Day 3: Zipf and Zipf's law

George Kingsley Zipf
1902–1950 (CC 4.0)

We will discuss the famous Zipf's law. In connection with this, we also discuss the problems of the quantitative structure and distribution of vocabulary.

George Kingsley Zipf (1902.01.07–1950.09.25) was an American linguist and philosopher who studied the statistical characteristics of various languages. Zip represented the German department at Harvard University and became a University Lecturer. Zip also conducted research in Chinese and demographics. Zipf's law is said to be able to explain the general characteristics of distribution of web access frequency, income, etc.

1 Zipf curves

"TOM SAWYER ABROAD" By Mark Twain in 1896
CHAPTER I. TOM SEEKS NEW ADVENTURES

DO you reckon Tom Sawyer was satisfied after all them adventures? I mean the adventures we had down the river, and the time we set the darky Jim free and Tom got shot in the leg. No, he wasn't. It only just p'isoned him for more. That was all the effect it had. You see, when we three came back up the river in glory, as you may say, from that long travel, and the village received us with a torchlight procession and speeches, and everybody hurrah'd and shouted, it made us heroes, and that was what Tom Sawyer had always been hankering to be.

For a while he WAS satisfied. Everybody made much of him, and he tilted up his nose and stepped around the town as though he owned it. Some called him Tom Sawyer the Traveler, and that just swelled him up fit to bust. You see he laid over me and Jim considerable, because we only went down the river on a raft and came back by the steamboat, but Tom went by the steamboat both ways. The boys envied me and Jim a good deal, but land! they just knuckled to the dirt before TOM.

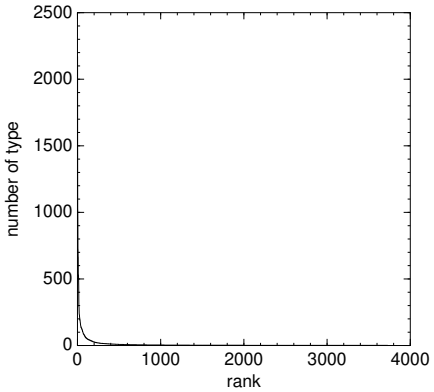

Figure 1: Word frequency and its ranking of "Tom Sawyer" by Mark Twain

Table 1: A list of the top 10 (left) and bottom 10 (left) in frequency of "Tom sawyer abroad."

Top 10	freq.	word	bottom 10	freq.	word
1	2172	and	3721	1	yahoos
2	1603	the	3722	1	yanked
3	1027	a	3723	1	yanking
4	1006	it	3724	1	yart
5	880	to	3725	1	yelpers
6	730	t	3726	1	yers
7	718	was	3727	1	yit
8	671	he	3728	1	yourn
9	660	of	3729	1	yuther
10	538	we	3730	1	zip

2 L shaped distribution

"Botchan" (Natsume Soseki, 1906) was divided into words, the frequency was calculated, and the most frequent ones were ranked and drawn. In any text, the frequency and rank of words is an L-shaped graph (Figure 2) (Mizutani 1975: 6).

> **"Botchan" Natsume Soseki 1906**
> 親譲りの無鉄砲で小供の時から損ばかりしている。小学校に居る時分学校の二階から飛び降りて一週間程腰を抜かした事がある。なぜそんな無闇をしたと聞く人があるかも知れぬ。別段深い理由でもない。新築の二階から首を出していたら、同級生の一人が冗談に、いくら威張っても、そこから飛び降りる事は出来まい。弱虫やーい。と囃したからである。小使に負ぶさって帰って来た時、おやじが大きな眼をして二階位から飛び降りて腰を抜かす奴があるかと云ったから、この次は抜かさずに飛んで見せますと答えた。
>
> He is reckless like his parents and has been losing money since he was a small child. When I was in elementary school, I once jumped from the second floor of the school and was knocked out for about a week. Some people might ask why I did it so recklessly. Not for any deeper reason. When I was sticking my head out of the second floor of a new building, one of my classmates fooled me and said that no matter how much I dared, I would never be able to jump from the window. The reason for this was because one of my classmates jokingly said, "You are a wimp!" When I came home after a servant gave me a piggyback ride, my father looked at me with big eyes and said, "Who would hurt one's back by jumping from the second floor? I replied, "Next time, I'll fly without hurting my back.

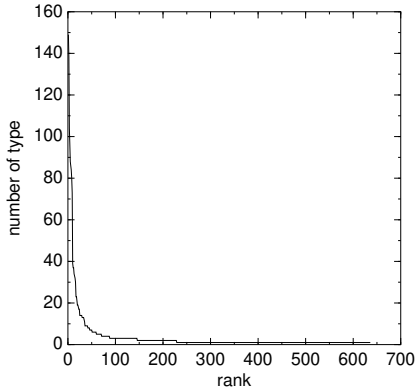

Figure 2: Word frequency and its ranking in the first 15 lines of "Bocchan"

List 1: Command list of word frequency in the first 15 lines of "Bocchan"

```
1 % head −15 bochan.txt| mecab | awk '{if␣($4)␣print␣$4}' | sort | uniq −c | sort −nr | nl |\
2       awk '{print␣$1,$2}' | graph −Tps −−x−label "rank" −−y−label "number␣of␣type" > t.eps
```

List 2: The top 10 and bottom 10 tokens extracted from the first 15 lines of "Bocchan"

```
1 % head −15 bochan.txt| mecab | awk '{if␣($4)␣print␣$4}␣' | sort | uniq −c | sort −nr | nl | head −10
2 % head −15 bochan.txt| mecab | awk '{if␣($4)␣print␣$4}␣' | sort | uniq −c | sort −nr | nl | tail −10
```

Q1: Examine Table 2 and discuss what words appear in the top and bottom ten places.

