Theoretical Population Biology 73 (2008) 543–551

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Faà di Bruno's formula and the distributions of random partitions in population genetics and physics

Fred M. Hoppe

Department of Mathematics and Statistics, McMaster University, 1280 Main St. W., Hamilton, Ontario, L8S 4K1, Canada

ARTICLE INFO

Article history:
Received 21 December 2007
Available online 29 March 2008

Keywords:
Composite function
Compound sampling models
Derivative
Ewens sampling formula
Faà di Bruno's formula
Fisher logarthimic series
Negative-binomial
Partition
Probability
Taylor series

ABSTRACT

We show that the formula of Faà di Bruno for the derivative of a composite function gives, in special cases, the sampling distributions in population genetics that are due to Ewens and to Pitman. The composite function is the same in each case. Other sampling distributions also arise in this way, such as those arising from Dirichlet, multivariate hypergeometric, and multinomial models, special cases of which correspond to Bose–Einstein, Fermi–Dirac, and Maxwell–Boltzmann distributions in physics. Connections are made to compound sampling models.

1. Introduction

The purpose of this paper is to point out a remarkable relationship[1] between the formula by Faà di Bruno (di Bruno, 1855) for the nth derivative of a composite function and some sampling formulas in population genetics developed more than a century later as well as well-known distributions in statistical mechanics. These formulas arise as probability distributions on partitions. We will show that di Bruno's formula reduces to these distributions in special cases, thereby providing a unifying connection among them.

Recall that a partition of n is a way of writing n as a sum of positive integers where the order of the summands does not matter. Thus if n is written as

$$n = \underbrace{1+\cdots+1}_{b_1\text{times}}\underbrace{+2+\cdots+2}_{b_2\text{times}}+\cdots+\underbrace{n}_{b_n\text{times}}$$

the corresponding partition is denoted as $b = (b_1, \ldots, b_n)$ where the notation reflects that integer i appears b_i times in the partition so that $b_1 + 2b_2 + \cdots + nb_n = n$ and $b_1 + b_2 + \cdots + b_n = k$ where k is the number of summands adding to n, that is the number of components in the partition. For instance, the integer 4 can be written as: 1+1+1+1; 1+1+2; 1+3; 2+2; 4. These correspond to the partitions: (4, 0, 0, 0); (2, 1, 0, 0); (1, 0, 1, 0); (0, 2, 0, 0); (0, 0, 0, 1) with values of k equal to 4, 3, 2, 2, 1, respectively. A random partition Π_n is a random quantity taking values in the set of all partitions of n and when n is arbitrary we denote it by Π.

One important random partition that occurs in population genetics is known as the Ewens Sampling Formula (Ewens, 1972)

$$\mathbb{P}[\Pi_n = b] = \frac{n!}{[\theta]^n}\theta^k \prod_{i=1}^{n} \frac{1}{i^{b_i} b_i!} \tag{1}$$

valid for any $\theta > 0$ where $[\theta]^n \equiv \theta(\theta+1)\cdots(\theta+n-1)$ is an ascending factorial, with $[\theta]^0 = 1$. Here a sample of size n is taken from a Poisson–Dirichlet population (Kingman, 1975; Watterson, 1976) $P = (P_1, P_2, \ldots), 0 < P_i < 1, i = 1, 2, \ldots, \sum_{i=1}^{\infty} P_i = 1$, representing the random frequencies of distinct species or alleles in a neutral population, and b_i counts the number of species represented i times in the sample. P can be represented in the form of a residual allocation model (Griffiths, 1980) in size-biased form $(P_{(1)}, P_{(2)}, \ldots)$ using an urn model (Hoppe, 1984, 1987)

$$\begin{cases} P_{(1)} = Z_1 \\ P_{(n)} = Z_n \prod_{i=1}^{n-1}(1 - Z_i), \quad n \geq 2 \end{cases} \tag{2}$$

where the $\{Z_i\}$ are independent identically distributed Beta$(1, \theta)$ random variables. The model (2) is known as the GEM distribution in genetics (Ewens, 1990).

E-mail address: hoppe@mcmaster.ca.

[1] A preliminary version of these ideas was presented at the Threads Colloquium, McMaster University, Sept. 15, 2006.

0040-5809/$ – see front matter
doi:10.1016/j.tpb.2008.03.003

A sampling formula due to Pitman (1992) also arises when sampling from (2) where the $\{Z_i\}$ are still independent, however not identically distributed, but rather as Beta$(1-\alpha, \theta+i\alpha)$ random variables, respectively, for some $0 < \alpha < 1, \theta + \alpha > 0$. The corresponding partition distribution is

$$\mathbb{P}[\Pi_n = b] = \frac{n!}{[\theta]^n} \prod_{j=0}^{k-1}(\theta + j\alpha) \prod_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i} b_i!}. \tag{3}$$

There is another parameter range for which (3) defines a distribution: $\alpha = -r, \theta = Mr$ where M is a positive integer and $r > 0$. Now the $\{Z_i\}$ are independent Beta$(1 + \frac{\theta}{M}, \frac{(M-i)\theta}{M})$ random variables and the residual allocation model has a finite number of types with

$$\begin{cases} P_{(1)} = Z_1 \\ P_{(n)} = Z_n \prod_{i=1}^{n-1}(1 - Z_i), \quad 2 \le n \le M-1 \\ P_{(M)} = 1 - P_{(1)} - \cdots - P_{(M-1)}. \end{cases} \tag{4}$$

The corresponding sampling formula (3) is then more commonly expressed as

$$\mathbb{P}[\Pi_n = b] = \frac{n![M]_k}{[\theta]^n} \prod_{i=1}^{n} \left(\frac{[\theta/M]^i}{i!} \right)^{b_i} \frac{1}{b_i!} \tag{5}$$

where $[M]_k = M(M-1)\ldots(M-k+1)$ is the descending factorial k times starting at M. Notice that if $k > M$ then $[M]_k = 0$ implying that this distribution concentrates on partitions with at most M parts. In genetic terms, the partitions involve at most M distinct alleles, with probability one. The model (4) describes a size-biased relabelling of a symmetric Dirichlet distribution $P = (P_1, \ldots, P_M)$, based on the order in which the different types enter a sample (Hoppe, 1987, page 132). The equivalent sampling formula (5) shows the multinomial roots of this case.

We now state the connection between di Bruno's formula and these partitions. Let $g(x), f(x)$ be suitably differentiable functions and consider the composite function $h(x) = g(f(x))$. Di Bruno's formula (di Bruno, 1855) states that the nth derivative of $h(x)$ is

$$h^{(n)}(x) = \sum \frac{n!}{b_1! \ldots b_n!} g^{(k)}(f(x)) \prod_{i=1}^{n} \frac{(f^{(i)}(x))^{b_i}}{i!^{b_i}} \tag{6}$$

where the sum is over all partitions $b = (b_1, \ldots, b_n)$ of the integer n and k is the number of parts in the partition.

All three three sampling formulas (1), (3) and (5) bear a striking resemblance to (6), yet a search of the relevant literature did not turn up the very attractive relationship between these two topics. Di Bruno's formula was mentioned, but only in the context of differentiation or composition of exponential generating functions (Pitman, 2006). We show in this paper that (6) reduces to these three partitions in special cases, as well as others that occur in physics. We also provide a probabilistic explanation using compounding processes.

For Ewens' sampling formula it is appropriate to take

$$g(x) = \mathrm{e}^{\theta x} \qquad f(x) = -\log(1-x).$$

Then as shown below, (6) simplifies to an identity comprising positive terms which can be identified as a probability distribution given by (1). The same occurs with respect to (3), for instance when $\theta > 0$, using

$$g(x) = \frac{1}{(\alpha(1-x))^{\frac{\theta}{\alpha}}} \qquad f(x) = 1 - \frac{(1-x)^{\alpha}}{\alpha}$$

and also for (5) with

$$g(x) = x^M \quad \text{and} \quad f(x) = \frac{1}{(1-x)^{\theta/M}}.$$

Remarkably, the composite function

$$h(x) = \frac{1}{(1-x)^{\theta}}$$

is the *same* in all three cases, meaning that different expansions of the same function yield these different sampling formulas.

