Invariant expectation values in the sampling of discrete frequency distributions

Paolo Rossi

Dipartimento di Fisica dell'Università di Pisa and I.N.F.N., Sezione di Pisa, Largo Bruno Pontecorvo 3, I-56127 Pisa, Italy
rossi@df.unipi.it

Invariant expectation values in the sampling of discrete frequency distributions

Paolo Rossi

Dipartimento di Fisica dell'Università di Pisa and I.N.F.N., Sezione di Pisa, Largo Bruno Pontecorvo 3, I-56127 Pisa, Italy
rossi@df.unipi.it

Abstract

The general relationship between an arbitrary frequency distribution and the expected value of the frequency distributions of its samples is discussed. A wide set of measurable quantities ("invariant moments") whose expected value does not in general depend on the size of the sample is constructed and illustrated by applying the results to the Ewens sampling formula. Invariant moments are especially useful in the sampling of systems characterized by the absence of an intrinsic scale. Distribution functions that may parametrize the samples of scale-free distributions are considered and their invariant expectation values are computed. The conditions under which the scaling limit of such distributions may exist are described.

Keywords: Sampling, frequency distributions, Ewens formula, negative binomial distributions, scaling

1. Introduction

In many interesting physical, biological and social phenomena, whenever no intrinsic scale for the relevant variables is present, the emergence of "scaling laws" is phenomenologically observed [1]. However, strictly speaking, a power law is not a proper way of fitting empirical data, since no choice of the exponent can keep the higher moments of a power law distribution from diverging, while every phenomenological distribution leads to finite values for all moments. This is not just a technicality: it is rather a reflection of the fact that the long tail of a power law distribution is in practice cut off by the existence of some "hidden" scale, irrelevant in the scaling region, but eventually forcing some upper limit on the variables describing the system. It would therefore be convenient to parametrize the data by means of more regular distribution functions, sufficiently damped for very large values of the variables, but admitting power law distributions as regular limits when the control parameter implementing the cutoff is sent to its limiting value.

A related issue concerns the effects of sampling, which may be non-trivial even when we restrict our attention to the expected values of the sampled variables. On average sampling does not affect the distributions of individual objects belonging to different kinds, but when we consider frequency distributions (that is the number of kinds that are represented k times in a given population) we cannot in general expect that the frequency distribution in the samples be the same as in the original population, even after averaging over many different samples, basically because the cutoff induced by sampling acts differently (and in general nontrivially) at different scales. It is therefore quite important to be able to extract from the frequency distribution of the samples some information reflecting directly some intrinsic property of the underlying distribution.

Our purpose is therefore threefold. First we discuss the general relationship existing between some arbitrary frequency distribution and the expected value of the frequency distributions of its samples, and construct observables whose expected values turn out to be independent of the sample size, and therefore coinciding with the value taken by the same observables in the full distribution.

Then we study classes of distributions whose samples preserve the functional dependence on the parameters present in the original distribution, establishing the connection between the (a priori unknown) values of the parameters of the distribution and the (empirically estimated) parameters of the sample distributions.

Finally we study the scaling limit of these distributions (when it exists), in order to explore the possibility of their use for the phenomenological description of systems that are theoretically expected to show scaling in the limit when all empirical cutoffs (including those induced by sampling) are going to disappear.

In Section 2 we establish the notation and the general framework of our analysis.

In Section 3 we construct a wide set of combinations of expected values that do not depend on the sample size.

In Section 4 we apply our approach to the popular Ewens sampling formula, showing that its features are consistent with the general pattern and computing its invariant expectation values.

In Section 5 we consider the limiting case of a small sampling applied to a large population.

In Section 6 we focus on the case when the original population and its samples are sufficiently large in comparison with typical frequency values, finding a useful mathematical relationship between the generating function of the expected values of the sample distributions and the generating function of the original distribution.

In Section 7 we analyze a class of distributions (the so-called negative binomial distributions) admitting a scaling limit and enjoying the property that the distribution of expected values of the samples has the same mathematical form as the original distribution. We also compute in a closed form the values ot the basic invariant expectation values for these distributions.

Finally in Section 8 we analyze the scaling limit itself and discuss the conditions under which one may expect this limit to be a sensible description of the original system.

Appendices are devoted to the proofs of some mathematical results and to discussing the issue of correlation between random samples.

2. The general framework

We consider a set of N objects ("individuals") belonging to S different kinds ("species"), and we assume that the set contains $\hat{N}_a$ objects of the a-th kind, subject to the constraint $\sum_a \hat{N}_a = N$.

A sample is a set of n objects, containing $\hat{n}_a$ objects of the a-th kind, subject to the constraint $\sum_a \hat{n}_a = n$.

The probability $P_{\{\hat{n}_a\}}$ of extracting a specific sample $\{\hat{n}_a\}$ from a given set $\{\hat{N}_a\}$ is obtained from the multivariate hypergeometric distribution

$$P_{\{\hat{n}_a\}} = \binom{N}{n}^{-1} \prod_{a=1}^{S} \binom{\hat{N}_a}{\hat{n}_a}.$$

One can easily compute the relevant expected values, obtaining in particular

$$\langle \hat{n}_a \rangle = \hat{N}_a \frac{n}{N} = n\, \hat{p}_a, \qquad\qquad \langle \hat{n}_a^2 \rangle - \langle \hat{n}_a \rangle^2 = \frac{N-n}{N-1} n \hat{p}_a (1 - \hat{p}_a)$$

where $\hat{p}_a \equiv \bar{N}_a / N$ is the probability of extracting an object of the a-th kind in a single extraction.

