Motivation [edit]

Zipf's law states that given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc. For example, in the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly 7% of all word occurrences (69,971 out of slightly over 1 million). True to Zipf's Law, the second-place word "of" accounts for slightly over 3.5% of words (36,411 occurrences), followed by "and" (28,852). Only 135 vocabulary items are needed to account for half the Brown Corpus.[4]

A "Startlingly Neat & Simple" Rule & Five Graphs About Patterns That Might Surprise You

February 17, 2015

By Plotly

(This article was first published on **Plotly**, and kindly contributed to R-bloggers)

George Zipf popularized an idea—Zipf's Law—that approximates populations of cities, distribution of money in counties, and how frequently words are used. Nobel Prize-winning columnist Paul Krugmans wrote of Zipf's Law that

Zipf Distribution

George Zipf popularized an idea- "Zipfs Law" - that approximates populations of cities, distribution of money in counties, and how frequently words are used.

Nobel Prize-winning columnist Paul Krugmans wrote of Zipfs Law that

> *the usual complaint about economic theory is that our models are oversimplified that they offer excessively neat views of complex, messy reality. [In the case of Zipfs law] the reverse is true: we have complex, messy models, yet reality is startlingly neat and simple.*