RESEARCH ARTICLE

The common patterns of abundance: the log series and Zipf's law [version 1; peer review: 3 approved]

Steven A. Frank

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, 92697-2525, USA

v1 **First published:** 25 Mar 2019, **8**:334 (https://doi.org/10.12688/f1000research.18681.1)
Latest published: 25 Mar 2019, **8**:334 (https://doi.org/10.12688/f1000research.18681.1)

Abstract

In a language corpus, the probability that a word occurs n times is often proportional to $1/n^2$. Assigning rank, s, to words according to their abundance, log s vs log n typically has a slope of minus one. That simple Zipf's law pattern also arises in the population sizes of cities, the sizes of corporations, and other patterns of abundance. By contrast, for the abundances of different biological species, the probability of a population of size n is typically proportional to $1/n$, declining exponentially for larger n, the log series pattern.

This article shows that the differing patterns of Zipf's law and the log series arise as the opposing endpoints of a more general theory. The general theory follows from the generic form of all probability patterns as a consequence of conserved average values and the associated invariances of scale.

To understand the common patterns of abundance, the generic form of probability distributions plus the conserved average abundance is sufficient. The general theory includes cases that are between the Zipf and log series endpoints, providing a broad framework for analyzing widely observed abundance patterns.

Keywords

scaling patterns, ecology, demography, linguistics, probability theory

Open Peer Review

Referee Status: ✓ ✓ ✓

	Invited Referees 1	2	3
version 1 published 25 Mar 2019	✓ report	✓ report	✓ report

1 **Luís M. A. Bettencourt**, University of Chicago, USA
Santa Fe Institute, USA

2 **Jose Lobo**, Arizona State University, USA

3 **Scott E. Page**, University of Michigan, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Steven A. Frank (safrank@uci.edu)

Author roles: Frank SA: Conceptualization, Formal Analysis, Funding Acquisition, Investigation, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The Donald Bren Foundation supports my research.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Frank SA. **The common patterns of abundance: the log series and Zipf's law [version 1; peer review: 3 approved]** F1000Research 2019, **8**:334 (https://doi.org/10.12688/f1000research.18681.1)

First published: 25 Mar 2019, **8**:334 (https://doi.org/10.12688/f1000research.18681.1)

Introduction

A few simple patterns recur in nature. Adding up random processes often leads to the bell-shaped normal distribution. Death and other failures typically follow the extreme value distributions.

Those simple patterns recur under widely varying conditions. Something fundamental must set the relations between pattern and underlying process. To understand the common patterns of nature, we must know what fundamentally constrains the forms that we see.

Without that general understanding, we will often reach for unnecessarily detailed and complex models of process to explain what is in fact some structural property that influences the invariant form of observed pattern.

We already understand that the central limit theorem explains the widely observed normal distribution[1]. Similar limit theorems explain why failure often follows the extreme value pattern[2,3].

The puzzles set by other commonly observed patterns remain unsolved. Each of those puzzles poses a challenge. The solutions will likely broaden our general understanding of what causes pattern. Such insight will help greatly in the big data analyses that play an increasingly important role in modern science.

Zipf's law is one of the great unsolved puzzles of invariant pattern. The frequency of word usage[4], the sizes of cities[5,6], and the sizes of corporations[7] have the same shape. On a log-log plot of rank versus abundance, the slope is minus one. For cities, the largest city would have a rank of one, the second largest city a rank of two, and so on. Abundance is population size.

The abundance of species is another great unsolved puzzle of invariant pattern. In an ecological community, the probability that a species has a population size of n individuals is proportional to p^n/n, the log series pattern[8]. Communities differ only in their average population size, described by the parameter, p. Actual data vary, but most often fit closely to the log series[9].

In this article, I show that Zipf's law and the log series arise as the opposing endpoints of a more general theory. That theory provides insight into the particular puzzles of Zipf's law and species abundances. The analysis also suggests deeper insights that will help to unify understanding of commonly observed patterns.

