Koherencja i prawo Zipfa

Plan prezentacji

1. Czym jest prawo zipfa
2. Prawo Benforda
3. Koherencja prawa zipfa

Czym jest prawo Zipfa?

Prawo Zipfa jest prawem empirycznym głoszące, że wiele rodzajów danych tworzonych przez ludzi lub odnoszących się do ich zachowań cechuje charakterystyczny rozkład wartości, w którym dystrybucja częstotliwości występowania poszczególnych wartości jest odwrotnie proporcjonalna do ich rangi statystycznej

Wspomnianą zależność o której mówi prawo Zipfa można matematycznie przedstawić wzorem:

$$r \times f = \mathbf{constans}$$

r - ranga częstotliwości występowania

f - częstotliwość występowania

Prawo Zipfa dla tekstów

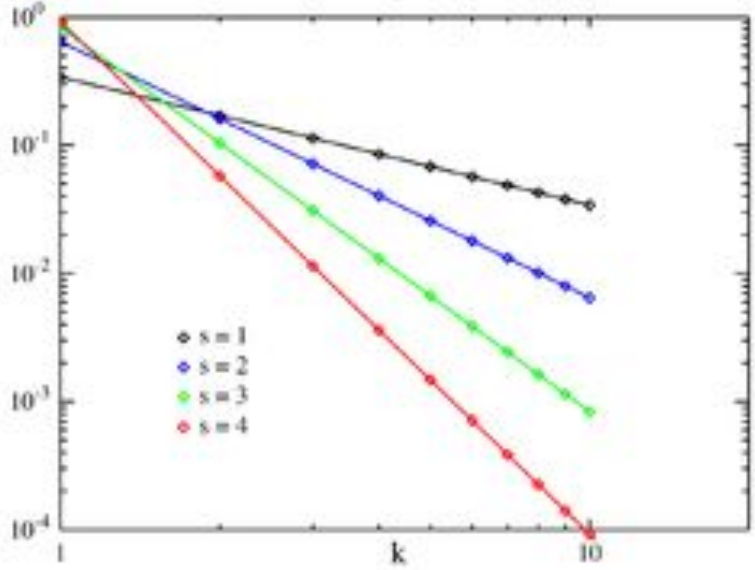
Zależność opisywaną przez prawo Zipfa można łatwo przedstawić na przykładzie testów. Przykładem obrazującym to prawo może być przykładowo **korpus Browna**. W korpusie tym mamy:

- Słowo ‘the’ stanowiące ok. 7% tekstu z 1 pozycją statystyczną
- Słowo “of” stanowiące ok. 3,5% tekstu z 2 pozycją statystyczną
- Słowo “a” stanowiące ok. 1,75% testu z 3 pozycją statystyczną

Jak widać każde kolejne najczęściej występujące słowo występuje dwukrotnie rzadziej niż poprzednie.

Prawo Zipfa

Prawo Zipfa dobrze ilustrują wykresy ze skalą logarytmiczną. Ze względu na zależność opisywaną przez prawo zipfa wykresy przedstawiające zależności dla niego zachodzące na skali logarytmicznej tworzą funkcje liniowe.



Występowanie prawa Zipfa

Istnieje wiele zjawisk do których aplikuje się prawo Zipfa. Kilka z nich to:

- Częstości występowania wyrażeń matematycznych w tekstach technicznych
- częstości występowania wysokości nut w tekstach muzycznych
- ranking wielkości miast
- rozkład wysokości dochodów osobistych

Prawo Benforda - szczególny przypadek prawa Zipfa

Jest to prawo określające prawdopodobieństwo występowania pierwszej cyfry w wielu danych:

- wartości stałych fizycznych
- wartości określające powierzchnię jezior
- kwoty w zeznaniach podatkowych

Prawo Benforda

W tabeli przedstawiono częstotliwości występowania cyfr systemu dziesiętnego na pierwszej pozycji.

Taka zależność jest dość powszechnie spotykana w życiu codziennym.

Pierwsza cyfra	Częstość
1	30,1%
2	17,6%
3	12,5%
4	9,7%
5	7,9%
6	6,7%
7	5,8%
8	5,1%
9	4,6%

Prawo Benforda

$$P(d) = \frac{\log_{10}\left(1 + \frac{1}{d}\right)}{\log_{10}(B)}$$

Gdzie:

