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.

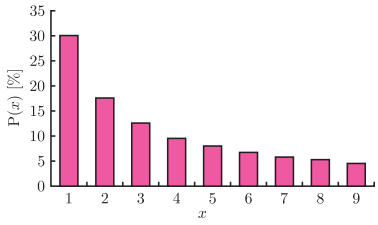
Prawo Benforda

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

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

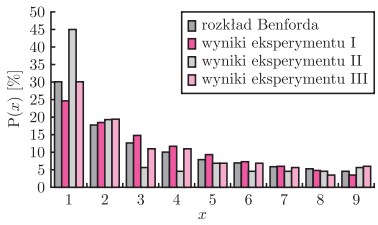
gdzie x oznacza pierwszą znaczącą cyfrę ($x = 1, 2, \ldots, 9$), natomiast $\mathrm{P}(x)$ oznacza prawdopodobieństwo, z jakim cyfra x będzie pierwszą cyfrą liczby.



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.

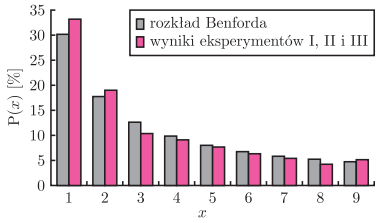
x	1	2	3	4	5	6	7	8	9
$\mathrm{P}(x)$ [%]	30,1	17,61	12,49	9,69	7,92	6,69	5,80	5,12	4,58



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.

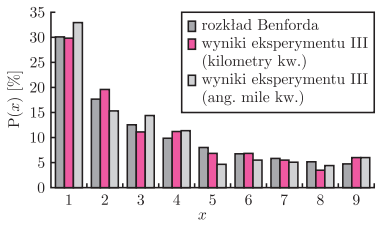
Uniwersalność prawa Benforda

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.

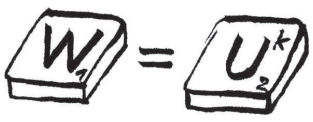
Literatura

[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*,

tj. najmniejsze w porządku słownikowym. Każde takie słowo pierwotne najmniejsze w klasie swoich obrotów cyklicznych nazwiemy właśnie słowem *pierwszym* (inna nazwa: słowo Lyndona). Kilka przykładów słów pierwszych: b, $aabab$, $aaaab$ oraz niepierwszych: $aabaab$, $abaab$. Zanim wyjaśnimy nieco tajemniczą nazwę rozważanej rodziny słów, przyjrzyjmy się następującej, równoważnej definicji słów pierwszych:

Mówimy, że słowo u jest nie większe leksykograficznie niż słowo v (co zapisujemy po prostu jako $u \leqslant v$), jeśli albo u jest prefiksem (czyli początkowym fragmentem) v, albo na pierwszej pozycji, na której u i v różnią się, w u występuje litera mniejsza niż w v. Oczywiście $u < v$, jeśli $u \leqslant v$ i $u \neq v$. Przykładowo: $aba < abaab$, $aabab < abaab$.

Twierdzenie 1. *Niepuste słowo u jest pierwsze wtedy i tylko wtedy, gdy każdy właściwy sufiks u (czyli końcowy fragment różny od u i od słowa pustego* 1*) jest leksykograficznie większy niż u.*

Dowód. ($\Leftarrow$) Załóżmy, że słowo u jest mniejsze leksykograficznie od wszystkich swoich właściwych sufiksów. Słowo u musi być pierwotne, gdyż w przeciwnym razie mielibyśmy $u = w^k$ dla $k \geqslant 2$ i sufiks u postaci w^{k-1} byłby mniejszy leksykograficznie niż u. Dalej, jeśli mamy $u = xy$ i $x, y \neq 1$, to $u < y$, a skoro y jest krótsze niż u, to u nie może być prefiksem y i mamy $u < yx$. To pokazuje, że dowolny obrót cykliczny u jest większy leksykograficznie niż u, czyli rzeczywiście u jest pierwsze.

1 oznacza słowo puste (czyli słowo zeroliterowe).

($\Rightarrow$) Załóżmy, że u jest pierwsze, i niech $u = xy$ dla $x, y \neq 1$; chcemy wykazać, że $u < y$. Dowód przeprowadzimy nie wprost. Gdyby zachodziło $y \leqslant u$ i y *nie* byłoby prefiksem u, to mielibyśmy także $yx < u$ i u nie byłoby pierwsze. Skoro tak, to y jest prefiksem u, ale przecież także i jego sufiksem, czyli $u = xy = yz$ dla pewnego z. Czytelnikowi zaznajomionemu z artykułem Wojciecha Plandowskiego taka równość może coś przypominać: otóż jest to dokładnie „główne” równanie, jakie zostało w nim rozwiązane! Korzystając z gotowego wyniku, mamy, że $x = (pq)^i$, $z = (qp)^i$ oraz $y = (pq)^j p$ dla pewnych słów p oraz $q \neq 1$, czyli $u = (pq)^k p$ dla $k = i + j \geqslant 1$. Dodajmy, że słowo p także musi być niepuste, gdyż w przeciwnym przypadku $j \geqslant 1$ (jako że $y \neq 1$) i $u = q^k$ dla $k \geqslant 2$, czyli u nie byłoby pierwotne ani, tym bardziej, pierwsze.

