frac{100}{k}\sum_{i=1}^{k}\left|\frac{c_i - \hat{c}_i}{\hat{c}_i}\right| = \frac{100}{k}\sum_{i=1}^{k}\left|\frac{n_i}{\hat{n}_i} - 1\right|$$

(4.5) $$M_2 = \frac{1}{k}\sqrt{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}$$

(4.6) $$M_3 = \sqrt{\frac{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}{k}}$$

(4.7) $$M_{4|} = \frac{100\sqrt{\sum_{i=1}^{k}(c_i - \hat{c}_i)^2}}{\sqrt{\sum_{i=1}^{k}\hat{c}_i^2}}$$

(4.8) $$M_5 = \sum_{i=1}^{k}|c_i - \hat{c}_i| = \frac{100\sum_{i=1}^{k}|n_i - \hat{n}_i|}{n}$$

Mierniki te są niezależne od wielkości zbioru n i przyjmują tym mniejsze wartości, im bardziej zgodne ze sobą są porównywane rozkłady częstości. Generalnie wskazują one na przeciętną wielkość różnic pomiędzy częstościami faktycznymi a częstościami teoretycznymi w danym teście.

Mierniki M2 oraz M3 wskazują na przeciętną różnicę pomiędzy wartościami empirycznymi a teoretycznymi porównywanych rozkładów. W literaturze preferuje się miernik M3, jakkolwiek bardziej naturalny wydaje się miernik M2. Miernik M3 jest większy od miernika M2 (w przypadku testów F1, D2, D3, L1 – trzykrotnie większy) i szacuje z dużym nadmiarem wielkość różnic pomiędzy porównywanymi rozkładami.

Mierniki M1 oraz M4 wskazują, jaka jest przeciętna różnica między wartościami empirycznymi a teoretycznymi w relacji do wartości teoretycznych rozkładu. Miernik M5 pokazuje, jaką część wszystkich obserwacji trzeba by zmienić, aby rozkłady empiryczne pokryły się z rozkładami teoretycznymi. Wielkości tych trzech mierników wyrażone są w procentach.

Szóstym miernikiem zgodności jest współczynnik korelacji linowej ***r*** pomiędzy empirycznymi i teoretycznymi częstościami.

(4.9) $$r = \frac{\sum_{i=1}^{k}(n_i - \overline{n})(\hat{n}_i - \overline{\hat{n}})}{\sqrt{\sum_{i=1}^{k}(n_i - \overline{n})^2 \sum_{i=1}^{k}(\hat{n}_i - \overline{\hat{n}})^2}}$$

W powyższym wzorze $\overline{n}$ oraz $\overline{\hat{n}}$ to średnie arytmetyczne z empirycznych i teoretycznych liczebności. Identyczne wartości współczynników korelacji uzyskuje się, jeżeli we wzorze (4.9) przyjmie się nie liczebności rozkładów, ale ich częstości lub prawdopodobieństwa. Współczynnika ***r*** nie można wyznaczyć dla testu L1, gdyż częstości teoretyczne w tym przypadku są identyczne i nie mają żadnej zmienności.

Druga grupa zawiera statystyki testów zgodności rozkładów. Uwzględniono tu 5 testów zgodności – chi-kwadrat, test ***z*** oraz trzy testy Kołmogorowa–Smirnowa.

Statystyki te dane są wzorami:

(4.10) $$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \hat{n}_i)^2}{\hat{n}_i}$$

(4.11) $$z_i = \frac{p_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)/n}} \qquad (i = 1,..., k)$$

(4.12) $$KS1 = D\sqrt{\frac{n^2}{2n}} \qquad D = \max_i \left| f_i - \hat{f}_i \right| \qquad (i = 1,..., k)$$

(4.13) $$KS2 = D\sqrt{n} \qquad D = \max_i \left| f_i - \hat{f}_i \right| \qquad (i = 1,..., k)$$

(4.14) $$KS3 = V_N * [\sqrt{N} + 0{,}155 + 0{,}24N^{-1/2}] \qquad gdzie \qquad (i = 1,..., k)$$
$$V_N = D_N^+ + D_N^- \quad D_N^+ = \sup_i \left[f_i - \hat{f}_i \right] \quad D_N^- = \sup_i \left[\hat{f}_i - f_i \right] \quad N = \frac{n^2}{2n}$$

Ostatnia grupa parametrów znajdująca się w tabeli sumarycznej odnosi się do całego analizowanego zbioru, a nie do poszczególnych testów F1,F2,....

Jest to sześć podstawowych statystyk opisowych – wartości maksymalne i minimalne zbioru, średnia arytmetyczna i współczynnik zmienności oraz współczynniki skośności (asymetrii) i spłaszczenia (kurtozy). Parametry te określone są następującymi wzorami:

(4.15) $$\max = \max_i \{x_i\} \qquad (i = 1,..., n)$$

(4.16) $$\min = \min_i \{x_i\} \qquad (i = 1,..., n)$$

(4.17) $$sred = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(4.18) $$wsp.V = \frac{sred}{s}100$$

(4.19) $$asym = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i - sred}{s}\right)^3$$

(4.20) $$kurt = \frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - sred}{s}\right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

W powyższych wzorach x_i oznacza *i*-tą realizację zmiennej, *n* jest ogólną liczbą obserwacji, natomiast s – odchyleniem standardowym danym wzorem:

(4.21) $$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - sred)^2}{n-1}}$$

4.6. Tabele z wynikami testu ***z***

Poza tabelą główną makro dostarcza drugiej tabeli ze szczegółowymi wynikami testu ***z*** (por. tab. 4.2). W górnej części tej tabeli przytoczono moduły wartości statystyki danej wzorem (4.11) dla różnych rozkładów cyfr F1, D2, D3, L1 oraz F2. Opuszczono natomiast wyniki dla testu F3 z uwagi na ich obszerność.

Statystyki przekraczające co do modułu wartości krytyczne, wynoszące 1,96 dla poziomu istotności α=0,05 oraz 2,58 dla poziomu istotności α=0,01, wskazują miejsca (cyfry i ich kombinacje) w których występuje ewentualna niezgodność rozkładów empirycznych z rozkładami wynikającymi z prawa Benforda. Tak więc w przypadku stwierdzenia, że porównywane rozkłady nie są zgodne powstaje konieczność lokalizacji tej niezgodności, do czego można wykorzystać właśnie statystyki ***z***.

W wierszu oraz w kolumnie tabeli 4.2 oznaczonej symbolem ***max*** podano maksymalne wartości tych statystyk. Są to najbardziej „podejrzane" kombinacje cyfr, które świadczą o niezgodności rozkładów empirycznych z rozkładami Benforda. Ponadto w ostatnich dwóch kolumnach górnej części tabeli przytoczono liczbę kombinacji cyfr, dla których stwierdzono istotne odchylenia pomiędzy częstościami empirycznymi a częstościami teoretycznymi.

Dolna część tabeli 4.2 posiada podobną konstrukcję, z tym że wartości statystyk ***z*** zamieniono na rangi, przyjmując następujące wartości graniczne:

1. ranga 0 dla której z<2 – sytuacje zgodne z rozkładami Benforda,
2. ranga 1 dla której z€<2 - 5> - sytuacje niezgodne z rozkładami Benforda,
3. ranga 2 dla której z€<5 – 8> - sytuacje o dużej niezgodności z rozkładami Benforda,
4. ranga 3 dla której z>8 – sytuacje o bardzo wysokiej niezgodności z rozkładami Benforda.

W ostatnich 4 kolumnach dolnej części tabeli 4.2 przytoczono liczbę przypadków z poszczególnymi rangami dla każdego rodzaju testu F1, D2, D3, L2 oraz F2. W ostatnich dwóch wierszach tabeli 4.2 podano sumaryczne liczby przypadków z wyróżnionymi rangami oraz ich rozkłady procentowe.

W przypadku niezgodności analizowanych rozkładów z rozkładami Benforda najbardziej wątpliwe są kombinacje cyfr z rangą „3", następnie z rangą „2" itd. Duża liczba sytuacji z wysokimi rangami sugeruje generalnie o istotnej niezgodności z rozkładami Benforda. W takim przypadku inne statystyki także powinny wskazywać na wyraźny stopień niezgodności porównywanych rozkładów.

Tab 4.2. Tabela z wartościami testu z

Nazwa:	**Wartość netto sprzedaży w okresie I-X 2008 wg faktur sprzedażowych**												
FK1	**Source:**	Firma usługowa zajmująca się transportem. Dane z programu księgowego Ramzes.											
Test *z*	0	1	2	3	4	5	6	7	8	9	max	α=0,05	α=0,01
F1	-	1,49	2,75	2,66	3,96	3,94	2,67	7,00	0,58	6,24	7,00	7	7
F2	1,26	3,25	6,27	1,09	1,18	11,85	1,55	0,02	1,85	3,63	11,85	4	4
F3	33,42	2,31	5,86	6,71	4,44	3,67	6,08	5,76	0,78	5,43	33,42	9	8
L1	18,61	9,75	1,34	2,35	3,00	10,40	0,54	11,47	3,91	11,56	18,61	8	6
F12–10	7,13	6,68	6,00	0,85	2,50	5,38	3,53	0,31	6,50	0,32	7,13	57	46
F12–20	2,48	0,95	0,34	1,73	0,30	0,84	1,84	3,85	2,95	0,52	3,85		
F12–30	1,01	2,55	1,44	2,34	6,65	1,85	2,21	0,29	2,48	0,66	6,65		
F12–40	2,44	0,62	0,89	1,51	2,28	2,43	1,66	4,45	0,16	0,15	4,45		
F12–50	9,51	0,00	0,72	0,36	2,42	4,69	0,35	1,38	1,17	1,15	9,51		
F12–60	4,99	1,11	2,38	0,21	1,32	5,40	2,89	3,88	2,81	4,81	5,40		
F12–70	3,48	2,89	4,32	3,48	2,62	1,84	4,91	0,15	3,46	5,24	5,24		
F12–80	3,33	15,21	3,30	3,77	2,62	10,05	3,99	5,59	3,73	4,37	15,21		
F12–90	3,43	4,69	5,20	5,01	4,25	5,35	3,85	4,97	3,46	3,39	5,35	1,96	2,58
max	33,4	15,2	6,3	6,7	6,6	11,8	6,1	11,5	6,5	11,6	33,4		

Test *z*	0	1	2	3	4	5	6	7	8	9	[0]	[1]	[2]	[3]
F1	**–**		**1**	**1**	**1**	**1**	**1**	**3**		**3**	**2**	**5**		**2**
F2		**1**	**3**			**3**				**1**	**6**	**2**		**2**
F3	**3**	**1**	**2**	**3**	**2**	**1**	**3**	**2**		**2**	**1**	**2**	**4**	**3**
L1	**3**	**3**		**1**	**1**	**3**		**3**	**1**	**3**	**2**	**3**		**5**
F12–10	**3**	**3**	**2**		**1**	**2**	**1**		**3**		**3**	**2**	**2**	**3**
F12–20	**1**							**1**	**1**		**7**	**3**		
F12–30		**1**		**1**	**3**		**1**		**1**		**5**	**4**		**1**
F12–40	**1**				**1**	**1**		**2**			**6**	**3**	**1**	
F12–50	**3**				**1**	**2**					**7**	**1**	**1**	**1**
F12–60	**2**		**1**			**2**	**1**	**1**	**1**	**2**	**3**	**4**	**3**	
F12–70	**1**	**1**	**2**	**1**	**1**		**2**		**1**	**2**	**2**	**5**	**3**	
F12–80	**1**	**3**	**1**	**1**	**1**	**3**	**1**	**2**	**1**	**2**	**0**	**6**	**2**	**2**
F12–90	**1**	**2**	**2**	**2**	**2**	**2**	**1**	**2**	**1**	**1**	**0**	**4**	**6**	
									Liczba	**129**	**44**	**44**	**22**	**19**
									Procent	100	**34,1**	**34,1**	**17,1**	**14,7**

4.7. Tabele robocze

W arkuszu BF_OUTPUT_H znajdują się także tabele robocze ze szczegółowymi wynikami obliczeń. Dla każdego z 6 uwzględnionych w analizie testów: F1, F2, F3, D2, D3 i L1 podawana jest identyczna tabela z formułami obliczeniowymi. Tabele te różnią się liczbą wierszy – dla testu F1 jest to 9 wierszy, dla testów D2, D3 i L1 – 10 wierszy, dla testu F2 – 90 wierszy oraz dla testu F3 – 900 wierszy (nie licząc 3 wierszy zawierających finalne wartości parametrów).

