Finite-sample frequency distributions originating from an equiprobability distribution

Thorsten Pöschel*
Humboldt-Universität zu Berlin, Charité, Institut für Biochemie, Monbijoustraße 2, Berlin D-10117, Germany

Jan A. Freund†
Humboldt-Universität zu Berlin, Institut für Physik, Invalidenstraße 110, Berlin D-10115, Germany
(Received 27 November 2001; published 9 August 2002)

Given an equidistribution for probabilities $p(i)=1/N$, $i=1,\ldots,N$. What is the expected corresponding rank ordered frequency distribution $f(i)$, $i=1,\ldots,N$, if an ensemble of M events is drawn?

DOI: 10.1103/PhysRevE.66.026103 PACS number(s): 02.50.Fz, 89.75.Da

I. INTRODUCTION

The probability $p(i)$ to draw an event i from a set of N possible events is defined as the limits

$$p(i)=\lim_{M\to\infty}\frac{M_i}{M} \quad \text{with} \quad \sum_{j=1}^{N} M_j=M, \tag{1}$$

where $f(i)\equiv M_i/M$ is the relative frequency to find the ith event M_i times in a randomly chosen sample of M. According to the law of large numbers the relative frequencies stochastically converge to the corresponding probabilities, hence, for very large sample size $M\gg N$ we can practically identify both values. For smaller sample size, however, the distribution of relative frequencies may deviate significantly from the probability distribution (e.g., [1–3] and many others). Assume further the equidistribution $p(i)=1/N$, then for very large sample size $M\gg N$ one expects that all or almost all of the N possible events are found in the sample and occur with approximately equal frequency, whereas in the opposite case $M\ll N$ almost all events occur only once or a few times in the sample, i.e., the latter frequency distribution will deviate significantly from the former one. From this simple argumentation one may conclude that the frequency distribution that one expects depends sensitively on the sample size M. This type of finite size effects is of major relevance for statistical analysis of DNA and other biosequences, e.g., [4–7]. In this paper we want to calculate the expected frequency distribution which one finds in dependence on the sample size M.

If we draw a frequency distribution of N different events there are $N!$ possibilities to arrange the events along the abscissa. An arrangement that leads to a decaying function for the frequencies or probabilities we call a Zipf order and the corresponding distribution a Zipf ordered distribution. To find a Zipf ordered frequency distribution that we have to expect if M events from an equidistribution are drawn we have to determine in dependence on M how many events (on average) are not drawn, i.e., are drawn zero times, how many are drawn once, twice, etc.

In the following section we derive the expectation value for the number $\langle K_i\rangle$ of those events which ocurr i times in the sample. The analytic expression is then used to infer the unknown number N of total events and to compare the theoretically expected Zipf ordered frequency distribution with the measured one. The results are useful in connection with entropy estimates computed from finite samples.

II. THE NUMBER K_i OF DIFFERENT EVENTS EACH OCCURRING EXACTLY i TIMES AND ITS EXPECTATION VALUE $\langle K_i\rangle$

The result of M subsequent drawings from a set of N different equiprobable events can be identified with randomly placing M indistinguishable balls in N indistinguishable urns, each having the same probability N^{-1}. Denoting the number of urns containing exactly i balls by k_i, a possible outcome can be shortly described by the vector $(k_0,k_1,\ldots,k_M)$; this is what we call a cluster configuration. The number of empty urns is given by k_0 and, consequently, the number of occupied urns by $N-k_0$. Any admissible cluster configuration obeys the following two conditions:

$$\sum_{i=0}^{M} k_i=N \quad \text{(total number of urns)}, \tag{2}$$

$$\sum_{i=0}^{M} k_i i=M \quad \text{(total number of balls)}. \tag{3}$$

We are interested in the stochastic variable K_i, denoting the number of urns each filled with exactly i balls, and its expectation value $\langle K_i\rangle$. Introducing for each $i=0,1,\ldots$ and each urn $j=1,\ldots,N$ its related indicator $I_i(j)$ by the following definition:

$$I_i(j)=\begin{cases}1 & \text{if urn } j \text{ contains exactly } i \text{ balls}\\ 0 & \text{else,}\end{cases} \tag{4}$$

the random variable K_i is related to the stochastic indicators by $K_i=\Sigma_{j=1}^{N} I_i(j)$. Due to the additivity of the expectation operator we find

