Correct ordering in the Zipf-Poisson ensemble

Justin S. Dyer and Art B. Owen

Author's Footnote:

Justin S. Dyer is a recent graduate of the Department of Statistics, Stanford University (email: jsdyer@stanfordalumni.org). Art B. Owen is Professor, Department of Statistics, Stanford University. Mailing address: Department of Statistics, Stanford University, Stanford, CA 94305. (email: owen@stanford.edu). This work was supported by NSF grant DMS-0906056 and by a National Science Foundation Graduate Research Fellowship.

Abstract

Rankings based on counts are often presented to identify popular items, such as baby names, English words or web sites. This article shows in some examples that the number of correctly identified items can be very small. We introduce a standard error versus rank plot to diagnose possible misrankings. Then to explain the slowly growing number of correct ranks, we model the entire set of count data via a Zipf–Poisson ensemble with independent $X_i \sim \mathrm{Poi}(Ni^{-\alpha})$ for $\alpha > 1$ and $N > 0$ and integers $i \geq 1$. We show that as $N \to \infty$, the first $n'(N)$ random variables have their proper order $X_1 > X_2 > \cdots > X_{n'}$ relative to each other, with probability tending to 1 for n' up to $(AN/\log(N))^{1/(\alpha+2)}$ for $A = \alpha^2(\alpha+2)/4$. We also show that the rate $N^{1/(\alpha+2)}$ cannot be achieved. The ordering of the first $n'(N)$ entities does not preclude $X_m > X_{n'}$ for some interloping $m > n'$. However, we show that the first n'' random variables are correctly ordered exclusive of any interlopers, with probability tending to 1 if $n'' \leq (BN/\log(N))^{1/(\alpha+2)}$ for any $B < A$.

1 Introduction

Rankings based on counted data are frequently quoted. We read of the most popular search queries in a given year, the most widely read or emailed articles from a web site, and the most frequently visited websites. Older examples include lists of word frequencies and popular baby names.

One issue that comes up with such data is how much credence to give to the observed rankings. If the process generating the rankings were to be repeated, some of the rankings might remain fixed while others would change. It may

Rank	Word	Count
1	the	6,187,267
2	be	4,239,632
3	of	3,093,444
4	and	2,687,863
5	a	2,186,369
6	in	1,924,315
7	to	1,620,850
8	have	1,375,636
9	it	1,090,186
10	to	1,039,323

Table 1: The top ten most frequent words from the British National Corpus, with their frequencies. Item 7 is the word 'to', used as an infinitive marker, while item 10 is 'to' used as a preposition. In "I went to the library to read." the first 'to' is a preposition and the second is an infinitive marker.

be clear that the highest ranked items are in their correct order and that the lower ranked ones are in a somewhat arbitrary order. But that leaves the issue of where to draw the line, as well as how to predict the amount of data needed in order to properly order the top n items.

The quantities we consider commonly show a power law decay. As a famous example, the model of Zipf (1949) has word frequency for the i'th most popular word falling off proportionally to $i^{-\alpha}$. Data usually show some deviations from a pure Zipf model: A few of the top items may be outliers, and there may also be some curvature in the log-log plot of item versus rank. The Zipf–Mandelbrot law for which the i'th frequency is proportional to $(i+k)^{-\alpha}$ where $k \geq 0$ is often a much better fit. That, and many other models are described in Popescu (2009, Chapter 9). For definiteness, we will focus on the Zipf model.

The usual model for sampling from long-tailed distributions assumes IID data (Johnson et al., 2005). However, in the applications we consider, IID sampling is unrealistic. For example, if one gathers a large sample of English text, the word 'the' will be the most frequent word with overwhelming probability. That one word is roughly 6% of written English (see Table 1) and no other word has such high probability. In IID sampling from a heavy tailed distribution, such very large counts would appear a random number of times. An IID sample of n items will have roughly Poi(1) of them larger than the $1-1/n$ quantile of the underlying distribution. Also, $\max_i X_i / \sum_i X_i$ will be very unstable under such sampling. In large samples of English text, by contrast, there will always be precisely one very popular word, it will always be 'the', and it will always be roughly 6% of the data.

Because IID sampling is unrealistic, we turn instead to modeling the entire

data set as just one realization of an ensemble. To this end, we propose a Zipf-Poisson ensemble, in which $X_i \sim \mathrm{Poi}(Ni^{-\alpha})$ independently for parameters $\alpha > 1$ and $N > 0$ and items indexed by integers $i \geq 1$. Compared to IID sampling, observation i in this model pertains to one specific real world entity, such as the word 'the' for $i = 1$ in the BNC. More general models replace the power law by other decreasing functions of i. But even this simple model sheds some light on the correctness of rankings. The number of correctly ranked items grows slowly in the limit as $N \to \infty$. We give precise upper and lower bounds below.

An outline of the paper is as follows. Section 2 shows some example data sets and gives a graphical summary which sheds light on how many order reversals there might be among the top few items. Section 3 presents our main result. For the Zipf-Poisson ensemble, the first $n = (AN/\log(N))^{1/(\alpha+2)}$ items will be in the correct order with probability tending to 1 as $N \to \infty$, if $A < \alpha^2(\alpha+2)/4$. There is a very sharp ordering threshold: if we remove the $\log(N)$, then the probability of correct ordering tends to 0 instead of 1. Section 3 also contains Lemma 1 which is of independent interest. It gives a Chernoff bound for the Skellam (1946) distribution: For $\lambda \geq \nu > 0$ we show that $\mathbb{P}(\mathrm{Poi}(\lambda) \leq \mathrm{Poi}(\nu)) \leq \exp(-(\sqrt{\lambda} - \sqrt{\nu})^2)$ where $\mathrm{Poi}(\lambda)$ and $\mathrm{Poi}(\nu)$ are independent Poisson random variables with the given means. Section 4 simulates a Zipf-Poisson ensemble and finds that the number of correctly ordered entities closely matches the asymptotic predictions. Section 5 presents our conclusions. When we have another decay rate, other than Zipf, then a numerical method described in Section 5 is very convenient to apply.

2 Examples and a graphical display

In this section we analyze a few data sets with items ranked by their counts. Our model for the data has $X_i \sim \mathrm{Poi}(N\lambda_i)$ independently for $i \geq 1$ where λ_i is some decreasing function of i, such as $\lambda_i = i^{-\alpha}$ for the Zipf-Poisson case that we focus on.