W kolejnych kolumnach tych tabel znajdują się następujące informacje:

1. identyfikator cyfry (lub ich kombinacji),
2. empiryczne liczebności obserwacji zaczynających się od danej cyfry (lub ich kombinacji) [Emp – L],
3. empiryczna częstość (%) tych obserwacji [Emp %],
4. częstość wynikająca z prawa Benforda [Benf %],
5. liczebność obserwacji wynikająca z prawa Benforda [Benf – L],
6. elementy składowe statystyki chi-kwadrat [chi],
7. elementy empirycznego rozkładu częstości w postaci prawdopodobieństwa [f(i)],
8. elementy teoretycznego rozkładu częstości w postaci prawdopodobieństwa [b(i)],
9. elementy składowe testu ***z*** [z],
10. elementy składowe dystrybuanty empirycznego rozkładu częstości [F(i)],
11. elementy składowe dystrybuanty teoretycznego rozkładu częstości [B(i)],
12. moduły różnic dystrybuant [F(i)-B(i)],
13. moduły dodatnich różnic dystrybuant [F(i)–B(i)],
14. moduły ujemnych różnic dystrybuant [F(i)–B(i)],
15. względne różnice pomiędzy częstościami rozkładów empirycznych oraz rozkładów Benforda [(E–B)/B],
16. kwadraty różnic pomiędzy częstościami rozkładów empirycznych oraz rozkładów Benforda [(E–B)^2],
17. kwadraty częstości rozkładów Benforda [B^2],
18. moduły różnic pomiędzy częstościami rozkładów empirycznych oraz rozkładów Benforda [(E–B) %].

W odrębnych (końcowych) wierszach podawane są sumy wartości znajdujących się w poszczególnych kolumnach oraz wartości wynikowych parametrów, które są przenoszone do omówionych poprzednio tabel sumarycznych.

Tab 4.3. Tabela robocza z obliczeniami dla testów F1 oraz D2

F1	Emp–L	Benf %	Benf–L	Emp %	F(i)	Chi	f(i)	B(i)	b(i)	abs(z)	[F(i)–B(i)]	[F(i)–B(i)]+	[F(i)–B(i)]–	(E–B)/B	(E–B)^2	B^2	[E–B]%	z
1	3107	30,10	3177	29,44	0,294	1,5	0,29	0,301	0,301	1,49	0,007		0,007	0,022	0,441	906,2	70,1	–1,49
2	1966	17,61	1858	18,63	0,481	6,2	0,19	0,477	0,176	2,75	0,004	0,004		0,058	1,038	310,1	107,5	2,75
3	1409	12,49	1319	13,35	0,614	6,2	0,13	0,602	0,125	2,66	0,012	0,012		0,069	0,734	156,1	90,4	2,66
4	1143	9,69	1023	10,83	0,722	14,1	0,11	0,699	0,097	3,96	0,024	0,024		0,118	1,297	93,9	120,2	3,96
5	945	7,92	836	8,95	0,812	14,3	0,09	0,778	0,079	3,94	0,034	0,034		0,131	1,073	62,7	109,3	3,94
6	638	6,69	707	6,05	0,872	6,7	0,06	0,845	0,067	2,67	0,027	0,027		0,097	0,422	44,8	68,6	–2,67
7	444	5,80	612	4,21	0,915	46,1	0,04	0,903	0,058	7,00	0,011	0,011		0,275	2,535	33,6	168,0	–7,00
8	553	5,12	540	5,24	0,967	0,3	0,05	0,954	0,051	0,58	0,013	0,013		0,024	0,015	26,2	13,1	0,58
9	349	4,58	483	3,31	1,000	37,1	0,03	1,000	0,046	6,24	0,034	0,034	0,007	0,277	1,610	20,9	133,9	–6,24
Suma:	10554	100	10554	100		132,6	1,0		1,0	7	2,46	3,48	2,94	1,1	9,166	1654,5	881,2	
						15,5			0,992	7	1,36	1,36	1,75	11,9	0,34	7,4	9791,1	
						8,6		20,89	7E-08		1,8	2,6	1,7		1,01			
						CHI			r	z	K–S1	K–S2	K–S3	M1	M2/M3	M4	M5	

D2	Emp–L	Benf %	Benf –L	Emp %	F(i)	Chi	f(i)	B(i)	b(i)	abs(z)	[F(i)–B(i)]	[F(i)–B(i)]+	[F(i)–B(i)]–	(E–B)/B	(E–B)^2	B^2	[E–B]%	z
0	1305	11,97	1263	12,36	0,124	1,4	0,124	0,120	0,120	1,26	0,004	0,004		0,033	0,158	143,2	41,90	1,26
1	1096	11,39	1202	10,38	0,227	9,3	0,104	0,234	0,114	3,25	0,006		0,006	0,088	1,009	129,7	106,00	–3,25
2	948	10,88	1149	8,98	0,317	35,0	0,090	0,342	0,109	6,27	0,025		0,025	0,175	3,609	118,4	200,50	–6,27
3	1067	10,43	1101	10,11	0,418	1,1	0,101	0,447	0,104	1,09	0,028		0,028	0,031	0,104	108,8	34,09	–1,09
4	1095	10,03	1059	10,38	0,522	1,2	0,104	0,547	0,100	1,18	0,025		0,025	0,034	0,119	100,6	36,35	1,18
5	1380	9,67	1020	13,08	0,653	126,8	0,131	0,644	0,097	11,85	0,009	0,009		0,353	11,614	93,5	359,67	11,85
6	939	9,34	985	8,90	0,742	2,2	0,089	0,737	0,093	1,55	0,005	0,005		0,047	0,194	87,2	46,48	–1,55
7	953	9,04	954	9,03	0,832	0,0	0,090	0,827	0,090	0,02	0,005	0,005		0,001	0,000	81,6	0,57	–0,02
8	978	8,76	924	9,27	0,925	3,1	0,093	0,915	0,088	1,85	0,010	0,010		0,058	0,260	76,7	53,79	1,85
9	793	8,50	897	7,51	1,000	12,1	0,075	1,000	0,085	3,63	0,028	0,010	0,028	0,116	0,972	72,2	104,06	–3,63
Suma:	10554	100	10554	100		192,2	1,0		1,0	4	2,06	2,91	2,77	0,94	18,038	1012,0	983,41	
						16,9			0,552	4	1,36	1,36	1,75	9,4	0,42	13,4	9834,1	
						11,4		1,87	0,049		1,5	2,1	1,6		1,34			
						CHI			r	z	K–S1	K–S2	K–S3	M1	M2/M3	M4	M5	

Tab 4.4. Tabela robocza z obliczeniami dla testów D3 oraz L1

D3	Emp – L	Benf %	Benf – L	Emp %	F(i)	Chi	f(i)	B(i)	b(i)	abs(z)	[F(i)–B(i)]	[F(i)–B(i)]+	[F(i)–B(i)] –	(E–B)/B	(E–B)^2	B^2	[E–B] %	z
0	2112	10,18	1074	20,02	0,200	1003,1	0,200	0,102	0,102	33,42	0,098	0,098		0,966	96,761	103,6	#####	33,42
1	998	10,14	1070	9,46	0,295	4,8	0,095	0,203	0,101	2,31	0,092	0,092		0,067	0,462	102,8	71,72	–2,31
2	884	10,10	1065	8,38	0,379	30,9	0,084	0,304	0,101	5,86	0,074	0,074		0,170	2,957	102,0	181,46	–5,86
3	854	10,06	1061	8,09	0,459	40,5	0,081	0,405	0,101	6,71	0,055	0,055		0,195	3,857	101,1	207,25	–6,71
4	920	10,02	1057	8,72	0,547	17,8	0,087	0,505	0,100	4,44	0,042	0,042		0,130	1,688	100,4	137,08	–4,44
5	1166	9,98	1053	11,05	0,657	12,1	0,111	0,605	0,100	3,67	0,052	0,052		0,107	1,148	99,6	113,04	3,67
6	862	9,94	1049	8,17	0,739	33,3	0,082	0,704	0,099	6,08	0,035	0,035		0,178	3,137	98,8	186,88	–6,08
7	868	9,90	1045	8,23	0,821	29,9	0,082	0,803	0,099	5,76	0,018	0,018		0,169	2,809	98,0	176,85	–5,76
8	1017	9,86	1041	9,64	0,917	0,5	0,096	0,902	0,099	0,78	0,016	0,016		0,023	0,051	97,3	23,86	–0,78
9	871	9,83	1037	8,25	1,000	26,5	0,083	1,000	0,098	5,43	0,098	0,098	0,000	0,160	2,472	96,6	165,92	–5,43
Suma:	10552	100	10552	100		1199,5	1,0		1,0	9	7,15	10,10	7,15	2,17	115,342	1000,1	#####	
						16,9			–	8	1,36	1,36	1,75	21,7	1,07	34,0	#####	
						70,9		#ARG!	–		5,3	7,4	4,1		3,40			
						CHI			r	z	K–S 1	K–S 2	K–S 3	M1	M2/M3	M4	M5	

L1	Emp – L	Benf %	Benf – L	Emp %	F(i)	Chi	f(i)	B(i)	b(i)	abs(z)	[F(i)–B(i)]	[F(i)–B(i)]+	[F(i)–B(i)] –	(E–B)/B	(E–B)^2	B^2	[E–B] %	z
1	755	10	1055	7,15	0,072	85,5	0,072	0,100	0,100	9,75	0,028		0,028	0,285	8,102	100,0	300,40	–9,75
2	1014	10	1055	9,61	0,168	1,6	0,096	0,200	0,100	1,34	0,032		0,032	0,039	0,154	100,0	41,40	–1,34
3	983	10	1055	9,31	0,261	5,0	0,093	0,300	0,100	2,35	0,039		0,039	0,069	0,471	100,0	72,40	–2,35
4	1148	10	1055	10,88	0,370	8,1	0,109	0,400	0,100	3,00	0,030		0,030	0,088	0,770	100,0	92,60	3,00
5	1376	10	1055	13,04	0,500	97,4	0,130	0,500	0,100	10,40	0,000		0,000	0,304	9,228	100,0	320,60	10,40
6	1072	10	1055	10,16	0,601	0,3	0,102	0,600	0,100	0,54	0,001	0,001		0,016	0,025	100,0	16,60	0,54
7	702	10	1055	6,65	0,668	118,3	0,067	0,700	0,100	11,47	0,032		0,032	0,335	11,212	100,0	353,40	–11,47
8	1176	10	1055	11,14	0,779	13,8	0,111	0,800	0,100	3,91	0,021		0,021	0,114	1,306	100,0	120,60	3,91
9	699	10	1055	6,62	0,846	120,4	0,066	0,900	0,100	11,56	0,054		0,054	0,338	11,404	100,0	356,40	–11,56
0	1629	10	1055	15,43	1,000	311,7	0,154	1,000	0,100	18,61	0,054	0,001	0,054	0,543	29,538	100,0	573,60	18,61
Suma:	10554	100	10554	100		762,1	1,0		1,0	8	3,95	5,58	4,06	2,13	72,208	1000,0	#####	
						16,9			–	7	1,36	1,36	1,75	21,3	0,85	26,9	#####	
						45,0			–		2,9	4,1	2,3		2,69			
						CHI			r	z	K–S1	K–S2	K–S3	M1	M2/M3	M4	M5	

4.8. Omówienie przykładu empirycznego

W przykładzie przeanalizowany został zestaw danych składający się z 10 554 rekordów. Stanowi on zbiór faktur sprzedażowych pewnej firmy działającej w branży transportowej. Wystawione one były w okresie I–X 2008.

