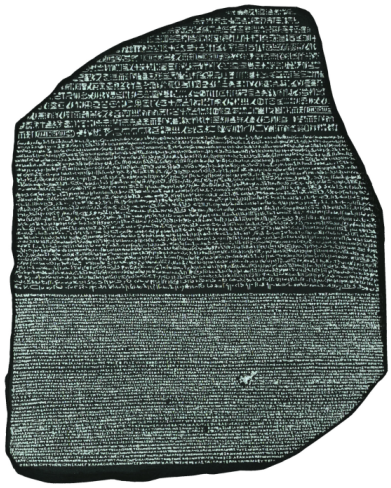
Pages 1–6
2023



Zipfian behaviour with missing data

Les Hatton and Greg Warr

31-Aug-2023 ...

Keywords: Notes

ABSTRACT

Zipfian power-law behaviour (with a droopy tail) follows naturally from CoHSI homogeneous theory, the backbone of the Fundamental Law of Inequality. The result is a power-law curve of frequency v. *rank*, degrading naturally into a drooping tail for sparsely-occupied ranks. In this paper we discuss the various ways in which the underlying dataset might be incomplete, for example with missing ranks or affected by noise such as distortions in the measure used to bin data samples and how we must interpret this.

CONTENTS

1 THE DERIVATION OF THE COHSI DISTRIBUTIONS

1.1 The Heterogeneous CoHSI Distribution

The standard formulation of Statistical Mechanics [1] for a simple heterogeneous system of the form of Fig. 1 where each component is represented by a string of beads of different colours and distinguishable order when we set as constraints the total information content (I) and the total number of tokens (T) and using Stirling's approximation for $\log(t_i!)$ reduces to [2]

Figure 1. Illustrating the CoHSI heterogeneous model. Seven components are shown as strings of tokens that are distinguishable by both their colour and their order but have no intrinsic meaning. In the case of the system of proteins, each string would correspond to a unique protein and each of the tokens would be an amino acid. Different colours would indicate different amino acids. On the other hand, if we applying this to software, each string would be a software component and the beads would be programming language tokens.

$$\log\Omega = T\log T - T - \sum_{i=1}^{M}\{t_i\log(t_i) - t_i\} + \alpha\{T - \sum_{i=1}^{M} t_i\} + \beta\{I - \sum_{i=1}^{M} I_i\} \quad (1)$$

Here I_i is the Hartley-Shannon Information content of the i^{th} string.

Applying the $\delta()$ variation to (1) to vary the string sizes t_i whilst keeping the a_i constant in the usual way [1] and simplifying gives

$$0 = -\sum_{i=1}^{M} \big(\log t_i + \alpha + \beta \frac{dI_i}{dt_i} \big) \delta t_i \qquad (2)$$

Calculating the I_i for the heterogeneous case is not trivial but can be done using a recurrence relation as detailed in [2]. For the homogeneous case however, it is much simpler as we shall now see.

1.2 The Homogeneous CoHSI Distribution

In contrast to the string of beads model of the heterogeneous CoHSI distribution, there is another way of arranging our beads amongst components [2]. We call this a homogeneous system. In such a system, the system can be represented by a number of components each appearing as a *bin*. A bin contains only one kind of token in *indistinguishable* order and no two bins have the same coloured bead. This is illustrated using beads of different colours in Fig. 2. This distribution encompasses a wide class of systems as different as word counts in textual documents and the distribution of elements in the universe [3]. In such systems, a heterogeneous definition of information for a bin is degenerate since there is only one kind of bead in each bin in contrast to the different colours of bead on the same string in the heterogeneous case. We need therefore a different definition for information content, which as we will show leads directly in our theory to an alternative proof of Zipf's law which is known to be present in many datasets, [4].

We anticipate that these distinctions between heterogeneous and homogeneous systems will lead to different information measures with consequently different properties. We should not be surprised by this as precisely the same occurs in physical systems where distinguishable order leads to Bose-Einstein statistics and indistinguishable order leads to Fermi-Dirac statistics [1].

Figure 2. Illustrating the CoHSI homogeneous model. In each bin, all of the tokens are identical (i.e. of the same colour). Each bin contains tokens of a different colour and the beads are by definition in no distinguishable order.

Whichever definition of Hartley-Shannon Information is used, we recall that the methodology simply tells us the most likely, or *canonical* distribution for discrete systems with the same fixed size and fixed information content, howsoever defined.