When we sort the observed counts, we get $X_{(1)} \geq X_{(2)} \geq \ldots$ where $X_{(i)}$ is the i'th (decreasing) order statistic of the data. The ranking question reduces to asking when $i = (i)$. A simple test statistic is

$$S_i = \frac{X_{(i)} - X_{(i+1)}}{\sqrt{X_{(i)} + X_{(i+1)}}}, \tag{2.1}$$

the estimated number of standard deviations under the Poisson assumption, between two consecutive counts. We find it useful to plot S_i versus i. When S_i is smaller than 2 or so, there is some doubt about the ordering at position i.

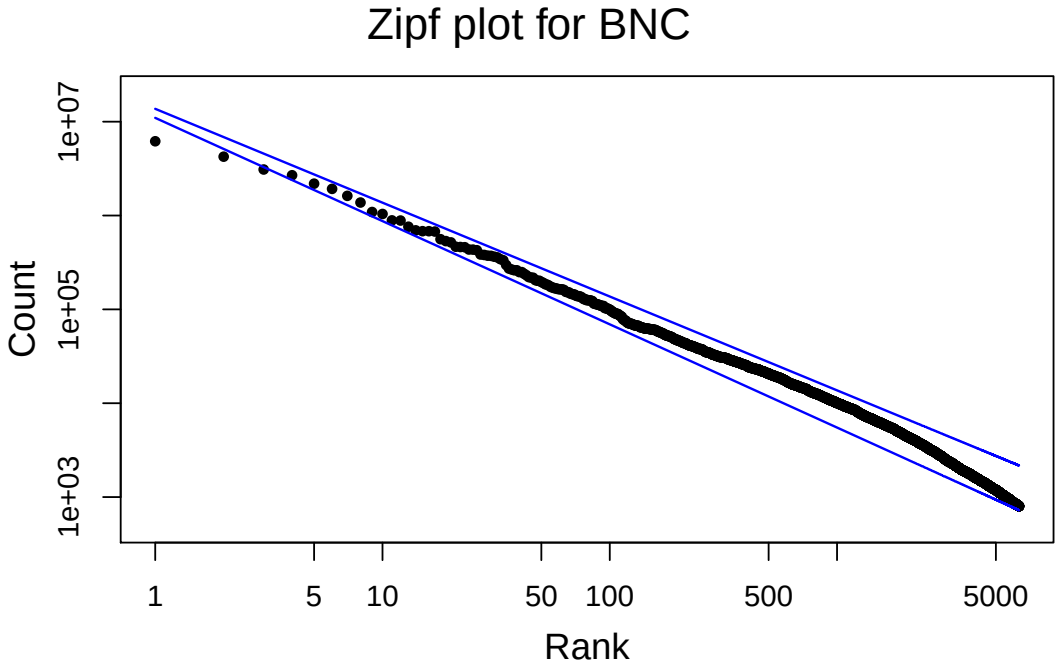

Figure 1: Zipf plot for the British National Corpus data. The reference line above the data has slope -1, while that below the data has slope -1.1.

2.1 British National Corpus

Our first example is based on approximately 100,000,000 words of the British National Corpus (BNC) (Aston and Burnard, 1998). Figure 1 plots the frequency of English words versus their rank on a log-log scale, for all words appearing at least 800 times in the BNC. The counts are from Kilgarrif (2006). These data have a nearly linear trend with a slope just steeper than -1. They are not perfectly Zipf-like, but the fit is extraordinarily good considering its parsimony: just one parameter α gives a good summary for nearly 100 million counted words. We see that the top three words are actually somewhat rarer than a Zipf model predicts, because the data curve slightly downward at the upper left of the plot. There is also clear evidence of a bend in the plot at the lower right, so a second parameter would improve the fit.

The most frequent word 'the' is much more frequent than the second most frequent word 'be'. The process generating this data clearly favors the word 'the' over 'be' and a p-value for whether these words might be equally frequent, using Poisson assumptions is overwhelmingly significant. Though the 9'th and 10'th words have counts that are within a few percent of each other, they too are significantly different, as judged by $(X_{(9)} - X_{(10)})/\sqrt{X_{(9)} + X_{(10)}} \doteq 34.9$, the number of estimated standard deviations separating them. The 500'th and 501'st most popular words are 'report' and 'pass' with counts of 20,660 and 20,633 respectively. These are not significantly different.

Figure 2 plots standard errors S_i from equation (2.1) for the first 100

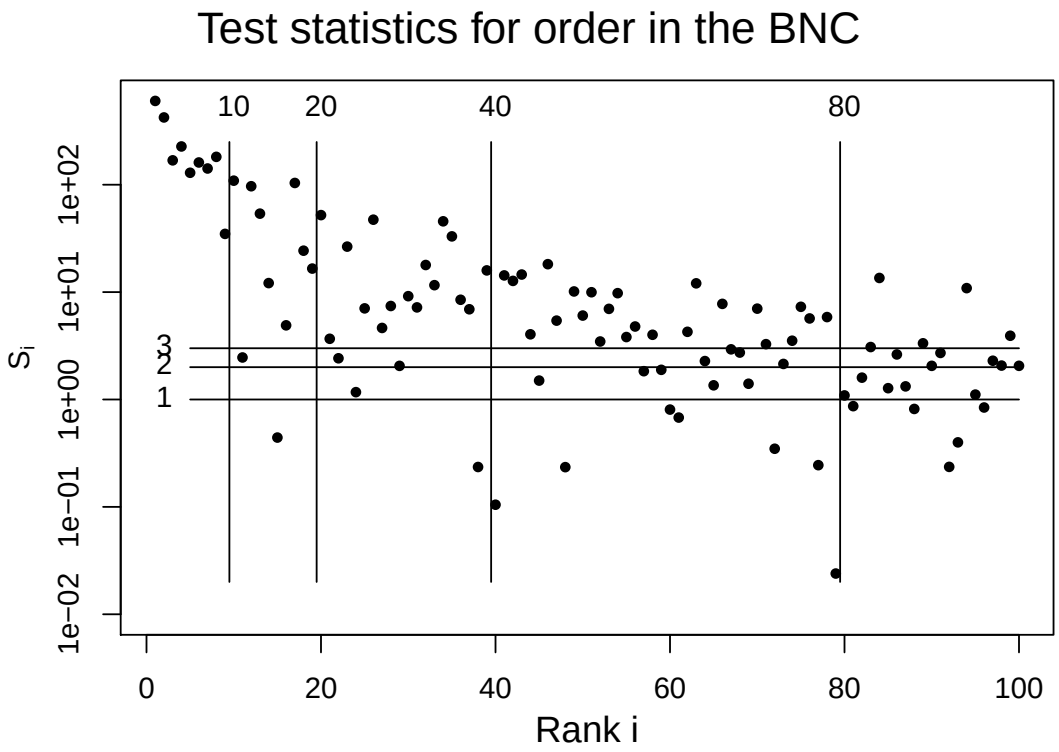

Figure 2: This figure plots $S_i = (X_{(i)} - X_{(i+1)})/\sqrt{X_{(i)} + X_{(i+1)}}$ versus $i = 1, \ldots, 100$ for the BNC data. The horizontal reference lines are drawn at 1, 2 and 3 standard errors. The vertical reference lines correspond to gaps between 10'th, 20'th, 40'th, 80'th observations and their respective predecessors.