*Email address: thorsten.poeschel@charite.de
URL: summa. physik.hu-berlin.de/~kies
†Email address: freund@physik.hu-berlin.de
URL: summa. physik.hu-berlin.de/~janf

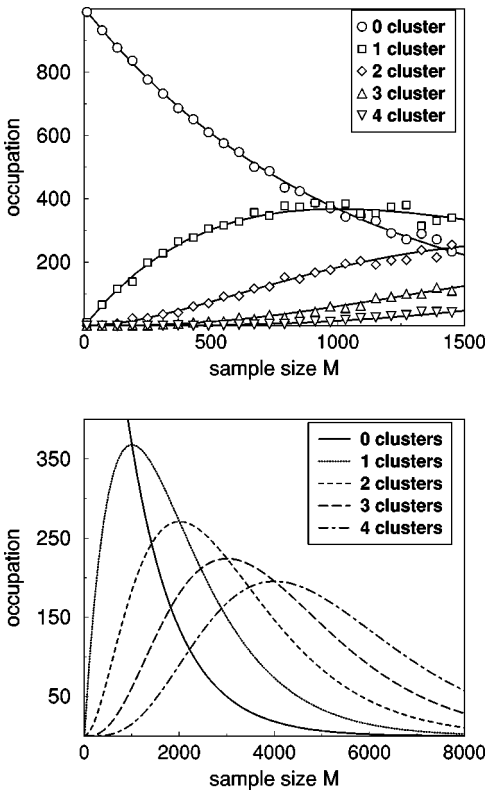

FIG. 1. Expectation values for clusters of different sizes over the sample size M taken from a set of $N=1000$ equidistributed different events. The lines show the theoretical result calculated from Eq. (6). The symbols show the cluster distribution found from a numerical simulation where M random integers have been drawn from the interval [1,1000]. The lower figure shows the same data for a larger range of sample sizes.

$$\langle K_i\rangle=\left\langle \sum_{j=1}^{N} I_i(j)\right\rangle=\sum_{j=1}^{N}\langle I_i(j)\rangle=N\langle I_i(j)\rangle, \tag{5}$$

where we have used that all $\langle I_i(j)\rangle$ are identical. The probability to find exactly i balls in any of the urns (here labeled j) and the remaining balls distributed arbitrarily among the remaining $N-1$ urns is the binomial distribution, hence,

$$\langle K_i\rangle=N\binom{M}{i}\frac{1}{N^i}\left(1-\frac{1}{N}\right)^{(M-i)}. \tag{6}$$

The expectation value $\langle K_i\rangle$ indicates how often events in a sample of size M occur exactly i times on an average. We call these occupation numbers i clusters. Obviously, for small $M\ll N$ nearly all of the N possible events are 0 clusters, i.e., they do not occur in our sample. As M increases the number of single occupations increases as well. For still growing M the number of multiple occupation becomes larger and, therefore, the number of 1 clusters decreases as more and more events occur multiple times in the sample. Figure 1 shows the occupation of the N possible different events as a function of the sample size M. The lines show the theoretical result Eq. (6) and the symbols in the left of Fig. 1 show the clusters as they have been found in sets of random numbers.

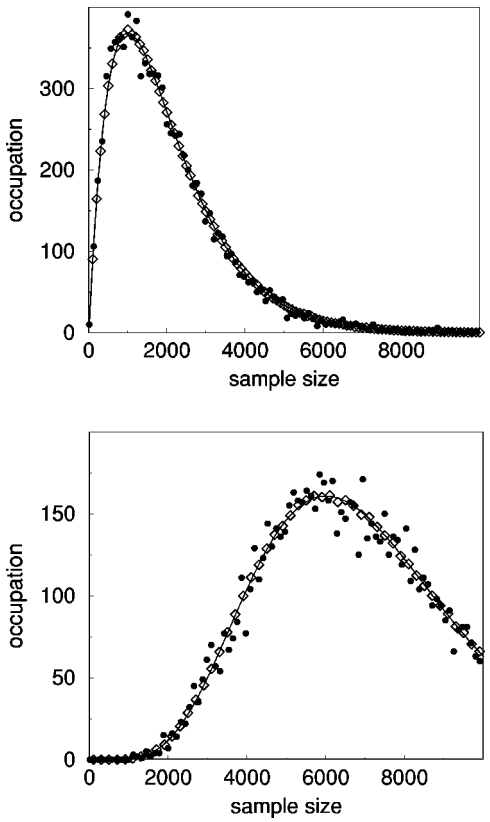

FIG. 2. If samples of M events are drawn independently, the resulting averaged cluster distribution converges to the theoretical value Eq. (6). The filled dots show the distribution of clusters of size 2 (top) and 6 (bottom) resulting from a single drawing. The diamonds are averaged cluster distributions over 100 independent drawings of random numbers. The solid lines are the theoretical values according to Eq. (6).