In homogeneous systems, we can envision that each bead or *token* carries a payload unique to its colour such that each bin contains *only tokens with the same payload, unique to that bin*. We don't need to know what this payload is and it is indeed irrelevant as the beads have no intrinsic meaning in this theory; the only property that differently coloured beads have is that of distinguishability. We represent this system by assembling beads of the same colour in the appropriate bin, Fig. 2. As mentioned earlier, we cannot simply follow the heterogeneous model by setting the size of the alphabet for each bin as each bin is degenerate - the alphabet $a_i = 1$ for all bins. Since log 1 = 0, this would give the total Hartley-Shannon Information as zero. Recall however that we are only looking for the total number of ways of arranging the beads in these bins so that each bin has beads of a unique colour without any regard to order.

Suppose then we have M bins such that the i^{th} bin contains t_i beads of unique colour b_i, where the total number of beads is $T = \sum_{i=1}^{M} t_i$. We will renumber them without loss of generality so that $t_1 \leq t_2 \leq .. \leq t_M$ so that i becomes synonymous with rank.

We proceed as follows. Imagine an infinite reservoir comprising equal numbers of each of the M colours. The numbers are equal because any colour has the same probability of being selected as any other colour. We want to fill all the bins using T beads drawn from this reservoir such that we finish up with a system where each of our M bins has its own unique colour as shown in Fig. 2. We need to know the total number of ways in which this is possible because this immediately tells us the Hartley-Shannon Information. To proceed, first select the bin numbered M and then fill it by selecting t_M beads of the same colour. Since we are selecting from M different colours and we have an infinite number of beads, the probability that we will achieve this selecting at random is $(1/M)^{t_M}$. For the second bin, we must then choose t_{M-1} beads of the same colour from one of the remaining $M-1$ colours, since no two bins can have the same colour, so the probability of filling this bin with the t_{M-1} beads of this colour from the remaining colours is $(1/(M-1))^{t_{M-1}}$, and so on.

The total number of ways N_h this can be done to populate all M bins is then given by this combined probability multiplied by the total number of ways in which T beads can be selected without constraint, which is $T!$. So we get,

$$N_h = T! \Big[\big(\frac{1}{M}\big)^{t_M} \times \big(\frac{1}{M-1}\big)^{t_{M-1}} \times .. \times \big(\frac{1}{1}\big)^{t_1} \Big] = T! \prod_{i=1}^{M} \big(\frac{1}{i}\big)^{t_i} \qquad (3)$$

Rewriting (3) then, the information content of this system is

$$\log N_h = \log T! + \sum_{i=1}^{M} t_i \log \big(\frac{1}{i}\big) = \log T! - \sum_{i=1}^{M} t_i \log i \qquad (4)$$

Following the heterogeneous development by folding this into the third term on the right hand side of (1) and applying the $\delta()$ operator using Stirling's approximation, the equivalent of (2) now gives

$$0 = -\sum_{i=1}^{M} \big(\log t_i + \kappa + \eta \log i \big) \delta t_i, \qquad (5)$$

leading to a homogeneous system pdf given by

$$t_i \sim i^{-\eta}, \qquad (6)$$

where η, κ are once again Lagrange undetermined multipliers.

There are notable differences between this and the heterogeneous case.

- No approximation is necessary for components, (i.e. bins rather than strings as there is no distinguishable order), when t_i is comparable to a_i, since this does not arise in the homogeneous case.
- This is a pure power-law at all values of t_i but arranged in order of *rank* since as we have already pointed out i is now synonymous with rank; this is in fact Zipf's law, except for bins which have the lowest

populations, in which case Stirling's law is not sufficiently accurate and the result is a natural droop in the tail of the distribution, i.e. the most sparsely-populated bins, as we shall see. The explicit appearance of rank is a direct result of CoHSI.

1.3 Comparing heterogeneous and homogeneous systems

Both heterogeneous and homogeneous systems share precisely the same underlying methodology, both are the overwhelmingly most likely distributions given their total size and their total Information. What is crucially different is the measure of Information used in each.

To re-iterate, heterogeneous distributions result when each component in a system comprises an *ordered* string of tokens (beads) with more than one type of token (colour of bead) represented. In contrast, each component of a homogeneous distribution is simply a bin containing all beads of the same type (colour) with no order implied. Each bin of course has a different colour.