Analiza polegała na przebadaniu zbioru sześcioma różnymi testami, biorącymi pod uwagę:

- pierwszą cyfrę znaczącą,
- drugą cyfrę znaczącą,
- trzecią cyfrę znaczącą,
- dwie pierwsze cyfry znaczące,
- trzy pierwsze cyfry znaczące,
- ostatnią cyfrę znaczącą.

Jako wynik analizy wyznaczona została tabela zbiorcza zawierająca zestawienia wszystkich wskaźników statystycznych dla wymienionych sześciu testów empirycznych.

Ważnym wskaźnikiem opisującym jakość badanych danych jest test ***z***. Umożliwia on badanie stopnia odchylenia wartości empirycznych od teoretycznych na każdym jednostkowym etapie analizy (dla testu F1: F1(1), F1(2), itd., dla testu F2: F2(10), F2(11) itd.).

Wyznaczenie największej wartości testu ***z*** w kontekście rozpatrywanego testu empirycznego umożliwia określenie najbardziej podejrzanego obszaru wewnątrz zbioru danych.

W przypadku rozpatrywanego przykładu empirycznego najwyższa wartość testu ***z*** wystąpiła dla F3(814) (w przypadku testu F3). W rozpatrywanym okresie zostało wystawionych 144 faktur dla których wartość rozpoczynała się od sekwencji „814". Teoretyczna wartość testu F3(814) powinna wynieść 5 przy zadanej wielkości zbioru danych.

Informacja ta powinna stanowić podstawę głębszej analizy faktur na poziomie szczegółowym (badanie przyczyny tak dużej rozbieżności).

Analityk przeprowadzając badanie powinien brać pod uwagę charakter danych. W przypadku danych finansowych zaobserwować można tendencję do zaokrąglania wartości po przecinku (do pełnej kwoty, 50 groszy itp.). Wyjaśnia to zaburzoną częstotliwość występowania cyfr 0 oraz 5 lub końcówek „50" oraz „00". Tendencję tą widać wyraźnie na wykresach testów F2 (dla wartości „5"), F3 (dla wartości „0") lub dla testu L1 (obie wartości: „0" oraz „5").

Rozdział 5
Analiza rozkładów cyfr znaczących na przykładzie danych finansowych

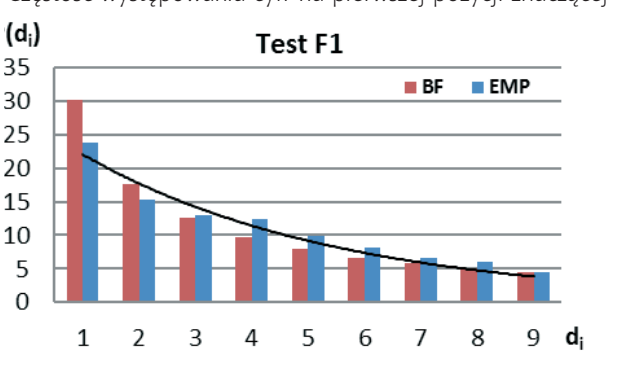
Dane finansowo-księgowe wykorzystywane do ewidencji kosztów oraz przychodów firmy mogą stać się przedmiotem manipulacji lub fałszerstwa. Znanym określeniem „kreatywna rachunkowość" opisuje się działanie polegające na zniekształcaniu obrazu stanu finansów danego przedsiębiorstwa w celu zatajenia pewnych informacji lub przedstawienia nieprawidłowych wyników. Weryfikację rzetelności danych finansowych zalicza się do głównych zadań audytu prowadzących do wykrycia nieprawidłowości oraz oszustw finansowych. Prawo Benforda może być z powodzeniem wykorzystywane do przeprowadzenia kompleksowego badania, którego wynikiem będzie nakreślenie obszarów wymagających głębszej analizy.

Za pomocą makra arkusza kalkulacyjnego, zaprezentowanego w rozdziale czwartym, przeanalizowano zbiór danych zawierający faktury zakupowe pewnej apteki działającej na terenie Krakowa. Zestaw składający się z 5816 obserwacji zawierał dokumenty generowane na przestrzeni lat 2006–2011. Poniżej przedstawiono wyniki analizy wartości netto faktur. Możliwe jest także przeanalizowanie wartości brutto lub VAT, w celu wychwycenia ewentualnych nieprawidłowości przy stawkach podatkowych.

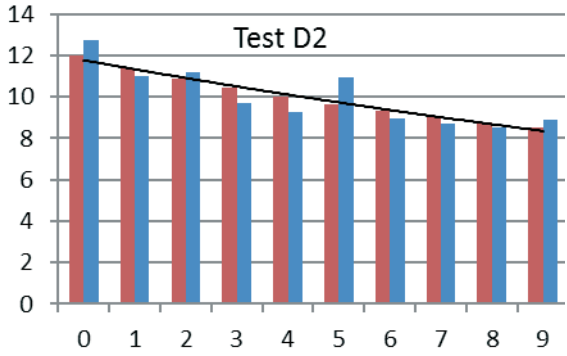
Analizę zbioru rozpocząć należy od przetestowania częstości występowania cyfr na pierwszych pozycjach znaczących. Na rysunku 5.1 przedstawiono graficzną prezentację wyników testu F1. Jak można zauważyć, wartości dla danych empirycznych różnią się od wzorca wyznaczonego przez prawo Benforda. Największym odchyleniem na pierwszej pozycji znaczącej odznacza się cyfra 1 (−6,3 %) oraz 4 (2,8 %), natomiast najmniejszym cyfra 9 (−0,1 %).

Przyglądając się bliżej częstościom cyfr występujących dokładnie na drugiej (rys. 5.2) oraz na trzeciej pozycji znaczącej (rys. 5.3) można dojść do wniosku, że nie odbiegają one w dużym stopniu od wyznaczonego wzorca. W przypadku testu D2, największe różnice zauważyć można dla cyfr 5 (1,3%), 4 (-0,8%), 0 (0,7%) oraz 3 (−0,7%). W teście D3 cyfra 0 przewyższa o 1,8% wartość teoretyczną testu, natomiast cyfra 5 występuje o 0,6% rzadziej niż to wynika z teoretycznego punktu widzenia.

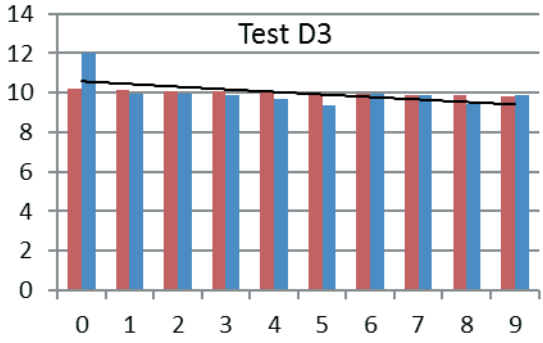
Rys. 5.1. Częstość występowania cyfr na pierwszej pozycji znaczącej



Rys. 5.2. Częstość występowania cyfr na drugiej pozycji znaczącej



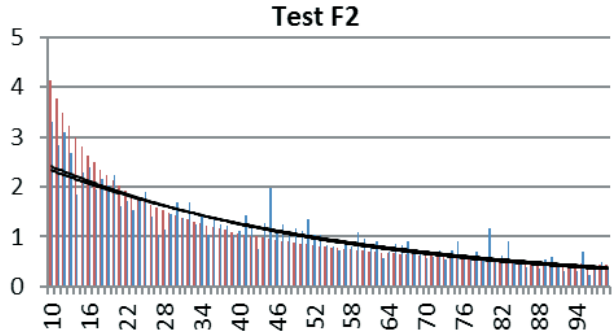
Rys. 5.3. Częstość występowania cyfr na trzeciej pozycji znaczącej



Wyniki testu F2 (rys. 5.4), badającego ciąg cyfr na pierwszych dwóch pozycjach znaczących, mogą wskazać obszary wymagające dalszej, bardziej szczegółowej i wnikliwej analizy. Częstość występowania wartości netto faktur zaczynających sie od sekwencji „45" jest niemal dwukrotnie większa,

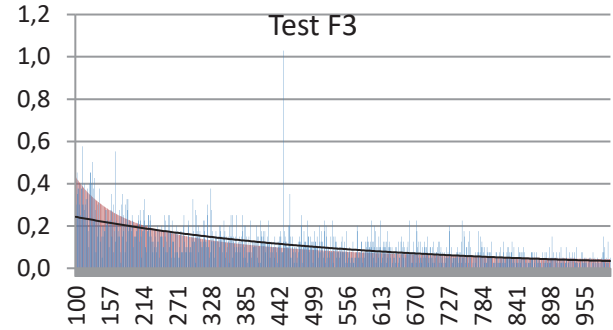
niż wynikałoby to z prawa Benforda. Naturalną rzeczą wydaje się zatem skupienie uwagi na tych właśnie dokumentach.

Rys. 5.4. Częstość występowania cyfr na pierwszych dwóch pozycjach znaczących



Kolejnym etapem analizy może być wykonanie testu F3. Opisane powyżej badanie wyłoniło faktury rozpoczynające się od ciągu cyfr „45". Przyglądając się wynikom testu F3 (rys. 5.5) można wyszczególnić kolejne dokumenty, których wartość rozpoczyna się od sekwencji „450". Tym sposobem zawęża się zbiór faktur jedynie do tych, które z największym prawdopodobieństwem mogły być obarczone błędem.

