

# Data from our man Zipf

## Principles of Complex Systems

### CSYS/MATH 300, Fall, 2010

Prof. Peter Dodds

Department of Mathematics & Statistics  
Center for Complex Systems  
Vermont Advanced Computing Center  
University of Vermont



The  
UNIVERSITY  
of VERMONT



COMPLEX SYSTEMS CENTER



# Outline

Data from our man  
Zipf

Zipf in brief

Zipf in brief

Zipfian empirics

Zipfian empirics

References

References



# George Kingsley Zipf:

## In brief:

- ▶ Zipf (田) (1902–1950) was a linguist at Harvard, specializing in Chinese languages.
- ▶ Unusual passion for statistical analysis of texts.
- ▶ Studied human behavior much more generally...

## Zipf in brief

Zipfian empirics

References

## Zipf's masterwork:

- ▶ “Human Behavior and the Principle of Least Effort”  
Addison-Wesley, 1949  
Cambridge, MA [2]
- ▶ Bonus field of study: Glottometrics. (田)
- ▶ Bonus ‘word’ word: Glossolalia. (田)



# Human Behavior/Principle of Least Effort:

Data from our man  
Zipf

Zipf in brief

Zipfian empirics

References

## From the Preface—

Nearly twenty-five years ago it occurred to me that we might gain considerable insight into the mainsprings of human behavior if we viewed it purely as a natural phenomenon like everything else in the universe, ...

## And—

... the expressed purpose of this book is to establish **The Principle of Least Effort** as the primary principle that governs our entire individual and collective behavior ...



# The Principle of Least Effort:

Data from our man  
Zipf

Zipf in brief

Zipfian empirics

References

## Zipf's framing (p. 1):

“... a person in solving his immediate problems will view these against the background of his probable future problems *as estimated by himself.*”

“... he will strive ... to minimize the *total work* that he must expend in solving *both* his immediate problems *and* his probable future problems.”

“[he will strive to] minimize the *probable average rate of his work-expenditure...*”



# Rampaging research

## Within Human Behavior and the Principle of Least Effort:

- ▶ City sizes
- ▶ # retail stores in cities
- ▶ # services (barber shops, beauty parlors, cleaning, ...)
- ▶ # people in occupations
- ▶ # one-way trips in cars and trucks vs. distance
- ▶ # new items by dateline
- ▶ weight moved between cities by rail
- ▶ # telephone messages between cities
- ▶ # people moving vs. distance
- ▶ # marriages vs. distance
- ▶ Observed **general dependency of 'interactions'** between **cities A and B** on  $P_A P_B / D_{AB}$  where  $P_A$  and  $P_B$  are population size and  $D_{AB}$  is distance between A and B.  $\Rightarrow$  'Gravity Law.'

Data from our man  
Zipf

Zipf in brief

Zipfian empirics

References



# Zipfian empirics:

- ▶ **vocabulary balance:**  $f \sim r^{-1} \rightarrow r \cdot f \sim \text{constant}$   
( $f$  = frequency,  $r$  = rank).

TABLE 2-1

Arbitrary Ranks with Frequencies  
in James Joyce's *Ulysses*  
(Hanley Index)

I Rank ( $r$ )	II Frequency ( $f$ )	III Product of I and II ( $r \times f = C$ )	IV Theoretical Length of <i>Ulysses</i> ( $C \times 10$ )
10	2,653	26,530	265,500
20	1,311	26,220	262,200
30	926	27,780	277,800
40	717	28,680	286,800
50	556	27,800	278,800
100	265	26,500	265,000
200	133	26,600	266,000
300	84	25,200	252,000
400	62	24,800	248,000
500	50	25,000	250,000
1,000	26	26,000	260,000
2,000	12	24,000	240,000
3,000	8	24,000	240,000
4,000	6	24,000	240,000
5,000	5	25,000	250,000
10,000	2	20,000	200,000
20,000	1	20,000	200,000
29,899	1	29,899	298,990



# Zipfian empirics:

- ▶  $f \sim r^{-1}$  for word frequency:

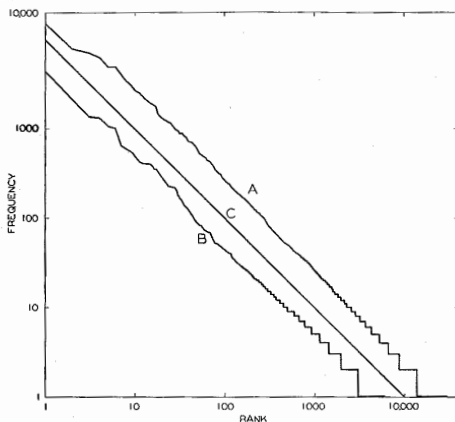


Fig. 2-1. The rank-frequency distribution of words. (A) The James Joyce data; (B) the Eldridge data; (C) ideal curve with slope of negative unity.





# Zipf's basic idea:

## Forces of Unification and Diversification:

- ▶ Easiest for the speaker to use just one word.
  - ▶ **Encoding is simple** but **decoding is hard**
- ▶ Zipf uses the analogy of tools: **one tool for all tasks**.
  
- ▶ Optimal for listener if all pieces of information correspond to different words (or morphemes).
- ▶ Analogy: a specialized tool for every task.
  - ▶ **Decoding is simple** but **encoding is hard**
  
- ▶ Zipf thereby argues for a tension that should lead to an uneven distribution of word usage.
- ▶ No formal theory beyond this...



# Zipfian empirics:

- ▶ Number of meanings  $m_r \propto f_r^{1/2}$  where  $r$  is rank and  $f_r$  is frequency.

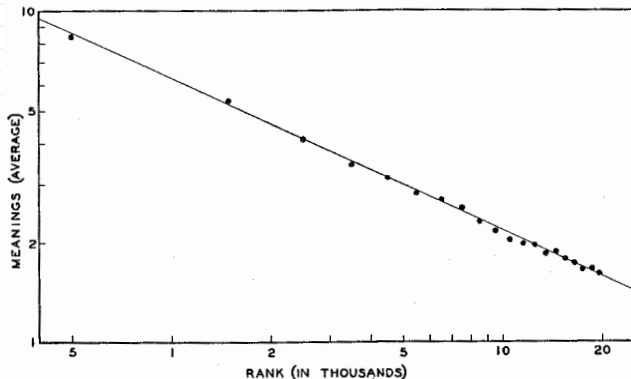


Fig. 2-2. The meaning-frequency distribution of words.



# Zipfian empirics:

- ▶ Article length in the Encyclopedia Britannica:

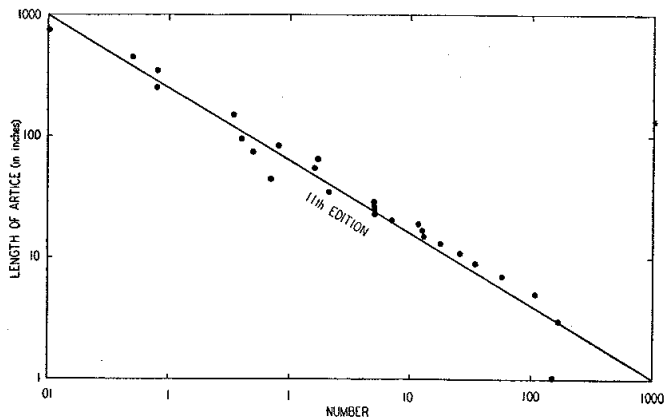


Fig. 5-3. The number of different articles of like length in samples of the 11th edition of the *Encyclopaedia Britannica*. Lengths in inches.

- ▶ (?) slope of  $-3/5$  corresponds to  $\gamma = 5/3$ .