Theory

The argument begins with the invariances that define alternative probability patterns[10,11]. To analyze the invariances of a probability distribution, note that we can write almost any probability distribution, q_z, as

$$q_z = ke^{-\lambda T_z}, \tag{1}$$

in which $T(z) \equiv T_z$ is a function of the variable z. The probability pattern, q_z, is invariant to a constant shift, $T_z \mapsto a + T_z$, because we can write the transformed probability pattern in Equation 1 as

$$q_z = k_a e^{-\lambda(a+T_z)} = ke^{-\lambda T_z},$$

with $k = k_a e^{-\lambda a}$. We express k in this way because k adjusts to satisfy the constraint that the total probability be one. In other words, conserved total probability implies that the probability pattern is shift invariant with respect to T_z.

Now consider the consequences if the average of some value over the distribution q_z is conserved. That constraint causes the probability pattern to be invariant to a multiplicative stretching (or shrinking), $T_z \mapsto bT_z$, because

$$q_z = ke^{-\lambda_b bT_z} = ke^{-\lambda T_z},$$

with $\lambda = \lambda_b b$. We specify λ in this way because λ adjusts to satisfy the constraint of conserved average value. Thus, invariant average value implies that the probability pattern is stretch invariant with respect to T_z.

Conserved total probability and conserved average value cause the probability pattern to be invariant to an affine transformation of the T_z scale, $T_z \mapsto a + bT_z$, in which "affine" means both shift and stretch.

The affine invariance of probability patterns with respect to T_z induces significant structure on the form of T_z and the associated form of probability patterns. Understanding that structure provides insight into probability patterns and the processes that generate them[10,12,13].

In particular, Frank & Smith[12] showed that the invariance of probability patterns to affine transformation, $T_z \mapsto a + bT_z$, implies that T_z satisfies the differential equation

$$\frac{\mathrm{d}T_z}{\mathrm{d}w} = \alpha + \beta T_z,$$

in which $w(z)$ is a function of the variable z. The solution of this differential equation expresses the scaling of probability patterns in the generic form

$$T_z = \frac{1}{\beta}(e^{\beta w} - 1), \tag{2}$$

in which, because of the affine invariance of T_z, I have added and multiplied by constants to obtain a convenient form, with $T_z \mapsto w$ as $\beta \mapsto 0$. With this expression for T_z, we may write probability patterns generically as

$$q_z = ke^{-\lambda(e^{\beta w}-1)/\beta}. \quad (3)$$

Turning now to the log series and Zipf's law, the relation $n = e^r$ between observed pattern, n, and process, r, plays a central role. Here, r represents the total of all proportional processes acting on abundance. A proportional process simply means that the number of individuals or entities affected by the process increases in proportion to the number currently present, n.

The sum of all of the proportional processes acting on abundance over some period of time is

$$r = \int_0^\tau m(t)\mathrm{d}t.$$

Here, $m(t)$ is a proportional process acting at time t to change abundance. The value of $r = \log n$ is the total of the m values over the total time, τ. For simplicity, I assume $n_0 = 1$.

Proportional processes are often discussed in terms of population growth[5,14]. However, many different processes act individually on the members of a population. If the number of individuals affected increases in proportion to population size, then the process is a proportional process.

Growth and other proportional processes often lead to an approximate power law, $q_n \approx kn^{-\rho}$. However, the exponent of a growth process does not necessarily match the values observed in the log series and Zipf's law. We need both the power law aspect of proportional process and something further to get the specific forms of those widely observed abundance distributions. That something further arises from conserved quantities and their associated invariances.

The log series and Zipf's law follow as special cases of the generic probability pattern in Equation 3. To analyze abundance, focus on the process scale by letting the variable of interest be $z \equiv r$, with the key scaling simply the process variable itself, $w(r) = r$. Then Equation 3 becomes

$$q_r\mathrm{d}r = ke^{-\lambda(e^{\beta r}-1)/\beta}\mathrm{d}r, \quad (4)$$

in which $q_r\mathrm{d}r$ is the probability of a process value, r, in the interval $r + \mathrm{d}r$. From the relation between abundance and process, $n = e^r$, we can change from the process scale to the abundance scale by the substitutions $r \mapsto \log n$ and $\mathrm{d}r \mapsto n^{-1}\mathrm{d}n$, yielding the identical probability pattern expressed on the abundance scale

$$q_n\mathrm{d}n = kn^{-1}e^{-\lambda(n^\beta-1)/\beta}\mathrm{d}n. \quad (5)$$

The value of k always adjusts to satisfy the constraint of invariant total probability, and the value of λ always adjusts to satisfy the constraint of invariant average value.