Rys. 5.5. Częstość występowania cyfr na pierwszych trzech pozycjach znaczących

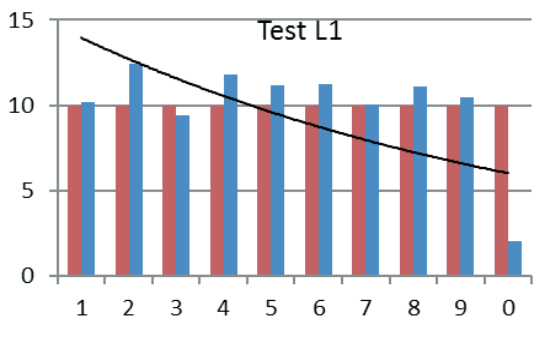


Przyglądając się dokładniej zbiorowi danych, a w szczególności fakturom wyłonionym dzięki testom F2 oraz F3, można wysunąć wnioski potwierdzające lub odrzucające nasze podejrzenia dotyczące manipulacji wartościami faktur. W przypadku przeanalizowanego zestawu danych zauważyć można, że faktury, których kwoty rozpoczynają się od sekwencji „450" stanowią pewien koszt stały, występujący cyklicznie, będący opłatą tej apteki względem biura rachunkowego świadczącego swoje usługi. Faktury wystawiane były

od początku badanego okresu, a zatem od 2006, aż do końca 2011 roku w miesięcznych odstępach. Większość firm korzysta z usług biur rachunkowych, które wystawiają regularne rachunki na z góry wynegocjowane kwoty zależne m.in. od ilości dokumentów danej firmy w ciągu miesiąca. Zatem w tym przypadku, częste występowanie faktur opiewających na stałe kwoty nie powinno wzbudzać podejrzeń. Przed wykonaniem analizy należy poznać specyfikę danych oraz obszar działalności firmy, aby umieć zinterpretować prawdopodobne anomalie.

Ostatni przeprowadzony test (L1) pokazuje częstości występowania cyfr na ostatniej pozycji znaczącej. Każda cyfra od 0 do 9 powinna występować z jednakowym prawdopodobieństwem. Na rysunku 5.6 można jednak zauważyć znaczne odchylenia pomiędzy wartościami empirycznymi a teoretycznymi. W szczególności cyfra 0 występuje stosunkowo rzadko na ostatniej pozycji znaczącej względem teoretycznego rozkładu (odchylenie wynosi 8,0 %). Natomiast cyfra 2 występuje zbyt często – różnica wynosi ok. 2,4 %.

Rys. 5.6. Częstość występowania cyfr na ostatniej znaczącej pozycji



Powyżej przedstawiona została interpretacja badania na podstawie analizy graficznej. Tabela 5.1 zawiera zestawienie mierników podobieństwa M1– M5 rozkładu empirycznego z rozkładem Benforda oraz testy zgodności: test chi-kwadrat oraz trzy wersje testów Kołmogorova–Smirnova.

Wyniki testu chi-kwadrat wskazują na brak zgodności rozkładu empirycznego z teoretycznym rozkładem Benforda we wszystkich testach z wyjątkiem F3, dla którego wartość (16,0) jest mniejsza od wartości krytycznej (16,9) na poziomie istotności równym 0,05. Wartość krytyczna dla testu F2 (16,9) jest niewiele mniejsza od wartości testu (20,8), a zatem w tym przypadku moglibyśmy zastanowić się nad przyjęciem hipotezy o zgodności dwóch rozkładów. W pozostałych testach wartość statystyki testowej

przekracza kilkakrotnie wartość krytyczną na poziomie istotności równym 0,05, co świadczyłoby o braku zgodności badanych rozkładów.

Przypatrując się bliżej wynikom testów Kołmogorova–Smirnowa możemy dojść do podobnych wniosków jak w przypadku testu chi-kwadrat. Jedynie dla testów F2 oraz F3 możemy mówić o zgodności rozkładu empirycznego z rozkładem Benforda zarówno na poziomie istotności 0,05, jak i 0,01. Wartości statystyk testowych w pozostałych testach przewyższają wartości krytyczne.

Zaprezentowane testy należą do najczęściej wykorzystywanych narzędzi weryfikacji zgodności dwóch rozkładów. Niestety, w przypadku prawa Benforda wydają się nie znajdować praktycznego zastosowania. Wynik testu chi-kwadrat zależy od liczby obserwacji, a zatem w przypadku dużych zbiorów danych, z jakimi mamy do czynienia dokonując analizy rozkładów cyfr, wartość statystyki będzie stosunkowo szybko wzrastać wraz ze wzrostem liczby obserwacji.

W przypadku badań przeprowadzanych za pomocą analizy Benforda, warto skupić się na wynikach testu ***z*** (tab. 5.2). Wyznacza on odchylenia pomiędzy rozkładem częstości występowania poszczególnych cyfr lub sekwencji cyfr na danych pozycjach a częstościami wzorcowymi, wynikającymi z prawa rozkładu cyfr.

W przypadku testu F1, największym odchyleniem odznacza się cyfra „1" (10,08) a najmniejszym „9" (0,18). Spostrzeżenia te potwierdza rysunek 5.1, na którym wyraźnie widać różnicę pomiędzy rozkładem empirycznym a rozkładem Benforda dla „1".

Test ***z*** wskazał niezgodność badanych rozkładów w przypadku siedmiu spośród dziewięciu badanych cyfr zarówno na poziomie istotności $\alpha = 0{,}05$ jak i $\alpha = 0{,}01$. Wyniki dla cyfr „3" oraz „9" są mniejsze od zadanych wartości krytycznych, a zatem wskazują na zgodność rozkładów.

Dla testu F2 największą różnicą odznacza sie cyfra „5" (3,01), natomiast dla testu F3 cyfra „0" (2,23). Wartości obu odchyleń sugerują niezgodność badanych rozkładów cyfr na poziomie istotności 0,05 (test F2 i F3) oraz 0,01 (test F2). Powyższe wyniki porównać można z rysunkami 5.2 i 5.3 (*częstość występowania cyfr na drugiej pozycji znaczącej)*.

Zgodnie z tym, co przedstawia rysunek 5.6, największym odchyleniem w przypadku testu L1 odznacza się cyfra „0" (19,59). Nie jest to jednak jedyne odchylenie, gdyż test ***z*** wskazuje na 5–6 obszarów niezgodności, w zależności od przyjętego poziomu istotności.

Test F2, badający rozkład cyfr na pierwszych dwóch pozycjach znaczących, wymaga wyznaczenia 90 statystyk testu z wynikających z liczby kom-

binacji cyfr, które mogą się pojawić. Jak ukazuje tabela 5.2, w przypadku tego testu mamy do czynienia z 27 (α=0,05) lub 15 (α=0,01) obszarami niezgodności. Z uwagi na konieczność wyznaczenia 900 statystyk testu ***z*** dla testu F3, wyniki nie zostały tu zaprezentowane.

Tab. 5.1. Testy zgodności – dane finansowe

Name	Purchase invoices – net													α=0,05	**1,36**	1,36	1,75	–	min [–] %
FV1	**Source:**	Confident												α=0,01	1,63	1,63	2	–	max [–] %
	n	**M1**	**M2**	**M3**	**M4**	**M5**	**r**	**p (r)**	**chi**	**chi–0,05**	**p(chi)**	**k**	**z – 0,05**	**z – 0,01**	**KS1**	**KS2**	**KS3**	**stat.**	**param.**
F1	5414	16,5	0,86	2,59	19,1	931,7	0,976	0	192,8	15,5	0	9	7	7	4,45	6,29	4,48	**max**	48 890
F12	5255	20,6	0,03	0,31	22,4	1003,3	0,934	0	342,1	112	0	90	24	11	3,82	5,4	4,11	**min**	–4 352,20
F123	3983	49,4	0	0,07	50,1	1671,9	0,692	0	1541,1	969,9	0	900	65	29	3,44	4,86	4,16	**avg:**	413
F2	5255	5,5	0,2	0,63	6,3	289,1	0,878	0	20,8	16,9	0,013	10	1	1	0,42	0,59	0,8	**st. dev.**	295,3
F3	3983	3,8	0,2	0,64	6,4	152,6	–	–	16	16,9	0,067	10	1	1	0,81	1,15	0,84	**kurt.**	1 178,90
L1	5416	17,1	0,88	2,79	27,9	928,4	–	–	422,3	16,9	0	10	6	6	4,16	5,88	4,16	**skew.**	30,4

*Tab. 5.2. Test **z** – dane finansowe*

Test z	0	1	2	3	4	5	6	7	8	9	max	α=0,05	α=0,01
F1	–	10,08	4,38	1,21	6,94	5,39	4,54	2,67	3,02	0,18	10,08	7	7
F2	1,8	1,36	0,73	1,51	1,91	3,01	0,45	1,29	0,44	1,4	3,01	1	1
F3	2,23	0,69	0,23	0,95	0,45	0,77	0,66	0,31	0,96	0,99	2,23	1	0
L1	19,59	0,43	6	1,43	4,5	2,87	3,1	0,2	2,74	1,2	19,59	6	5
F12–10	3,42	3,96	1,87	2,56	5,2	2,62	1,49	2,83	1,27	1,35	5,2	27	15
F12–20	0,31	2,26	1,44	1,93	0,52	0,82	1,58	3,44	2,5	0,31	3,44		
F12–30	1,25	0,19	1,97	0,51	0,83	1,27	1,07	0,42	0,5	0,59	1,97		
F12–40	0,65	2,58	0,9	1,24	2,65	7,73	1,47	2,21	1,8	2,4	7,73		
F12–50	3,45	3,76	1,53	1,67	0,28	0,4	0,1	1,11	0,44	3,59	3,76		
F12–60	1,95	1,58	2,35	0,34	0,08	2,19	2,3	2,07	0,11	1,75	2,35		
F12–70	0,41	0,33	1,51	0,18	1,68	3,03	0,59	0,48	1,66	0,11	3,03		
F12–80	6,08	0,03	0,84	3,94	0,54	0,1	0,42	0,36	1,08	0,73	6,08		
F12–90	1,58	0,73	0,88	0,17	1,18	2,5	2,11	0,43	0,85	0,34	2,5	1,96	2,58
max	19,6	4	6	3,9	5,2	7,7	3,1	3,4	2,7	3,6	19,6		

Zakończenie

W monografii przedstawiono metody i narzędzia weryfikacji rzetelności danych źródłowych wykorzystujące prawa rozkładu cyfr znaczących. Metody te można stosować dla danych spełniających określone warunki. Jednym z nich jest ograniczenie analizy do danych w skali ilorazowej o dużym zakresie zmienności.

Autorzy niniejszej pracy postawili sobie zadanie sprawdzenia, czy można opracować podobne procedury weryfikacji danych dla potrzeb e-learningu. Głównie chodzi tu o ocenę wiarygodności danych uzyskiwanych w badaniach ankietowych na temat opinii użytkowników o kursach e-learnigowych. Podobny problem występuje przy ewaluacji standardowych zajęć dydaktycznych, gdzie z założenia przyjmuje się, że wszystkie pozyskane opinie są rzetelne i wiarygodne. Generalnie problem jest znacznie szerszy i dotyczy wszelkich badań ankietowych. Można tu sformułować postulat utworzenia odrębnej dziedziny wiedzy, którą można określić mianem ankietometrii. Jednym z jej zadań byłby pomiar rzetelności danych ankietowych, a także sposoby zwiększenia stopnia ich wiarygodności.