Miłośnicy algorytmów tekstowych (takich jak algorytm wyszukiwania wzorca KMP) rozpoznają zapewne w tym równaniu i jego rozwiązaniu związek między prefikso-sufiksami słowa a jego okresami.

No to teraz pójdzie już z górki. Słowo u jest pierwsze, więc rozważając jego dwa wybrane obroty cykliczne, otrzymujemy: $u = (pq)^k p = p(qp)^k < (qp)^k p$ oraz $u = (pq)^k p < p(pq)^k$. Na mocy pierwszej nierówności mamy $pq < qp$, natomiast z drugiej, po usunięciu początkowego p otrzymujemy $qp < pq$. Te dwie nierówności dają nam oczekiwaną sprzeczność. $\triangle$

Otrzymana sprzeczność dowodzi, że słowa pierwsze są bezokresowe (nie mają żadnego nietrywialnego okresu).

Nie koniec na tym – istnieje jeszcze inna, równoważna, choć nieco bardziej egzotyczna definicja słów pierwszych. Jej uzasadnienie pozostawiamy Czytelnikowi, nam i tak będzie ona potrzebna jedynie jako własność tej rodziny słów.

Twierdzenie 2. *Słowo u jest słowem pierwszym wtedy i tylko wtedy, gdy jest jednoliterowe albo jest postaci $u = xy$, przy czym $x < y$ i x, y są słowami pierwszymi.*

Po tym wstępie możemy już wreszcie zdradzić, skąd wzięła się nazwa rozważanej rodziny słów. Chodzi mianowicie o następującą analogię do liczb pierwszych: każdą liczbę naturalną można przedstawić jednoznacznie (oczywiście z dokładnością do kolejności) jako iloczyn liczb pierwszych, a każde słowo w pewnym sensie jednoznacznie jako sklejenie słów pierwszych.

Twierdzenie 3. *Dowolne słowo u można przedstawić jednoznacznie jako sklejenie pewnej liczby słów pierwszych $u = l_1 l_2 \dots l_k$ dla $l_1 \geqslant l_2 \geqslant \dots \geqslant l_k$.*

Dowód. Czy każde słowo można w ogóle przedstawić jako sklejenie (jakichkolwiek) słów pierwszych? Można, wystarczy podzielić je na pojedyncze litery, które, jak by nie patrzeć, są słowami pierwszymi. Teraz, dopóki w bieżącym rozkładzie występują dwa kolejne słowa pierwsze l_i, l_{i+1}, takie że $l_i < l_{i+1}$, dopóty możemy takie słowa sklejać w jedno słowo pierwsze $l_i l_{i+1}$ na mocy Twierdzenia 2. Powtarzając tę operację do skutku, znajdziemy jakieś przedstawienie u w postaci sklejenia nierosnącego ciągu słów pierwszych.

No dobrze, a czemu takie sklejenie jest tylko jedno? Załóżmy, że byłyby dwa różne, $u = l_1 l_2 \dots l_k = l'_1 l'_2 \dots l'_m$. Możemy założyć, że $l_1 \neq l'_1$, w przeciwnym przypadku możemy tę parę słów wykreślić z rozkładu. Niech więc, bez straty ogólności, będzie $|l_1| > |l'_1|$. Wówczas musi istnieć rozkład: $l_1 = l'_1 l'_2 \dots l'_i \alpha$, przy czym α jest *niepustym* prefiksem l'_{i+1}. No i teraz możemy zapisać ciąg nierówności, który prowadzi do sprzeczności, a którego uzasadnienie pozostawiamy Czytelnikowi (jedyne nietrywialne miejsce polega na zastosowaniu Twierdzenia 1):

$|u|$ oznacza długość słowa u, czyli liczbę liter w tym słowie.

Polecamy Czytelnikowi sprawdzenie Twierdzenia 3 na jakimś przykładzie, np. dla słowa $abaababaabaab$.

$$l'_1 < l_1 < \alpha \leqslant l'_{i+1} \leqslant l'_1. \quad \triangle$$

To w takim razie już „wszystko” jasne, poznaliśmy jakieś mniej lub bardziej interesujące własności słów pierwszych, w tym wyjaśniliśmy ich nazwę, można by właściwie zakończyć niniejszy artykuł. Niewykluczone jednak, że Czytelnikowi, który dobrnął do tego miejsca, pozostał po lekturze pewien niedosyt. Chciałoby się, żeby te słowa pierwsze miały jakieś *naprawdę* fajne własności (tak jak liczby pierwsze), a może nawet żeby miały jakiś