For $\beta = 1$, we obtain the log series distribution

$$q_n = kn^{-1}e^{-\lambda n}, \quad (6)$$

replacing $n - 1$ by n in the exponential term which, because of affine invariance, describe the same probability pattern. The log series is often written with $e^{-\lambda} = p$, and thus $q_n = kp^n/n$. One typically observes discrete values $n = 1, 2, \ldots$. The Supplemental material for this article[15] shows the relation between discrete and continuous distributions[16] and the domain of the variables. The continuous analysis here is sufficient to understand pattern.

For $\beta \to 0$, we have $(n^\beta - 1)/\beta \to \log n$, which yields

$$q_n = \lambda n^{-(1+\lambda)} \quad (7)$$

for $n \geq 1$. If we constrain average abundance, $\langle n \rangle$, with respect to this distribution, then

$$\lambda = \frac{1}{1 - 1/\langle n \rangle}.$$

For any average abundance that is finite and not small, $\lambda \to 1$, which is Zipf's law.

Equation 5 provides a general expression for abundance distributions. The log series and Zipf's law set the endpoints of $\beta = 1$ and $\beta \to 0$. We can understand the differences between abundance distributions in terms of the parameter β by writing the distribution in the generic form of Equation 1, with the defining affine invariant scale

$$T_n = \frac{\log n}{\lambda} + \frac{n^\beta - 1}{\beta}. \quad (8)$$

This scale expresses the invariances that define the pattern. At the Zipf's law endpoint, $\beta \to 0$, the scale becomes $2 \log n = 2r$, when satisfying the constraint that the average abundance, $\langle n \rangle$, is sufficiently large.

In this case, with affine invariant scale $T_n = 2r$, neither addition nor multiplication of process value, $r \mapsto a + br$, alters the pattern. We could have started with this affine invariance, and derived the probability pattern from the invariance properties[10,11].

For the log series endpoint, $\beta = 1$, the affine invariant scale is

$$T_n = \frac{1}{\lambda}\log n + n.$$

The dominant aspect of the scale changes with n. For small abundances, the logarithmic scale $r = \log n$ dominates, and for large abundances, the linear scale $n = e^r$ dominates. Many common probability patterns change their scaling with magnitude[13,17].

For log series patterns, the dominance of scale at small magnitude by r corresponds to affine invariance with respect to r. At larger abundances, the dominance by the effectively linear scale, n, corresponds to invariance to a shift in process $r \mapsto a + r$, but not to a multiplication of process, $r \mapsto br$, because $e^{br} = n^b$ is a power transformation of abundance. Linear scales are not

invariant to power transformations. Once again, we could have derived the pattern from the invariances.

In Equation 8, intermediate values of β combine aspects of Zipf's law and the log series. The closer β is to one of the endpoints, the more the invariance characteristics of that endpoint dominate pattern.

Conclusion

This analysis shows how two great and seemingly unconnected puzzles solve very simply in terms of a single continuum between alternative invariances. This approach reveals the simple invariant structure of many common probability patterns.

Data availability

Underlying data

All data underlying the results are available as part of the article and no additional source data are required.

Extended data

Zenodo: Supplemental Material for "The common patterns of abundance: the log series and Zipf's law". https://doi.org/10.5281/zenodo.2597895[15].

Grant information

The Donald Bren Foundation supports my research.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

I completed this work while on sabbatical in the Theoretical Biology group of the Institute for Integrative Biology at Eidgenössische Technische Hochschule (ETH) Zürich.