Automatyczne przeniesienie procedur opartych na prawie Benforda jest w tym przypadku nieuzasadnione, gdyż dane ankietowe na ogół są pomiarami na skalach słabszych (rangowa, nominalna, dychotomiczna). Tym niemniej można tu podjąć próby opracowania narzędzi zbliżonych do metod przedstawionych w niniejszej monografii.

Dla przykładu, jeżeli w pytaniach ankiety ewaluacyjnej stosowana jest skala nominalna:

[1] zdecydowanie tak,
[2] raczej tak,
[3] raczej nie,
[4] zdecydowanie nie,

oraz z wcześniejszych badań dysponuje się szacunkami pozwalającymi ustalić kolejność tych wariantów wg częstości ich pojawiania się np. [2], [1], [3],

[4], to można zaproponować następującą procedurę weryfikacji rzetelności ankiet.

1. W każdym pytaniu kontrolnym wykorzystanym do weryfikacji ankiet zmienia się numerację wariantów przypisując wariantom częstszym niskie numery i *vice versa.* W powyższym przykładzie numeracja wariantów byłaby następująca: [1] raczej tak, [2] zdecydowanie tak, [3] raczej nie, [4] zdecydowanie nie;
2. Dla każdej ankiety wyznacza się parametr będący iloczynem numerów oznaczających warianty wskazane przez ankietowanych [X]. Minimalna wartość tego parametru równa jest 1 i oznacza sytuację, gdy w danej ankiecie na wszystkie pytania kontrolne zawsze wskazywano warianty najczęściej wybierane, którym przypisano numer [1]. Wartość maksymalna, to iloczyn numerów oznaczających najrzadziej spotykane warianty, np. przy 10 pytaniach mających po 4 warianty jest to 4^10=1 048 576;
3. Dokonuje się analizy rozkładu wartości parametru X. Należy oczekiwać, że częściej będą się pojawiały ankiety z mniejszymi wartościami iloczynu X niż większymi;
4. Zbiór ankiet można uznać za wiarygodny, jeżeli rozkład parametru X ma charakter monotoniczny, jeżeli natomiast wartości tego parametru wskazują na istnienie odrębnych skupisk, to podzbiór lub podzbiory ankiet o najwyższych wartościach iloczynu X należy wyeliminować z badań jako mało wiarygodne.

Zaproponowana metoda pozwala na przeprowadzanie analiz o podobnym charakterze, jak w przypadku prawa Benforda. Z pomiarów w skali nominalnej uzyskuje się skalę o dużym zakresie zmienności, a przedmiotem analizy jest parametr, który z definicji częściej przyjmuje raczej niskie wartości niż wysokie.

Jeżeli opisana wyżej metoda sprawdzi się w praktyce, to można ją stosować we wszystkich badaniach ankietowych i sondażowych, Skuteczność metody można zweryfikować na dowolnych danych ankietowych uzyskiwanych w drodze powtarzanych cyklicznie sondaży, na podstawie których można ustalić kolejność wariantów pytań wg częstości ich pojawiania się.

Spis tabel

Spis rysunków

Aneks

A1. Wykaz ważniejszych prac związanych z prawem Benforda opublikowanych przed 1970 rokiem oraz w latach 2009–2010

Lp.	Rok	Tytuł
1	1881	Newcomb, S. *Note on the frequency of use of the different digits in natural numbers*. American Journal of Mathematics 4(1), 39–40.
2	1912	Poincaré, H. *Répartition des décimales dans une table numérique*. p. 313–320 in: Calcul des Probabilités, Gauthier–Villars, Paris.
3	1916	Weyl, H. *Über die Gleichverteilung von Zahlen mod Eins*. Mathematische Annalen 77, 313–352.
4	1917	Franel, J. *A propos des tables de logarithmes*. Festschrift der Naturforschenden Gesellschaft in Zürich, Vierteljahrsschrift 62, 286–295.
5	1920	Boring, E.G. *The logic of the normal law of error in mental measurement*. American Journal of Psychology 31, 1–33. ISSN:0002–9556.
6	1938	Benford, F. *The law of anomalous numbers*. Proceedings of the American Philosophical Society 78, 551–572.
7	1939	Lévy, P. *L'addition des variables aléatoires définies sur une circonférence*. Bull. Soc. Math. France 67, 1–41.
8	1944	Goudsmit, S.A. and Furry, W.H. *Significant figures of numbers in statistical tables*. Nature 154(3921), 800–801. ISSN:0028–0836.
9	1945	Furry, W.H. and Hurwitz, H. *Distribution of numbers and distribution of significant figures*. Nature 155(3924), 52–53.
10	1946	Furlan, L.V. *Das Harmoniegesetz der Statistik: Eine Untersuchung ueber die metrische Interdependenz der sozialen Erscheinungen*. Basel, Verlag fuer Recht und Gesellschaft AG, xiii:504p.
11	1948	Geiringer, H. *Review of L.V. Furlan's book*. Journal of the American Statistical Association 43, 325–328.
12	1948	Hsü, E.H. *An Experimental Study on "Mental Numbers" and a New Application*. The Journal of General Psychology 38, 57–67.
13	1950	Moser, L. and Macon, N. *On the distribution of first digits of powers*. Scripta Mathematica 16, 290–292.
14	1952	Tsuji, M. *On the uniform distribution of numbers mod 1*. Journal of the Mathematical Society of Japan 4(3/4), 313–322.
15	1956	Herzel, A. *Sulla distribuzione delle cifre iniziali die numeri statistici*. Atti XV e XVI Riunione sci., Roma 1956, 205–228.
16	1956	Wallis, W.A. and Roberts, H.V. *First digits of statistical tables*. Example 331, p. 331–332 in: Statistics: A New Approach, The Free Press of Glencoe, Illinois, USA.
17	1957	Gini, C. *Sulla frequenza delle cifre iniziali dei numeri osservati*. Bull. Inst. Internat. Stat., 29th session, 35(2), 57–76.
18	1961	Cigler, J. and Helmberg, G. *Neuere Entwicklungen der Theorie der Gleichverteilung*. Jahresbericht der Deutschen Mathematiker Vereinigung 64, 1–50.
19	1961	Pinkham, R.S. *On the Distribution of First Significant Digits*. Annals of Mathematical Statistics 32(4), 1223–1230. ISSN:0003–4851.

Lp.	Rok	Tytuł
20	1961	Wouk, A. *On digit distributions of random variables*. J. Soc. Indust. Appl. Math. 9(4), 597–603.
21	1963	Weaver, W. *The distribution of first significant digits*. p. 270–277 in: *Lady Luck: The Theory of Probability*, Doubleday Anchor Series, New York. Republished by Dover, 1982.
22	1964	Cigler, J. *Methods of summability and uniform distribution mod 1*. Compositio Mathematica 16, 44–51.
23	1965	Good, I.J. *Letter to the Editor*. The American Statistician 19(3), 43.
24	1965	Konheim, A.G. *Mantissa distribution*. Mathematics of Computation 19, 143–144.
25	1966	Flehinger, B.J. *On the Probability that a Random Integer has Initial Digit A*. American Mathematical Monthly 73(10), 1056–1061. ISSN:0002–9890.
26	1967	Duncan, R.L. *An application of uniform distributions to the Fibonacci numbers*. Fibonacci Quarterly 5, 137–140.
27	1968	Adhikari, A.K. and Sarkar, B.P. *Distributions of most significant digit in certain functions whose arguments are random variables*. Sankhya –The Indian Journal of Statistics Series B, no. 30, 47–58.
28	1968	Shenton, L.R. *Periodicity and density of modified Fibonacci sequences*. Fibonacci Quarterly 6(2), 109–116.
29	1969	Adhikari, A.K. *Some Results on Distribution of Most Significant Digit*. Sankhya –The Indian Journal of Statistics Series B, 31 (Dec), 413–420. ISSN:0581–5738.
30	1969	Bumby, R. and Ellentuck, E. *Finitely additive measures and the first digit problem*. Fundamenta Mathematicae 65, 33–42.
31	1969	Duncan, R.L. *Note on the initial digit problem*. Fibonacci Quarterly 7(5), 474–475.
32	1969	Fairthorne, R.A. *Progress in Documentation – Empirical Hyperbolic Distribtuions (Bradford–Zipf–Mandelbrot) for Bibliometric Description and Prediction*. Journal of Documentation 25(4), 319–343; reprinted 2005 in Journal of Documentation 61(2), 171–193. ISSN:0022–0418.
33	1969	Kuipers, L. *Remark on a paper by R.L. Duncan concerning the uniform distribution mod 1 of the sequence of the logarithms of the Fibonacci numbers*. Fibonacci Quarterly 7, 465–466, 473.
34	1969	Raimi, R.A. *The Peculiar Distribution of First Digits*. Scientific American 221(6), 109–119. ISSN:0036–8733.
35	1969	Raimi, R.A. *On Distribution of First Significant Figures*. American Mathematical Monthly 76(4), 342–348. ISSN:0002–9890.
36	1970	Brown, J.R. and Duncan, R.L. *Modulo one uniform distribution of the sequence of logarithms of certain recursive sequences*. Fibonacci Quarterly 8, 482–486. ISSN:0015–0517.
37	1970	Hamming, R. *On the distribution of numbers*. Bell Syst. Tech. J. 49(8), 1609–1625.
LATA 2009–2010		
555	2009	Abumrad, J. and Krulwich, R. *From Benford to Erdős*. WNYC Radiolab, 9 October.
556	2009	Aldous, D. and Phan, T. *When Can One Test an Explanation? Compare and Contrast Benford's Law and the Fuzzy CLT*. Class project report, Statistics Department, UC Berkeley.
557	2009	Altamirano, C. and Robledo, A. *Generalized Thermodynamics underlying the laws of Zipf and Benford*. p. 2232–2237 in: J. Zhou (ed.), Complex Sciences 2009, Part II, LNICST 5. ISSN:1867–8211.
558	2009	Battersby, S. *Statistics hint at fraud in Iranian election*. New Scientist 24 June.
559	2009	Beber, B. and Scacco, A. *The Devil is in the Digits*. Washington Post, 20 June.
560	2009	Beer, T.W. *Terminal digit preference: beware of Benford's law*. Journal of Clinical Pathology 62(2), 192.
561	2009	Bialik, C. *Rise and Flaw of Internet's Election-Fraud Hunters*. The Wall Street Journal, 1 July.