# Zipfian empirics:

- ▶ Population size of districts:

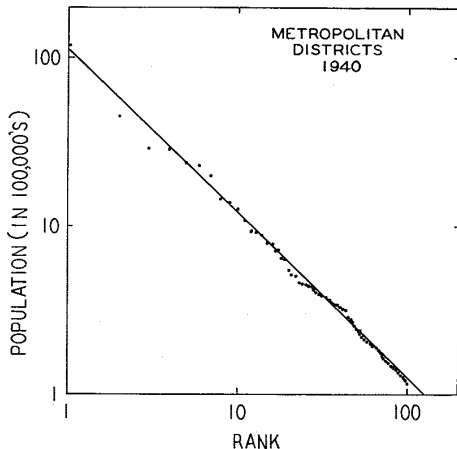


Fig. 9-2. Metropolitan districts. One hundred largest in the U. S. A. in 1940, ranked in the order of decreasing population size.

- ▶  $\alpha = 1$  corresponds to  $\gamma = 1 + 1/\alpha = 2$ .



# Zipfian empirics:

- ▶ Number of employees in organizations

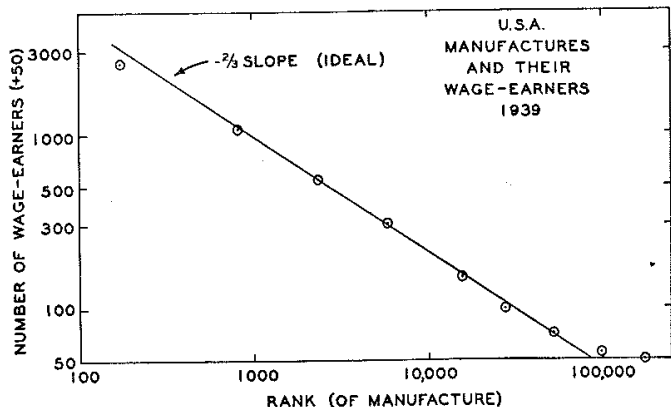


Fig. 9-8. Manufactures and their wage earners in the U. S. A. in 1939, with the manufactures ranked in the order of their decreasing number of wage earners.

- ▶  $\alpha = 2/3$  corresponds to  $\gamma = 1 + 1/\alpha = 5/2$ .



## Zipfian empirics:

- ▶ # news items as a function of population  $P_2$  of location in the Chicago Tribune
- ▶  $D$  = distance,  $P_1$  = Chicago's population
- ▶ Solid line = +1 exponent.

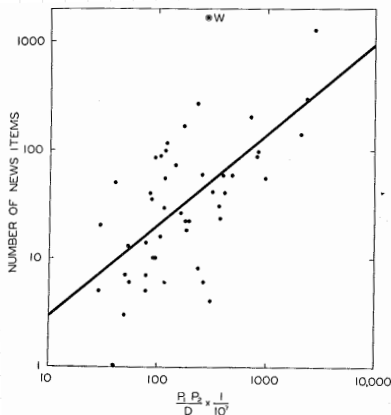


Fig. 9-10. Number of different news items in *The Chicago Tribune* (*W* is the dateline of Washington, D. C.).



## Zipfian empirics:

- ▶ # obituaries in the New York Times for locations with population  $P_2$ .
- ▶  $D$  = distance,  $P_1$  = New York's population
- ▶ Solid line = +1 exponent.

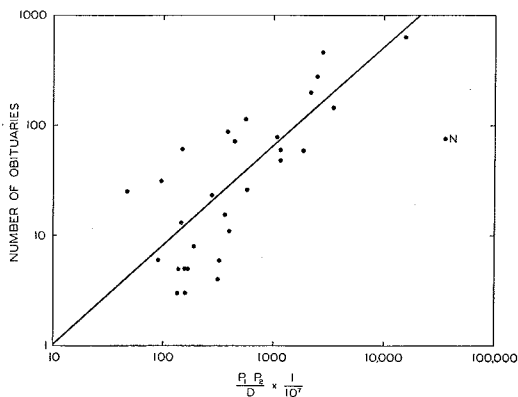


Fig. 9-11. Number of obituaries in *The New York Times* ( $N$  represents Newark, New Jersey).



# Zipfian empirics:

- ▶ Movement of stuff between cities
- ▶  $D$  = distance,  $P_1$  and  $P_2$  = city populations.
- ▶ Solid line = +1 exponent.