A previous version of this article is available on arXiv: https://arxiv.org/abs/1812.09662

References

1. Fischer H: **A History of the Central Limit Theorem: From Classical to Modern Probability Theory.** Springer, New York, 2011.
Publisher Full Text
2. Kotz S, Nadarajah S: **Extreme Value Distributions: Theory and Applications.** World Scientific, Singapore, 2000.
Publisher Full Text
3. Coles S: **An Introduction to Statistical Modeling of Extreme Values.** Springer, New York, 2001.
Publisher Full Text
4. Zipf GK: **The Psycho-biology of Language.** Houghton Mifflin, Boston, 1935.
Reference Source
5. Gabaix X: **Zipf's law for cities: an explanation.** *Q J Econ.* 1999; **114**(3): 739–767.
Reference Source
6. Arshad S, Hu S, Ashraf BN: **Zipf's law and city size distribution: a survey of the literature and future research agenda.** *Physica A: Stat Mech Appl.* 2018; **492**: 75–92.
Publisher Full Text
7. Axtell RL: **Zipf distribution of U.S. firm sizes.** *Science.* 2001; **293**(5536): 1818–1820.
PubMed Abstract | Publisher Full Text
8. Fisher RA, Corbet AS, Williams CB: **The relation between the number of species and the number of individuals in a random sample of an animal population.** *J Anim Ecol.* 1943; **12**(1): 42–58.
Publisher Full Text
9. Baldridge E, Harris DJ, Xiao X, *et al.*: **An extensive comparison of species-abundance distribution models.** *PeerJ.* 2016; **4**: e2823.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Frank SA: **Common probability patterns arise from simple invariances.** *Entropy.* 2016; **18**(5): 192.
Publisher Full Text
11. Frank SA: **Measurement invariance explains the universal law of generalization for psychological perception.** *Proc Natl Acad Sci U S A.* 2018; **115**(39): 9803–9806.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Frank SA, Smith E: **A simple derivation and classification of common probability distributions based on information symmetry and measurement scale.** *J Evol Biol.* 2011; **24**(3): 469–484.
PubMed Abstract | Publisher Full Text
13. Frank SA: **How to read probability distributions as statements about process.** *Entropy.* 2014; **16**: 6059–6098.
Publisher Full Text
14. Gibrat R: **Les Inégalités Économiques.** Librairie du Recueil Sirey, Paris. 1931.
Reference Source
15. Frank SA: **Supplemental Material for "The common patterns of abundance: the log series and Zipf's law".** 2019.
http://www.doi.org/10.5281/zenodo.2597895
16. Au C, Tam J: **Transforming variables using the Dirac generalized function.** *Am Stat.* 1999; **53**(3): 270–272.
Publisher Full Text
17. Frank SA: **The invariances of power law size distributions [version 2; peer review: 2 approved].** *F1000Res.* 2016; **5**: 2074.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Referee Status: ✔ ✔ ✔

Version 1

Referee Report 17 April 2019

https://doi.org/10.5256/f1000research.20456.r46235

✔ **Scott E. Page**
Department of Political Science, Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI, USA

I found this article to be fascinating and elucidating but also a bit frustrating to read. The central claim of the article is that one can construct a family of distributions such that Zipf's law and species abundance are the endpoints of a more general process.

For that result to be interesting, the result has to be **for a general process** and not **for a family of distributions.**

The latter is easy. I just say, "here is a family of distributions, f(x) = x^{-a}" and then say that at one endpoint a=1 I have a species area law and at the other endpoint a=2, I get Zipf's law.

The contribution of the article lies in convincing us that the paper has done something other than an elaborate change of variables that simply restates that result through obfuscation.

So what does the paper do? The paper shows that if we restrict attention to probability patterns (by the way, it would be nice if "pattern" were formally defined) that are invariant to affine transformations then we have a specific form given by equation (3).

Given the form in equation (3), the author then claims that n represents pattern and r represents process. This needs to be elucidated.

For the main result, once we have invariance to affine transformation we get the differential equation with

dT_z/dw = \alpha + \beta T_z

From here, why doesn't it just follow that if \beta = 0, we have something that's going to fall off with a common invariant scale and for \beta = 1 the invariant scale changes with n.