It may be useful to consider also the limiting values of these expressions for small samples $\hat{n}_a \ll \hat{N}_a$. In this limit the probability of a specific sample is well approximated by the multinomial distribution

$$P_{\{\hat{n}_a\}} = n! \prod_{a=1}^{S} \frac{1}{\hat{n}_a!} (\hat{p}_a)^{\hat{n}_a}.$$

A frequency distribution is a set of values $\{N_k\}$, where N_k is the number of kinds such that for each of them there are k objects in the original set. According to the definition, the following conditions must be satisfied:

$$\sum_{k=1}^{N} N_k = S, \qquad\qquad \sum_{k=1}^{N} k\, N_k = N.$$

The frequency distribution of a sample is a set of values $\{n_l\}$, satisfying the conditions

$$\sum_{l=0}^{n} n_l = S, \qquad\qquad \sum_{l=1}^{n} l\, n_l = n.$$

Notice that the frequency distribution of a sample formally includes the (unobservable) value n_0, corresponding to the number of kinds, present in the original set, which are not represented in the sample.

It is in principle possible to compute the probability of any sample distribution $\{n_l\}$ as a function of a given set $\{N_k\}$. To this purpose it is convenient to define the intermediate variables N_{kl}, representing the (random) number of kinds characterized by k objects in the original set and by l ($l \leq k$) objects in the sample. The variables N_{kl} are constrained by the conditions:

$$\sum_{l=0}^{n} N_{kl} = N_k, \qquad\qquad \sum_{k=1}^{N} N_{kl} = n_l.$$

The probability $P_{\{N_{kl}\}}$ of a specific configuration $\{N_{kl}\}$ follows from the general probability formula [2]:

$$P_{\{N_{kl}\}} = \binom{N}{n}^{-1} \prod_{k=1}^{N} \left[N_k! \prod_{l=0}^{k} \frac{1}{N_{kl}!} \binom{k}{l}^{N_{kl}} \right],$$

subject to the constraint $\sum_{l=0}^{n} N_{kl} = N_k$.

The probability $P_{\{n_l\}}$ of finding a frequency distribution $\{n_l\}$ in a sample is obtained by summing the probabilities $P_{\{N_{kl}\}}$ over all configurations satisfiying the constraint $\sum_{k=1}^{N} N_{kl} = n_l$. The corresponding multivariate generating function can be defined as

$$\varepsilon^{(n)}(\{t_l\}) \equiv \sum_{\{n_l\}} P_{\{n_l\}} \prod_{l=0}^{n} t_l^{n_l} = \sum_{\{N_{kl}\}} P_{\{N_{kl}\}} \prod_{k=1}^{N} \prod_{l=0}^{k} t_l^{N_{kl}}.$$

It is also possible (and it will be quite convenient) to define a cumulative generating function $E(x;\{t_l\})$ for the probability of finding the frequency distributions $P_{\{n_l\}}$ for samples of all possible sizes :

$$E(x;\{t_l\}) \equiv \sum_{n=0}^{N} \binom{N}{n} \varepsilon^{(n)}(\{t_l\}) x^n = \sum_{\{N_{kl}\}} \prod_{k=1}^{N} \Big(N_k! \prod_{l=0}^{k} \frac{1}{N_{kl}!} \Big[\binom{k}{l} x^l t_l \Big]^{N_{kl}} \Big) = \prod_{k=1}^{N} \Big[\sum_{l=0}^{k} \binom{k}{l} x^l t_l \Big]^{N_k},$$

where we used the explicit expression of $P_{\{N_{kl}\}}$ and all the relevant constraints.

The expected values $\langle n_l \rangle$ can be computed starting from the above expressions and from the relationship

$$\langle n_l \rangle = \sum_{k=1}^{N} \langle N_{kl} \rangle = \sum_{k=1}^{N} \sum_{\{N_{jm}\}} N_{kl} P_{\{N_{jm}\}}.$$

Straightforward manipulations lead to the results [2]

$$\langle N_{kl} \rangle = N_k \frac{\binom{k}{l}\binom{N-k}{n-l}}{\binom{N}{n}}, \qquad\qquad \langle n_l \rangle = \frac{\sum_{k=1}^{N} N_k \binom{k}{l}\binom{N-k}{n-l}}{\binom{N}{n}}.$$

It is easy to check that the following relationships are satisfied:

$$\sum_{l=0}^{n} \langle n_l \rangle = \sum_{k=1}^{N} N_k = S, \qquad\qquad \sum_{l=0}^{n} l\, \langle n_l \rangle = \Big(\sum_{k=1}^{N} k\, N_k \Big) \frac{n}{N} = n.$$