Interestingly, for (3) the choice of pair (g, f) depends on whether $-\alpha < \theta < 0, \theta = 0$, or $\theta > 0$, which may shed some light on the distinction or origin of these partitions in specific applications. Moreover, the distributions for sampling with replacement (multinomial), or sampling without replacement (multivariate hypergeometric) from a finite population can also be obtained from (6). Together with (5), special cases of these are the familiar Maxwell–Boltzman, Bose–Einstein, and Fermi–Dirac distributions of statistical mechanics in physics.

In this paper we make explicit, and explore, this fascinating connection. On the one hand, we have probability distributions arising from a probabilistic sampling process from a population. On the other hand, we have an expansion of a composite function that gives a probability distribution over partitions of n that has this nice probabilistic interpretation. The similarity between di Bruno's formula and partitions in biology suggests that techniques and results from one area may be of value in studying the other.

2. Ewens' sampling formula

For arbitrary $\theta > 0$ let

$$g(x) = \mathrm{e}^{\theta x} \quad \text{and} \quad f(x) = -\log(1-x)$$

so that

$$g^{(k)}(x) = \theta^k \mathrm{e}^{\theta x}, \quad k = 0, 1, \ldots \quad \text{and}$$
$$f^{(i)}(x) = \frac{(i-1)!}{(1-x)^i}, \quad i = 1, 2, \ldots.$$

Hence

$$h(x) \equiv g(f(x)) = \frac{1}{(1-x)^{\theta}} \quad \text{and} \quad h^{(n)}(x) = \frac{[\theta]^n}{(1-x)^{\theta+n}},$$
$$n = 0, 1, \ldots.$$

These formulas, substituted into (6), lead to

$$\frac{[\theta]^n}{(1-x)^{\theta+n}} = \sum n! \frac{\theta^k}{(1-x)^{\theta}} \prod_{i=1}^{n} \left(\frac{(i-1)!}{(1-x)^i} \right)^{b_i} \frac{1}{i!^{b_i} b_i!}. \tag{7}$$

Observe that $\frac{(i-1)!}{i!} = \frac{1}{i}$ and $\prod_{i=1}^{n} \left(\frac{1}{(1-x)^i} \right)^{b_i} = \frac{1}{(1-x)^{\sum i b_i}} = \frac{1}{(1-x)^n}$, which shows that there is a common factor $\frac{1}{(1-x)^{\theta+n}}$ involving x on both sides of (7) that can be cancelled, leaving the identity

$$[\theta]^n = \sum n! \theta^k \prod_{i=1}^{n} \frac{1}{i^{b_i} b_i!}.$$

We divide both sides by $[\theta]^n$ giving

$$1 = \sum \frac{n!}{[\theta]^n} \theta^k \prod_{i=1}^{n} \frac{1}{i^{b_i} b_i!} \tag{8}$$

where the sum is over all partitions $b = (b_1, \ldots, b_n)$ of the integer n and $k \equiv b_1 + \cdots + b_n$ is the number of terms in the partition, which identifies each term on the right side of (8) as describing a probability distribution over partitions of n

$$\mathbb{P}[\Pi_n = b] = \frac{n!}{[\theta]^n} \theta^k \prod_{i=1}^{n} \frac{1}{i^{b_i} b_i!}.$$

This is (1) and therefore we obtain Ewens' Sampling Formula in a purely analytical manner as the partition that arises from an identity obtained by di Bruno's formula, although not in its sampling probabilistic context.

The coefficient of θ^k in the expansion of $[\theta]^n$ is the absolute value of a Stirling number of the first kind, denoted as $|S_n^k|$. From (8) we can read off

$$|S_n^k| = \sum_{b:b_1+\cdots+b_n=k} n! \prod_{i=1}^{n} \frac{1}{i^{b_i} b_i!}$$

a formula going back to Cauchy (Shepp and Lloyd (1966) and Kaucký (1985)). According to Ewens (2004) the distribution of the random number of terms K_n in the partition described by (1) is given by

$$\mathbb{P}[K_n = k] = |S_n^k| \frac{\theta^k}{[\theta]^n}$$

which thus also follows directly from the identity (8).

To summarize, we have shown that the expansion in powers of x of the function $h(x) = \frac{1}{(1-x)^\theta}$ when expressed as a composite function to which di Bruno's formula is applied leads to an identity that is the distribution (1). We next turn to different pairs of functions which give the other sampling formulas mentioned.

3. Pitman's formula

3.1. $0 < \alpha < 1, \theta > 0$

For $0 < \alpha < 1, \theta > 0$ consider the pair of functions

$$g(x) = \frac{1}{(\alpha(1-x))^{\frac{\theta}{\alpha}}} \quad \text{and} \quad f(x) = 1 - \frac{(1-x)^\alpha}{\alpha} \tag{9}$$

so that

$$g^{(k)}(x) = [\theta/\alpha]^k \frac{1}{\alpha^{\theta/\alpha}} (1-x)^{\frac{\theta}{\alpha}} \quad \text{and} \quad f^{(i)}(x) = \frac{[1-\alpha]^{i-1}}{(1-x)^{i-\alpha}}.$$

The parameter α cancels out in the composite function giving the same h as in Ewens' case

$$h(x) \equiv g(f(x)) = \frac{1}{(1-x)^\theta}.$$

Now (6) becomes

$$\begin{aligned}
\frac{[\theta]^n}{(1-x)^{\theta+n}} &= \sum n! [\theta/\alpha]^k \frac{1}{\alpha^{\theta/\alpha}(1-f(x))^{\frac{\theta}{\alpha}+k}} \\
&\quad \times \prod_{i=1}^{n} \left(\frac{[1-\alpha]^{i-1}}{(1-x)^{i-\alpha}} \right)^{b_i} \frac{1}{i!^{b_i} b_i!} \\
&= \sum n! [\theta/\alpha]^k \frac{1}{\alpha^{\theta/\alpha} (\frac{(1-x)^\alpha}{\alpha})^{\frac{\theta}{\alpha}+k}} \\
&\quad \times \prod_{i=1}^{n} \left(\frac{[1-\alpha]^{i-1}}{i!} \right)^{b_i} \left(\frac{1}{(1-x)^{i-\alpha}} \right)^{b_i} \frac{1}{b_i!} \\
&= \sum n! [\theta/\alpha]^k \frac{\alpha^k}{(1-x)^{\theta+\alpha k}} \\
&\quad \times \prod_{i=1}^{n} \left(\frac{[1-\alpha]^{i-1}}{i!} \right)^{b_i} \frac{1}{b_i!} \frac{1}{(1-x)^{n-\alpha k}}.
\end{aligned}$$

The factor $(1-x)^{\theta+n}$ cancels from the denominator of both sides and $[\theta/\alpha]^k \alpha^k$ reduces to $\prod_{j=0}^{k-1}(\theta + j\alpha)$, leaving

$$[\theta]^n = n! \sum \prod_{j=0}^{k-1} (\theta + j\alpha) \prod_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i} b_i!}. \tag{10}$$

Finally, as before, we divide both sides by $[\theta]^n$ to produce the equivalent identity

$$1 = \frac{n!}{[\theta]^n} \sum \prod_{j=0}^{k-1} (\theta + j\alpha) \prod_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i} b_i!}. \tag{11}$$

Again, this identifies each term on the right side of (11) as describing a probability distribution, namely (3). Although $f(x)$ is not defined for $\alpha = 0$, nonetheless, the partition distribution described by (11) does makes sense if we let $\alpha \to 0$ whereby Ewens formula arises in this limit.