The conclusion of the paper needs to be expanded. As a reader, I need a richer explanation for how the approach "reveals the simple invariant structure" of common probability distributions. In the conclusion, we should be given more intuition for how the holding the average abundance constant drives the results. Also, it would be nice to have more insight into what would cause a system to have more or fewer proportional processes acting on it.

Quibble:
Why isn't r a function of tau—the period of time?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Complexity

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 17 April 2019

https://doi.org/10.5256/f1000research.20456.r46236

Jose Lobo
School of Sustainability, Arizona State University, Tempe, AZ, USA

1. The manuscript addresses the relationship between two probability distributions that, although originated in specific research domains, have gone on to be widely used as representations of growth processes.

2. The mathematical derivations are clear.

3. The conclusion that "two great and seemingly unconnected puzzles solve very simply in terms of a single continuum between alternative invariances. This approach reveals the simple invariant structure of many common probability patterns." clearly follows from the exposition and is a useful contribution.

4. The usefulness and scope of the conclusion would be strengthened if the author considered another distribution which arises often in the explorations of growth processes: the log normal.

5. It would also strengthen the usefulness of the manuscript if the author were to expand on "this approach reveals the simple invariant structure of many common probability patterns.", in particular recapitulating what is the invariant structure.

6. Having connected two widely used distributions, what sort of research questions can now be addressed? How can the invariant structure linking two distributions be used in contexts other than Zipf's law or species distributions?

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Urban economics, growth models, analysis of statistical distributions as models of change.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 15 April 2019

https://doi.org/10.5256/f1000research.20456.r46232

Luís M. A. Bettencourt [1,2]
[1] Department of Ecology and Evolution, University of Chicago, Chicago, IL, USA
[2] Santa Fe Institute, Santa Fe, NM, USA

This manuscript approaches the origins of two particularly important distributions describing abundances in biological and social populations from the perspective of mathematical invariances of their mathematical forms.

The author shows in this way that Fisher's log series distribution and Zipf's "law" can arise in different

limits of the same parameter, characterizing a family of affine transformations that includes translations and scale transformations of growth rates.

The mathematical derivation is clear and elegant, so that the manuscript makes an important contribution to formal models deriving these abundance distributions.

What I think would improve the manuscript is greater contact with other methods for the derivation of these same distributions of abundance and an expanded discussion of limits.

Specifically:

1. The relationship between population dynamics and invariances of the abundance (or rate) distributions could be made a little more explicit: Population dynamics models (in analogy to other dynamical systems) are mappings, tracing explicit variable transformation over time, such as changes of “position” (translations, r-> a + r), or dilations (r -> b r). Asking for invariances of distributions under these dynamical transformations is the usual way to derive the distributions as steady state abundances. Power laws, such as Zipf’s law, are invariant under (stochastic) dilations, for example, while Fisher’s log series are invariant under other simple types of population dynamics (as in Volkov et at[1]). I’d appreciate a bit more discussion bridging these two approaches.

2. As, the author shows the derivation of Zipf’s law requires not only a parameter choice (beta -> 0) but also the limit of the average abundance -> infinity. Without the latter, the power law exponent won’t be Zipf’s. In dynamical derivations of Zipf’s law one asks instead that geometric random motion of the population abundances, is subjected to a (“reflecting”) boundary condition for small population sizes that stops them from getting too small, as in [5]. Under what circumstances are these two additional requirements (besides transformational invariances under multiplicative growth) equivalent? They seem to have a different character as one is a limit, while the other a boundary condition—is the limiting condition on the average the most general condition?

3. It would be interesting to describe the conditions (in terms of beta and any limits or time dependence on averages) for deriving the third distribution often invoked to describe the same abundances, the log-normal, in terms of the reasoning about invariances advanced here. This is discussed to some extent in previous work by the author, in reference [12]. I think its inclusion and discussion would benefit the current manuscript.

References

1. Volkov I, Banavar JR, Hubbell SP, Maritan A: Patterns of relative species abundance in rainforests and coral reefs.*Nature*. 2007; **450** (7166): 45-9 PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
No source data required

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Population Dynamics, Dynamical Systems, Ecology and Evolution, Statistical Mechanics, Complex Systems.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.