In order to fully appreciate the relevance of considerations based on the expected values we must evaluate the variance of the frequency distribution of the samples. Taking second derivatives of the generating function $E(x; t_l)$ one obtains:

$$\langle n_l^2\rangle - \langle n_l\rangle^2 = \sum_{k,k'} N_k N_{k'} \binom{k}{l}\binom{k'}{l}\left[\frac{\binom{N-k-k'}{n-2l}}{\binom{N}{n}} - \frac{\binom{N-k}{n-l}}{\binom{N}{n}}\frac{\binom{N-k'}{n-l}}{\binom{N}{n}}\right] + \sum_k N_k \binom{k}{l}\left[\frac{\binom{N-k}{n-l}}{\binom{N}{n}} - \binom{k}{l}\frac{\binom{N-2k}{n-2l}}{\binom{N}{n}}\right].$$

Notice that in the large N limit the term quadratic in N_k is depressed by a power of $1/N$. This observation suggests that important limits of the above results may be obtained when considering large populations.

3. Invariant expectation values

It is important to be able to define a set of expectation values that are independent of the size of the sample, and therefore may reflect directly the properties of the original frequency distribution.

We consider the following combinations of expected values:

$$\langle m^{(n)}_{\{p_i\}}\rangle \equiv \binom{n}{P}^{-1} \sum_{\{q_i\}} \prod_{i=1}^{I} \Big[\binom{q_i}{p_i}\frac{\partial}{\partial t_{q_i}}\Big] \varepsilon^{(n)}(\{t_l\}|_{\{t_l=1\}},$$

where p_i are I arbitrary positive integer numbers, subject only to the constraint that $P \equiv \sum_i p_i \le n$.

The definition of the quantities appearing in the r.h.s. implies that the derivatives with respect to t_{q_i} are the joint factorial moments of the distribution; therefore we are dealing with weighted combinations of joint factorial moments. When some of the indices p_i are equal to one, the expected values may be expressed in terms of a combination of lower rank moments ($I' < I$).

It is possible to recognize that the quantities $\langle m^{(P)}_{\{p_i\}}\rangle$ are related (up to a trivial combinatorial factor taking into account the existence of n_p coincident values of the indices p_i) to the probability of finding the configurations $\{p_i\}$ in the sample containing P elements, and are therefore strictly connected with the probabilities $P_{\{n_p\}}$.

Exploiting the properties of the generating function $E(x; \{t_l\})$ we prove in Appendix A that

$$\langle m^{(n)}_{\{p_i\}}\rangle = \langle m^{(N)}_{\{p_i\}}\rangle \equiv M_{\{p_i\}}$$

for all sets $\{p_i\}$ such that $P \le n$. Hence the expected values of the nontrivial invariant moments $m_{\{p_i\}}$ evaluated for samples of arbitrary size $n \ge P$, coincide with the moments $M_{\{p_i\}}$ of the original frequency distribution. If the original set was generated by a random process, the $M_{\{p_i\}}$ will also be expected values. Recalling that $\varepsilon^{(N)}(\{t_l\} \equiv \prod_k t_k^{N_k}$ we may now generate a representation of all $P_{\{n_p\}}$ in terms of N_k, without making use of the coefficients N_{kl}.

The properties of the binomial coefficients make it possible to invert the relationship between invariant moments and joint factorial moments, thus finding that

$$\Big[\prod_{i=1}^{I}\frac{\partial}{\partial t_{q_i}}\Big]\varepsilon^{(n)}(\{t_l\}|_{\{t_l=1\}} = \sum_{\{p_i\}}\prod_{i=1}^{I}(-1)^{p_i-q_i}\binom{p_i}{q_i}\langle\binom{n}{P}m^{(n)}_{\{p_i\}}\rangle = \sum_{\{p_i\}}\prod_{i=1}^{I}(-1)^{p_i-q_i}\binom{p_i}{q_i}\binom{n}{P}M_{\{p_i\}}.$$

The basic invariant moments are

$$m_p^{(n)} = \binom{n}{p}^{-1}\sum_{q=p}^{n}\binom{q}{p}n_q.$$

According to the inversion formula

$$\langle n_l\rangle = \frac{\partial\varepsilon^{(n)}}{\partial t_l}|_{\{t_m=1\}} = \sum_{p=l}^{n}(-1)^{p-l}\binom{p}{l}\binom{n}{p}\langle m_p^{(n)}\rangle = \sum_{p=l}^{n}(-1)^{p-l}\binom{p}{l}\binom{n}{p}M_p.$$

One may define generating functions for the expected values of n_l and of the basic invariant moments:

$$f^{(n)}(t) \equiv \sum_{l=0}^{n} \langle n_l \rangle t^l, \qquad g^{(n)}(z) \equiv f^{(n)}\big(1+\frac{z}{n}\big) = \sum_{p=0}^{n} \binom{n}{p} M_p \Big(\frac{z}{n}\Big)^p.$$

Notice that a special case of the above formula is

$$F(t) \equiv \sum_{k=0}^{N} N_k t^k, \qquad G(z) \equiv F\big(1+\frac{z}{N}\big) = \sum_{p=0}^{N} \binom{N}{p} M_p \Big(\frac{z}{N}\Big)^p.$$

It is immediate to recognize that $g^{(n)}(0) = G(0) = M_0 \equiv S$, and $\frac{dg}{dz}^{(n)}(0) = \frac{dG}{dz}(0) = M_1 \equiv 1$.