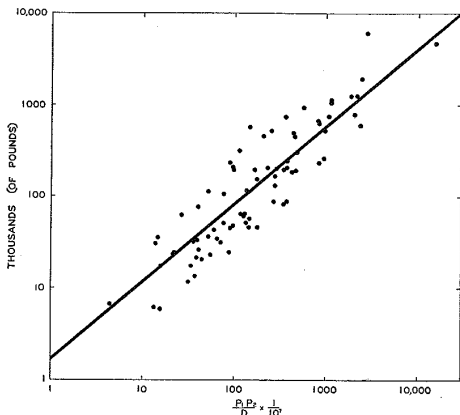


Fig. 9-14. Railway express. The movement by weight (less carload lots) between 13 arbitrary cities in the U. S. A., May 1939.





# Zipfian empirics:

- ▶ Length of trip versus frequency of trip.
- ▶ Solid line =  $-1/2$  exponent corresponds to  $\gamma = 2$ .

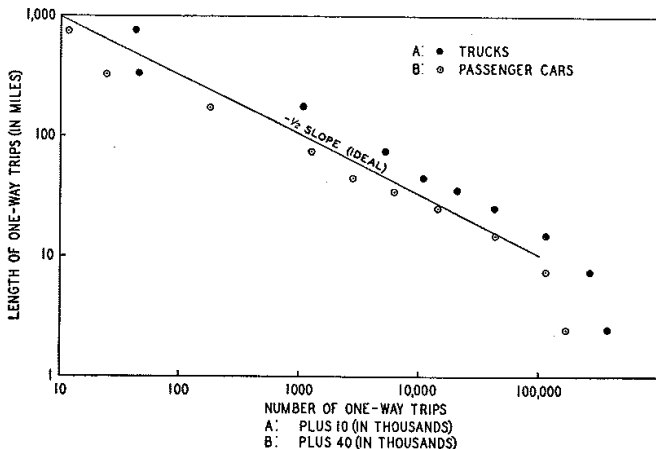


Fig. 9-19. Trucks and passenger cars: the number of one-way trips of like length.



# Zipfian empirics:

- ▶ The probability of marriage?
- ▶  $\gamma = 1$ ?

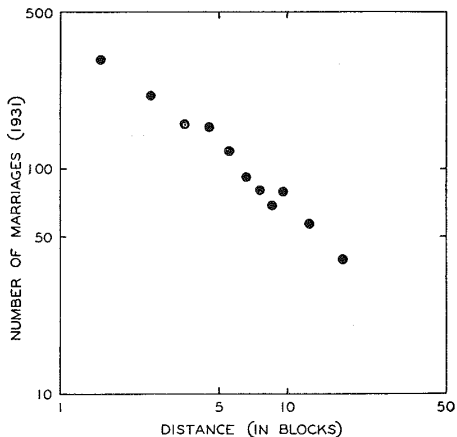


Fig. 9-22. Number of marriage licenses issued to 5,000 pairs of applicants living within Philadelphia in 1931 and separated by varying distances (the data of J. H. S. Bossard).



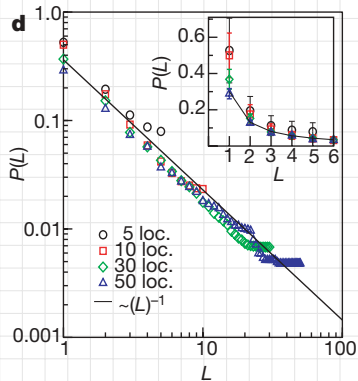
# Recent Zipf action:

Data from our man  
Zipf

Zipf in brief

Zipfian empirics

References



- ▶ Probability of people being in certain locations follows a Zipfish law...
- ▶ From González et al., Nature (2008)  
“Understanding individual human mobility patterns” [1]



# References I

- [1] M. C. González, C. A. Hidalgo, and A.-L. Barabási.  
Understanding individual human mobility patterns.  
[Nature](#), 453:779–782, 2008. [pdf](#) (⊞)
- [2] G. K. Zipf.  
Human Behaviour and the Principle of Least-Effort.  
Addison-Wesley, Cambridge, MA, 1949.