Lp.	Rok	Tytuł
562	2009	Bonache, A., Moris, K. and Maurice, J. *Risque associé à l'utilisation de la loi de Benford pour détecter les fraudes dans le secteur de la mode*. Munich Personal RePEc Archive (MPRA) Paper No. 15352, posted 26 May 2009.
563	2009	Bradley, J.R. and Farnsworth, D.L. *What is Benford's Law?* Teaching Statistics 31(1), 2–6.
564	2009	Burns, B. *Sensitivity to statistical regularities: People (largely) follow Benford's law*. p. 2872–2877 in: Proceedings of CogSci 2009, Amsterdam, The Netherlands.
565	2009	Chou, M.C, Kong, Q., Teo, C.P., Wang, Z. and Zheng, H. *Benford's Law and Number Selection in Fixed-Odds Numbers Game*. Journal of Gambling Studies.
566	2009	Ciofalo, M. *Entropy, Benford's first digit law, and the distribution of everything*. Dipartimento di Ingegneria Nucleare, Università degli Studi di Palermo, Italy.
567	2009	Docampo, S., del Mar Trigo, M., Aira, M., Cabezudo, B .and Flores-Moya, A. *Benford's law applied to aerobiological data and its potential as a quality control tool*. Aerobiologia 25, 275–283. ISSN:0393–5965.
568	2009	Dorfleitner, G. and Klein, C. *Psychological barriers in European stock markets: Where are they?* Global Finance Journal 19(3), 268–285.
569	2009	Fewster, R.M. *A simple Explanation of Benford's Law*. American Statistician 63(1), 20–25.
570	2009	Gauvrit, N. and Delahaye, J.P. *Scatter and regularity imply Benford's Law ... and more*. arXive preprint.
571	2009	Gauvrit, N. and Delahaye, J.P. *Loi de Benford générale (General Benford Law)*. Mathématiques et sciences humaines/ Mathematics and Social Sciences 186, 5–15.
572	2009	Gonzalez-Garcia, J. and Pastor, G. *Benford's Law and Macroeconomic Data Quality*. International Monetary Fund Working Paper WP/09/10, Statistics Department, January 2009.
573	2009	Günnel, S. and Tödter, K.H. *Does Benford's law hold in economic research and forecasting?* Empirica 36, 273–292.
574	2009	Hayes, S. *Benford's law in relation to terminal digit preference*. Journal of Clinical Pathology 62(6), 575–576.
575	2009	Hürlimann, W. *Generalizing Benford's law using power laws: application to integer sequences*. International Journal of Mathematics and Mathematical Sciences, Article ID 970284, 10 p.
576	2009	Intilla, F. *New Pattern Found in Prime Numbers*. Mathematics, News & Press – A Blog by F. Intilla; posted May 10.
577	2009	Janvresse, E. *Quel est le début de ce nombre?* Images des Mathématiques, 26 December.
578	2009	Jolissaint, P. *Loi de Benford, relations de récurrence et suites équidistribuées II*. Elem. Math. 64 (1), 21–36.
579	2009	Judge, G. and Schechter, L. *Detecting problems in survey data using Benford's law*. J. Human Resources 44, 1–24.
580	2009	Kafri, O. *Entropy Principle in Direct Derivation of Benford's Law*. arXiv:0901.3047v2.
581	2009	Lee, J., Cho, W.K. and Judge, G. *Stigler's approach to recovering the distribution of first significant digits in natural data sets*. CUDARE Working Paper 1072, University of California, Berkeley.
582	2009	Lesser, L. and Glickman, M. *Using magic in the teaching of probability and statistics*. Model Assisted Statistics and Applications 4, 265–274.
583	2009	Lin, F., Chang, C. and Wu, S. A *study on the relationship between related party transactions and monthly sales in Taiwan's publicly issued companies*. To appear in: Journal of the Chinese Institute of Industrial Engineers.
584	2009	Lin, F., Guan, L. and Fang, W. *Heaping In Reported Earnings: Evidence from Monthly Financial Reports of Taiwanese Firms*.

Lp.	Rok	Tytuł
585	2009	Luque, B. and Lacasa, L. *The first–digit frequencies of prime numbers and Riemann zeta zeros*. Proc. Royal Soc. A, published online 22Apr09.
586	2009	Mebane, W.R., Jr. *Note on the presidential election in Iran, June 2009*. updated notes on author's website.
587	2009	Morrow, J. *Benford's Law, Families of Distributions and a Test Basis*. unpublished manuscript.
588	2009	Ni, D., Wei, L. and Ren, Z. *Benford's Law and β-Decay Half-Lives*. Commun. Theor. Phys. 51, 713–716.
589	2009	Nigrini, M.J. and Miller, S.J. *Data Diagnostics Using Second-Order Tests of Benford's Law*. Auditing: A Journal of Practice & Theory 28(2), 305–324.
590	2009	Perone, C.S. *An analysis of Benford's Law applied to Twitter*. Pyevolve, 11 August.
591	2009	Phillips, T. *Simon Newcomb and "Natural Numbers" (Benford's Law)*. American Mathematical Society Feature Column, Monthly Essays on Mathematical Topics.
592	2009	Ravikumar, B. *A simple multiplication game and its analysis*. Accepted for publication in the International Journal of Combinatorial Number Theory.
593	2009	Rehmeyer, J. *Statistical tests suggestive of fraud in Iran's election*. ScienceNews, 10 July.
594	2009	Roman, M. and Robert, C. *Le premier chiffre à gauche*. e-print, University of Paris V.
595	2009	Romero-Rochin, V. *A derivation of Benford's Law ... and a vindication of Newcomb*. Preprint arXiv:0909.3822.
596	2009	Roukema, B.F. *Benford's Law Anomalies in the 2009 Iranian presidential election*. Preprint arXiv:0906.2789.
597	2009	Ryder, P. *The Relationship Between the Newcomb–Benford Law and the Distribution of Rational Numbers*. Zeitschrift für Naturforschung 64a, 615–617.
598	2009	Ryder, P. *Multiple origins of the Newcomb–Benford law: rational numbers, exponential growth and random fragmentation*. Staats- und Universitätsbibliothek Bremen, Germany.
599	2009	Shao, L. and Ma, B.Q. *First Digit Distribution of Hadron full width*. Modern Physics Letters A, 24(40), 3275–3282.
600	2009	Steutel, F. *De wet van Benford*. STAtOR 10(1), 22–23.
601	2009	Szpiro, G. *Neues aus dem Reich der Primzahlen*. Neue Zürcher Zeitung, 27 Mai.
602	2009	Tödter, K.H. *Benford's Law as an Indicator of Fraud in Economics*. German Economic Review 10(3), p 339–351 (2009).
603	2009	Trono, J.A. *Discovering more properties of the Fibonacci sequence*. Journal of Computing Sciences in Colleges 24(5), 130–135. ISSN:1937–4771.
604	2009	Wagon, S. *Benford's Law and Data Spread*. Wolfram Online Demonstrations Projects.
605	2009	Weisstein, E.W. *Benford's Law*. MathWorld (A Wolfram Web Resource).
606	2010	Aldous, D. and Phan, T. *When Can One Test an Explanation? Compare and Contrast Benford's Law and the Fuzzy CLT*. The American Statistician 64(3), 221–227.
607	2010	Berger, A. and Hill, T.P. *Fundamental Flaws in Feller's Classical Derivation of Benford's Law*. University of Alberta preprint.
608	2010	Courtland, R. *Curious mathematical law is rife in nature*. Issue 2782, New Scientist, 14 October.
609	2010	Farkas, J. and Gyürky, G. *The significance of using the Newcomb–Benford law as a test of nuclear half-life calculations*. Acta Physica Polonica B 41, 1213–1221. ISSN:PL 0587–4254.
610	2010	Kaynar, B., Berger, A., Hill, T. and Ridder A. *Finite-state Markov Chains Obey Benford's Law*. math arXiv.
611	2010	Morrison, K.E. *The Multiplication Game*. Mathematics Magazine 83, 100–110.

Lp.	Rok	Tytuł
612	2010	Nillsen, R. *Randomness and Recurrence in Dynamical Systems: a real analysis approach*. Carus Monograph #31, Mathematical Association of America.
613	2010	Oleksy, M. *Data Mining und Benford's Law als Controllinginstrumente*. Band 45, Wismarer Schriften zu Management und Recht, Europäischer Hochschulverlag, Bremen. ISSN:978–3867414–40.
614	2010	Pan, D., Shao, L. and Ma, B.Q. *Benford's Law in Statistical Physics*. Wolfram Online Demonstrations Projects.
615	2010	Ross, K. *Benford's Law, a growth industry*. Accepted for publication in the American Mathematical Monthly..
616	2010	Sambridge, M, Tkalčić, H. and Jackson, A. *Benford's law in the Natural Sciences*. Geophysical Research Letters (to appear).
617	2010	Shao, L. and Ma, B.Q. *Empirical mantissa distributions of pulsars*. Astroparticle Physics 33, 255–262.
618	2010	Shao, L. and Ma, B.Q. *The significant digit law in statistical physics*. Physica A 389, 3109–3116.
619	2010	Shao, L. and Ma, B.Q. *First-digit law in nonextensive statistics*. Physical Review E 82, 041110.
620	2010	Strzałka, D. *On some properties of Benford's law*. Journal of the Korean Math. Soc. 47, 1055–1075.
621	2010	Szewczak, Z.S. *A limit theorem for random sums modulo 1*. Statistics & Probability Letters.

Źródło: www.benfordonline.net.

A2. Streszczenia artykułów znajdujących się na witrynie www.benfordonline.net i opublikowanych przed 1970 rokiem

S. Newcomb (1881)

Note on the frequency of use of the different digits in natural numbers
American Journal of Mathematics 4(1), p. 39–40.
INTRODUCTION: That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones. The first significant figure is oftener 1 than any other digit, and the frequency diminishes up to 9. The question naturally arises whether the reverse would be true of logarithms. That is, in a table of anti-logarithms, would the last part be more used than the first, or would every part be used equally? The law of frequency in the one case may be deduced from. that in the other. The question we have to consider is, what is the probability that if a natural number be taken at random its first significant digit will be n, its second n', etc.
Subject Area(s): General Interest.

E.G. Boring (1920)

The logic of the normal law of error in mental measurement
American Journal of Psychology 31, p. 1–33.
INTRODUCTION: No amount of practically successful "mental measurement" in laboratories, school-systems, factories or the army can relieve us, if we do not wish to waste time, of the necessity of stopping, every so often, to take account of first principles. Psychophysics, with more than half a century of history to its credit, has repeatedly found the need to eliminate its logically unfit and reorganize its forces: it is a long cry from the principle of the just noticeable difference to the principle of the psychometric function. The mental test as a new- comer had first to prove its right to consideration. Now that it has been accepted it must pass under the critical eye and learn to conform. And what needs to be said, in way of ad- monition, applies especially to the mental test, although for no other reason than that the mental test is the lustiest form of mental measurement that one meets today. It, especially, merits a discriminating encouragement.
Subject Area(s): Psychology.

F. Benford (1938)

The law of anomalous numbers
Proceedings of the American Philosophical Society 78, p. 551–572.
ABSTRACT: It has been observed that the first pages of a table of common logarithms show more wear than do the last pages, indicating that more used numbers begin

with the digit 1 than with the digit 9. A compilation of some 20,000 first digits taken from widely divergent sources shows that there is a logarithmic distribution of first digits when the numbers are composed of four or more digits. An analysis of the numbers from different sources shows that the numbers taken from unrelated subjects, such as a group of newspaper items, show a much better agreement with a logarithmic distribution than do numbers from mathematical tabulations or other formal data. There is here the peculiar fact that numbers that individually are without relationship are, when considered in large groups, in good agreement with a distribution law–hence the name " Anomalous Numbers." A further analysis of the data shows a strong tendency for bodies of numerical data to fall into geometric series. If the series is made up of numbers containing three or more digits the first digits form a logarithmic series. If the numbers contain only single digits the geometric relation still holds but the simple logarithmic relation no longer applies. An equation is given showing the frequencies of first digits in the different orders of numbers 1 to 10, 10 to 100, etc. The equation also gives the frequency of digits in the second, third * * place of a multi-digit number, and it is shown that the same law applies to reciprocals. There are many instances showing that the geometric series, or the logarithmic law, has long been recognized as a common phenomenon in factual literature and in the ordinary affairs of life. The wire gauge and drill gauge of the mechanic, the magnitude scale of the astronomer and the sensory response curves of the psychologist are all particular examples of a relationship that seems to extend to all human affairs. The Law of Anomalous Numbers is thus a general probability law of widespread application.
Subject Area(s): Applied Mathematics, General Interest, Statistics.