If we draw only once a sample of size M from a set of N possible events and calculate the occupation numbers (the cluster frequencies) the cluster distribution itself is fluctuating (Fig. 2, filled dots). Averaging the cluster distribution over a number of independent selections each of size M the cluster distribution converges to the theoretical curve predicted by Eq. (6). Figure 2 shows the cluster distribution as found from a random sample of size M out of $N=1000$ allowed events together with cluster distributions averaged over 100 independent drawings.

III. ZIPF ORDERED FREQUENCY DISTRIBUTION

Equation (6) allows to determine the expectation value of the number of different events N^* in dependence on the total number of drawn events M, which is simply related to the expected probability to find a cluster of size zero,

$$N^*=N-\langle K_0\rangle, \tag{7}$$

from which we compute

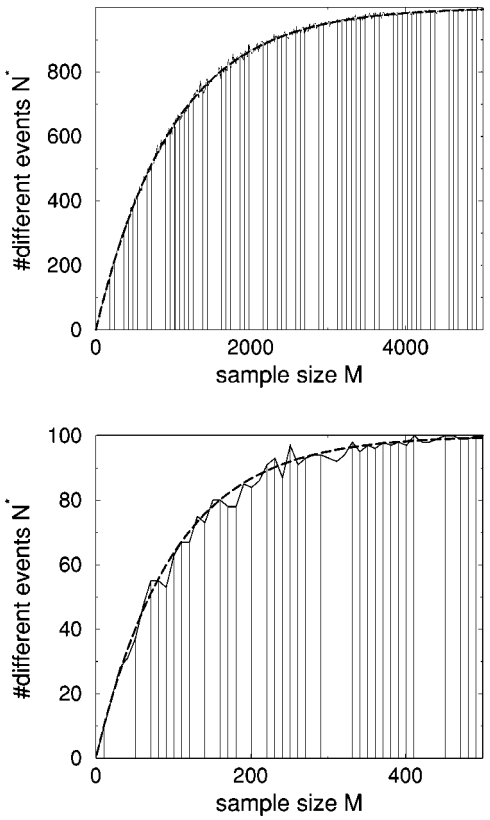

FIG. 3. Number of different events N^* in a sample of size M drawn from a equidistribution $p_i = 1/N$. The dashed line shows the analytical result Eq. (8), the histograms show the results of a computer simulation. top: $N = 1000$, bottom: $N = 100$.

$$\frac{N^*}{N} = 1 - \left(1 - \frac{1}{N}\right)^M = 1 - \exp\left[M \ln\left(1 - \frac{1}{N}\right)\right]$$
$$= 1 - \exp\left(-\frac{M}{N}\right)\exp\left[-O\left(\frac{M}{N^2}\right)\right]. \quad (8)$$

From Eq. (8) we see that for all $M \ll N^2$ Eq. (8) can be approximated to very good accuracy by

$$\frac{N_a^*}{N} \approx 1 - \exp\left(-\frac{M}{N}\right), \quad (9)$$

a result which has been found before by computer simulations [8]. The maximal *absolute* deviation is $N^* - N_a^* = 1/e \approx 0.37$ (for $N = M = 1$) that falls rapidly to $1/(2e) \approx 0.18$ as N goes to infinity. In Fig. 3 the analytical result (8) is compared with a simulation. The histograms in the figure show the results of a single realization. If we average the numerical results over several runs the numerical curve falls together with the analytical one.