4. Application to the Ewens sampling formula

The multivariate Ewens distribution [3, 4], called in genetics the Ewens sampling formula, describes a specific probability for the partition of n into parts, and found its main applications in the context of the neutral theory of evolution and in the unified neutral theory of biodiversity [5, 6]. Since the Ewens formula and its possibile generalizations have been the subject of a wide and still growing literature [7, 8, 9, 10], it may be interesting to apply the results presented in Section 3 to this specific instance. In our notation the Ewens probability distribution takes the form

$$P_{\{n_l\}} = \frac{1}{\aleph_n} \prod_{l=1}^{n} \frac{1}{n_l!} \Big(\frac{\theta}{l}\Big)^{n_l},$$

where $\aleph_n \equiv \frac{\Gamma(\theta+n)}{n!\,\Gamma(\theta)}$, $0 < \theta < \infty$ and $\sum l\, n_l = n$.

The joint factorial moments of the Ewens distribution are easily computed [11] and one can show that

$$\prod_{i=1}^{I} \Big[\frac{\partial}{\partial t_{q_i}}\Big] \sum_{\{n_l\}} P_{\{n_l\}} \prod_{l=0}^{n} t_l^{n_l} = \frac{\aleph_{n-Q}}{\aleph_n} \prod_i \Big(\frac{\theta}{q_i}\Big), \qquad Q \equiv \sum_i q_i \le n.$$

We are then left with the task of computing the summations appearing in the equation

$$\langle m^{(n)}_{\{p_i\}} \rangle = \binom{n}{P}^{-1} \sum_{\{q_i\}} \prod_{i=1}^{I} \binom{q_i}{p_i} \frac{\aleph_{n-Q}}{\aleph_n} \prod_i \Big(\frac{\theta}{q_i}\Big) = \frac{P!\,(n-P)!\,\Gamma(\theta)}{\Gamma(\theta+n)} \prod_{i=1}^{I} \Big(\frac{\theta}{p_i}\Big) \sum_{\{q_i\}} \prod_{i=1}^{I} \binom{q_i-1}{p_i-1} \frac{\Gamma(\theta+n-Q)}{\Gamma(\theta)\,(n-Q)!}.$$

We prove in Appendix B that

$$\sum_{\{q_i \ge p_i\}} \prod_{i=1}^{I} \binom{q_i-1}{p_i-1} \frac{\Gamma(\theta+n-Q)}{\Gamma(\theta)\,(n-Q)!} = \frac{\Gamma(\theta+n)}{\Gamma(\theta+P)\,(n-P)!},$$

hence

$$\langle m^{(n)}_{\{p_i\}} \rangle = \frac{P!\,\Gamma(\theta)}{\Gamma(\theta+P)} \prod_{i=1}^{I} \Big(\frac{\theta}{p_i}\Big) \equiv \frac{1}{\aleph_P} \prod_{i=1}^{I} \Big(\frac{\theta}{p_i}\Big),$$

showing that the expected values of the invariant moments of the Ewens distribution are independent of the sample size and are related to the probability of the configuration $\{p_i\}$ in the sampling of P elements.

We stress that invariant moments, because of their independence from the size of the sample, may become a highly valuable tool in testing the applicability of the Ewens distribution (and of the conceptual assumptions underlying its derivation) to the interpretation of actual empirical data.

5. Large population and small samples

A significant simplification occurs when $N \to \infty$ while all other variables are kept finite. Setting $\tilde{x} \equiv Nx$ and $t_0 = 1$ in the cumulative generating function and taking the large N limit we obtain

$$\tilde{E}(\tilde{x};\{t_l\}) \equiv 1 + \sum_{n=1}^{\infty} \tilde{\varepsilon}^{(n)}(\{t_l\}) \frac{\tilde{x}^n}{n!} = \prod_{k=1}^{\infty} \Big[1 + \sum_{l=1}^{k} \binom{k}{l} \Big(\frac{\tilde{x}}{N}\Big)^l t_l\Big]^{N_k} \to \prod_{k=1}^{\infty} \Big[1 + \sum_{l=1}^{\infty} \Big(\frac{k\,\tilde{x}}{N}\Big)^l \frac{t_l}{l!}\Big]^{N_k}.$$

We now define (for n, l different from zero) the following set of coefficients:

$$c^{(n)}(\{n_l\}) \equiv (-1)^{s-1}(s-1)! \prod_{l=1}^{n} \frac{1}{n_l!} \Big(\frac{1}{l!}\Big)^{n_l},$$

where $s = \sum_l n_l$ and $n = \sum_l l\, n_l$.

Notice that the definition of $\tilde{E}$ implies that

$$\ln \tilde{E}(\tilde{x};\{t_l\}) \equiv \sum_{n=1}^{\infty} \Big[\sum_{\{n_l\}} c^{(n)}(\{n_l\}) \prod_{l=1}^{n} \Big(\tilde{\varepsilon}^{(l)}(\{t_l\})\Big)^{n_l}\Big] \tilde{x}^n.$$

It is also possible to recognize that, under the same assumptions,

$$\ln \tilde{E}(\tilde{x};\{t_l\}) = \sum_{n=1}^{\infty} \Big[\sum_{\{n_l\}} c^{(n)}(\{n_l\}) \prod_{l=1}^{n} t_l^{n_l}\Big] \tilde{m}_n^{(N)} x^n,$$

where we introduced the large N limit of the basic invariant moments: $\tilde{m}_p^{(N)} \equiv \sum_q N_q (q/N)^p$.