2000		2005		2010	
Jacob	34,454	Jacob	25,794	Jacob	21,875
Michael	32,016	Michael	23,774	Ethan	17,866
Matthew	28,563	Joshua	23,208	Michael	17,133
Joshua	27,522	Matthew	21,440	Jayden	17,030
Christopher	24,916	Ethan	21,293	William	16,870

Table 2: Five most popular boys names for several years. Adjacent years show more overlap than years 5 apart do.

consecutive word comparisons. Looking at the reference lines it is clear that the first 10 words are ordered with high confidence. One of the rankings between 10 and 20 has a very small S_i corresponding to a possible transposition. There are a handful of potential misorderings between ranks 20 and 40 and they become more common afterwards.

2.2 U.S. baby names

The Social Security Administration has published baby name data for the U.S at `http://www.ssa.gov/oact/babynames/`. For every year from 1880 through 2010 inclusive, the data lists the number of babies born with each given name. There are separate data for boys and girls. When fewer than 5 babies have a particular name, those data are not revealed.

Some of the most popular names for 2000, 2005 and 2010 are listed in Tables 2 and 3. Many patterns are evident in this data. One can compare how popularity of names evolves over time at possibly different rates for boys and girls. For instance, name popularity is much more changeable now than it was in the 1930s.

For our purposes we are interested in the reliability of the rankings. From one point of view, we have a near census containing almost all of the names, so the observed ranks are correct for those years. There were exactly 3744 Brian's registered in 2010, placing Brian exactly 100'th. But modeling the underlying process is more useful. The process might change slowly from one year to the next, but the noise (which we model by Poisson sampling) in one year should be unrelated to the next or previous year's noise.

Figure 3 plots the birth data for 2010. From the Zipf plot we see quite clear curvature for the popular names followed by very straight lines for both boys and girls. From the standard error plot we see that the first two differences (hence the first three names) are very reliably determined for both boys and girls. After that, many of the girls' rankings are well separated while fewer of the boys rankings are.

2000		2005		2010	
Emily	25,949	Emily	23,907	Isabella	22,731
Hannah	23,066	Emma	20,318	Sophia	20,477
Madison	19,965	Madison	19,549	Emma	17,179
Ashley	17,991	Abigail	15,738	Olivia	16,860
Sarah	17,677	Olivia	15,685	Ava	15,300

Table 3: Five most popular girls names for several years. Adjacent years show more overlap than years 5 apart do.

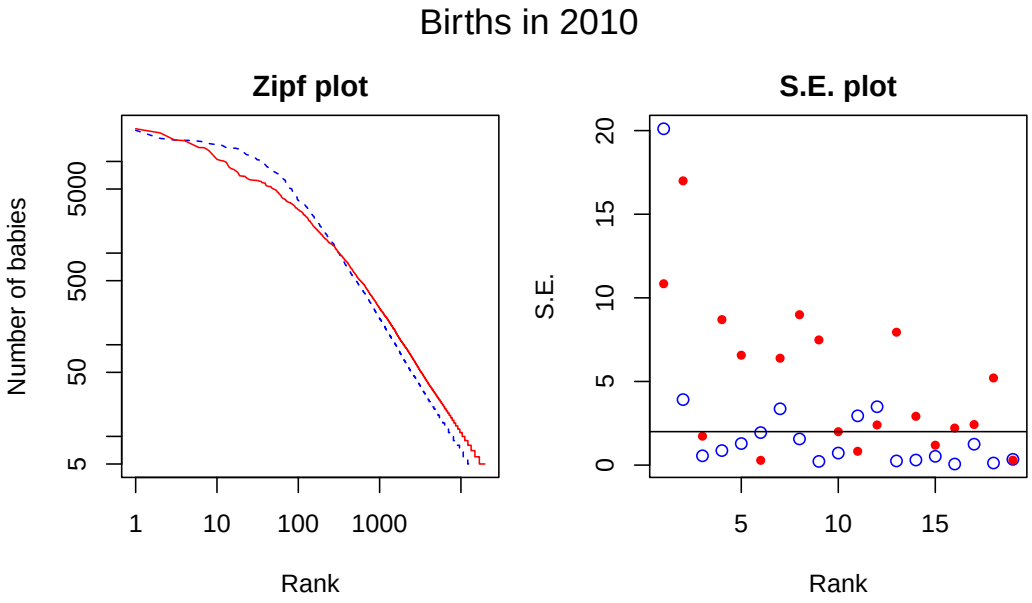

Figure 3: This figure shows most popular U.S. baby names for 2010. The left panel is a Zipf plot (boys dashed line, girls connected) and the right panel shows the S.E.s (boys as circles, girls as dots). The reference line in the right panel is at 2 standard errors.

Site	Unique visitors	Views
facebook.com	8.8×10^8	1.0×10^{12}
youtube.com	8.0×10^8	1.0×10^{11}
yahoo.com	5.9×10^8	7.7×10^{10}
live.com	4.9×10^8	8.4×10^{10}
msn.com	4.4×10^8	2.0×10^{10}
wikipedia.org	4.1×10^8	6.0×10^9
blogspot.com	3.4×10^8	4.9×10^9
baidu.com	3.0×10^8	1.1×10^{11}
microsoft.com	2.5×10^8	2.5×10^9
qq.com	2.5×10^8	3.9×10^{10}
bing.com	2.3×10^8	9.5×10^9
ask.com	1.9×10^8	2.0×10^9
adobe.com	1.6×10^8	1.0×10^9
taobao.com	1.6×10^8	1.1×10^{10}
twitter.com	1.6×10^8	5.9×10^9
youku.com	1.4×10^8	4.0×10^9
soso.com	1.4×10^8	3.6×10^9
wordpress.com	1.3×10^8	9.6×10^8
sohu.com	1.2×10^8	5.8×10^9
hao123.com	1.2×10^8	6.5×10^9

Table 4: Top 20 most popular websites according to Google's doubleclick ad planner http://www.google.com/adplanner/static/top1000/. The list "excludes adult sites, ad networks, domains that don't have publicly visible content or don't load properly, and certain Google sites." For instance, google.com does not appear in this list.