Q2: Discuss what changes would occur in the frequency of the upper and lower tenth positions if the amount of text increased.

Q3: Discuss what would happen to the shape of the plot and the letters in the upper and lower tens if the calculations were done with letters instead of words.

Table 2: A list of the top 10 (left) and bottom 10 (left) in frequency by morphological analysis of the first 15 lines of "Bocchan".

Top 10	freq.	word	bottom 10	freq.	word
1	149	て (te)	627	1	ざあざあ (zaazaa)
2	138	た (ta)	628	1	さす (sasu)
3	104	の (no)	629	1	ごまかす (gomakasu)
4	97	だ (da)	630	1	ぐらぐら (guragura)
5	88	は (wa)	631	1	ぐう (guu)
6	86	が (ga)	632	1	ぐいぐい (guigui)
7	84	に (ni)	633	1	ぎゅうぎゅう (gyuugyuu)
8	80	と (to)	634	1	がた (gata)
9	73	を (wo)	635	1	うん (un)
10	39	から (kara)	636	1	あんな (anna)

List 3: A method for calculating the frequency rankings of characters found in "Botchan."

```
% head -513 bochan.txt | texcount -freq - | tail -1892 | nl | head -1891 | awk '{print␣$1,$3}' | \
    graph -TX
```

3 Zipf's law

Read the following sentences and discuss the later questions.

Zipf's law

What is Zipf's law? Suppose the frequency of a word is tabulated and the word is ranked k. It is a rule of thumb that if the frequency of a word is ranked k, then the frequency of that word is approximately $1/k$ of the frequency of the first-ranked word. For example, if the frequency of the first-ranked word is 100, then The frequency of the second-ranked word is 50, and The frequency of the third word is 33, the frequency of the fourth word is The frequency of the fourth word is 25, and so on. and so on. It is known that there is a law between rank and frequency

Zipf's law is expressed by the following equation.

$$f_r \simeq \frac{c}{r} \quad (r = 1, 2, 3, ...n)$$

Where, the frequency, rank, and constant are f, r, c respectively.

The theoretical explanation has not yet been successful. Zipf's law is not limited to language, and it has been found that it applies to various relationships between frequency and rank. The probability distribution (discrete distribution) that follows this is called the Zipf distribution. At the early stage of the research, Zipf himself used the term "Principle of Least Effort" to explain it (Zipf 1935, 1949).

Q4: What does "rule of thumb" mean?

Q5: What statistical test can be used to find out that a distribution approximates a probability distribution of Zipf's law?

Q6: Explain what "discrete" means in discrete distribution using specific examples.

Q7: Discuss why the relationship between the rank and frequency of use of a word is called the "law of least effort."

Q8: It has been shown that Zipf's law applies to a wide variety of frequency-rank relationships, not just language. Discuss what else it applies to besides language.

Q9: Actually, $f_r \simeq \frac{c}{r}$ $(r = 1, 2, 3, ..n)$ does not apply to the previous "Bochan" data. Discuss what alternative ideas are possible.

4 When Zipf's law appears

It has been confirmed that Zipf's law be true to the following phenomena. (Wikipedia.org)

1. frequency of word usage
2. Frequency of access to web pages
3. Population of a city
4. income of the top 3% of people
5. frequency of musical notes
6. amount of gene expression in a cell
7. size of earthquakes
8. size of fragments when solids break

Q10: For each of the above, discuss why it follows Zipf's Law.

5 Related concepts

Zipf's law is a kind of power law. Zipf distribution has the same form as Pareto distribution (continuous distribution) by variable transformation. It is a discrete form of the Pareto distribution. Pareto principle is a special case of the Pareto distribution and is also related to the 80-20 law. It is also called "rank-size rule."

References

Mizutani, Sizuo (1975) "Mijikai sakuhin no goi no ryōteki kōzō: Shōwa shoki ryūkōka no chōsa kara 1 / On the statistical structure of vocabulary in short works: From the survey of Japanese popular songs from the 1930s, (1)", *Mathematical Linguistics*, Vol. 72, pp. 1–12.

Zipf, George Kingsley (1935) *The Psycho-Biology of Language*, Boston-Cambridge Mass.: Houghton Mifflin.

——— (1949) *Human Behavior and The Principle of Least Effort, An Introduction to Human Ecology*: Addison-Wesley Press Inc.