Mathematically, the difference can be seen by comparing equations (2), (5). For convenience, they are repeated here in the form in which they are solved as heterogeneous and homogeneous respectively.

$$0 = \left(\log t_i + \alpha + \beta \frac{dI_i}{dt_i} \right) \tag{7}$$

$$0 = \left(\log t_i + \kappa + \eta \log i \right) \tag{8}$$

They look very similar but this similarity conceals a fundamental difference. *The solution of (7) leads to frequency v.* ***length*** *(t_i) relationship, whereas the solution of (8) leads to frequency v.* ***rank*** *relationship.*

First we note that this is not an artificial distinction, it emerges naturally from the mathematics. It turns out that both (7) and (8) are dominated by power-laws parameterised in the case of heterogeneous systems by the Lagrange undetermined constants α, β and the homogeneous case by the Lagrange undetermined constants κ, η, but the "x-axis" in the former is length and in the latter is rank.

Why does this matter? Well, consider what happens when our simple data model breaks down under real-world conditions of missing or distorted data.

1.3.1 Missing data for the i^{th} rank or component In the heterogeneous case, if we have no data for the i^{th} component, i.e. no values of t_i, then a gap quite naturally appears in the plot, since its "x-axis" is t_i. Dealing with gaps is a routine procedure in mathematics using interpolation via collocation (honouring existing data points using splines of various kinds) or data fitting using some kind of metric such as least squares. For power-laws specifically when we are considering a general "x-axis" rather than using the rank, some form of non-linear collection or binning is often used [4]. There are many options.

In the homogeneous case, its not immediately obvious how to deal with a missing bin because the "x-axis" is rank. If a bin is empty for some reason, rank ordering means that all the bins are shuffled up so that only non-empty bins contribute to the functional shape. However, the argument used above in deriving equation (8) *assumes that the M bins completely characterise all possibilities.* In other words it assumes that the data are complete. If they are not complete, then for consistency with the heterogeneous model, we should leave gaps in the frequency v. rank distribution and use the "x-axis", in this case protein multiplicity, as a proxy for rank number.

The essence of applying the homogeneous case to real-world systems is to find a criterion to distinguish in which bin a data point is to be placed. The only requirements are that each bin contains an unordered collection of objects with the same property *and* that no two bins overlap *and* that this completely describes all data points. This is expanded at length in [3] where many examples of such mappings are shown, for example with lunar craters, ranges of diameters are used leading inevitably to the frequency v. rank Zipfian distribution. Sometimes these examples have a known set of values for binning, for example if contiguous k-tuples are used to break up the genome, we know that there are exactly 4^k possibilities since there are 4 nucleotides. Other times, we may not know the maximum bin number because it is defined by the data.

To illustrate what happens when this mapping has gaps, i.e. unpopulated or missing bins, consider the example of horizontal gene transfer [5]. This is an example of a homogeneous system where the criterion determining the colour of a bead is the number of times an exact copy of the same protein occurs in more than one species (or its equivalent) somewhere in the three domains of life (Archaea, Bacteria and Eukaryota) or viruses. This is known as the *protein multiplicity*. If we focus on a single bin, each bead corresponds to a particular protein which happens to be copied the same number of times around the phylogenetic tree of life or viruses. First we should note that the proteins quite likely have nothing else in common. Why should they? There is no physico-chemical property which they would obviously have in common.

It should be obvious that for low multiplicity, the corresponding bin will likely have many proteins and for high multiplicity, many fewer. Amongst millions of proteins known to be distributed around the phylogenetic tree and viruses, it is pretty likely the same one will occur say twice, whilst it seems likely that proteins occurring say 10000 times elsewhere are much rarer. As we increase multiplicity from 2, sooner or later we will encounter a bin which is empty. In fact as the multiplicity further increases, more and more bins are likely to be empty.

Consider Fig. 3. This shows five generations of the Trembl protein database (ftp://uniprot.org), versions 15-07 to 21-03 with the protein multiplicity plotted against the multiplicity rather than the rank. As can be seen (and robustly supported statistically), these form the characteristic straight line of a power-law as this is a log-log cumulative complementary distribution function. However looking down the individual lines as we get to sparsely occupied bins, gaps are increasingly obvious. However as the "x-axis" here is multiplicity, the gaps are explicitly plotted. This is consistent with the way we treat heterogeneous distributions for which the "x-axis" is length.

We should note three things. First of all, the first empty bin occurs later and later as the Trembl distribution grows in size. Second, up to the number of this bin, multiplicity and rank are synonymous. Third, an empty bin in one version might not be an empty bin in a later or even earlier version of the database. Consider Table 1, where we can see all these factors.