2.3 Web site popularity

Table 4 lists the top 20 web sites ranked by popularity, according to Google's doubleclick ad planner. It is taken from a list of the top 1000 web sites as measured by estimated unique users, with some websites excluded from the tally. There is also an estimated number of views. This data was gathered for July 2011. Each count is only reported to one or two significant figures.

Figure 4 shows the Zipf and S.E. plots for the number of total visitors. The number of visitors is rounded to only two significant figures. For example adobe.com, taobao.com and twitter.com are all given as having 1.6×10^8 unique visitors. The plotted standard errors are based on jittering the data to avoid trivial zeros. Assuming that the data were properly sorted before being rounded to 2 significant figures, the jittering proceeds by adding $\mathbf{U}(-1/2, 1/2) \times 10^6$ random variables to those three values and then sorting them to restore their order. The other values were similarly jittered according

Figure 4: Zipf and S.E. plot for number of unique visitors. The visitor counts have been jittered as described in the text. There is a horizontal reference line at 2 standard errors.

to their measured precision. The first 50 or so rankings are reliable under a Poisson noise model.

Figure 5 shows the Zipf and S.E. plots for the total number of page views These data are also jittered to mitigate their rounding. Most of the first 130 or so rankings are reliable under a Poisson noise model.

3 Some theory

In our examples, the likely number of correctly ordered items ranges from a handful for baby names, to a few dozen for words, to perhaps a hundred for web sites as measured by number of views. The total number of babies born is just under 2×10^6 each, for boys and girls. There were about 10^8 words in the BNC and there are about 2×10^{12} total web views for the 1000 sites listed. For data of this sort the number of reliably estimated rankings grows with sample size but the growth is slow.

In this section we give some theory to explain this slow growth. We use the Zipf-Poisson ensemble because we can get sharp rates for it. Data sets often have some curvature followed by a lengthy linear portion on the log-log plot of counts versus ranks. That linearity corresponds to the Zipf-Poisson model and so that model sheds light on data from more general mechanisms.

Our main theoretical result is the following:

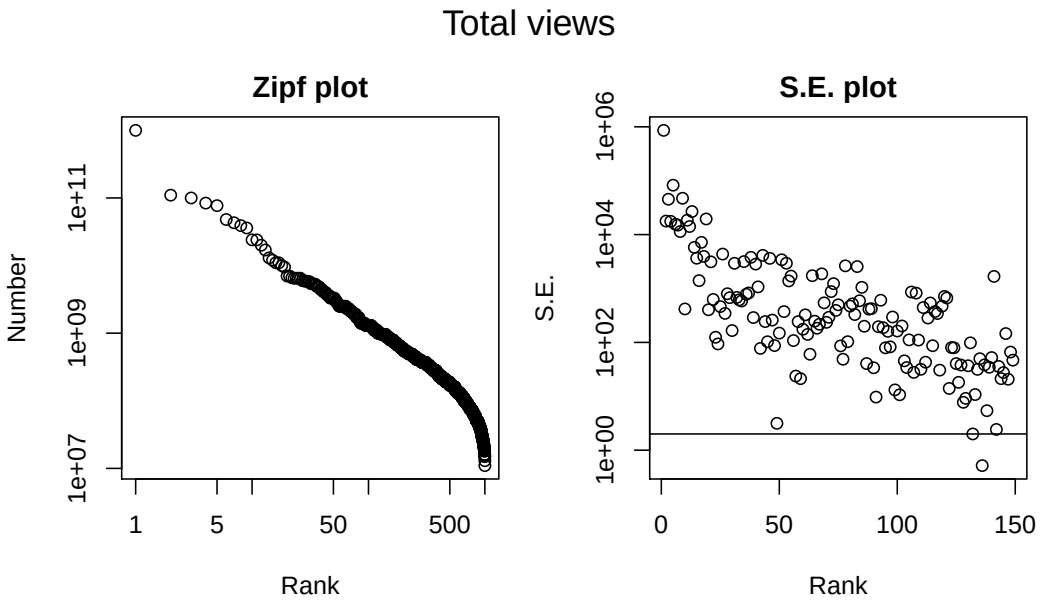

Figure 5: Zipf and S.E. plot for total views. The view counts have been jittered as described in the text. There is a horizontal reference line at 2 standard errors.

Theorem 1. *Let X_i be sampled from the Zipf–Poisson ensemble with parameter $\alpha > 1$. If $n = n(N) \le (AN/\log(N))^{1/(\alpha+2)}$ for $A = \alpha^2(\alpha+2)/4$, then*

$$\lim_{N\to\infty} \mathbb{P}\big(X_1 > X_2 > \cdots > X_n\big) = 1. \tag{3.1}$$

If $n = n(N) \le (BN/\log(N))^{1/(\alpha+2)}$ for $B < A$, then

$$\lim_{N\to\infty} \mathbb{P}\big(X_1 > X_2 > \cdots > X_n > \max_{i>n} X_i\big) = 1. \tag{3.2}$$

If $n = n(N) \ge CN^{1/(\alpha+2)}$ for any $C > 0$, then

$$\lim_{N\to\infty} \mathbb{P}\big(X_1 > X_2 > \cdots > X_n\big) = 0. \tag{3.3}$$

Equation (3.1) states that the top $n' = \lfloor (AN/\log(N))^{1/(\alpha+2)} \rfloor$ entities, with $A = \alpha^2(\alpha+2)/4$, are correctly ordered among themselves with probability tending to 1 as $N \to \infty$. From $\alpha > 1$ we have $A > 3/4$. Equation (3.3) shows that we cannot remove $\log(N)$ from the denominator, because the first $CN^{1/(\alpha+2)}$ entities will fail to have the correct joint ordering with a probability approaching 1 as $N \to \infty$.

Equation (3.1) leaves open the possibility that some entity beyond the n'th manages to get among the top n' entities due to sampling fluctuations. Those entities each have only a small chance to be bigger than $X_{n'}$, but there are infinitely many of them. Equation (3.2) shows that with probability tending to

1, the first $n'' = \lfloor (BN/\log(N))^{1/(\alpha+2)} \rfloor$ entities are the correct first n'' entities in the correct order. The limit holds for any $B < A$. That is, there is very little scope for interlopers.

3.1 Some useful inequalities

The proof of Theorem 1 makes use of some bounds on Poisson probabilities and the gamma function, collected here.