3.2. $0 < \alpha < 1, -\alpha < \theta < 0$

For $0 < \beta < 1, 0 < \alpha < 1$ let

$$g(x) = 1 - (1-x)^\beta \quad \text{and} \quad f(x) = 1 - (1-x)^\alpha$$

so that

$$g^{(k)}(x) = \frac{\beta[1-\beta]^{k-1}}{(1-x)^{k-\beta}} \quad \text{and} \quad f^{(i)}(x) = \frac{\alpha[1-\alpha]^{i-1}}{(1-x)^{i-\alpha}}$$

$$h(x) = 1 - (1-x)^{\alpha\beta} \quad \text{and} \quad h^{(n)}(x) = \frac{\alpha\beta[1-\alpha\beta]^{n-1}}{(1-x)^{n-\alpha\beta}}$$

Di Bruno's formula becomes

$$\begin{aligned}
\frac{\alpha\beta[1-\alpha\beta]^{n-1}}{(1-x)^{n-\alpha\beta}} &= \sum n! \frac{\beta[1-\beta]^{k-1}}{(1-x)^{\alpha k-\alpha\beta}} \\
&\quad \times \prod_{i=1}^{n} \left(\frac{\alpha[1-\alpha]^{i-1}}{(1-x)^{i-\alpha}} \right)^{b_i} \frac{1}{i!^{b_i} b_i!} \\
&= \sum n! \frac{\beta[1-\beta]^{k-1}}{(1-x)^{\alpha k-\alpha\beta}} \alpha^k \\
&\quad \times \prod_{i=1}^{n} \left(\frac{[1-\alpha]^{i-1}}{(1-x)^{i-\alpha}} \right)^{b_i} \frac{1}{i!^{b_i} b_i!}
\end{aligned}$$

which simplifies (by canceling a common factor $\frac{\alpha\beta}{(1-x)^{n-\alpha\beta}}$ on both sides and absorbing the remaining α^{k-1} into the ascending factorial $[1-\beta]^{k-1}$) to

$$[1-\alpha\beta]^{n-1} = \sum n! \prod_{j=1}^{k-1} (j\alpha - \alpha\beta) \prod_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i} b_i!}. \tag{12}$$

Multiplication of both sides of (12) by $-\alpha\beta$ and setting $\theta = -\alpha\beta$ leads to

$$[\theta]^n = n! \sum \prod_{j=0}^{k-1} (\theta + j\alpha) \prod_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i} b_i!} \tag{13}$$

which is (10) except that θ is negative. However, the only negative factors in (13) are θ on each side, corresponding to the first term in each ascending factorial. Although $\theta < 0$, when we divide both sides of (12) by $[\theta]^n$ the resulting (11) still defines a bonafide probability distribution on partitions of n since θ cancels in the numerator and denominator, leaving all terms positive.

3.3. $0 < \alpha < 1, \theta = 0$

The remaining case of Pitman's formula is $0 < \alpha < 1, \theta = 0$ for which a third pair of functions (g, f) is required. Let

$$g(x) = -\log(1-x) \quad \text{and} \quad f(x) = 1 - (1-x)^\alpha$$

and then

$$g^{(k)}(x) = \frac{(k-1)!}{(1-x)^k} \quad \text{and} \quad f^{(i)}(x) = \frac{\alpha[1-\alpha]^{i-1}}{(1-x)^{i-\alpha}}$$

$$h(x) = -\alpha \log(1-x) \quad \text{and} \quad h^{(n)}(x) = -\alpha \frac{(n-1)!}{(1-x)^n}.$$

Di Bruno's formula gives

$$\frac{\alpha(n-1)!}{(1-x)^n} = \sum n! \frac{(k-1)!}{(1-x)^{\alpha k}} \prod_{i=1}^{n} \left(\frac{\alpha[1-\alpha]^{i-1}}{(1-x)^{i-\alpha}} \right)^{b_i} \frac{1}{i!^{b_i} b_i!}$$

which simplifies to

$$1 = n\sum\prod_{j=1}^{k-1}(j\alpha)\prod_{i=1}^{n}\frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i}b_i!}$$

which is (11) for $\theta = 0$ (there is a common factor θ in numerator and denominator that cancels before setting θ to 0 in (11)).

4. Symmetric dirichlet partitions

For $\theta > 0$ and integer $M \geq 1$, let

$$g(x) = x^M \quad \text{and} \quad f(x) = \frac{1}{(1-x)^{\theta/M}}$$

so that

$$g^{(k)}(x) = [M]_k x^{M-k}, \quad k = 0, 1, \ldots, M \quad \text{and}$$
$$f^{(i)}(x) = \frac{[\theta/M]^i}{(1-x)^{\theta/M+i}}, \quad i = 0, 1, \ldots$$

where $[M]_k = M(M-1)\ldots(M-k+1)$ is the descending factorial k times starting at M. As in the two previous cases h again has the form

$$h(x) \equiv g(f(x)) = \frac{1}{(1-x)^{\theta}}.$$

Di Bruno's formula (6) becomes

$$\frac{[\theta]^n}{(1-x)^{\theta+n}} = \sum n![M]_k \frac{1}{\left((1-x)^{\frac{\theta}{M}}\right)^{M-k}} \times \prod_{i=1}^{n}\left(\frac{[\theta/M]^i}{(1-x)^{\frac{\theta}{M}+i}}\right)^{b_i}\frac{1}{i!^{b_i}b_i!}$$
$$= \sum \frac{n![M]_k}{(1-x)^{\theta-\frac{k\theta}{M}}}\prod_{i=1}^{n}\left(\frac{[\theta/M]^i}{i!}\right)^{b_i}\frac{1}{b_i!}\frac{1}{(1-x)^{\frac{k\theta}{M}+n}}.$$

Once more, the factor involving x can be eliminated after which we may divide both sides by $[\theta]^n$, leaving the identity

$$1 = \frac{n!}{[\theta]^n}\sum[M]_k\prod_{i=1}^{n}\left(\frac{[\theta/M]^i}{i!}\right)^{b_i}\frac{1}{b_i!} \tag{14}$$

which identifies another random partition Π_n of n. The individual terms in the sum of this identity determine the probability distribution (5).

4.1. Probabilistic interpretation I

We will rewrite (5) in order to interpret it as a more familiar object.

$$\mathbb{P}[\Pi_n = b] = n!\frac{M!}{(M-k)!}\frac{1}{\frac{(\theta+n-1)!}{(\theta-1)!}}\frac{1}{\prod_{i=1}^{n}b_i!}\prod_{i=1}^{n}\binom{\theta/M+i-1}{\theta/M-1}^{b_i}$$
$$= \frac{M!}{k!(M-k)!}\frac{1}{\frac{(\theta+n-1)!}{n!(\theta-1)!}}\frac{k!}{\prod_{i=1}^{n}b_i!}\prod_{i=1}^{n}\binom{\theta/M+i-1}{\theta/M-1}^{b_i}$$
$$= \frac{\binom{M}{k}\binom{k}{b_1,\ldots,b_n}\prod_{i=1}^{n}\binom{\theta/M+i-1}{\theta/M-1}^{b_i}}{\binom{\theta+n-1}{\theta-1}} \tag{15}$$

where $\binom{k}{b_1,\ldots,b_n} = \frac{k!}{\prod_{i=1}^{n}b_i!}$ is a multinomial coefficient. Such a partition arises from a random sample of size n taken from a symmetric Dirichlet population $P = (P_1, \ldots, P_M)$ described by the density

$$d(p_1, \ldots, p_{M-1}) = \frac{\Gamma(\theta)}{\Gamma(\theta/M)^M}\prod_{i=1}^{M-1}p_i^{\theta/M-1}p_M^{\theta/M-1},$$
$$0 \leq p_i \leq 1, p_1 + \cdots + p_{M-1} \leq 1$$

where $p_M = 1 - p_1 - p_2 - \cdots - p_{M-1}$.

