Data from our man Zipf
Principles of Complex Systems
CSYS/MATH 300, Spring, 2013 | #SpringPoCS2013

Prof. Peter Dodds
@peterdodds
Department of Mathematics & Statistics | Center for Complex Systems | Vermont Advanced Computing Center | University of Vermont

Licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.

Outline

Zipf in brief

Zipfian empirics

Yet more Zipfian Empirics

References

George Kingsley Zipf:

In brief:

- Zipf (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's masterwork:

- "Human Behavior and the Principle of Least Effort" Addison-Wesley, 1949 Cambridge, MA [3]

- Bonus field of study: Glottometrics. (⊞)
- Bonus 'word' word: Glossolalia. (⊞)

Human Behavior/Principle of Least Effort:

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:

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. ⇒ 'Gravity Law.'
Zipfian empirics:

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

\[ \text{Table 2-1} \]

<table>
<thead>
<tr>
<th>Rank</th>
<th>Frequency</th>
<th>Product</th>
<th>Theoretical Length of Vocab.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>363</td>
<td>3630</td>
<td>51528</td>
</tr>
<tr>
<td>20</td>
<td>1722</td>
<td>34444</td>
<td>162328</td>
</tr>
<tr>
<td>30</td>
<td>755</td>
<td>22650</td>
<td>371402</td>
</tr>
<tr>
<td>40</td>
<td>377</td>
<td>14480</td>
<td>288500</td>
</tr>
<tr>
<td>50</td>
<td>154</td>
<td>7700</td>
<td>278370</td>
</tr>
<tr>
<td>100</td>
<td>86</td>
<td>8600</td>
<td>278300</td>
</tr>
<tr>
<td>200</td>
<td>42</td>
<td>4200</td>
<td>278200</td>
</tr>
<tr>
<td>300</td>
<td>25</td>
<td>2500</td>
<td>252400</td>
</tr>
<tr>
<td>400</td>
<td>16</td>
<td>1600</td>
<td>220200</td>
</tr>
<tr>
<td>500</td>
<td>10</td>
<td>1000</td>
<td>195300</td>
</tr>
</tbody>
</table>

\[ f \sim r^{-1} \text{ for word frequency:} \]

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... (more later\(^\text{[1]}\))

\[ (?) \text{ slope of } -3/5 \text{ corresponds to } \gamma = 5/3. \]

\[ \alpha = 1 \text{ corresponds to } \gamma = 1 + 1/\alpha = 2. \]
Zipfian empirics:

- Number of employees in organizations
  - \( \alpha = 2/3 \) 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.

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.

Zipfian empirics:

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

Zipfian empirics:

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

Zipfian empirics:

- The probability of marriage?
  - \( \gamma = 1 \)?
Comment #60 in Math and the City (III) by Strogatz, NYT:

George Kingsley Zipf was my teacher at Harvard...He had given a class project where we were to see if Chemical Company when ranked by the number of different chemicals they produced, followed his Law of Least Effort. I missed turning in my assignment due to the accidental death of my father...When I returned from the funeral I was given a message to call Dr. Zipf immediately. I did and when I explained why I was late turning in the data. He said, "Well, your father's gone and I (Zipf) have no pipeline to God, I expect the data will be on my desk tomorrow morning!!!" My mother, sister and extended family spread huge books of trade magazines on the kitchen and dining room tables and furiously went to work...We worked until late in the night and finished the project...I flew to Harvard the next morning and angrily gave the hundreds of "three by five cards" to Zipf. He said, "Thank you." Years later, I wondered whether his "meaness" had really been his way of helping me and my family to take our minds of our grief that day and concentrate on finishing my assignment. In my youth I thought not, but now as I approach 80, I like to think his seemingly burlful attitude was really an act of kindness.

—Jim Terry

Recent Zipf action:

- Probability of people being in certain locations follows a Zipfiah law...

Bonus: Marta’s talk (III) at UVM’s 2011 TEDx event “Big Data, Big Stories.” (II)

Zipfian empirics:

Table 2-3

<table>
<thead>
<tr>
<th>F</th>
<th>Calculated F₀</th>
<th>n₀</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12,124</td>
<td>4,678</td>
</tr>
<tr>
<td>2</td>
<td>35,212</td>
<td>6,496</td>
</tr>
<tr>
<td>3</td>
<td>11,103</td>
<td>4,486</td>
</tr>
<tr>
<td>4</td>
<td>23,073</td>
<td>4,578</td>
</tr>
<tr>
<td>5</td>
<td>18,300</td>
<td>4,486</td>
</tr>
<tr>
<td>6</td>
<td>15,940</td>
<td>4,578</td>
</tr>
<tr>
<td>7</td>
<td>14,248</td>
<td>4,578</td>
</tr>
<tr>
<td>8</td>
<td>13,103</td>
<td>4,486</td>
</tr>
<tr>
<td>9</td>
<td>12,344</td>
<td>4,578</td>
</tr>
<tr>
<td>10</td>
<td>11,795</td>
<td>4,486</td>
</tr>
<tr>
<td>11</td>
<td>11,303</td>
<td>4,486</td>
</tr>
<tr>
<td>12</td>
<td>10,904</td>
<td>4,486</td>
</tr>
<tr>
<td>13</td>
<td>10,515</td>
<td>4,486</td>
</tr>
<tr>
<td>14</td>
<td>10,146</td>
<td>4,486</td>
</tr>
<tr>
<td>15</td>
<td>9,787</td>
<td>4,486</td>
</tr>
<tr>
<td>20</td>
<td>8,708</td>
<td>4,486</td>
</tr>
<tr>
<td>30</td>
<td>7,529</td>
<td>4,486</td>
</tr>
<tr>
<td>50</td>
<td>6,350</td>
<td>4,486</td>
</tr>
</tbody>
</table>

References

Fig. 2-3. The number-frequency relationship of words. (A) Homer’s Iliad; (B) James Joyce’s Ulysses.

Fig. 2-4. The inward-frequency relationship. The number of different spots of like size (Δr, ng) between the expiration of worth amounting time in James’ Ulysses.

Zipfian empirics:
Zipfian empirics:

- Article length in the Encyclopedia Brittanica

Fig. 5-2. The number of different articles of like length in sampler of the 14th edition of the Encyclopedia Brittanica. Lengths in octos.

Zipfian empirics:

- Data from our man Zipf

Zipfian empirics:

- Table 6-1

The number of different genera of like number of different species of the flora of India (after J. C. Willis).

<table>
<thead>
<tr>
<th>No. of Genera</th>
<th>No. of Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>573</td>
<td>1</td>
</tr>
<tr>
<td>176</td>
<td>2</td>
</tr>
<tr>
<td>85</td>
<td>3</td>
</tr>
<tr>
<td>49</td>
<td>4</td>
</tr>
<tr>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>6</td>
</tr>
</tbody>
</table>

Fig. 9-9. Gross receipts, number of full-time employees, and total payroll of service establishments in the U.S.A. in 1959 ranked in the order of their decreasing number of members as in Fig. 9-4 supra.

Zipfian empirics:

- Data from our man Zipf

Zipfian empirics:

- Data from our man Zipf
Zipfian empirics:

- # species per genera:

![Graph showing the relationship between number of species and number of genera.](image)

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

Zipfian empirics:

![Graph showing the relationship between distance and number of families moving.](image)

References I

Least effort and the origins of scaling in human language.

Understanding individual human mobility patterns.

Human Behaviour and the Principle of Least-Effort.
Addison-Wesley, Cambridge, MA, 1949.