For the wide range of practical interest, $5/8 \leqslant N_a^*/M < 1$, from Eq. (9) we may approximate the entropy of the distribution if we know the number of different events N^* contained in a sample of size M:

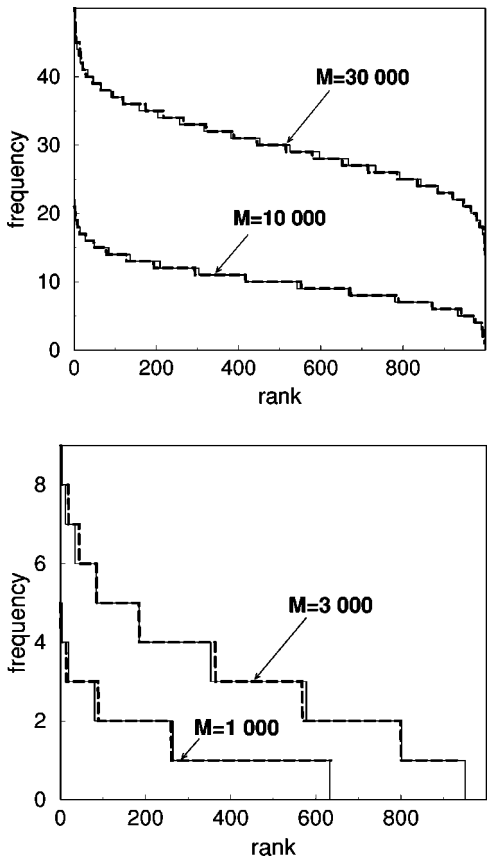

FIG. 4. Zipf ordered frequency distributions calculated from samples of size M from a equiprobability distribution of $N = 1000$ different random numbers (dashed lines). The theoretical curve due to Eq. (11) is drawn with solid lines.

$$S = \log_2 N \approx \log_2\left(\frac{[3M + \sqrt{3M(8N^* - 5M)}]M}{12(M - N^*)}\right). \quad (10)$$

We want to compile the results from the preceding section to find the desired Zipf ordered frequency distribution. Equation (6) tells how many, in average, events do not occur in the sample (drawn zero times), how many are drawn once, twice, etc. This yields directly the Zipf ordered frequency distribution

$$f(i) = \begin{cases} 0 & \text{for} \quad N \geqslant i > N - \langle K_0 \rangle \\ 1 & \text{for} \quad N - \langle K_0 \rangle \geqslant i > N - \langle K_0 \rangle - \langle K_1 \rangle \\ \cdots & \\ j & \text{for} \quad N - \sum_{k=0}^{j-1} \langle K_k \rangle \geqslant i > N - \sum_{k=0}^{j} \langle K_k \rangle. \end{cases} \quad (11)$$

Using Stirling's formula to expand the expressions in Eq. (6) the analytical result Eq. (11) can be written easily in elementary functions.

Figure 4 shows Zipf ordered frequency distributions calculated from a sample of random numbers (dashed lines) together with the theoretical distributions due to Eq. (11) (solid lines). The combinatorial theory derived in the preced-

ing section predicts the Zipf ordered frequency distribution that results from an equidistribution with good accuracy.

IV. DISCUSSION

Based on combinatorial considerations we calculated the expectation values to find k_i events *exactly i* times in a sample of size M which have been drawn from an equidistribution. These cluster probabilities allow to estimate the total number of events N, given the number of *different* events found in a sample of size M is known. Moreover, the full Zipf ordered frequency distribution could be constructed. By numerical simulations it has been demonstrated that the analytically derived values coincide with experimental results, i.e., with cluster distributions and Zipf ordered frequency distributions originating from finite samples of random numbers.

ACKNOWLEDGMENTS

Discussions with Werner Ebeling and Rosa Ramírez are acknowledged.

[1] A. O. Schmitt, H. Herzel, and W. Ebeling, Europhys. Lett. **28**, 303 (1993).
[2] T. Pöschel, W. Ebeling, and H. Rosé, J. Stat. Phys. **80**, 1443 (1995).
[3] H. Herzel, A. O. Schmitt, and W. Ebeling, Chaos, Solitons Fractals **4**, 97 (1994).
[4] P. Allegrini, M. Buiatti, P. Grigolini, and B. West, Phys. Rev. E **58**, 3640 (1998).
[5] P. Bernaola-Galván, I. Grosse, P. Carpena, J. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).
[6] D. Holste, I. Grosse, and H. Herzel, Phys. Rev. E **64**, 041917 (2001).
[7] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, M. Simons, and H. E. Stanley, Phys. Rev. E **47**, 3730 (1993).
[8] R. Ramírez (private communication).