To see that (15) results from such a sample, consider a random sample of size n taken from a population given by (5), let $X_j = \#\{j : \text{type } j \text{ is in the sample}\}$, and let $x = (x_1, \ldots, x_M)$ be a set of (ordered) occupancy numbers. Then the distribution of $X = (X_1, \ldots, X_M)$ is given by the expectation of a multinomial with random probabilities

$$\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M]$$
$$= \mathbb{E}\left[\binom{n}{x_1,\ldots,x_M}\prod_{j=1}^{M}P_j^{x_j}\right]$$
$$= \binom{n}{x_1,\ldots,x_M}\int\cdots\int\prod_{j=1}^{M-1}p_j^{x_j}p_M^{x_M}\frac{\Gamma(\theta)}{\Gamma(\theta/M)^M} \times \prod_{j=1}^{M-1}p_j^{\theta/M-1}p_M^{\theta/M-1}dp_1\ldots dp_{M-1}$$
$$= \binom{n}{x_1,\ldots,x_M}\int\cdots\int\prod_{j=1}^{M-1}p_j^{x_j+\theta/M-1}p_M^{x_M+\theta-1} \times \frac{\Gamma(\theta)}{\Gamma(\theta/M)^M}\mathrm{d}p_1\ldots\mathrm{d}p_{M-1}$$

where the integral is over the simplex $\{0 \leq p_i \leq 1, 1 \leq i \leq M-1, p_1 + \cdots p_{M-1} \leq 1\}$. Therefore

$$\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M] = \binom{n}{x_1,\ldots,x_M}\frac{\Gamma(\theta)}{\Gamma(\theta/M)^M} \times \frac{\prod_{j=1}^{M}\Gamma(\theta/M+x_j)}{\Gamma(\theta+n)}.$$

For a partition b of the integer n, consider all sets of occupancy numbers $\{x_1, \ldots, x_M\}$ satisfying $b_i = \#\{j : x_j = i\}$, $1 \leq i \leq n$, so that b_i counts the number of times integer i appears among the $\{x_j\}$. The distribution of X then induces a probability distribution on the set of all partitions b of the integer n. Let Π_n represent the corresponding random partition. We compute the probability distribution of Π_n by counting how many samples result in a specified $b = (b_1, \ldots, b_n)$. Suppose that b has k parts. First select which of the M types will be represented in such a sample. There are $\binom{M}{k}$ choices. Next decide which of the k types will be represented b_1 times, b_2 times, $\ldots$, b_n times. Since $b_1 + \cdots + b_n = k$ there are $\binom{k}{b_1,\ldots,b_n}$ selections. This counts the number of samples with unordered occupancy numbers $\{x_1, \ldots, x_M\}$ and since each gives the same partition b it follows that

$$\mathbb{P}[\Pi_n = b] = \binom{M}{k}\binom{k}{b_1,\ldots,b_n}\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M]$$
$$= \binom{M}{k}\binom{k}{b_1,\ldots,b_n}\binom{n}{x_1,\ldots,x_M}\frac{\Gamma(\theta)}{\Gamma(\theta/M)^M} \times \frac{\prod_{j=1}^{M}\Gamma(\theta/M+x_j)}{\Gamma(\theta+n)}$$
$$= \binom{M}{k}\binom{k}{b_1,\ldots,b_n}\frac{n!}{\prod_{j=1}^{M}x_j!}\frac{\Gamma(\theta)}{\Gamma(\theta/M)^M} \times \frac{\prod_{j=1}^{M}\Gamma(\theta/M+x_j)}{\Gamma(\theta+n)}.$$

But $\prod_{j=1}^{M} x_j! = \prod_{i=1}^{n} i!^{b_i}$, $\prod_{j=1}^{M} \Gamma(\theta/M + x_j) = \prod_{i=1}^{n} \Gamma(\theta/M + i)^{b_i}$, and $\frac{\Gamma(\theta/M+i)}{i!\Gamma(\theta/M)} = \binom{\theta/M+i-1}{\theta/M-1}$ which recovers (15) and shows that another expansion of the same composite function $h(x) = \frac{1}{(1-x)^\theta}$ now produces the sampling formula from a symmetric Dirichlet.

In the special case $\theta = M$, (5) simplifies to the partition from the familiar uniform or Bose–Einstein distribution.

4.2. Probabilistic interpretation II

Let

$$G(x) = x^M \quad \text{and} \quad F(x) = \frac{(1-\phi)^{\theta/M}}{(1-\phi x)^{\theta/M}}$$

where $0 < \phi < 1$, with composite function

$$H(x) = G(F(x)) = \frac{(1-\phi)^\theta}{(1-\phi x)^\theta}.$$

Di Bruno's formula applied to the pair (G, F) still leads to the partition (5).

G and F are probability generating functions although G corresponds to a degenerate random variable concentrated at the value M. Thus consider a population comprised of a fixed number M of species having X_j individuals of species j or, equivalently, sample a random number X_j of species j where the X_j are independent identically distributed random variables with p.g.f. $F(x)$ and probability distribution

$$\mathbb{P}[X = i] = \binom{\theta/M + i - 1}{\theta/M - 1} \phi^i (1-\phi)^{\theta/M}.$$

The sum $N = \sum_{j=1}^{M} X_j$ has p.g.f. H and (negative-binomial) distribution

$$\mathbb{P}[N = n] = \binom{\theta + n - 1}{\theta - 1} \phi^n (1-\phi)^\theta.$$

The conditional distribution of the X_j given N is well-known and given by

$$\begin{aligned}\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M | N = n] &= \frac{\prod_{j=1}^{M} \binom{\theta/M+x_j-1}{\theta/M-1} \phi^{x_j} (1-\phi)^{\theta/M}}{\binom{\theta+n-1}{\theta-1} \phi^n (1-\phi)^\theta} \\ &= \frac{n!\Gamma(\theta)}{\Gamma(\theta+n)} \prod_{j=1}^{M} \frac{\Gamma(\theta/M + x_j)}{x_j!\Gamma(\theta/M)}\end{aligned}$$

and now follow the argument round (15).

5. Symmetric multivariate hypergeometric

For integers $M \geq 1$, $C \geq 1$ let

$$g(x) = x^M \quad \text{and} \quad f(x) = x^C$$

so that $h(x) = x^{MC}$. This gives

$$g^{(k)}(x) = \begin{cases} [M]_k x^{M-k} & \text{if } k \leq M \\ 0 & \text{if } k > M \end{cases}$$

$$f^{(i)}(x) = \begin{cases} [C]_i x^{C-i} & \text{if } i \leq C \\ 0 & \text{if } i > C \end{cases}$$

and (6) becomes

$$[MC]_n x^{MC-n} = \sum n! [M]_k x^{MC-Mk} \prod_{i=1}^{n} \left(\frac{[C]_i x^{C-i}}{i!} \right)^{b_i} \frac{1}{b_i!}$$

which reduces to

$$1 = \frac{n!}{[MC]_n} \sum [M]_k \prod_{i=1}^{n} \frac{([C]_i)^{b_i}}{i!^{b_i} b_i!}$$

and the corresponding partition distribution is

$$\mathbb{P}[\Pi_n = b] = \frac{n!}{[MC]_n} [M]_k \prod_{i=1}^{n} \frac{([C]_i)^{b_i}}{i!^{b_i} b_i!}. \tag{16}$$

If $i > C$ the product on the right side of (16) is zero unless the corresponding $b_i = 0$. As a result, the only partitions b that contribute to the sum in (16) must be restricted to be of the form $b = (b_1, b_2, \ldots, b_C, 0, 0, \ldots, 0)$. Since b_i is the number of times that i is represented in the partition, this means that any integer can be represented at most C times. Similarly, if a partition has $k > M$ parts then $[M]_k = 0$, meaning this partition does not contribute to the sum and as a result $b_1 + b_2 + \cdots + b_n \equiv b_1 + b_2 + \cdots + b_C \leq M$ so the probability distribution (16) concentrates on partitions with at most M parts.