Let $Y \sim \mathrm{Poi}(\lambda)$. Shorack and Wellner (1986, page 485) have the following exponential bounds

$$\mathbb{P}(Y \geq t) \leq \Big(1 - \frac{\lambda}{t+1}\Big)^{-1} \frac{e^{-\lambda}\lambda^t}{t!} \quad \text{for integers } t \geq \lambda, \quad \text{and} \tag{3.4}$$

$$\mathbb{P}(Y \leq t) \leq \Big(1 - \frac{t}{\lambda}\Big)^{-1} \frac{e^{-\lambda}\lambda^t}{t!} \quad \text{for integers } t < \lambda. \tag{3.5}$$

Klar (2000) shows that (3.4) holds for $t \geq \lambda - 1$. Equation (3.4) holds for real valued $t \geq \lambda$ and equation (3.5) also holds for real valued $t < \lambda$. In both cases we interpret $t!$ as $\Gamma(t+1)$.

A classic result of Teicher (1955) is that

$$\mathbb{P}(Y \leq \lambda) \geq \exp(-1) \tag{3.6}$$

when $Y \sim \mathrm{Poi}(\lambda)$. If $Y \sim \mathrm{Poi}(\lambda)$, then

$$\sup_{-\infty<t<\infty} \left| \mathbb{P}\left(\frac{Y-\lambda}{\sqrt{\lambda}} \leq t \right) - \Phi(t) \right| \leq \frac{0.8}{\sqrt{\lambda}}, \tag{3.7}$$

where Φ is the standard normal CDF. Equation (3.7) follows by specializing a Berry-Esseen result for compound Poisson distributions (Michel, 1993, Theorem 1) to the case of a Poisson distribution.

We will also use Gautschi's (1959) inequality on the Gamma function,

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+1)^{1-s} \tag{3.8}$$

which holds for $x > 0$ and $0 < s < 1$.

3.2 Correct relative ordering, equation (3.1)

The difference of two independent Poisson random variables has a Skellam (1946) distribution. We begin with a Chernoff bound for the Skellam distribution.

Lemma 1. *Let $Z = X - Y$ where $X \sim \mathrm{Poi}(\lambda)$ and $Y \sim \mathrm{Poi}(\nu)$ are independent and $\lambda \geq \nu$. Then*

$$\mathbb{P}(Z \leq 0) \leq \exp\big(-(\sqrt{\lambda} - \sqrt{\nu})^2\big). \tag{3.9}$$

Proof. Let $\varphi(t) = \lambda e^{-t} + \nu e^{t}$. Then φ is a convex function attaining its minimum at $t^* = \log(\sqrt{\lambda/\nu}) \geq 0$, with $\varphi(t^*) = 2\sqrt{\lambda\nu}$. Using the Laplace transform of the Poisson distribution

$$m(t) \equiv \mathbb{E}(e^{-tZ}) = e^{\lambda(e^{-t}-1)} e^{\nu(e^{t}-1)} = e^{-(\lambda+\nu)} e^{\varphi(t)}.$$

For $t \geq 0$, Markov's inequality gives $\mathbb{P}(Z \leq 0) = \mathbb{P}(e^{-tZ} \geq 1) \leq \mathbb{E}(e^{-tZ})$. Taking $t = t^*$ yields (3.9). □

Lemma 2. *Let X_i be sampled from the Zipf–Poisson ensemble. Then for $n \geq 2$,*

$$\mathbb{P}(X_1 > X_2 > \cdots > X_n) \geq 1 - n \exp\Big(-\frac{N\alpha^2}{4}\, n^{-\alpha-2}\Big). \tag{3.10}$$

Proof. By Lemma 1 and the Bonferroni inequality, the probability that $X_{i+1} \geq X_i$ holds for any $i < n$ is no more than

$$\sum_{i=1}^{n-1} \exp\big(-\big(\sqrt{\lambda_i} - \sqrt{\lambda_{i+1}}\big)^2\big) = \sum_{i=1}^{n-1} \exp\big(-N\big(\sqrt{\theta_i} - \sqrt{\theta_{i+1}}\big)^2\big). \tag{3.11}$$

For $x \geq 1$, let $f(x) = x^{-\alpha/2}$. Then $|\sqrt{\theta_i} - \sqrt{\theta_{i+1}}| = |f(i) - f(i+1)| = |f'(z)|$ for some $z \in (i, i+1)$. Because $|f'|$ is decreasing, (3.11) is at most $n \exp(-N f'(n)^2)$, establishing (3.10). □

Now we can establish the first claim in Theorem 1.

Corollary 1. *Let X_i be sampled from the Zipf–Poisson ensemble. Choose $n = n(N) \geq 2$ so that $n \leq (AN/\log(N))^{1/(\alpha+2)}$ holds for all large enough N where $A = \alpha^2(\alpha+2)/4$. Then*

$$\lim_{N\to\infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n) = 1.$$

Proof. Let $n = (A_N N/\log(N))^{1/(\alpha+2)}$ for $A_N \leq A$. Then from equation (3.10) we need to show that $n \exp\big(-(N\alpha^2/4)n^{-\alpha-2}\big) \to 0$ as $N \to \infty$. The logarithm of this quantity is

$$\begin{aligned}
&\log(n) - \frac{N\alpha^2}{4}\, n^{-\alpha-2} \\
&= \frac{1}{\alpha+2}\Big(\log(A_N) + \log(N) - \log\log(N)\Big) - \frac{N\alpha^2}{4}\frac{\log(N)}{A_N N} \\
&= \log(N)\Big(\frac{1}{\alpha+2} - \frac{\alpha^2}{4A_N}\Big) - \frac{\log\log(N)}{\alpha+2} + \frac{A_N}{\alpha+2} \\
&\leq \log(N)\Big(\frac{1}{\alpha+2} - \frac{\alpha^2}{4A}\Big) - \frac{\log\log(N)}{\alpha+2} + \frac{A}{\alpha+2}.
\end{aligned}$$

This tends to $-\infty$ as needed if and only if $1/(\alpha+2) - \alpha^2/(4A) \leq 0$, which corresponds to $A \leq \alpha^2(\alpha+2)/4$. □

3.3 Correct absolute ordering, equation (3.2)

For the second claim in Theorem 1 we need to control the probability that one of the entities X_i from the tail given by $i > n$, can jump over one of the first n entities. Lemma 3 bounds the probability that an entity from the tail of the Zipf–Poisson ensemble can jump over a high level τ.