Table 1. First missing bin Trembl releases.

Trembl version	Total proteins with 1 copy	First missing bin	Notes
7-01	5,980	40	-
13-01	51,709	106	-
15-07	215,916	122	Populated in 13-01
18-02	909,066	162	Populated in 13-01, missing in 15-07
21-03	1,995,3545	237	-

It is clear that this is a missing data problem which we can solve by following the example of the heterogeneous distribution and extending the homogeneous case to missing bins assuming that the criterion used to distinguish the bins, which in this case represent protein multiplicity, *is synonymous with rank* and leaving the gaps in the frequency v. rank plot.

We can see this in plots of the full datasets. When we plot this data in extended homogeneous form with gaps corresponding to missing bins included, we get Fig. 3. This is a nearly 4 decadal power-law strongly supported by both necessary and sufficient statistical arguments [5]. However if we plot this as pure rank order assuming the data is complete i.e. no missing bins, the points all shuffle up to give Fig. 4. This is identical with Fig. 3 up to the first missing bin and thereafter as more and more bins are missing, an accelerated droop can be seen reducing the power-law by an entire decade.

To summarise this section, it seems clear then following the heterogeneous model, that assuming protein multiplicity as a proxy for rank is the consistent approach when we have missing data.

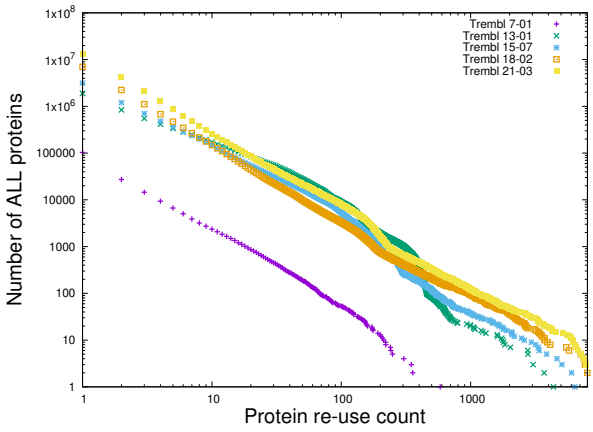


Figure 3. By multiplicity.

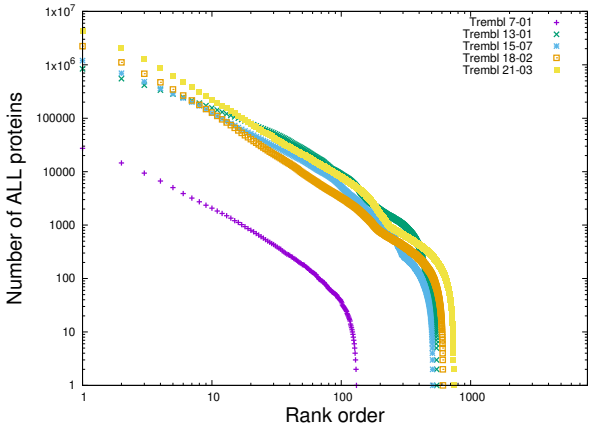


Figure 4. By rank.

1.4 Potential data distortions

Data distortion occurs when we incorrectly measure either length in the case of heterogeneous systems or mis-categorize a bead in homogeneous systems placing it in the wrong bin. Let's give an example of this in a heterogeneous system.

We know with a very high degree of precision [2], that if we plot the frequency of software components of different lengths *measured in programming language tokens*, we will get a multi-decade high precision power-law exactly as predicted by the application of CoHSI (Conservation of Hartley-Shannon Information). Fig. 5 illustrates.

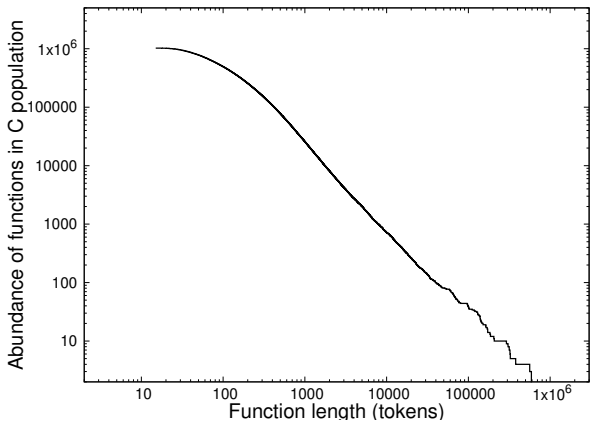


Figure 5. A population of software illustrating the emphatic power-law of function lengths when measured in programming language tokens.