5.1. Probabilistic interpretation I

Consider a finite population of size MC comprised of M subpopulations each of size C, or equivalently, consider M boxes labelled $1, 2, \ldots, M$ each filled with C balls. Take a simple random sample of size n without replacement from this population and let $X \equiv (X_1, \ldots, X_M)$ represent the ordered occupancy numbers. The distribution of X is given by

$$\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M] = \frac{\prod_{j=1}^{M} \binom{C}{x_j}}{\binom{MC}{n}}. \tag{17}$$

Let Π_n denote the random variable describing the partition induced by X. By symmetry, each sample of size n with the same partition b has the same probability. Consider then a partition with k parts (in this case representative subpopulations). Each sample with k parts can be obtained by first selecting the k subpopulations to be involved, in $\binom{M}{k}$ ways, then determining the number of individuals to be chosen from each of the k subpopulations, in $\binom{k}{b_1, \ldots, b_n}$ ways, and finally, by selecting the individuals from each subpopulation, in $\prod_{i=1}^{n} \binom{C}{i}^{b_i}$ ways. As there are altogether $\binom{MC}{n}$ ways of selecting n objects from MC this results in

$$\mathbb{P}[\Pi_n = b] = \frac{\binom{M}{k} \binom{k}{b} \prod_{i=1}^{n} \binom{C}{i}^{b_i}}{\binom{MC}{n}}. \tag{18}$$

The right hand sides of (16) and (18) can be shown to be the same and thus di Bruno's formula for the choice $g(x) = x^M, f(x) = x^C$ leads to the partition distribution obtained from a sample without replacement from a finite population of size MC that is comprised of M subpopulations each of size C. The random partition Π_n counts how many subpopulations are represented b times, without regard to which ones they are.

Kingman (1980) has cited an example by Watterson of the partition where

$$b_1 = n, \quad b_2 = b_3 = \cdots = b_n = 0 \quad \text{with probability 1} \tag{19}$$

which cannot be expressed as a certain mixture. It is easily checked that (19) is a special case of (16) where $C = 1$ and our analysis thus positions Watterson's example as arising from sampling without replacement, a connection that is obscured when $C = 1$.

In the special case $C = 1$, (5) also simplifies to the partition from the familiar Fermi–Dirac distribution.

5.2. Probabilistic interpretation II

It is not hard to check that di Bruno's formula applied to the functions

$$G(x) = x^M \quad \text{and} \quad F(x) = (1-\phi+\phi x)^C$$

where $0 < \phi \leq 1$ also leads to (16) because the dependence on ϕ cancels from both sides of (6). This suggests another probabilistic interpretation. Again consider a finite population of size MC comprised of M subpopulation each of size C but now instead of taking a simple random sample of size n from the overall population, sample individually a random number of individuals from each subpopulation according to a binomial distribution with C trials and success probability ϕ. (Note. This is equivalent to selecting each of the MC individuals independently with probability ϕ. The difference between this sampling and that in the previous subsection is that the size of the sample was fixed at n earlier but is random here.)

In place of (17) consider the conditional probability

$$\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M | N = n] = \frac{\prod_{j=1}^{M} \binom{C}{x_j} \phi^{x_j}(1-\phi)^{C-x_j}}{\binom{MC}{n} \phi^n (1-\phi)^{MC-n}} = \frac{\prod_{j=1}^{M} \binom{C}{x_j}}{\binom{MC}{n}}$$

where $N = \sum_{j=1}^{M} X_j$ which is the same probability distribution as (17) but now conditional on the total sample size, giving another probabilistic model whose partition distribution is (16).

6. Symmetric multinomial

In this last example, $M \geq 1$ is an integer and $\theta > 0$. Let

$$g(x) = x^M \quad \text{and} \quad f(x) = \mathrm{e}^{\theta x}$$

resulting in $h(x) = \mathrm{e}^{M\theta x}$. This gives

$$g^{(k)}(x) = \begin{cases} [M]_k x^{M-k} & \text{if } 1 \leq k \leq M \\ 0 & \text{if } k > M \end{cases}$$

$$f^{(i)}(x) = \theta^i \mathrm{e}^{\theta x} \quad \text{if } i = 1, 2, \ldots$$

$$h^{(n)}(x) = (M\theta)^n \mathrm{e}^{M\theta x} \quad \text{if } n = 1, 2, \ldots$$

and (6) becomes

$$(M\theta)^n \mathrm{e}^{M\theta x} = \sum n! [M]_k \mathrm{e}^{(M\theta - k\theta)x} \prod_{i=1}^{n} \left(\frac{\theta^i \mathrm{e}^{\theta x}}{i!} \right)^{b_i} \frac{1}{b_i!}$$

which simplifies to

$$1 = \frac{n!}{M^n} \sum [M]_k \prod_{i=1}^{n} \frac{1}{i!^{b_i} b_i!}$$

and the corresponding partition distribution is

$$\mathbb{P}[\Pi_n = b] = \frac{n!}{M^n} [M]_k \prod_{i=1}^{n} \frac{1}{i!^{b_i} b_i!}. \tag{20}$$

As for the case of sampling from a symmetric Dirichlet population, if $k > M$ then $[M]_k = 0$ so the distribution concentrates on partitions with at most M parts.

6.1. Probabilistic interpretation I

For a probabilistic interpretation of (20) suppose that a sample $X \equiv (X_1, \ldots, X_M)$ of size n is taken from a symmetric multinomial population $P \equiv (P_1, \ldots, P_M)$ where P is multinomial with M cells and equally likely success probability vector $p \equiv (p_1, \ldots, p_M) = (\frac{1}{M}, \ldots, \frac{1}{M})$. Then

$$\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M] = \binom{n}{x_1, \ldots, x_M} \prod_{j=1}^{M} \left(\frac{1}{M} \right)^{x_j} = \binom{n}{x_1, \ldots, x_M} \frac{1}{M^n} \tag{21}$$

and so for a partition b with occupancy numbers $x_1, \ldots, x_M$ where $b_1 + \cdots + b_n = k$

$$\mathbb{P}[\Pi_n = b] = \binom{M}{k} \binom{k}{b_1 \ldots b_n} \mathbb{P}[X_1 = x_1, \ldots, X_M = x_M] = \frac{[M]_k}{k!} \frac{k!}{\prod_{i=1}^{n} b_i!} \frac{n!}{\prod_{j=1}^{M} x_j!} \frac{1}{M^n} = \frac{n!}{M^n} [M]_k \prod_{i=1}^{n} \frac{1}{i!^{b_i} b_i!} \tag{22}$$

which shows that (20) arises from sampling a symmetric multinomial population or the familiar Maxwell–Boltzman distribution.

6.2. Probabilistic interpretation II

Apply di Bruno's formula to the functions

$$G(x) = x^M \quad \text{and} \quad F(x) = \mathrm{e}^{\theta(x-1)}$$

to also obtain (20). Consider an infinite population with M types and take a Poisson, mean θ, number of individuals X_j independently from each type. With $N = \sum_{j=1}^{M} X_j$ consider

$$\mathbb{P}[X_1 = x_1, \ldots, X_M = x_M | N = n] = \prod_{j=1}^{M} \frac{\frac{\mathrm{e}^{-\theta} \theta^{x_j}}{x_j!}}{\frac{\mathrm{e}^{-\theta M} (\theta M)^n}{n!}} = \frac{1}{M^n} \frac{1}{\prod_{j=1}^{M} x_j!}$$

which is the same as (21) and as in (22) leads to

$$\mathbb{P}[\Pi = b | N = n] = \frac{n!}{M^n} [M]_k \prod_{i=1}^{n} \frac{1}{i!^{b_i} b_i!}$$

which is the same probability distribution as (20) but now conditional on the total sample size.