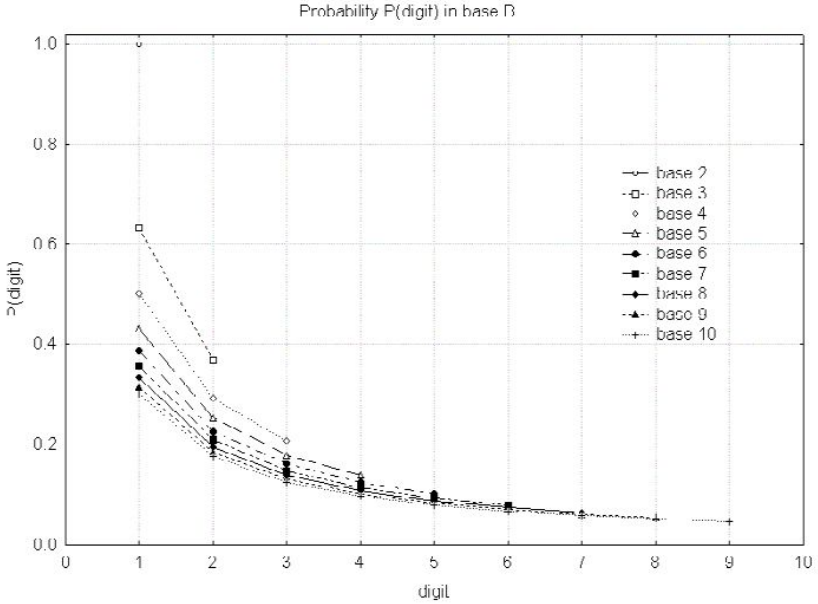
d - konkretna cyfra z określonego systemu liczbowego

B - baza systemu liczbowego

P(d) - prawdopodobieństwo pojawienia się cyfry z przedziału <1; B - 1> jako pierwszej

Prawo Benforta

Analizując prawo Benforta można dojść do ciekawego spostrzeżenia. W miar wzrostu wartości podstawy systemu B prawdopodobieństwo, że pierwsza znacząca cyfra będzie równa *d* stopniowo maleje.



Koherencja prawa Zipfa

Jeżeli zbiór cechuje się prawem Zipfa to podzbiór tego zbioru nie koniecznie cechuje się prawem Zipfa.

Podobnie dla unii dwóch zbiorów posiadających prawo Zipfa

$$x(k) = x_M / k$$

x - wartość

k - ranga

x_m - maksymalna wartość

Wzory

$$k' = k - k^*$$

$$x(k') = \frac{x'_M}{[(k'/k^*) + 1 + (1/k^*)]} \cong \frac{x'_M}{[(k'/k^*) + 1]}$$

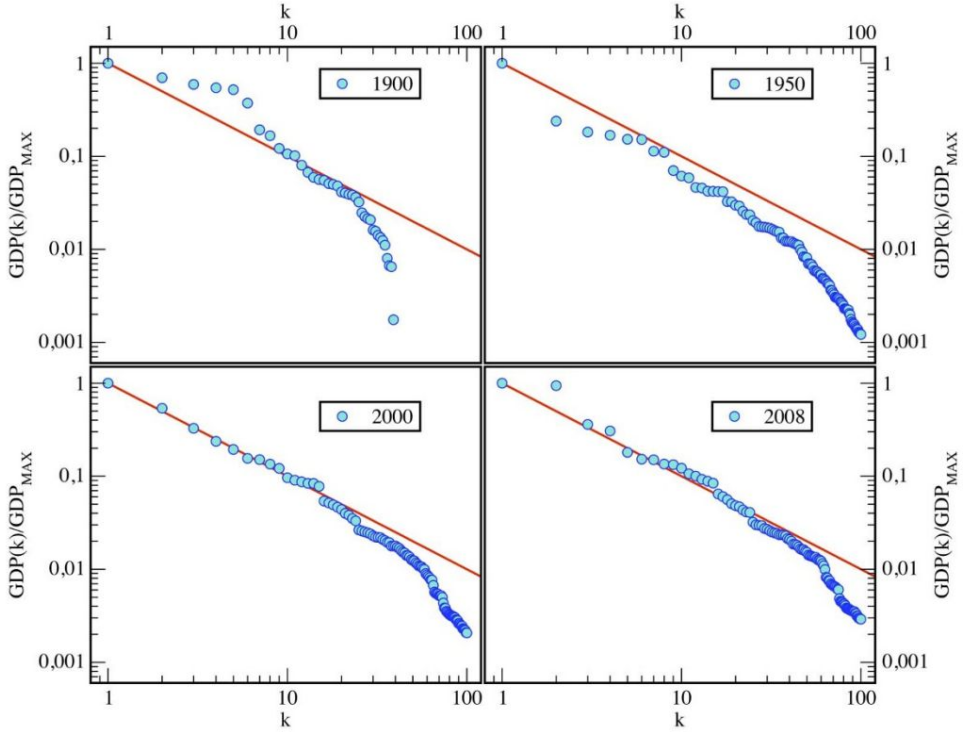
Ratio w całym zbiorze $x(k+1)/x(k) = k/(k+1),$

