Fenomen rozkładu Benforda

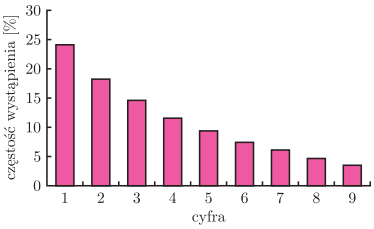
Jakub SZYMANOWSKI[*]

Większość osób świadomych powiązań między światem matematyki a rzeczywistością zgodzi się, że na każdym kroku spotykamy się z rachunkiem prawdopodobieństwa. Oprócz niektórym dobrze znanych zagadnień związanych z grami losowymi pewne prawidłowości probabilistyczne możemy spotkać również w bardziej niespodziewanych miejscach.

Przypadek czy coś więcej?

Wykonamy trzy eksperymenty na różnych, niepowiązanych seriach danych. W każdym z eksperymentów wyznaczymy częstość występowania każdej cyfry na najbardziej znaczącej pozycji w pewnym zbiorze wartości liczbowych. Wyniki przedstawimy za pomocą tabeli i wykresu.

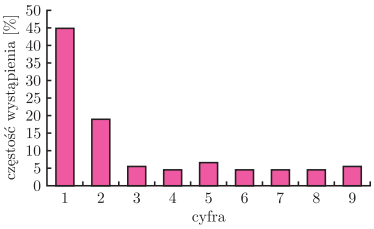
Eksperyment I. Ze zbioru liczb naturalnych z zakresu od 1 do 9999 losujemy liczbę p, wykorzystując generator liczb losowych o rozkładzie równomiernym. Następnie z zakresu liczb naturalnych od 1 do p losujemy, również wykorzystując rozkład równomierny, liczbę r. Całą tę operację powtarzamy 100 000 razy, otrzymując w ten sposób listę R wszystkich wylosowanych liczb r. Dla każdej cyfry wyznaczamy (procentowo) jej częstość występowania na najbardziej znaczącej pozycji w elementach listy R – przybliżone wyniki prezentuje poniższa tabela i wykres na rysunku 1.



Rys. 1. Wyniki eksperymentu I.

cyfra	1	2	3	4	5	6	7	8	9
częstość wystąpienia [%]	24,27	18,40	14,61	11,65	9,32	7,46	6,08	4,74	3,47

Eksperyment II. W drugim eksperymencie posłużymy się układem okresowym pierwiastków chemicznych, a dokładniej, jednym z parametrów każdego pierwiastka – masą atomową. Podobnie jak w eksperymencie pierwszym, interesuje nas jedynie pierwsza cyfra każdej liczby reprezentującej masę atomową. Częstość występowania wszystkich cyfr na tej pozycji (w przybliżeniu) można odczytać z poniższej tabeli i wykresu na rysunku 2.

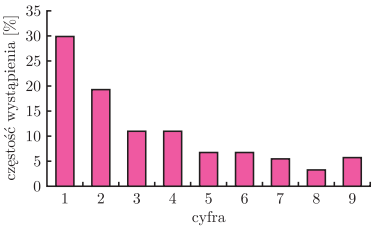


Rys. 2. Wyniki eksperymentu II.

Zbiór danych oparto o układ okresowy pierwiastków odkrytych do roku 1938 zatwierdzony przez IUPAC (ang. *International Union of Pure and Applied Chemistry*) `http://www.iupac.org`

cyfra	1	2	3	4	5	6	7	8	9
częstość wystąpienia [%]	44,94	19,10	5,62	4,49	6,74	4,49	4,49	4,49	5,62

Eksperyment III. Ostatni eksperyment ma charakter geograficzny – posłużymy się tutaj powierzchnią w km^2 wszystkich państw świata. Znów badamy tylko częstość występowania poszczególnych cyfr na najbardziej znaczącej pozycji; przybliżone wyniki zawarte zostały w tabeli i na rysunku 3.



Rys. 3. Wyniki eksperymentu III.

cyfra	1	2	3	4	5	6	7	8	9
częstość wystąpienia [%]	29,96	19,41	10,97	10,97	6,75	6,75	5,49	3,38	5,91

Pomysły na eksperymenty II i III zaczerpnięte zostały z publikacji F. Benforda *The law of anomalous numbers* [1]. W eksperymencie II posłużono się tylko tymi pierwiastkami, które odkryte zostały przed publikacją F. Benforda. Eksperyment I przeprowadzony został zgodnie z pomysłem autora.