Comparing the two results we conclude that, for each value of $n > 0$,

$$\sum_{\{n_l\}} c^{(n)}(\{n_l\}) \prod_{l=1}^{n} \Big(\tilde{\varepsilon}^{(l)}(\{t_l\})\Big)^{n_l} = \Big[\sum_{\{n_l\}} c^{(n)}(\{n_l\}) \prod_{l=1}^{n} t_l^{n_l}\Big] \tilde{m}_n^{(N)}.$$

These equations allow in principle for the recursive determination of all $\tilde{\varepsilon}^{(n)}(\{t_l\}$ in terms of $\{\tilde{m}_p^{(N)}\}$ (with $p \leq n$), starting from the initial condition $\tilde{\varepsilon}^{(1)} = t_1$. Higher rank invariant moments ($I > 1$) in the large N limit become polynomials in the basic moments.

However one must keep in mind that, when the set $\{N_k\}$ is not fixed, but generated by a probability distribution (as in the case of the Ewens formula), the expected values of the products of basic moments appearing in the l.h.s. do not coincide with the products of the expected values.

6. Large populations and large samples

When $k, l \ll N, n$ one may systematically exploit the property that, for small a and b,

$$\binom{N-a}{n-b} \to \frac{\rho^b (1-\rho)^{b-a}}{\rho^n (1-\rho)^{N-n}} \qquad\qquad \rho \equiv \frac{n}{N}.$$

Expressing N and n in terms of N_{kl} one may then obtain

$$P_{\{N_{kl}\}} \to \prod_{k=1}^{N} \Big[N_k! \prod_{l=0}^{k} \frac{1}{N_{kl}!} P_{kl}^{N_{kl}}\Big] \qquad\qquad P_{kl} \equiv \binom{k}{l} \rho^l (1-\rho)^{k-l}.$$

As shown in Appendix C, in this limit the constraint $\sum l\, n_l = n$ becomes irrelevant, and expected values of products of N_{kl} with different values of the index k factorize into products of expected values computed for each separate value of k.

We can therefore compute directly the generating function for a fixed sample size, generalizing the multivariate multinomial distribution:

$$\varepsilon^{(n)}(\{t_l\}) = \prod_{k=1}^{N} \sum_{\{N_{kl}\}} \left[N_k! \prod_{l=0}^{k} \frac{1}{N_{kl}!} (P_{kl}\, t_l)^{N_{kl}} \right] = \prod_{k=1}^{N} \Big[\sum_{l=0}^{k} P_{kl}\, t_l \Big]^{N_k}.$$

The consistency of the approximation is verified by observing that $\varepsilon^{(n)}(1) = 1$, because of the property that $\sum_{l=0}^{k} P_{kl} = 1$.

The expected value of n_l turns out to be:

$$\langle n_l \rangle = \sum_{k=1}^{N} N_k P_{kl},$$

and one may check that the conditions on $\sum_l \langle n_l \rangle$ and on $\sum_l l\, \langle n_l \rangle$ are still satisfied.

We can also calculate the variance of the frequency distribution of the samples when $k, l \ll N, n$:

$$\langle n_l^2 \rangle - \langle n_l \rangle^2 = \sum_{k=1}^{N} N_k P_{kl} (1 - P_{kl}).$$

The above expression is always smaller than $\langle n_l \rangle$ and as a consequence deviations from the mean become unimportant for sufficiently large values of $\langle n_l \rangle$.

In the same limit we may derive a notable relationship between the generating function of the original frequency distribution and the generating function of the expected values of its samples. In fact we may recognize that for sufficiently large N and n

$$g^{(n)}(z) = \sum_{p=0}^{\infty} M_p \frac{z^p}{p!} = G(z).$$

Since in general $f^{(n)}(t) = g^{(n)}\big(n(t-1)\big)$ and $F(t) = G\big(N(t-1)\big)$, it is then easy to check that in the limit under consideration

$$f^{(n)}(t) = F(1 - \rho + \rho\, t),$$

As a direct consequence of these results, whenever the (size-independent) function $\gamma(z) \equiv G(z) - G(0)$ can be cast into a form exhibiting no explicit parametric dependence on N, the expected values $\langle n_l \rangle$ can be obtained from N_k by the replacement $N \to n$.

Notice that in the limit $k, l \ll N, n$ the definition of the basic invariant moments $m_p^{(n)}$ simplifies to

$$m_p^{(n)} \to \frac{p!}{n^p} \sum_{l=p}^{n} n_l \binom{l}{p}.$$

It is worth analyzing in this limit the explicit expressions for the basic invariant moment $m_2^{(n)}$:

$$\langle m_2^{(n)} \rangle = M_2 \to \frac{1}{N^2} \sum_{k=1}^{N} k(k-1) N_k = \sum_{a=1}^{S} \big(\frac{\hat{N}_a}{N}\big)^2 - \frac{1}{N} = \sum_{a=1}^{S} \langle \big(\frac{\hat{n}_a}{n}\big)^2 \rangle - \frac{1}{n} \equiv \frac{1}{\alpha}.$$

As shown in Appendix D the above results may be used also in order to parametrize the expected value of the estimate for the correlation between samples under the assumption of independent random sampling.