Ratio w podzbiorze $(k^* + k')/(k^* + k' + 1)$

Nie są takie same

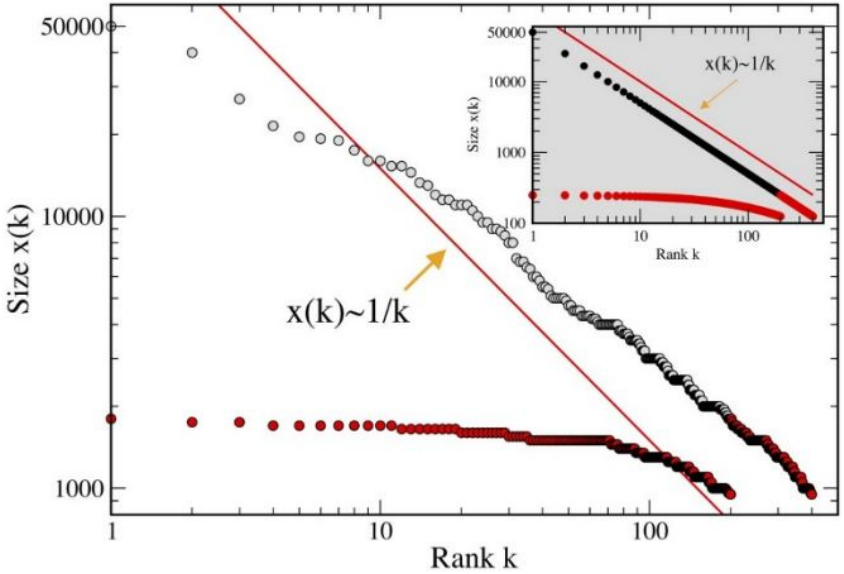
Przykłady braku koherencji

- PKB 100 krajów
- Dla małej liczby krajów, nie ma dużej zbieżności
- Wraz z globalizacją i przyrostem liczby krajów, prawo Zipfa zaczyna obowiązywać



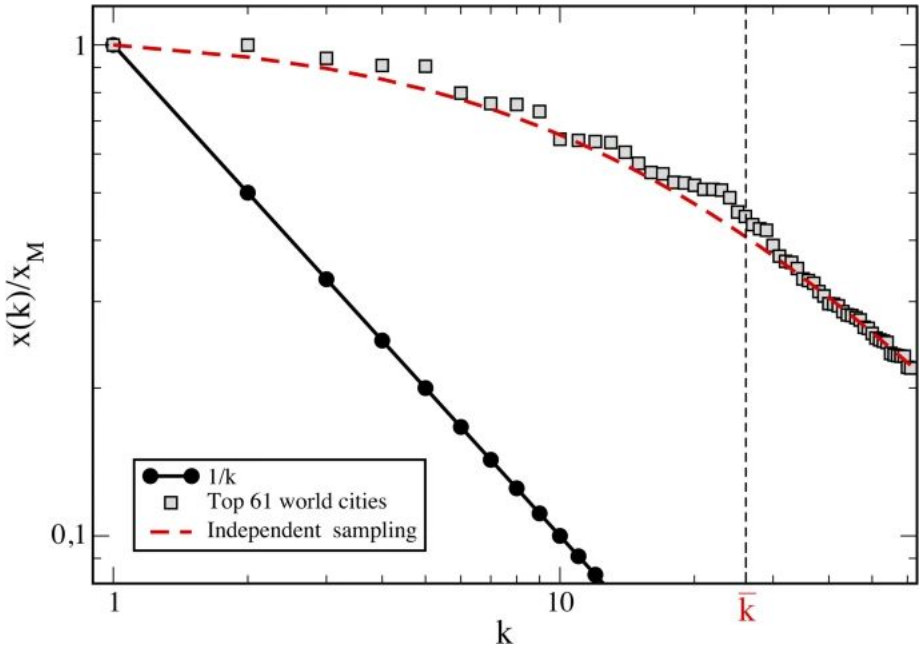
Najbogatsi ludzie w USA

- Przykład braku koherencji prawa Zipfa dla dwóch zbiorów
- Zbiór pierwszych 195 najbogatszych ludzi świata, oznaczonych na szaro, spełnia prawo Zipfa
- Zbiór drugich 195 najbogatszych ludzi świata, oznaczonych na czerwono, nie spełnia prawa Zipfa



Rozmiar miast w świecie

- Pokazuje odchylenie od prawa Zipfa
- Udowadnia, że niekoniecznie każdy zbiór danych będzie reprezentował prawo Zipfa



Dziękujemy za uwagę

Ireneusz Biela
Karol Brzegowy
Dominik Czerniak
Marek Darocha