7. Compound sampling

The choices of functions (g, f) are not unique to the partitions they generate. The pairs for the Ewens, Pitman ($\theta > 0$), and symmetric Dirichlet partitions in Sections 2–4 were chosen for simplicity of their composite functions and because they produce the same composite function $h(x) = \frac{1}{(1-x)^\theta}$ in all cases. This function can be differentiated directly to reveal its nth derivative without recourse to di Bruno's formula which therefore provides three different expansions of the *same* function, each recognizable as a distinct partition distribution.

Each function is a power series with non-negative coefficients and in the second of the probabilistic interpretations in Sections 4–6, g and f were modified, maintaining their functional structure, into probability generating functions. The corresponding partitions produced by di Bruno's formula were interpreted as the partitions that arise when a random sample is taken from a population comprised of a finite number M of types, whereby independent

identically distributed numbers of individuals $X_1, \ldots, X_M$ are sampled from each type.

A similar interpretation can be made for the formulas in Sections 2 and 3, except that the number of types will be infinite and M replaced by a random variable before the X_j are determined. For instance, in the derivation of the identity yielding Ewens' formula, replace $g(x) = e^{\theta x}$ and $f(x) = -\log(1-x)$ with

$$G(x) = e^{\lambda(x-1)} \quad \text{and} \quad F(x) = \frac{c - \log(1-\phi x)}{c - \log(1-\phi)} \tag{23}$$

respectively, where $0 < \phi < 1, c \geq 0$. G is the p.g.f. of a Poisson random variable with mean λ and F is the p.g.f of a random variable X with distribution

$$\mathbb{P}[X = i] = \begin{cases} \dfrac{c}{c - \log(1-\phi)} & i = 0 \\ \dfrac{1}{c - \log(1-\phi)} \dfrac{\phi^i}{i} & i = 1, 2, \ldots. \end{cases} \tag{24}$$

The special case $c = 0$ when there is no mass at the origin is Fisher's logarithmic series distribution.

The new composite function becomes

$$H(x) = G(F(x)) = e^{\lambda} e^{\lambda \frac{c - \log(1-\phi x)}{c - \log(1-\phi)}} = \left(\frac{1-\phi}{1-\phi x}\right)^{\theta} \tag{25}$$

where we have set

$$\theta = \frac{\lambda}{c - \log(1-\phi)}. \tag{26}$$

Apply di Bruno's formula to the pair (G, F) and Ewens' formula still results.

The occurrence of Ewens' formula in this special case of compound sampling when $c = 0$ is not new and di Bruno's formula in this special case is thus a manifestation of a result of Watterson (Section 2.3 of Watterson (1974)) who considers a population comprised of a Poisson number of species K for which, given $K = k$, the number X_j of individuals of species j are conditionally independently identically distributed according to the logarithmic distribution (24). Let N denote the total number of individuals, b_i the number of species of type i, and Π be the random variable representing the random partition of the integer N. Watterson's result states that conditional on $N = n$ the distribution Π is given by Ewens' formula (1) where $\theta = \frac{-\lambda}{\log(1-\phi)}$ as in (26) when $c = 0$. We note that, unlike in the random partition Π_n defined previously for fixed n, here n is replaced with a negative-binomial random variable N with "success" probability $1 - \phi$ having a distribution given by

$$\mathbb{P}[N = n] = \binom{\theta + n - 1}{n} \phi^n (1-\phi)^{\theta} \tag{27}$$

and whose p.g.f. is the composite function (25) in di Bruno's formula.

Next, we show that Watterson's result holds even when $c \neq 0$ for the p.g.f. in (23) following which we confirm an analogous result for the Pitman formulas. This will require deriving new sampling formulas when some species have 0 representatives and then taking the marginal distributions.

Let $\Pi^\star \equiv (B_0, \Pi)$ denote the partition of N that includes the (random) number B_0 of the K species having 0 observations. Thus, a typical value of $\Pi^\star$ may be denoted by $b^\star = (b_0, b_1, \ldots, b_n) \equiv (b_0, b)$ where $b = (b_1, \ldots, b_n)$ is the usual partition involving only observed species. As before write $k = b_1 + \cdots + b_n$. We first derive the conditional distributions $\mathbb{P}[\Pi^\star = (b_0, b_1, \ldots, b_n) | N = n]$.

Theorem 1. *The conditional distribution of the partition $\Pi^\star$ of the compound sampling process described by the model (23) is given by:*

$$\mathbb{P}[\Pi^\star = b^\star | N = n] = \frac{e^{-\theta c}(\theta c)^{b_0}}{b_0!} \frac{n!}{[\theta]^n} \theta^k \prod_{i=1}^{n} \frac{1}{i^{b_i} b_i!}. \tag{28}$$

Proof. From (23) and (25)

$$\begin{aligned} \mathbb{P}[K = k + b_0] &= \frac{e^{-\lambda}\lambda^{k+b_0}}{(k+b_0)!}, \quad k + b_0 \geq 1 \\ \mathbb{P}[N = n] &= \frac{[\theta]^n (1-\phi)^{\theta} \phi^n}{n!}, \quad n \geq 1 \end{aligned} \tag{29}$$

with X as in (24), since

$$\begin{aligned} \mathbb{P}[\Pi^\star = b^\star | K = k + b_0] &= \binom{k+b_0}{b^\star} \mathbb{P}[X = 0]^{b_0} \prod_{j=1}^{k} \mathbb{P}[X = x_j] \\ &= \binom{k+b_0}{b^\star} \mathbb{P}[X = 0]^{b_0} \prod_{i=1}^{n} (\mathbb{P}[X = i])^{b_i} \end{aligned} \tag{30}$$

where b is the partition induced by $\{x_1, \ldots, x_k\}$, then

$$\begin{aligned} \mathbb{P}[\Pi^\star = b^\star | N = n] &= \frac{\mathbb{P}[\Pi^\star = b^\star, N = n]}{\mathbb{P}[N = n]} \\ &= \frac{\mathbb{P}[\Pi^\star = b^\star]}{\mathbb{P}[N = n]} \text{ because } \{N = n\} \text{ is redundant in the numerator} \\ &= \frac{\mathbb{P}[\Pi^\star = b^\star, K = k + b_0]}{\mathbb{P}[N = n]} \text{ because } \{K = k + b_0\} \text{ is redundant} \\ &= \mathbb{P}[\Pi^\star = b^\star | K = k + b_0] \frac{\mathbb{P}[K = k + b_0]}{\mathbb{P}[N = n]} \\ &= \frac{(k+b_0)!}{b_0! \prod_{i=1}^{n} b_i!} \left(\frac{c}{c - \log(1-\phi)}\right)^{b_0} \prod_{i=1}^{n} \left(\frac{\phi^i}{i(c - \log(1-\phi))}\right)^{b_i} \\ &\quad \times \frac{e^{-\lambda}\lambda^{k+b_0}}{(k+b_0)!} \frac{n!}{[\theta]^n (1-\phi)^{\theta} \phi^n} \end{aligned}$$

from (24), (29) and (30), which simplifies to (28) using (26).

Corollary 1. *If Π is the partition corresponding to the observed species obtained from the compound sampling process (23) then the marginal conditional distribution $\mathbb{P}[\Pi = b | N = n]$ is given by Ewens sampling formula and conditionally on $N = n, B_0$ has a Poisson distribution with mean θc and is independent of Π.*

Proof. This follows immediately from the factorization on the right side of (28).

We next obtain conditional distributions $\mathbb{P}[\Pi^\star = b^\star | N = n]$ and $\mathbb{P}[\Pi = b | N = n]$ that will correspond to the Pitman formulas. First we require the corresponding G, F.