7. A class of distributions and its properties

As mentioned in the Introduction, distributions found in samples may often correspond to systems whose asymptotic ($N \to \infty$) distribution is expected to obey a scaling law. However the exponent of the scaling law will in general be nontrivial, in contrast with the prediction offered by the simplest neutral models. An example of empirical and theoretical evidence for nontriviality is offered by surname frequency distributions (see Ref. [12] for a recent review), recalling that surnames are expected to mimick the behavior of selectively neutral alleles. it is therefore especially interesting to consider parametrizations that may reflect notriviality of exponents, and in particular the class of negative binomial distributions [13], which can be obtained starting from the generating function

$$F_c(t) = \frac{N}{x} \frac{(1-x)^{1-c}}{c} \left[1 - (1-xt)^c\right] = \sum_{k=1}^{\infty} \frac{N}{x} \frac{(1-x)^{1-c}}{\Gamma(1-c)} \frac{\Gamma(k-c)}{k!} (xt)^k,$$

where $0 < x < 1$ and the parameter c is assumed to vary in the range $0 \leq c < 1$.

The asymptotic behaviour of the distribution for large k is easily obtained by observing that in this limit

$$\frac{\Gamma(k-c)}{k!} \to \frac{1}{k^{1+c}}, \qquad N_k \to \frac{N}{x} \frac{(1-x)^{1-c}}{\Gamma(1-c)} \frac{x^k}{k^{1+c}}.$$

We can now compute the generating function $f_c(t)$ for the expected values of the frequencies found in the samples according to the general rule previously discussed, and obtain

$$f_c(t) = f_c(0) + \frac{n}{y} \frac{(1-y)^{1-c}}{c} \left[1 - (1-yt)^c\right],$$

where we have defined $y = \frac{\rho x}{1-x+\rho x}$.

The distribution of the samples has the same form as the original distribution, once the replacements $N \to n$ and $x \to y$ have been performed, and therefore we obtain the asymptotic behaviour

$$n_l \to \frac{n}{y} \frac{(1-y)^{1-c}}{\Gamma(1-c)} \frac{y^k}{k^{1+c}}.$$

It is possible to define a combination of parameters independent of the size of the sample:

$$\beta = N \frac{1-x}{x} = n \frac{1-y}{y}.$$

It is useful to represent x and y in a form showing explicitly their dependence on the dimension of the sample and on the invariant parameter β:

$$x = \frac{N}{\beta + N}, \qquad y = \frac{n}{\beta + n}.$$

It is now possible to evaluate the expected value of the invariant moments from the expression

$$\gamma_c(z) \equiv G_c(z) - G_c(0) = \frac{\beta}{c} \left[1 - \left(1 + \frac{z}{\beta}\right)^c\right],$$

showing no explicit parametric dependence on N; we therefore obtain (for $p \neq 0$)

$$M_p = \beta^{1-p} \frac{\Gamma(p-c)}{\Gamma(1-c)} \to \frac{\beta^{1-p}}{\Gamma(1-c)} \frac{p!}{p^{1+c}}, \qquad \frac{1}{M_2} \equiv \alpha = \frac{\beta}{1-c}.$$

The limit of the above results when $c \to 0$ is smooth, and it corresponds to the Fisher distribution [14], such that

$$F_0(t) = -\beta \ln(1-xt), \qquad N_k = \beta \frac{x^k}{k}.$$

In the same limit one obtains:

$$f_0(t) = \beta \ln\Big(\frac{\beta+N}{\beta+n}\Big) - \beta\ln(1-yt), \qquad\qquad n_l = \beta\frac{y^l}{l}.$$

The generating function of the invariant moments is obtained from $\gamma_0(z) = -\beta\ln(1+z/\beta)$, and as a consequence the expected values of the invariant moments ($p \neq 0$) are exactly $M_p = (p-1)!\,\beta^{1-p}$.

It is worth noticing the peculiar relationship $\beta = \alpha$ (where α was defined in Section 6). By comparing these results with the large θ limit of the invariant moments of the Ewens distribution we may easily check Watterson's [15] and Hubbell's [5] observation that in this limit θ is strictly connected to Fisher's α.

8. The scaling limit

Let us now consider very large systems, and assume that we can gather information only through the sampling of n objects belonging to the system, with n large but not necessarily comparable to N.

The analysis of the invariant moments may then allow us to check the applicability of a phenomenological description of the samples based on some distribution falling into the classes discussed in the previous Sections.

In the case of a positive response to the check it is then possible to find numerical estimates of the parameter β and of the exponent c. Such estimates are clearly meaningful only if β does not turn out to be significantly greater than n.