This is not the case however if we use source lines of code to measure program size. Source lines of code are the conventional way of measuring program size and have been used for decades by computer scientists because of their visual appeal and easy measurement. They do not however map onto any information model and their definition varies between programming languages [6].

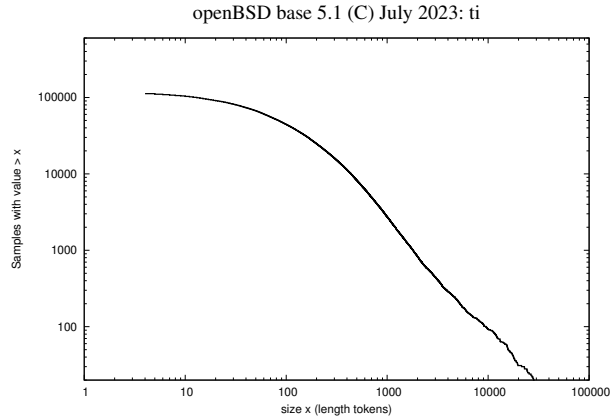


Figure 6. By tokens.

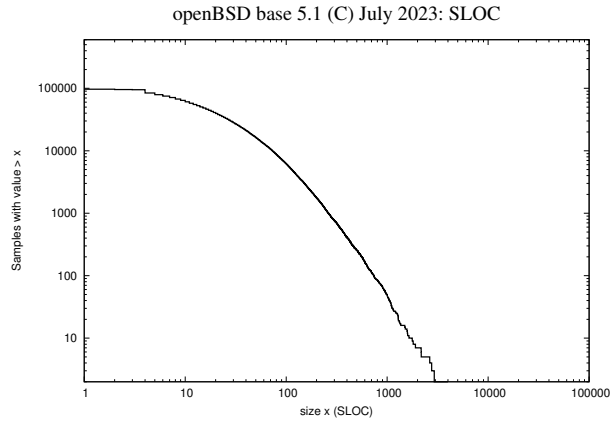


Figure 7. By source line of code.

Exactly the same data is shown measured in source lines of code as Fig. 7. The continuous curvature of this, a characteristic of log-normal behaviour, is just about evident when compared with the straightness of Fig. 6, but a detailed analysis by [7] on the same software distribution concludes that it is probably log-normal or double Pareto.

In other words, by distorting the length measurement, the high-precision power-law which occurs when using a measurement derived from Information Theory, that of language tokens, degrades into a more complex distribution when a more arbitrary length measurement with no relationship to Information Theory is used. Whether this is log-normal or double Pareto is irrelevant, the departure from pure power-law is an artifact of using the wrong measurement system.

2 CONCLUSIONS

Two distributions, one of frequency v. length (heterogeneous) and one of frequency v. rank (homogeneous) naturally emerge from CoHSI theory. With real data we must of course consider the effects of missing or corrupted data. In this paper we have demonstrated the effects of missing bins on a homogeneous system taken from the complete known set of proteins and how the theory has to be modified to deal with this.

REFERENCES

[1] Mike A. Glazer and Justin S. Wark. *Statistical Mechanics. A survival guide*. OUP, 2001.

[2] Les Hatton and Gregory Warr. Strong evidence of an information theoretical conservation principle linking all discrete systems. *R.Soc. open sci*, 6(191101), 11 2019.

[3] L. Hatton and G.W. Warr. *Exposing Nature's Bias: the Hidden Clockwork behind Society, Life and the Universe*. Bluespear Publishing, 2022. isbn 978-1-908-42204-0.

[4] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351, 2006.

[5] Les Hatton and Greg Warr. CoHSI IV; Unifying Horizontal and Vertical gene transfer - is mechanism irrelevant ? *arXiv*, 11 2018.

[6] Les Hatton. Dependable software depends on dependable measurement. *IEEE Computer*, XX(X):XX–XX, 2023.

[7] Israel Herraiz, Jesus M. Gonzalez-Barahona, and Gregorio Robles. Towards a theoretical model for software growth. In *Fourth International Workshop on Mining Software Repositories (MSR'07:ICSE Workshops 2007)*, pages 21–21, 2007.