Lemma 3. *Let X_i for $i \geq 1$ be from the Zipf–Poisson ensemble with parameter $\alpha > 1$. If $\tau \geq \lambda_n$ then*

$$\mathbb{P}\Big(\max_{i>n} X_i > \tau\Big) \leq \frac{N^{1/\alpha}}{\alpha} \frac{\tau+1}{\tau+1-\lambda_n} \frac{\tau^{-1/\alpha}}{\tau-1/\alpha}. \tag{3.12}$$

Proof. First, $\mathbb{P}(\max_{i>n} X_i > \tau) \leq \sum_{i=n+1}^{\infty} \mathbb{P}(X_i > \tau)$ and then from (3.4)

$$\mathbb{P}\Big(\max_{i>n} X_i > \tau\Big) \leq \Big(1 - \frac{\lambda_n}{\tau+1}\Big)^{-1} \sum_{i=n+1}^{\infty} \frac{e^{-\lambda_i}\lambda_i^{\tau}}{\Gamma(\tau+1)}.$$

Now $\lambda_i = Ni^{-\alpha}$. For $i > n$ we have $\tau > \lambda_i = Ni^{-\alpha}$. Over this range, $e^{-\lambda}\lambda^{\tau}$ is an increasing function of λ. Therefore,

$$\begin{aligned}
\sum_{i=n+1}^{\infty} e^{-\lambda_i}\lambda_i^{\tau} &\leq \int_n^{\infty} e^{-Nx^{-\alpha}} (Nx^{-\alpha})^{\tau} \,\mathrm{d}x \\
&\leq \frac{N^{1/\alpha}}{\alpha} \int_0^{Nn^{-\alpha}} e^{-y} y^{\tau-1/\alpha-1} \,\mathrm{d}y \\
&\leq \frac{N^{1/\alpha}}{\alpha} \Gamma(\tau - 1/\alpha).
\end{aligned}$$

As a result

$$\mathbb{P}\Big(\max_{i>n} X_i > \tau\Big) \leq \frac{N^{1/\alpha}}{\alpha} \frac{\tau+1}{\tau+1-\lambda_n} \frac{\Gamma(\tau-1/\alpha)}{\Gamma(\tau+1)}.$$

Now

$$\frac{\Gamma(\tau-1/\alpha)}{\Gamma(\tau+1)} = \frac{\Gamma(\tau+1-1/\alpha)}{\Gamma(\tau+1)} \frac{1}{\tau-1/\alpha} < \frac{\tau^{-1/\alpha}}{\tau-1/\alpha}$$

by Gautschi's inequality (3.8), with $s = 1 - 1/\alpha$, establishing (3.12). □

For an incorrect ordering to arise, either an entity from the tail exceeds a high level, or an entity from among the first n is unusually low. Lemma 4 uses a threshold for which both such events are unlikely, establishing the second claim (3.2) of Theorem 1.

Lemma 4. *Let X_i for $i \geq 1$ be from the Zipf–Poisson ensemble with parameter $\alpha > 1$. Let $n(N)$ satisfy $n \geq (AN/\log(N))^{1/(\alpha+2)}$ for $0 < A < A(\alpha) = \alpha^2(\alpha+2)/4$. Let $m \leq (BN/\log(N))^{1/(\alpha+2)}$ for $0 < B < A$. Then*

$$\lim_{N\to\infty} \mathbb{P}\Big(\max_{i>n} X_i \geq X_m\Big) = 0. \tag{3.13}$$

Proof. For any threshold τ,

$$\mathbb{P}\Big(\max_{i>n} X_i \geq X_m\Big) \leq \mathbb{P}\Big(\max_{i>n} X_i > \tau\Big) + \mathbb{P}(X_m \leq \tau). \tag{3.14}$$

The threshold we choose is $\tau = \sqrt{\lambda_m \lambda_n}$ where $\lambda_i = \mathbb{E}(X_i) = Ni^{-\alpha}$.

Write $n = (A_N N/\log(N))^{1/(\alpha+2)}$ and $m = (B_N N/\log(N))^{1/(\alpha+2)}$ for $0 < B_N < B < A_N < A < A(\alpha)$. Then $\tau = \sqrt{\lambda_m \lambda_n} = N(C_N N/\log(N))^{-\alpha/(\alpha+2)}$ where $C_N = \sqrt{A_N B_N}$. Therefore

$$\tau = O\big(N^{2/(\alpha+2)}(\log(N))^{\alpha/(\alpha+2)}\big).$$

By construction, $\tau > \lambda_n$ and so by Lemma 3

$$\mathbb{P}\Big(\max_{i>n} X_i > \tau\Big) \leq \frac{N^{1/\alpha}}{\alpha} \frac{\tau+1}{\tau+1-\lambda_n} \frac{\tau^{-1/\alpha}}{\tau - 1/\alpha}.$$

Because $\lambda_n/\tau = (B_N/A_N)^{\alpha/(2\alpha+4)}$, we have $(\tau+1)/(\tau+1-\lambda_n) = O(1)$. Therefore

$$\mathbb{P}\Big(\max_{i>n} X_i > \tau\Big) = O(N^{1/\alpha}\tau^{-1/\alpha-1}) = O\big(N^{-1/(\alpha+2)}(\log(N))^{(\alpha+1)/(\alpha+2)}\big)$$

and so the first term in (3.14) tends to 0 as $N \to \infty$.

For the second term in (3.14), notice that X_m has mean $\lambda_m > \tau$ and standard deviation $\sqrt{\lambda_m}$. Letting $\rho = \alpha/(\alpha+2)$ and applying Chebychev's inequality, we find that

$$\begin{aligned}
\mathbb{P}(X_m \leq \tau) &\leq \frac{\lambda_m}{(\tau - \lambda_m)^2} \\
&= \frac{B_N^\rho}{(B_N^\rho - C_N^\rho)^2} N^{-2/(\alpha+2)} (\log(N))^{-\rho} \\
&\leq \frac{1}{(A^{\rho/2} - B^{\rho/2})^2} N^{-2/(\alpha+2)} (\log(N))^{-\rho} \\
&\to 0
\end{aligned}$$

as $N \to \infty$. □

3.4 Limit to correct ordering, equation (3.3)

While we can get $(AN/\log(N))^{1/(\alpha+2)}$ entities properly ordered, there is a limit to the number of correctly ordered entities. We cannot get above $CN^{1/(\alpha+2)}$ correctly ordered entities, asymptotically. That is, the logarithmic factor cannot be removed. We begin with a lower bound on the probability of a wrong ordering for two consecutive entities.