When $0 < \alpha < 1, \theta > 0$, in place of the functions g, f in (9) use the p.g.f.

$$G(x) = \frac{(1-\phi)^r}{(1-\phi x)^r} \quad \text{and} \quad F(x) = \frac{1 - c(1-\tau x)^{\alpha}}{1 - c(1-\tau)^{\alpha}} \tag{31}$$

respectively, where $0 < c \leq 1, 0 < \tau < 1, r > 0$ so

$$\begin{aligned} H(x) &= \frac{(1-\phi)^r}{(1-\phi F(x))^r} \\ 1 - \phi F(x) &= \frac{1 - c(1-\tau)^{\alpha} - \phi}{1 - c(1-\tau)^{\alpha}} + \frac{\phi c (1-\tau x)^{\alpha}}{1 - c(1-\tau)^{\alpha}}. \end{aligned}$$

To recover the structure needed to explicitly differentiate H and obtain an identity, we need to set $\phi = 1 - c(1-\tau)^{\alpha}$ simplifying H to

Table 1
Sampling formulas and probability generating functions

Sampling formula	p.g.f. of K	p.g.f. of X	p.g.f. of N
Ewens	$e^{\lambda(x-1)}$	$\frac{c-\log(1-\phi x)}{c-\log(1-\phi)}$	$\frac{(1-\phi)^\theta}{(1-\phi x)^\theta}$
Pitman $0<\alpha<1, \theta>0$	$\frac{(1-\phi)^r}{(1-\phi x)^r}$	$\frac{1-c(1-\tau x)^\alpha}{1-c(1-\tau)^\alpha}$	$\frac{(1-\tau)^\theta}{(1-\tau x)^\theta}$
$0<\alpha<1, -\alpha<\theta<0$	$\frac{1-d(1-\phi x)^\beta}{1-d(1-\phi)^\beta}$	$\frac{1-c(1-\tau x)^\alpha}{1-c(1-\tau)^\alpha}$	$\frac{1-dc^\beta(1-\tau x)^{\alpha\beta}}{1-dc^\beta(1-\tau)^{\alpha\beta}}$
$0<\alpha<1, \theta=0$	$\frac{d-\log(1-\phi x)}{d-\log(1-\phi)}$	$\frac{1-c(1-\tau x)^\alpha}{1-c(1-\tau)^\alpha}$	$\frac{d-\log c-\alpha\log(1-\tau x)}{d-\log c-\alpha\log(1-\tau)}$
Dirichlet	x^M	$\frac{(1-\phi)^{\theta/M}}{(1-\phi x)^{\theta/M}}$	$\frac{(1-\phi)^\theta}{(1-\phi x)^\theta}$
Multinomial	x^M	$e^{\theta(x-1)}$	$e^{\theta M(x-1)}$
Multivariate hypergeometric	x^M	$(1-\phi+\phi x)^C$	$(1-\phi+\phi x)^{CM}$

$$H(x)=\frac{(1-\phi)^r}{(c(1-\tau x)^\alpha)^r}=\frac{(c(1-\tau)^\alpha)^r}{(c(1-\tau x)^\alpha)^r}=\frac{(1-\tau)^\theta}{(1-\tau x)^\theta}$$

where $\theta=\alpha r$. Application of di Bruno's formula leads to (11).

For $0<\alpha<1$, $-\alpha<\theta<0$ use the p.g.f.

$$G(x)=\frac{1-d(1-\phi x)^\beta}{1-d(1-\phi)^\beta}\quad\text{and}\quad F(x)=\frac{1-c(1-\tau x)^\alpha}{1-c(1-\tau)^\alpha}\tag{32}$$

where $0<c,d,\tau,\phi\le 1$ so that

$$H(x)=\frac{1-d(1-\phi F(x))^\beta}{1-d(1-\phi)^\beta}$$

$$1-\phi F(x)=\frac{1-c(1-\tau)^\alpha-\phi}{1-c(1-\tau)^\alpha}+\frac{\phi c(1-\tau x)^\alpha}{1-c(1-\tau)^\alpha}.$$

Again, require $\phi=1-c(1-\tau)^\alpha$, giving

$$H(x)=\frac{1-dc^\beta(1-\tau x)^{\alpha\beta}}{1-dc^\beta(1-\tau)^{\alpha\beta}}$$

and di Bruno's formula leads to (11) with $\theta=-\alpha\beta$. Note that $\tau=1$ or $\phi=1$ is permitted here but not in the previous case nor in the next one following.

Here, for the final case, $0<\alpha<1$, $\theta=0$ take

$$G(x)=\frac{d-\log(1-\phi x)}{d-\log(1-\phi)}\quad\text{and}\quad F(x)=\frac{1-c(1-\tau x)^\alpha}{1-c(1-\tau)^\alpha}\tag{33}$$

where $0<\tau,\phi<1$, $0<c\le 1$, $d\ge 0$ with $\phi=1-c(1-\tau)^\alpha$ and then

$$H(x)=\frac{d-\log c-\alpha\log(1-\tau x)}{d-\log c-\alpha\log(1-\tau)}$$

so again di Bruno's formula leads to (11).

Notice that the function F is the same in all three cases. For $\theta>0$, G is the p.g.f. of a negative-binomial random variable; for $-\alpha<\theta<0$ it has the same form as F; for $\theta=0$, G is the p.g.f. of a random variable distributed with Fisher's logarithmic series distribution except allowing a non-zero constant term for $c\ne 1$. For the Ewens case, the Pitman with $\theta>0$, and the Dirichlet, the p.g.f. of N is negative-binomial. Table 1 lists the p.g.f. versions of the composite functions to which di Bruno's formula was applied above, including those described previously in Sections 4–6 where the probabilistic interpretations already anticipate compound sampling except that the number of species is constant at M. Note that $\theta=\alpha r$ in the first Pitman case while $\theta=-\alpha\beta$ in the second.

Theorem 2. *The conditional distributions of the partitions $\Pi^\star$ of the compound processes described by the models (31)–(33) are given by:*

$$\mathbb{P}[\Pi^\star=b^\star|N=n]$$
$$=\frac{[\frac{\theta}{\alpha}+k]^{b_0}(1-c)^{b_0}c^{\frac{\theta}{\alpha}+k}}{b_0!}\frac{n!}{[\theta]^n}\prod_{j=1}^{k-1}(\theta+j\alpha)$$
$$\times\prod_{i=1}^{n}\frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i}b_i!}.\tag{34}$$

Proof. For the case (31), note that

$$\mathbb{P}[K=k+b_0]=\frac{(1-\phi)^r\phi^{k+b_0}[r]^{k+b_0}}{(k+b_0)!},\quad k+b_0\ge 1$$
$$\mathbb{P}[N=n]=\frac{(1-\tau)^\theta\tau^n[\theta]^n}{n!},\quad n\ge 1\tag{35}$$

while if X is a generic random variable with p.g.f. F in (31), then

$$\mathbb{P}[X=i]=\begin{cases}\dfrac{1-c}{1-c(1-\tau)^\alpha} & i=0\\ \dfrac{c}{1-c(1-\tau)^\alpha}\dfrac{\tau^i\alpha[1-\alpha]^{i-1}}{(i-1)!} & i\ge 1.\end{cases}\tag{36}$$

Now follow the proof of Theorem 1 to obtain

$$\mathbb{P}[\Pi^\star=b^\star|N=n]$$
$$=\mathbb{P}[\Pi^\star=b^\star|K=k+b_0]\frac{\mathbb{P}[K=k+b_0]}{\mathbb{P}[N=n]}$$
$$=\frac{(k+b_0)!}{b_0!\prod_{i=1}^n b_i!}\left(\frac{1-c}{\phi}\right)^{b_0}\left(\frac{\alpha c}{\phi}\right)^k\tau^n\prod_{i=1}^n\frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i}}$$
$$\times\frac{(1-\phi)^r\phi^{k+b_0}[r]^{k+b_0}}{(k+b_0)!}\frac{n!}{(1-\tau)^\theta\tau^n[\theta]^n}$$
$$=\frac{[r]^{k+b_0}(1-c)^{b_0}(\alpha c)^k}{b_0!}\frac{(1-\phi)^r}{(1-\tau)^\theta}\frac{n!}{[\theta]^n}\prod_{i=1}^n\frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i}b_i!}$$

from (30), (35) and (36). Since $1-\phi=c(1-\tau)^\alpha$ and $\theta=r\alpha$ the fraction $\frac{(1-\phi)^r}{(1-\tau)^\theta}=c^r$. Substitute $[r]^{k+b_0}=[r+k]^{b_0}[r]^k$ and combine $[r]^k$ with α^k as $[r]^k\alpha^k=\alpha^k\prod_{j=0}^{k-1}(r+j)=\prod_{j=0}^{k-1}(\theta+j\alpha)$ to arrive at

$$\frac{[r+k]^{b_0}(1-c)^{b_0}c^{r+k}}{b_0!}\frac{n!}{[\theta]^n}\prod_{j=1}^{k-1}(\alpha r+j\alpha)\prod_{i=1}^n\frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i}b_i!}$$

and use $\theta=\alpha r$ to complete the proof of (34).