Under these assumptions, we can infer a description of the original system, and in the case $N \gg n$ such a description will correspond to computing the limit $x \to 1$ of the previous results. As a consequence, at least for observable (i.e. not too large) values of k, the original distribution is expected to be well described by the scaling form

$$N_k \to \frac{N^c\beta^{1-c}}{\Gamma(1-c)}\frac{1}{k^{1+c}}.$$

Appendix A. Expected values of the invariant moments

We consider the cumulative generating function for the expected value of a given invariant moment (setting $P = \sum p_i$) for samples of all possible sizes:

$$\sum_{n=0}^{N} \binom{N}{n}\binom{n}{P}\langle m^{(n)}_{\{p_i\}}\rangle x^n \equiv \sum_{\{q_i\}}\prod_{i=1}^{I}\Big[\binom{q_i}{p_i}\frac{\partial}{\partial t_{q_i}}\Big]E(x;\{t_l\})|_{\{t_l=1\}}.$$

We now observe that, due to the property that

$$\ln E(x;\{t_l\}) = \sum_{k=1}^{N} N_k \ln\Big[\sum_{l=0}^{k}\binom{k}{l}t_l x^l\Big],$$

the derivatives appearing in the above defined cumulative generating function, computed at $t_l = 1$, can be expressed as summations (over $\{k_i\}$ indices) of products of N_{k_i} times a universal x-dependent factor

$$\Big[\prod_{i=1}^{I}\binom{k_i}{q_i}\Big]x^Q(1+x)^{N-K},$$

where $Q = \sum_i q_i$ and $K = \sum_i k_i$. However the following identity holds:

$$\binom{q_i}{p_i}\binom{k_i}{q_i} = \binom{k_i}{p_i}\binom{k_i-p_i}{q_i-p_i}.$$

As a consequence the cumulative generating function is proportional to the factor

$$\Big[\prod_{i=1}^{I}\binom{k_i}{p_i}\Big]x^P(1+x)^{N-K}\sum_{\{q_i\}}\Big[\prod_{i=1}^{I}\binom{k_i-p_i}{q_i-p_i}x^{Q-P}\Big]=\Big[\prod_{i=1}^{I}\binom{k_i}{p_i}\Big]x^P(1+x)^{N-P}.$$

The summations over the indices $\{k_i\}$ may now be formally performed, and, by matching the coefficient of x^N in the two sides of the equation, the result can be easily recognized to coincide with the expected value of the invariant moment computed for the original distribution times the combinatorial factor $\binom{N}{P}$.

In conclusion we find

$$\sum_{n=0}^{N}\binom{N}{n}\binom{n}{P}\langle m_{\{p_i\}}^{(n)}\rangle x^n=\binom{N}{P}\langle m_{\{p_i\}}^{(N)}\rangle x^P(1+x)^{N-P}.$$

Expanding the r.h.s. in powers of x and noticing that $\binom{N}{n}\binom{n}{P}=\binom{N}{P}\binom{N-P}{n-P}$ we finally obtain

$$\sum_{n=0}^{N}\binom{N}{P}\binom{N-P}{n-P}\langle m_{\{p_i\}}^{(n)}\rangle x^n=\sum_{n=0}^{N}\binom{N}{P}\binom{N-P}{n-P}\langle m_{\{p_i\}}^{(N)}\rangle x^n,$$

implying immediately that, as long as $n\geq P$,

$$\langle m_{\{p_i\}}^{(n)}\rangle=\langle m_{\{p_i\}}^{(N)}\rangle\equiv M_{\{p_i\}}.$$

Appendix B. Proof of a combinatorial formula

The identity

$$\sum_{\{q_i\geq p_i\}}\prod_{i=1}^{I}\binom{q_i-1}{p_i-1}\frac{\Gamma(\theta+n-Q)}{\Gamma(\theta)\,(n-Q)!}=\frac{\Gamma(\theta+n)}{\Gamma(\theta+P)\,(n-P)!},$$

where $Q=\sum_i q_i$ and $P=\sum_i P_i$, can be proven by recalling that for real numbers α and positive integers $q\geq p$

$$(1-x)^{-\alpha}=\sum_{m=0}^{\infty}\frac{\Gamma(\alpha+m)}{\Gamma(\alpha)\,m!}x^m,\qquad\qquad(1-x)^{-p}=\sum_{q=p}^{\infty}\binom{q-1}{p-1}x^{q-p}.$$

Hence expanding in powers of x the two sides of the identity

$$\Big[\prod_{i=1}^{I}(1-x)^{-p_i}\Big](1-x)^{-\theta}=(1-x)^{-(\theta+P)}$$

and exchanging the order of summations in the l.h.s. we get

$$\sum_{\{q_i\geq p_i\}}\prod_{i=1}^{I}\binom{q_i-1}{p_i-1}\frac{\Gamma(\theta+m+P-Q)}{\Gamma(\theta)\,(m+P-Q)!}=\frac{\Gamma(\theta+P+m)}{\Gamma(\theta+P)\,m!}.$$

The desired result is then obtained by setting $m=n-P$.