S.A. Goudsmit and W.H. Furry (1944)

Significant figures of numbers in statistical tables
Nature 154(3921), p. 800–801.
INTRODUCTION: It is a well-known fact that most numbers in statistical tables start with a small digit. For example, in population tables almost one third of the entries begin with the digit 1. The same holds true for most tables of the type occurring in the World's Almanac.
Subject Area(s): Applied Mathematics.

W.H. Furry and H. Hurwitz (1945)

Distribution of numbers and distribution of significant figures
Nature 155(3924), p. 52–53.
INTRODUCTION: In an earlier note it is shown that the distribution of first digits of the numbers in a table will obey a logarithmic law provided that a certain sum can be replaced by an integral. We shall here discuss this approximation in more detail.
Subject Area(s): Probability Theory.

L. Moser and N. Macon (1950)

On the distribution of first digits of powers
Scripta Mathematica 16, p. 290–292.
INTRODUCTION: In this note, we will prove that every finite sequence of digits appears as the first digits of some power of 2, and will generalize this result in several directions.
Subject Area(s): Number Theory.

R.S. Pinkham (1961)

On the Distribution of First Significant Digits
Annals of Mathematical Statistics 32(4), p. 1223–1230.
INTRODUCTION: This paper is a theoretical discussion of why and to what extent the so called "abnormal law" must hold.
Subject Area(s): Analysis, Probability Theory.

W. Weaver (1963)

The distribution of first significant digits
p. 270–277 in: *Lady Luck: The Theory of Probability*, Doubleday Anchor Series, New York. Republished by Dover, 1982.
COVERTEXT: "Should I take my umbrella?" "Should I buy insurance?" "Which horse should I bet on?" Every day – in business, in love affairs, in forecasting the weather or the stock market – questions arise which cannot be answered by a simple "yes" or "no". Many of these questions involve probability. Probabilistic thinking is as crucially important in ordinary affairs as it is in the most abstruse realms of science. This book is the best nontechnical introduction to probability ever written.
Reference Type: Book Chapter.
Subject Area(s): Probability Theory.

R.L. Duncan (1967)

An application of uniform distributions to the Fibonacci numbers
Fibonacci Quarterly 5, p. 137–140.
INTRODUCTION: Let p_n be the number of digits of the n-th Fibonacci number, and $\xi = (1+\sqrt{5})/2$. The object of this note is to show that both the upper and lower bounds in $(p_n - 1)/\log \xi \leqslant n - 1 \leqslant p_n/\log \xi$ are attained for a set of values n having positive density.
Subject Area(s): Number Theory.

A.K. Adhikari and B.P. Sarkar (1968)

Distributions of most significant digit in certain functions whose arguments are random variables

Sankhya – The Indian Journal of Statistics Series B, no. 30, p. 47–58.

SUMMARY: It is empirically well established that in large collections of numbers the proportions of entries with the most significant digit A is $\log_{10}(A+1)/A$. The property of the most significant digit has been studied in the present paper. It has been proved that when random numbers or their reciprocals are raised to higher and higher powers, they have log distributions of most significant digit in the limit. The property is also demonstrated in the limit by the products of random numbers as the number of terms in the product becomes higher and higher. The property is not, however, demonstrated by higher roots of the random numbers or their reciprocals in the limit. In fact there is a concentration at some particular digit. It has been shown that if X has log distribution of the most significant digit, so does 1/X and CX, C being any constant, under stronger conditions.

Subject Area(s): Probability Theory.

L.R. Shenton (1968)

Periodicity and density of modified Fibonacci sequences

Fibonacci Quarterly 6(2), p. 109–116.

INTRODUCTION: Periodicity of the last digit (or last two digits and so on) in a Fibonacci sequence has been discussed by Geller, use being made of a digital computer, and solved theoretically by Jarden. We may regard this as a periodic property of the right–most significant digit(s). There is a similar property for truncated Fibonacci sequences, the truncation being carried out prior to addition and on the right. Although this seems to be a somewhat artificial procedure it is the arithmetic involved on digital computers working in "floating point".

Subject Area(s): Number Theory.

A.K. Adhikari (1969)

Some Results on Distribution of Most Significant Digit

Sankhya – The Indian Journal of Statistics Series B, 31 (Dec), p. 413–420.

SUMMARY: This paper finds the distribution of the most significant digit of some functions of random variables $X_1, X_2, \ldots, X_n$, where these variables are independent and distributed uniformly in (0, 1). The probability that the most significant digit of Y_n is A (A=1, ... , 9) has been found, where Y_n is defined as the products of the reciprocals of n such random variables. It has been shown that this probability tends to $\log_{10}(A+1)/A$ as n tends to infinity. Similarly if Z_n is defined as $Z_n=X_1/X_2/\ldots/X_{n+1}$, it has been proved that the probability distribution of the most significant digit of Z_n also tends to $\log_{10}(A+1)/A$ as n tends to infinity. More generally, it is found that if $V_1, V_2, \ldots, V_n$ are defined as $V_1=B/X, \ldots, V_n=V_{n-1}/X_n$ where B is any random variable defined on the positive axis of the real line, the probability distribution of the most significant digit tends to $\log_{10}(A+1)/A$ as n tends to infinity.

Subject Area(s): Probability Theory.

R. Bumby and E. Ellentuck (1969)

Finitely additive measures and the first digit problem
Fundamenta Mathematicae 65, p. 33–42.
ABSTRACT: The first significant digit conjecture is stated as follows: The proportion of physical constants whose first significant digit lies between 1 and n, where $1 \leqslant n \leqslant 9$, is $\log_{10}(n+1)$. In this connection the authors define various sets of finitely additive set functions defined on P(N), the power set of N, where N is the set of natural numbers, in order to find a ``reasonable'' class of measures for which the first significant digit conjecture for natural numbers would be probabilistically verified. $\mathcal{M}$ is the set of non–atomic measures, i.e., those which satisfy the properties (i) $\mu(A \cup B)=\mu(A)+\mu(B)$ for $A,B \subset N$, $A \cap B= \emptyset$, (ii) $\mu(N)=1$, (iii) $\mu(\{n\})=0$ for all $n \in N$. T consists of the translation invariant measures, which satisfy the additional property that $\mu(A)=\mu(A+1)$ for $\mu \in T$ and for all $A \subset N$. If C is any class of measures and $A \subset N$, C(A) is defined to be the set $\{ \mu(A) \mid \mu \in C\}$. The authors prove that if P is the set of natural numbers having first significant digit equal to 1, T(P) is the entire interval $[0,1]$, thereby showing that translation invariant measures are too general to settle the first significant digit problem. The authors then proceed to extend the measures contained in $\mathcal{M}$ and T to the class S of sparse sets, which are the sets A of positive real numbers having the property that the cardinality of the set $A \cap [n,n+1)$ is bounded for all $n \in N$. R, the class of scale invariant measures, is defined to consist of those $\mu \in T$ for which $\mu(A)= \alpha\, \mu(\alpha\, A)$ for every $A \in S$ and every $\alpha >0$. In other words, ''thinning'' the set A by multiplying each element in it by α has the ''reasonable'' effect of multiplying its measure by $1/\alpha$. A somewhat technical theorem is proved which immediately implies that if R is restricted to P(N) and if P_n is the set of natural numbers whose first significant digit lies between 1 and n, then for $1 \leqslant n \leqslant 9$, $S(P_n)$ is the singleton $\log_{10}(n+1)\}$ which verifies the conjecture for any $\mu \in R$.
Subject Area(s): Measure Theory.

R.L. Duncan (1969)

Note on the initial digit problem
Fibonacci Quarterly 7(5), p. 474–475.
INTRODUCTION: The initial digit problem is concerned with the distribution of the first digits which occur in the set of all positive integers. If A is the set of all positive integers with initial digit a, then the asymptotic density of A, if it exists, would provide a suitable answer to the question "What is the probability that an integer chosen at random has initial digit equal to a?". However, it is easily shown that the asymptotic density doesn't exist. The purpose of this note is to show that the logarithmic density of A exists and is equal to $\log (1+1/a)$, where log x is the common logarithm.
Subject Area(s): Number Theory.

R.A. Fairthorne (1969)

Progress in Documentation – Empirical Hyperbolic Distribtuions (Bradford–Zipf-Mandelbrot) for Bibliometric Description and Prediction
Journal of Documentation 25(4), p. 319–343; reprinted 2005 in Journal of Documentation 61(2), 171–193.
ABSTRACT: Since 1960, and especially during the past three years, many papers have appeared about particular manifestations and applications of a certain class of empirical laws to a field that may be labelled conveniently 'Bibliometrics'. This term, resuscitated by Alan Pritchard (see page 348), denotes, in my paraphrase, quantitative treatment of the properties of recorded discourse and behaviour appertaining to it.
Subject Area(s): Statistics.

L. Kuipers (1969)

Remark on a paper by R.L. Duncan concerning the uniform distribution mod 1 of the sequence of the logarithms of the Fibonacci numbers
Fibonacci Quarterly 7, p. 465–466, 473.
INTRODUCTION: In the following we present a short proof of a theorem by RL Duncan.
Subject Area(s): Analysis, Number Theory.

R.A. Raimi (1969)

The Peculiar Distribution of First Digits
Scientific American 221(6), p. 109–119.
INTRODUCTION: In numbers that appear in tables of constants, lists of street addresses and similar tabulations the first digit of the number is 1 almost three times more often than one would expect. Why?
Subject Area(s): General Interest.

J.R. Brown and R.L. Duncan (1970)

Modulo one uniform distribution of the sequence of logarithms of certain recursive sequences
Fibonacci Quarterly 8, p. 482–486.
INTRODUCTION: The purpose of this paper is to show that the sequence (ln V_n) is uniformly distributed mod 1, where (V_n) is defined by a linear recurrence $V_{n+k} = a_{k-1}V_{n+k-1} + \ldots + a_0 V_n$, $n \geqslant 1$, the initial terms $V_1, V_2, \ldots, V_k$ being given positive numbers.
Subject Area(s): Number Theory.

R. Hamming (1970)

On the distribution of numbers

Bell Syst. Tech. J. 49(8), p. 1609–1625.

ABSTRACT: This paper examines the distribution of the mantissas of floating point numbers and shows how the arithmetic operations of a computer transform various distributions toward the limiting distribution $r(x) = 1/(x \ln b)$, $1/b \leqslant x \leqslant 1$, where b is the base of the number system. The paper also gives a number of applications to hardware, software, and general computing which show that this distribution is not merely an amusing curiosity. A brief examination of the distribution of exponents is included.

Subject Area(s): Applied Mathematics.