Lemma 5. *Let X_i be from the Zipf–Poisson ensemble with $\alpha > 1$. Suppose that $AN^{1/(\alpha+2)} \le i < i+1 \le BN^{1/(\alpha+2)}$ where $0 < A < B < \infty$. Then for large enough N,*

$$\mathbb{P}(X_{i+1} \ge X_i) \ge \frac{1}{3}\,\Phi\biggl(-\alpha\,\frac{A^{\alpha/2}}{B^{\alpha+1}}\biggr).$$

Proof. First $\mathbb{P}(X_{i+1} \ge X_i) \ge \mathbb{P}(X_{i+1} > \lambda_i)\mathbb{P}(X_i \le \lambda_i) \ge \mathbb{P}(X_{i+1} > \lambda_i)/e$ using Teicher's inequality (3.6). Next

$$\mathbb{P}(X_{i+1} > \lambda_i) = 1 - \mathbb{P}(X_{i+1} \le \lambda_i) \ge \Phi\biggl(\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}}\biggr) - \frac{0.8}{\sqrt{\lambda_{i+1}}}.$$

Now,

$$\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}} = \sqrt{N}\,\frac{(i+1)^{-\alpha} - i^{-\alpha}}{\sqrt{(i+1)^{-\alpha}}} = -\alpha\sqrt{N}\,\frac{(i+\eta)^{-\alpha-1}}{\sqrt{(i+1)^{-\alpha}}}$$

for some $\eta \in (0,1)$. Applying the bounds on i,

$$\frac{\lambda_{i+1} - \lambda_i}{\sqrt{\lambda_{i+1}}} \ge -\alpha\sqrt{N}\frac{(N^{1/(\alpha+2)}A)^{\alpha/2}}{(N^{1/(\alpha+2)}B)^{\alpha+1}} = -\alpha\frac{A^{\alpha/2}}{B^{\alpha+1}}.$$

Finally, letting $N \to \infty$ we have $\lambda_{i+1} \to \infty$ and so $0.8/\sqrt{\lambda_{i+1}}$ is eventually smaller than $(1-e/3)\Phi(-\alpha A^{\alpha/2}B^{-\alpha-1})$. Letting $\theta = -\alpha A^{\alpha/2}B^{-\alpha-1}$ we have, for large enough N,

$$\mathbb{P}(X_{i+1} \ge X_i) \ge \Bigl(\Phi(\theta) - \Bigl(1 - \frac{e}{3}\Bigr)\Phi(\theta)\Bigr)\frac{1}{e} = \frac{1}{3}\Phi(\theta). \quad \square$$

To complete the proof of Theorem 1 we establish equation (3.3). For n beyond a multiple of $N^{1/(\alpha+2)}$, the reverse orderings predicted by Lemma 5 cannot be avoided.

Corollary 2. *Let X_i be sampled from the Zipf–Poisson ensemble. Suppose that $n = n(N)$ satisfies $n \ge CN^{1/(\alpha+2)}$ for $0 < C < \infty$. Then*

$$\lim_{N\to\infty} \mathbb{P}(X_1 > X_2 > \cdots > X_n) = 0.$$

Proof. Let $p \in (0,1)$ be a constant such that $\mathbb{P}(X_{i+1} \ge X_i) \ge p$ holds for all large enough N and $(C/2)N^{1/(2+\alpha)} \le i < i+1 \le CN^{1/(2+\alpha)}$. For instance Lemma 5 shows that $p = \Phi(-\alpha(C/2)^{\alpha/2}/C^\alpha)/3 = \Phi(-\alpha(2C)^{-\alpha/2})/3$ is such a constant. Then

$$\mathbb{P}(X_1 > X_2 > \cdots > X_n) \le \prod_i{}^{*}\, \mathbb{P}(X_i > X_{i+1}) \tag{3.15}$$

holds where $\prod^*$ is over all odd integers $i \in [(C/2)N^{1/(\alpha+2)}, CN^{1/(\alpha+2)})$. There are roughly $CN^{1/(\alpha+2)}/4$ odd integers in the product. For large enough N, the right side of (3.15) is below $(1-p)^{CN^{1/(\alpha+2)}/5} \to 0$. $\square$

3.5 Summary

Theorem 1 makes three claims which we have proved as follows. First, equation (3.1) on correct ordering of the n most popular items within themselves, follows from Corollary 1. Combining that corollary with Lemma 4 to rule out interlopers, establishes the second claim, given by (3.2), in which the first n items are correctly identified and ordered. The third claim (3.3), showing the necessity of the logarithmic factor, follows from Corollary 2.

4 Monte Carlo

In this section we sample the Zipf-Poisson ensemble by Monte Carlo and compare the observed rankings to the ones predicted by the model. The parameters we use are modeled loosely on the BNC data.

We will use a value of α close to 1.1 because that value is in the range considered most reasonable for English text and it is near the slope seen in Figure 1. We also need a value for N. Let $T = \sum_{i=1}^{\infty} X_i$ be the total count. Then $\mathbb{E}(T) = \sum_{i=1}^{\infty} Ni^{-\alpha} = N\zeta(\alpha)$ where $\zeta(\cdot)$ is the Riemann zeta function. If we choose $\alpha = \alpha_* \doteq 1.106$ then we find that $\zeta(\alpha_*) = 10$ and so, for convenience of illustration, we choose $N = 10^8/10 = 10^7$ for our simulation.

Theorem 1 has the top $n' = (A(\alpha)N/(\log(N)))^{1/(\alpha+2)}$ entities correctly ordered among themselves with probability tending to 1. For our BNC-inspired model we get $n' = (A(\alpha_*)N_*/(\log(N_*)))^{1/(\alpha_*+2)} \doteq 72.08$.

Some simulated results are shown in Figure 6. The number of correctly ordered items ranged from 69 to 153 in those 1000 simulations. The number was only smaller than 72 for 2 of the simulated cases.

In our simulation, the first rank error to occur was usually a transposition between the n'th and $n+1$'st entity. This happened 982 times. There were 7 cases with a tie between the n'th and $n+1$'st entity. The remaining 11 cases all involved the $n+2$'nd entity getting ahead of the n'th. As a result, we see that interlopers are very rare, as quantified by Lemma 4 in Section 3.

For the BNC data we expect lots of transpositions to set in after rank 72 simply from the Zipf-Poisson sampling. We also saw that a few misorderings were likely before 72 due to some lack of fit of that model.

5 Discussion

We have seen graphically and statistically that rankings based on count data get the entire first n entities in proper order with high probability, for a value n that grows slowly with the total count N.

Our Poisson model for counts does not apply to Zipf plots for quantities like river lengths or sizes of ore bodies where the data are not counts.