Przyjrzyjmy się wynikom przeprowadzonych eksperymentów. Wykresy są niczym innym, jak empirycznie wyznaczonymi funkcjami gęstości prawdopodobieństwa dla następującego zadania: losujemy liczbę z danego zbioru i pytamy, z jakim prawdopodobieństwem jej pierwszą cyfrą będzie 1, z jakim 2, itd. Wykonaliśmy eksperymenty zupełnie niepowiązane, wyznaczone zaś funkcje gęstości prawdopodobieństwa wydają się podejrzanie podobne... Widać, że (na ogół) im niższa cyfra, tym bardziej prawdopodobne jest jej wystąpienie na początku losowej liczby ze zbioru danych.

Tablice logarytmiczne

Charakterystyczna zależność, jaką udało nam się spostrzec, została po raz pierwszy odnotowana w 1881 roku przez kanadyjskiego astronoma i matematyka Simona Newcomba. Przebywając w bibliotece United States Naval Observatory, Newcomb zauważył, że strony tablic logarytmicznych są brudniejsze na początku i coraz czystsze na dalszych kartkach. Wywnioskował, że korzystający z tablic

[*]inżynier Działu Systemów Statkowych w firmie IT-REM Sp. z o.o.

logarytmicznych częściej szukają liczb rozpoczynających się od niższych cyfr – te znajdują się na początku tablic. Swoje odkrycie (bez dowodu ogólnej prawidłowości) opublikował na stronach *American Journal of Mathematics*. Jego artykuł [4] nie spotkał się jednak z szerokim zainteresowaniem i niezwykle ciekawe zjawisko zostało zapomniane na 57 lat.

W 1938 roku Frank Benford, inżynier General Electric, nie zdając sobie sprawy z istnienia pracy Newcomba, dokonał tego samego odkrycia na podstawie stanu czystości tablic logarytmicznych. Zafascynowany tym zjawiskiem Benford zaczął sprawdzać, czy jego teoria znajduje potwierdzenie również w innych zbiorach danych, m.in. w powierzchniach rzek, liczbach drukowanych w gazetach, czy nawet cenach. Wyniki swoich badań przedstawił w artykule [1] wydrukowanym w *Proceedings of the American Philosophical Society*. Podobnie jak w artykule Newcomba, formalny dowód nie został przedstawiony.

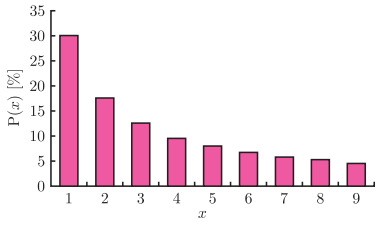
Prawo Benforda

W ten sposób świat dowiedział się o niezwykłej prawidłowości, która obecnie nosi nazwę *prawa Benforda*, *rozkładu Benforda* lub *prawa pierwszych* (*znaczących*) *cyfr*.

Dyskretny rozkład Benforda opisany jest zależnością

$$\mathrm{P}(x) = \log_{10}\left(1 + \frac{1}{x}\right), \tag{1}$$

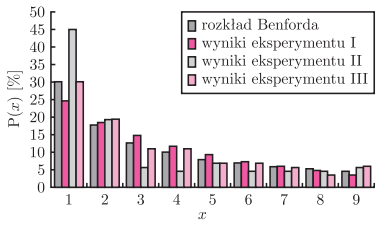
gdzie x oznacza pierwszą znaczącą cyfrę ($x = 1, 2, \ldots, 9$), natomiast $\mathrm{P}(x)$ oznacza prawdopodobieństwo, z jakim cyfra x będzie pierwszą cyfrą liczby.



Rys. 4. Funkcja gęstości prawdopodobieństwa rozkładu Benforda.

Przybliżone prawdopodobieństwo wystąpienia poszczególnych cyfr na najbardziej znaczącej pozycji przedstawia poniższa tabela, a funkcję gęstości prawdopodobieństwa – rysunek 4.

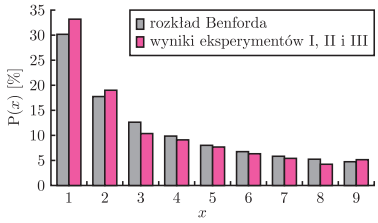
x	1	2	3	4	5	6	7	8	9
$\mathrm{P}(x)$ [%]	30,1	17,61	12,49	9,69	7,92	6,69	5,80	5,12	4,58



Rys. 5. Porównanie wyników eksperymentów z rozkładem Benforda.