Bibliografia

1. Adamic A., Huberman B.A., *Zipf's Law and the Internet*, Glottometrics, 3/2002, p. 143–150.
2. Aida M., Takahashi N., Abe T., *A proposal of Dual Zipfian Model for describing HTTP access trends and its application to address cache design*, IEICE Transactions on Communications, 81(7)/1998, p. 1475–1485.
3. Altman A., *Zipfian linquistics*, Glottometrics, 3/2002, p. 19–26.
4. Benford F., *The Law of Anomalous Numbers*, Proceedings of the American Philosfical Society, 78/1938, p. 551–572.
5. Carslaw C., *Anomalies in Income Numbers: Evidence of Goal Oriented Behavior*, The Accounting Review 63(2)/1988, p. 321–327.
6. Dembowska M., *Nauka o informacji naukowej (informatologia). Organizacja i problematyka badań w Polsce,* Instytut informacji naukowej, technicznej, ekonomicznej IINTE, Warszawa, 1991.
7. Farbaniec M., Grabiński T., Zabłocki B., Zając W., *Analiza wpływu przekształceń matematycznych na zbiory o zadanym rozkładzie cyfr, Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne*, Wydawnictwo Wyższej Szkoły Handlowej, Kielce, 2011.
8. Farbaniec M., Grabiński T., Zabłocki B., Zając W., *Application of the first digit law in credibility evaluation of the financial-accounting data based on particular cases,* Revizor – Časopis za teoriju i praksu. Auditor – Jornal for theory and practice, IEF Institut za Ekonomiku i Finansije, Belgrad, 2012.
9. Furlan L.V., *Das Harmoniesgestez der Statistik,Eine Untersuchung uber die metrische Interdependenz der sozialeen Erscheinungen*, Bael, Switzerland, Verlag fur Recht und Geselschaft, XIII/1948.
10. Gabaix X., *Zipf's Law and the growth of cities*, American Economic Review, 89/1999, p. 129–132.
11. Giles D.E.A., *Benford's Law and Naturally Occurring Prices in Certain e-Bay Auctions*, Applied Economics Letters, 14/2007, p. 157–161.
12. Goudsmith S.A., Furry W.H., *Significant figure of numbers in statistical tables*, Nature, 154/1944, p. 800–801.
13. Grabiński T., *Metody taksonometrii*, Akademia Ekonomiczna w Krakowie, Kraków 1991.
14. Grabiński T., *Analiza taksonometryczna krajów Europy w ujęciu regionów*, Akademia Ekonomiczna w Krakowie, Kraków 2003.
15. Grabiński T., *Propozycje w zakresie porządkowania diagramu Jana Czekanowskiego*, w pr. zbior. *Studia z zakresu metod ilościowych w ekonomii, demografii i socjologii*, Prace Komisji Socjologicznej PAN, O. Kraków, nr 40/1977, Wrocław–Warszawa–Kraków–Gdańsk.
16. Ha L.Q., Sicilia-Garcia E.I., Ming J., Smith F.J., *Extension of Zipf's Law to Words and Phrases*, The Association for Computational Linguistics, A Digital Archive of Research Papers.
17. Hill T.P., *A Statistical Derivation of the Significant-Digit Law*, Statistical Science 10(4)/1996, p. 354–363.
18. Hill T.P., *The first digital phenomenon*, American Scientist, 86/1998, p. 358–363.
19. Hill T.P., *Base–Invariance Implies Benford's Law*, Proceedings of the American Mathematical Society 123(3)/1995, p. 887–895.
20. Hill T.P., *The Significant-Digit Phenomenon*, American Mathematical Monthly 102(4)/1995, p. 322–327.
21. Hurlimann W., *A generalized Benford Law and its applications*, Advances and Applications in Statistics, 3/2003, p. 217–228.
22. Hurlimann W., *Generalizing Benford's Law Using Power Law: Application to Integer Sequences*, International Journal of Mathematics and Mathematical Sciences, 2009.
23. Kafri O., *Sociological inequality and the second law*, https://arxiv.org/abs/0805.3206.
24. Kafri O., *The second law as a cause of the evolution*, https://arxiv.org/abs/0711.4507.

25. Kantorovich A.V., Miller S.J., *Benford's law, values of L-functions and the 3x+1 problem*, Acta Arithmetica, 120(3)/2005, p. 269–297.
26. Koshy T., *Fibonacci and Lucas Numbers with Applications*, Wiley 2001.
27. Kuiper N.H., *Alternative proof of a theorem of Birnhaum and Pyke*, Annals of Mathematical Statistics 30/1959, p. 251–252.
28. Lee J., Tam Cho W.K., Judge G.G., *Stigler's approach to recovering the distribution of first significant digits in natural data sets*, Statistical and Probability Letters, 80/2010, p. 82–88.
29. Leijenhorst van D.C., Weide van der T.P., *A formal derivation of Heaps' Law*, Information Science, 170/2005, p. 263–272.
30. Ley E., *On the Peculiar Distribution of the U.S. Stock Indexes' Digits*, The American Statistician, 50(4)/1996, p. 311–313.
31. Loannides Y.M., Overman H.G., *Zipf's law for cities: an empirical examination*, Regional Science and Urban Economics 2002.
32. Lotka A.J., *The frequency distribution of scientific productivity*, Journal of the Washington Academy of Science, 16/1926, p. 317–323.
33. Luque B., Lacasa L., *The first-digit frequencies of prime numbers and Riemann zeta zeros*, Proceedings of the Royal Society A, 465/2009, p. 2197–2216.
34. Newcomb S., *Note on the frequency of use of the different digits in natural numbers*, American Journal of Mathematics 4(1)/1881, p. 39–40.
35. Nigrini M.J., *I've got your number: How a mathematical phenomenon can help CPAs uncover fraud and Rother irregulaties*, AICPA Journal of Accountancy Online Journal, 5/1999.
36. Nigrini M.J., Miller S., College W., *Data diagnostics using second order tests of Benford`s Law*, A Journal of Practice and Theory, 2009.
37. Nigrini M.J., *Can Benford's law be used in forensic accounting?*, The Balance Sheet, VI/1993, p. 7–8.
38. Nigrini M.J., *A taxpayer compliance application of Benford's law*, Journal of the American Taxation Association 18(1)/1996, p. 72–91.
39. Nigrini M.J., *The use of Benford's Law as an aid in analytical procedures*, Auditing – A Journal of Practice & Theory 16(2)/1997, p. 52–67.
40. Nigrini M.J., Miller S.J., *Benford's law applied to hydrological data – results and relevance to other geophysical data*, Math. Geol., 39/2007, p. 469–490.
41. Nigrini M.J., Mittermaier L.J., *Numerology for Accountants*. Journal of Accountancy, November 1998, p. 15.
42. Nigrini M.J., *Using digital frequencies to detect fraud*, Fraud Magazine, The White Paper Index 8(2)/1994, p. 3–6.
43. Nigrini M.J., *Adding value with digital analysis*, The Internal Auditor 56(1)/1999, p. 21–23.
44. Nigrini M.J., *Program_Details_2009.doc*, DATAS 2009.
45. Pacioli L., *Summa de Arithmetica Geometria, Proportioni et Proportionalita,* 1494.
46. Paul D.B., Baker, J.M., *The Design for the Wall Street Journal – based CSR Corpus*, Proc. ICSLP, 1992, p. 899–902.
47. Pietronero L., Tossati E., Tossati V., Vespignani A., *Explaining the uneven distribution of numbers in nature: the laws of Benford and Zipf*, Physica A, 293/2001, p. 297–304.
48. Pinkham R.S., *On the Distribution of First Significant Digits*, Annals of Mathematical Statistics 32(4)/1961, p. 1223–1230.
49. Raimi R.A., *The Peculiar Distribution of First Digits,* Scientific American 221(6)/1969, p. 109–119.
50. Raimi R.A., *On Distribution of First Significant Figures*, American Mathematical Monthly 76(4)/1969, p. 342–348.
51. Ratajewski J., *Wybrane problemy metodologiczne informologii nauki (informacji naukowej),* Prace Naukowe UŚ, Katowice, 1994.
52. Reed W.J., *The Pareto, Zipf and other power laws*, Economic Letters, 74/2001, p. 15–19.
53. Rousseau R., Kingsley G., *Zipf: life, ideas, his law and informetrics*, Glottometrics, 3/2002, p. 11–18.

54. Schurger K., *Extensions of Black-Scholes processes and Benford's law*, Stochastic Processes and Their Applications, vol.118, 7/2008, p. 1219–1243.
55. Sordylowa B., *Informacja naukowa w Polsce. Problemy teoretyczne, źródła, organizacja,* Ossolineum, Wrocław, 1987.
56. Stigler G.J., *The distribution of leading digits in statistical tables*, Chicago, 1945.
57. Stigler G.J., *The Economics of Information*, Journal of Political Economy, 69/1961.
58. Ścibor E., *Informacja naukowa w Polsce: tradycja i współczesność*, Olsztyn, 1998.
59. Thomas J.K., *Unusual patterns in reported earnings*, The Accounting Review, 64/1989, p. 773–787.
60. Washington L.C., *Benford's Law for Fibonacci and Lucas Number*, Fibonacci Quarterly, 19/1981, p. 175–177.
61. Zipf G.K., *To honor*, Glottometrics, 3/2002.
62. Zipf G.K., *National Unity and Disunity*, The Nation as a Bio-Social Organism, Principia Press, Bloomingotn Indiana, Princeton Press, 1941.
63. Żmigrodzki Z. i in. (red.), *Informacja naukowa: rozwój, metody, organizacja,* Wyd. SBP, Warszawa, 2006.

Netografia

http://www.aclweb.org/anthology/C/C02/C02-1117.pdf.
http://www.aicpa.org/pubs/jofa/may1999/ nigrini.htm.
http://www.benfordonline.net.
http://www.dataminingalapolonaise.wordpress.com/2010/01/06/prawo-benforda.
http://www.en.wikipedia.org/wiki/Benford's_law.
http://www.eskimo73.republika.pl/download/manual 30 pl.pdf.
http://www.eskimo73.republika.pl/maczek.html.
http://www.ezrstats.com.
http://www.fq.math.ca.
http://www.hmetic.ch.
http://www.intuitor.com/statistics/Benford's%20Law.html.
http://www.ipipan.waw.pl/~ldebowsk/uslugi/index.html.
http://www.math.gatech.edu/~hill/publications/cv.dir/1st-dig.pdf.
http://www.mathpages.com/HOME/kmath302/kmath302.htm.
http://www.mathworld.wolfram.com/BenfordsLaw.html.
http://www.mccombs.utexas.edu/faculty/jonathan.koehler/docs/sta309h/Benford_1998.pdf.
http://www.motherfunctor.org.
http://www.neuroinf.pl/Members/danek/swps/2008/Article.2008-05.../getFile.
http://www.newscientist.com/article/dn14461-five-scientific-discoveries-that-got-the-wrong- name.html? DCMP=ILC-hmts&nsref=news10_head_dn14461.
http://www.nigrini.com.
http://www.pl.wikipedia.org/wiki/Rozkład_Benforda.
http://www.portal.wsiz.rzeszow.pl/plik.aspx?id=2618.
http://www.rexswain.com/benford.html.
http://www.skg.pl/acl.
http://www2.statistics.com/resources/software/commercial/e/Ezrstat.php.
http://www.statsoft.pl.
http://www.tphill.net.