For the second parameter range (32) we use

$$\mathbb{P}[K=k+b_0]=\frac{d}{1-d(1-\phi)^\beta}\frac{\phi^{k+b_0}\beta[1-\beta]^{k+b_0-1}}{(k+b_0)!},$$
$$k+b_0\ge 1$$
$$\mathbb{P}[N=n]=\frac{dc}{1-dc(1-\tau)^{\alpha\beta}}\frac{\tau^n\alpha\beta[1-\alpha\beta]^{n-1}}{n!},\quad n\ge 1$$

and then decompose $\mathbb{P}[\Pi^\star=b^\star|N=n]$ in the same way to get

$$\mathbb{P}[\Pi^\star=b^\star|N=n]$$
$$=\frac{(k+b_0)!}{b_0!\prod_{i=1}^n b_i!}\left(\frac{1-c}{\phi}\right)^{b_0}\left(\frac{\alpha c}{\phi}\right)^k\tau^n$$

$$\times \prod_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i}} \frac{\mathrm{d}\phi^{k+b_0}\beta[1-\beta]^{k+b_0-1}}{(1-d(1-\phi)^{\beta})(k+b_0)!}$$
$$\times \frac{n!(1-dc^{\beta}(1-\tau)^{\alpha\beta})}{dc^{\beta}\tau^{n}\alpha\beta[1-\alpha\beta]^{n}}$$
$$= \frac{[1-\beta]^{k+b_0-1}(1-c)^{b_0}\alpha^{k-1}c^{k-\beta}}{b_0!} \frac{n!}{[1-\alpha\beta]^{n-1}}$$
$$\times \prod_{i=1}^{n} \frac{([1-\alpha]^{i-1})^{b_i}}{i!^{b_i}b_i!}$$

using $1-d(1-\phi)^{\beta} = 1-dc^{\beta}(1-\tau)^{\alpha\beta}$. We express $[1-\beta]^{k+b_0-1}$ as $[k-\beta]^{b_0}[1-\beta]^{k-1}$ and multiply $[1-\beta]^{k-1}$ by α^{k-1} to get $\prod_{j=1}^{k-1}(\theta+j\alpha)$ in the numerator, where $\theta = -\alpha\beta$. The ascending factorial in the denominator is $\prod_{j=1}^{n-1}(\theta+j)$ and if each of these factorials in the numerator and denominator is multiplied by θ then the lower limit in both products starts at 0 and we have derived (34).

Finally, for the last case (33) we have

$$\mathbb{P}[K = k+b_0] = \frac{\phi^{k+b_0}}{(k+b_0)(d-\log(1-\phi))}, \quad k+b_0 \geq 1$$
$$\mathbb{P}[N = n] = \frac{\alpha\tau^{n}}{n(d-\log c-\alpha\log(1-\tau))}, \quad n \geq 1$$

and this time express the factor $(k+b_0-1)!$ in the numerator as $(k-1)![k]^{b_0}$ when simplifying to arrive at (34).

We next derive the conditional distributions $\mathbb{P}[\Pi = b|N = n]$ corresponding to the partitions of the observed species. These are obtained as the marginal distributions of $\mathbb{P}[\Pi^{\star} = b^{\star}|N = n]$ by summing over all b_0.

Corollary 2. *If Π is the partition corresponding to the observed species obtained from the compound sampling processes (31)–(33), then the marginal conditional distribution $\mathbb{P}[\Pi = b|N = n]$ is given by (3).*

Proof. We have expressed Theorem 2 in such a way that (3) appears as a factor in (34). To obtain the marginal conditional distribution we sum on b_0. But

$$\sum_{b_0=0}^{\infty} \frac{[\frac{\theta}{\alpha}+k]^{b_0}(1-c)^{b_0}c^{\frac{\theta}{\alpha}+k}}{b_0!} = 1$$

in view of the expansion $\sum_{i=0}^{\infty}\frac{[A]^i}{i!}t^i = \frac{1}{(1-t)^A}$ leaving (3).

When $c = 1$ Theorem 2 is not applicable because $b_0 = 0$ and $\Pi^{\star}$ becomes Π. In this case, Corollary 2 obtains directly and the analysis simplifies considerably.

8. Conclusion

We have shown that Faà di Bruno's formula for the derivative of a composite function simplifies to a probability distribution on partitions of the integers in some cases. Among the distributions that can be obtained are those arising in population genetics and in statistical mechanics. In the cases of the Ewens, Pitman, and symmetric Dirichlet sampling formulas, the composite function is the same. The key to the use of the formula lies in finding pairs of functions whose derivatives and the derivative of the composite function can be computed directly. We have also shown that when the composite function is interpreted as a compound probability generating function then the partition coincides with the corresponding conditional partition for selecting a random number of individuals from a finite or random number of species.

Acknowledgments

I am grateful to Warren Ewens for providing critical comments on an early draft of this paper. The author was supported by NSERC Discovery Grant.

References

di Bruno Cabaliere, F.F., 1855. Sullo sviluppo delle Funzioni. Annali di Scienze Matem. e Fisiche 6, 479–480.
Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3, 87–112.
Ewens, W.J., 1990. Population genetics theory – the past and the future. In: Lessard, S. (Ed.), Mathematical and Statistical Developments of Evolutionary Theory. Kluwer, Amsterdam, pp. 177–227.
Ewens, W.J., 2004. Mathematical Population Genetics I. Theoretical Introduction, second ed. Springer, NY.
Griffiths, R., 1980. Unpublished Note. Monash University.
Hoppe, F.M., 1984. Pólya-like urns and the Ewens sampling formula. J. Math. Biol. 20, 91–94.
Hoppe, F.M., 1987. The sampling theory of neutral alleles and an urn model in population genetics. J. Math. Biol. 25, 123–159.
Kaucký, J., 1985. Kombinatoricke Identity. Nakladatelství ČSAV, Praha.
Kingman, J.F.C., 1975. Random discrete distributions. J. Roy. Statist. Soc. B 37, 1–22.
Kingman, J.F.C., 1980. The mathematics of genetic diversity, in: CBMS-NSF Regional Conference Series in Mathematics, SIAM, vol. 34, Philadelphia, PA.
Pitman, J., 1992. The two-parameter generalization of Ewens' random partition structure. In: Unpublished Note. University of California, Berkeley.
Pitman, J., 2006. Combinatorial Stochastic Processes. In: Lecture Notes in Mathematics, vol. 1875. Springer, Berlin.
Shepp, L., Lloyd, S., 1966. Ordered cycle lengths in a random permutation. Amer. Math. Soc. 121, 340–357.
Watterson, G.A., 1974. Models for logarithmic species abundance distributions. Theor. Popul. Biol. 6, 217–250.
Watterson, G.A., 1976. The stationary distribution of the infinitely-many neutral alleles diffusion model. J. Appl. Probab. 13, 639–651.