Możemy teraz porównać wyniki przeprowadzonych przez nas eksperymentów z zależnością (1) – graficznie przedstawia to rysunek 5.

Skoro prawo Benforda działa dla trzech niezależnych zbiorów danych, to powinno działać również wtedy, gdy rozpatrzymy wyniki wszystkich eksperymentów jednocześnie, co pokazuje rysunek 6. Korzystając z mocniej zróżnicowanych danych, otrzymaliśmy wyniki bardziej zbliżone do przewidywań teoretycznych.



Rys. 6. Porównanie wyników eksperymentów z rozkładem Benforda.

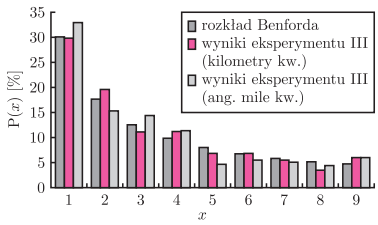
Uniwersalność prawa Benforda

Ważnym pytaniem jest, czy prawo Benforda jest uniwersalne: czy uzyskalibyśmy taki sam rozkład prawdopodobieństwa, gdybyśmy przeskalowali dane w zbiorze testowym? Na przykład, czy w eksperymencie III rozkład zmieni się, jeśli zastosujemy inne jednostki powierzchni, na przykład jardy, stopy lub mile kwadratowe?

Mówimy, że rozkład prawdopodobieństwa $\mathrm{P}(X)$ zmiennej losowej X jest niezmienniczy względem skalowania, jeżeli dla dowolnej liczby dodatniej α zachodzi równość $\mathrm{P}(X) = \mathrm{P}(\alpha \cdot X)$.

W 1961 roku Roger Pinkham stwierdził, że jeżeli prawo Benforda rzeczywiście występuje, to powinno mieć własność uniwersalności – wyniki powinny być takie same, niezależnie od tego, jakie miary stosujemy w danym zagadnieniu (zob. [5]).

Sprawdźmy to zatem, modyfikując eksperyment III: powierzchnie państw przeliczamy z kilometrów kwadratowych na angielskie mile kwadratowe i sprawdzamy częstotliwość występowania cyfr na najbardziej znaczącej pozycji.



Rys. 7. Porównanie wyników eksperymentów z rozkładem Benforda.

Z wykresu na rysunku 7 widać, że skalowanie danych prawie nie wpłynęło na rozkład prawdopodobieństwa. Niewielkie rozbieżności wynikają z faktu, iż dane te nie tworzą idealnego rozkładu Benforda.

Czy prawo Benforda działa zawsze?

Przytoczone eksperymenty pokazują, że prawo Benforda sprawdza się (z większą lub mniejszą dokładnością) dla wyników działań na liczbach naturalnych, parametrów pierwiastków chemicznych i danych geograficznych. Dodatkowo w artykule Benforda [1] można znaleźć szereg innych zbiorów danych, w których

Rozwiązanie zadania M 1298.
Liczba $a^2+b^2+c^2$ jest nieparzysta, a więc postaci $2p+1$. Wybierzmy $d=p$. Wtedy $a^2+b^2+c^2+d^2=(p+1)^2$. Pozostaje więc wykazać, że liczba p jest nieparzysta.

Liczby a, b, c są nieparzyste, a więc liczby a^2, b^2, c^2 dają z dzielenia przez 4 resztę 1. Wobec tego $2p+1 = a^2+b^2+c^2 \equiv 3 \pmod 4$, skąd wynika, że liczba p jest nieparzysta.

odnajdujemy tę prawidłowość. Możemy się zatem pokusić o pytanie, czy rozkład Benforda działa dla każdych zebranych danych liczbowych? Odpowiedź, oczywiście, brzmi: nie!

W eksperymencie III posłużyliśmy się danymi geograficznymi: powierzchnią w km^2 wszystkich państw świata. Badamy zatem dane, na które ma wpływ wiele czynników. Powierzchnia poszczególnych państw jest bardzo zróżnicowana – od Rosji o powierzchni 17 075 400 km^2 po Watykan – 0,44 km^2.

Jeżeli za bardzo zawęzimy zakres danych, okaże się, że prawo Benforda nie ma dla nich zastosowania. Na przykład, badając długości samochodów osobowych lub wysokość dorosłej żyrafy stwierdzimy, że niewiele z nich zaczyna się od cyfry 1. Wynika to z faktu, iż wartości tych danych są silnie ograniczone innymi czynnikami. Mało która żyrafa, zwłaszcza dorosła, mierzy poniżej 2 metrów.