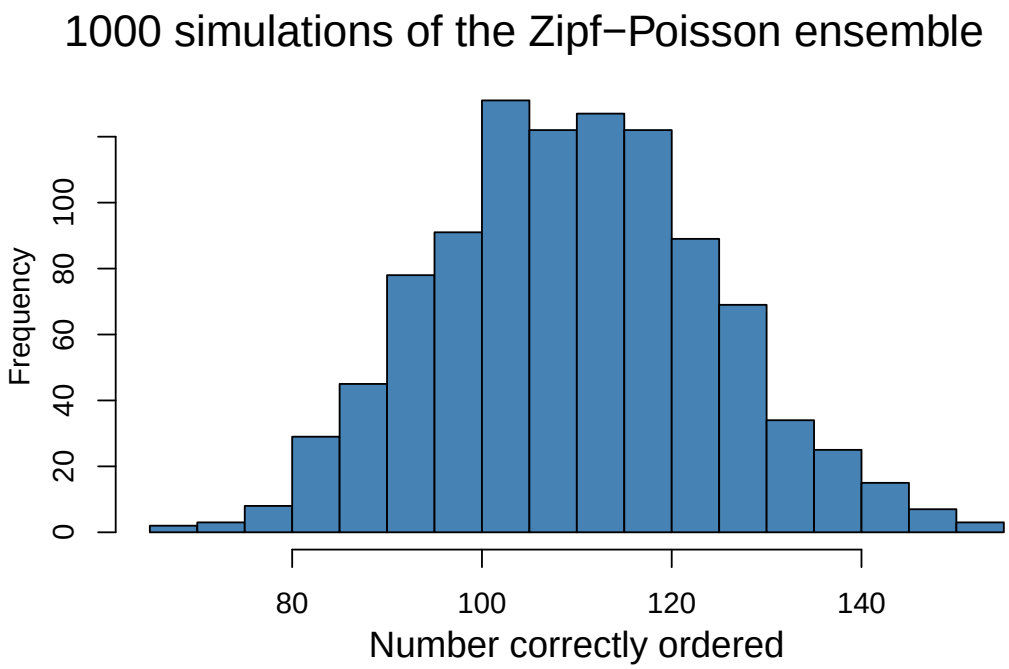

Figure 6: The Zipf–Poisson ensemble with $N = 10^7$ and $\alpha = 1.106$ was simulated 1000 times. The histogram shows the distribution of the number of correctly ordered words.

Our transition point is at $n' = (\alpha^2(\alpha+2)N/(4\log(N))^{1/(\alpha+2)}$ and estimating N from $T = \sum_i X_i$ leads to the estimate

$$\hat{n} = \left(\frac{\alpha^2(\alpha+2)T/\zeta(\alpha)}{4\log(T/\zeta(\alpha))} \right)^{\frac{1}{\alpha+2}}.$$

The threshold n' uses some slightly conservative estimates to get a rate in N. For the Zipf–Poisson ensemble with $N = 10^7$ and $\alpha = 1.106$ we can use (3.10) of Lemma 2 directly to find

$$1 - \mathbb{P}(X_1 > X_2 > \cdots > X_{72}) \le \sum_{i=1}^{71} \exp(-N(i^{-\alpha/2} - (i+1)^{-\alpha/2})^2) \doteq 0.0199.$$

We get a bound of 1% by taking $n = 70$ and a bound of 5% by taking $n = 76$. The formula for $\hat{n}$ comes remarkably close to what we get working directly with equation (3.10).

The Skellam bounds do not assume a Zipf rate for the Poisson means. Therefore we can use them to generalize the computation above. For example, with a Zipf–Mandelbrot–Poisson ensemble having $X_i \sim \mathrm{Poi}(N(i+k)^{-\alpha})$ we can still apply equation (3.10) to show that the probability of an error among

the first n ranks is at most

$$p(n; N, \alpha, k) = \sum_{i=1}^{n-1} \exp\Big(-N\Big((i+k)^{-\alpha/2} - (i+k+1)^{-\alpha/2}\Big)^2\Big). \tag{5.1}$$

A conservative estimate of the number of correct positions in the Zipf–Mandelbrot–Poisson ensemble is

$$n' = \max\{n \geq 1 \mid p(n; N, \alpha, k) \leq 0.01\} \tag{5.2}$$

with $n' = 0$ if $p(1; N, \alpha, k) > 0.01$. We can estimate N by $T/\zeta(\alpha, k-1)$ where $T = \sum_i X_i$ and $\zeta(\alpha, h) = \sum_{\ell=0}^{\infty} (\ell + h)^{\alpha}$ is the Hurwitz zeta function.

Equation (5.2) is conservative because it stems from the Bonferroni inequality, and does not adjust for two or more order relations being violated. It will be less conservative for small target probabilities like 0.01 than for large ones where adjustments are relatively more important.

A small number of correct unique words can correspond to a reasonably large fraction of word usage. The BNC is roughly 6.2% 'the' and the top 72 words comprise about 45.3% of the corpus.

For large N, the top $n_\epsilon = N^{1/(\alpha+2)-\epsilon}$ entities get properly ordered with very high probability for $0 < \epsilon < 1/(\alpha+2)$. The tail beyond n_ϵ accounts for a proportion of data close to $\zeta(\alpha)^{-1} \int_{n_\epsilon}^{\infty} x^{-\alpha}\, \mathrm{d}x = O(n_\epsilon^{-\alpha+1}) = O(N^{(1-\alpha)/(\alpha+2)+\epsilon'})$ for $\epsilon' = \epsilon(\alpha - 1)$. Taking small ϵ and recalling that $\alpha > 1$ we find that the fraction of data from improperly ordered entities vanishes in the Zipf–Poisson ensemble. When α is just barely larger than 1 the rate may be slow.

We frequently see an apparent slope in the data that corresponds to $\alpha < 1$. For example this happens with the web data for unique visitors (but not total views). The Zipf rule requires $\alpha > 1$ to be summable. Websites beyond the 1000'th probably show a steepening of the slope. For the baby name data, which includes even very small counts, we see a turning point to steeper and nearly linear slopes.

References

G. Aston and L. Burnard. *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press, Edinburgh, 1998.

W. Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *J. Math. Phys.*, 38:77–81, 1959.

N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate discrete distributions*. John Wiley & Sons, Hoboken, NJ, third edition, 2005.

A. Kilgarrif. BNC word frequency list. http://www.kilgarriff.co.uk/bnc-readme.html, 2006.

B. Klar. Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences*, 14:161–171, 2000.

R. Michel. On Berry–Esseen results for the compound Poisson distribution. *Insurance: Mathematics and Economics*, 13:35–37, 1993.

I.-I. Popescu. *Word frequency studies*. Mouton de Gruyter, Berlin, 2009.

G. R. Shorack and J. A. Wellner. *Empirical Processes With Applications to Statistics*. Wiley, New York, 1986.

J. G. Skellam. The frequency distribution of the difference between two Poisson variates belonging to different populations. *Journal of the Royal Statistical Society: Series A*, 109(3):296, 1946.

H. Teicher. An inequality on Poisson probabilities. *The Annals of Mathematical Statistics*, 26:147–149, 1955.

G. K. Zipf. *Human behavior and the principle of least effort*. Addison-Wesley Press, New York, 1949.