Appendix C. Fluctuations of the sample size in the unconstrained large N limit

The multivariate generating function $\varepsilon(\{t_l\})$ was computed in Section 6 after relaxing the constraint $\sum_l l\, n_l = n$. In order to show that the constraint is automatically satisfied in the large N limit we construct a generating function for the expected value of the powers of $\nu = \sum_l l\, n_l$ in the large N distribution:

$$\eta(w) \equiv \sum_\nu P_{(\nu)} w^\nu = \sum_{\{N_{kl}\}} P_{\{N_{kl}\}} w^{\sum l\, n_l} = \prod_{k=1}^{N} \Big(\sum_{\{N_{kl}\}} N_k! \prod_{l=0}^{k} \frac{1}{N_{kl}!} (P_{kl} w^l)^{N_{kl}} \Big),$$

and applying the multinomial formula we obtain

$$\eta(w) = \prod_{k=1}^{N} (\sum_{l=0}^{k} P_{kl} w^l)^{N_k} = \prod_{k=1}^{N} (1 - \rho + \rho\, w)^{k N_k} = ((1 - \rho + \rho\, w)^N.$$

Expanding the result in powers of w we immediately obtain $P_{(\nu)} = P_{N\nu}$.
Since ν is distributed according to the binomial distribution, the relevant expected values are

$$\langle \frac{\nu}{N} \rangle = \rho \equiv \frac{n}{N}, \qquad \langle \frac{\nu^2}{N^2} \rangle - \langle \frac{\nu}{N} \rangle^2 = \frac{1}{N} \rho(1 - \rho).$$

Hence fluctuations of ν/N around ρ vanish like $1/N$ in the large N limit.

Appendix D. Correlation between samples

An important test of randomness in sampling is offered by the estimate of the correlation between two different samples. Let us consider two random samples, characterized by the sets of values $\{\hat{n}_a\}$ and $\{\hat{m}_a\}$ and by their sizes n and m. The index a labels different kinds, as in Section 2. The correlation between the two samples is

$$C = \frac{\sum_{a=1}^{S} \hat{n}_a \hat{m}_a}{\sqrt{\sum_{a=1}^{S} \hat{n}_a^2} \sqrt{\sum_{a=1}^{S} \hat{m}_a^2}}.$$

Replacing $\hat{n}_a$ and $\hat{n}_a^2$ with their expected values, computed in Section 2, we obtain (in the large N limit)

$$\sum_{a=1}^{S} \langle \hat{n}_a \rangle \langle \hat{m}_a \rangle = nm \sum_{a=1}^{S} \hat{p}_a^2,$$

$$\sum_{a=1}^{S} \langle \hat{n}_a^2 \rangle = n^2 (\sum_{a=1}^{S} \hat{p}_a^2 + \frac{1}{n} - \frac{1}{N}), \qquad \sum_{a=1}^{S} \langle \hat{m}_a^2 \rangle = m^2 (\sum_{a=1}^{S} \hat{p}_a^2 + \frac{1}{m} - \frac{1}{N}).$$

By making use of the results presented in Section 6 we can now express the expected value of the correlation between samples in the form

$$\langle C \rangle = \frac{\frac{1}{\alpha} + \frac{1}{N}}{\sqrt{\frac{1}{\alpha} + \frac{1}{n}} \sqrt{\frac{1}{\alpha} + \frac{1}{m}}}.$$

For samples of equal size n the expected value of the correlation takes the form $\langle C \rangle = \frac{n}{\alpha + n} \frac{\alpha + N}{N}$.

References

[1] M.E.J. Newman, Power laws, Pareto distributions and Zipf's law, Contemporary Physics 46 (2005) 323-351

[2] D. Zelterman, *Discrete Distributions* (Chapter 6), Wiley (2004)

[3] W. J. Ewens, The Sampling Theory of Selectively Neutral Alleles, Theoretical Population Biology 3 (1972) 87-112

[4] S. Karlin and J. McGregor, Addendum to a Paper of W. Ewens, Theoretical Population Biology 3 (1972) 113-116

[5] S. P. Hubbell *The Unified Neutral Theory of Biodiversity and Biogeography*, Princeton University Press (2001)

[6] J. Rosindell, S.P. Hubbell and R.S. Etienne, *The Unified Neutral Theory of Biodiversity and Biogeography* at Age Ten, Trends in Ecology and Evolution 26 (2011) 340-348

[7] R.C. Griffiths, S. Lessard, Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles, Theoretical Population Biology 68 (2005) 167-177

[8] R.S. Etienne, A new sampling formula for neutral biodiversity, Ecology Letters 8 (2005) 253-260

[9] S. Lessard, An Exact Sampling Formula for the Wright-Fisher model and a Solution to a Conjecture About the Finite-Island Model, Genetics 177 (2007) 1249-1254

[10] A. Lambert, Species abundance distributions in neutral models with immigration or mutation and general lifetimes, Journal of Mathematical Biology 63 (2011) 57-72

[11] N. L. Johnson, S. Kotz, N. Balakrishnan, *Discrete Multivariate Distributions* (Chapter 41), Wiley (1997)

[12] P. Rossi, Surname distribution in population genetics and in statistical physics, to appear in Physics of Life Reviews (2013)

[13] J.M. Hilbe, *Negative Binomial Regression*, Cambridge University Press (2007)

[14] R.A. Fisher, A.S. Corbet, C.B. Williams, The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, The Journal of Animal Ecology 12 (1943) 42-58

[15] G.A. Watterson, Models for the Logarithmic Species Abundance Distributions, Theoretical Population Biology 6 (1974) 217-250