Może warto zatem pamiętać o prawie Benforda, rzucając sześcienną kostką do gry? Niestety, także nie. Każda liczba oczek ma takie samo prawdopodobieństwo wylosowania. Powtarzając wielokrotnie losowanie, uzyskamy rozkład prawdopodobieństwa zbliżony do równomiernego.

W 1995 roku amerykański profesor matematyki z Georgia Institute of Technology, Theodore P. Hill, przedstawił dowód prawa Benforda na łamach magazynu *Statistical Science* w tekście *A statistical derivation of the significant-digit law* [3].

Tylko ciekawostka?

Prawo Benforda jest samo w sobie bardzo ciekawym zjawiskiem, a w niektórych dziedzinach ma zastosowanie praktyczne. Służy jako narzędzie do sprawdzania poprawności obliczeń, prawdziwości danych statystycznych czy wykrywania oszustw w zeznaniach podatkowych i rozliczeniach finansowych.

Za pomocą prawa Benforda sprawdza się dokładność działania modeli matematycznych opisujących ewolucję danych z różnych dziedzin, na przykład modeli zmian populacji. Dla danych wejściowych spełniających prawo Benforda powinniśmy otrzymać dane wyjściowe, które również tę zależność spełniają. Jeżeli tak nie jest, oznacza to, że zastosowany model (algorytm) zakłócił „naturalny” rozkład danych.

Najpopularniejszym zastosowaniem prawa Benforda jest sprawdzanie poprawności zeznań podatkowych i rozliczeń. Okazuje się, że fałszerze bardzo często wybierają liczby rozpoczynające się od 4, 5 i 6 zamiast od 1, 2 i 3! Stąd, jeśli rozkład częstości występowania cyfr na pierwszych pozycjach nie jest zbliżony do rozkładu Benforda, to sprawdzający powinien zwrócić na to rozliczenie większą uwagę. Z całą pewnością o prawie Benforda nie wiedział skarbnik stanu Arizona, James Nelson, którego fałszerstwa na kwotę bliską 2 mln dolarów zostały wykryte przy zastosowaniu prawa pierwszych cyfr.

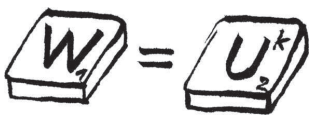
Literatura

[1] F. Benford, *The law of anomalous numbers*, Proc. Amer. Philos. Soc. 78 (1938), 551–572.
[2] T.P. Hill, *The first digit phenomenon*, Amer. Scientist 86 (1998), 358–363.
[3] T.P. Hill, *A statistical derivation of the significant-digit law*, Statist. Sci. 10 (1995), 354–363.
[4] S. Newcomb, *Note on the frequency of use of the different digits in natural numbers*, Amer. J. Math. 4 (1881), 39–40.
[5] R.S. Pinkham, *On the distribution of first significant digits*, Ann. Math. Statist. 32 (1961), 1223–1230.

Słowa pierwsze

Jakub RADOSZEWSKI

W numerze 10/2010 *Delty* pojawił się artykuł Wojciecha Plandowskiego, w którym autor po ciężkich bojach pokazuje rozwiązanie pewnego konkretnego typu równania na słowach. Mogłoby się wydawać: udało się, sprawa skończona. Tymczasem przy okazji w artykule pojawia się definicja i kilka ważnych własności słów pierwotnych, a stąd już tylko mały krok do innej ciekawej rodziny słów, mianowicie do słów pierwszych. To dobry pretekst, by coś o nich opowiedzieć.



Obrót cykliczny słowa polega na przerzuceniu dowolnej (w tym zerowej) liczby liter z początku słowa na koniec, np. wszystkimi obrotami cyklicznymi słowa *aba* są: *aba*, *baa* oraz *aab*.

Przypomnijmy, że słowo *pierwotne* to takie, które nie jest potęgą (u^k dla $k \geqslant 2$) żadnego niepustego słowa. Znamy już kilka własności takich słów, w szczególności to, że każde słowo w przedstawia się jednoznacznie w postaci $w = u^k$, gdzie u jest pierwotne; w tym artykule będzie nam wygodnie nazwać u *pierwiastkiem pierwotnym* słowa w. Dalej, wiemy, że każdy obrót cykliczny słowa pierwotnego jest pierwotny, a dodatkowo wszystkie takie obroty stanowią różne słowa. Wśród tych obrotów jedno słowo jest, w szczególności, najmniejsze *leksykograficznie*,