PROCEEDINGS OF THE ROYAL SOCIETY A

Proc. R. Soc. A (2006) **462**, 1181–1195
doi:10.1098/rspa.2005.1622
Published online 17 January 2006

Binomial subsampling

By Carsten Wiuf[1,*] and Michael P. H. Stumpf[2]

[1] *Bioinformatics Research Center, University of Aarhus, Høegh Guldbergsgade 10, 8000 Aarhus C, Denmark*
[2] *Division of Molecular Biosciences, Imperial College London, Wolfson Building, London SW7 2AZ, UK*

In this paper, we discuss statistical families $\mathcal{P}$ with the property that if the distribution of a random variable X is in $\mathcal{P}$, then so is the distribution of $Z\sim\mathrm{Bi}(X, p)$ for $0\le p\le 1$. (Here we take $Z\sim\mathrm{Bi}(X, p)$ to mean that given $X=x$, Z is a draw from the binomial distribution $\mathrm{Bi}(x, p)$.) It is said that the family is closed under binomial subsampling. We characterize such families in terms of probability generating functions and for families with finite moments of all orders we give a necessary and sufficient condition for the family to be closed under binomial subsampling. The results are illustrated with power series and other examples, and related to examples from mathematical biology. Finally, some issues concerning inference are discussed.

Keywords: biological networks; binomial distribution; closure property; power series; sampling

1. Introduction

Many examples in statistics relate to the problem of ‘thinning’ a random variable X in the sense that X itself is not observed, rather a ‘distorted’—or sampled—version of X, Z is observed. Z is influenced by the nature of X, the means of observation and stochastic errors. Z might have a functional relationship to X, $Z=g(X)$, or a stochastical relationship, $Z\sim F(X)$, where F is a distribution function depending on the value of X. Statistically, X can be thought of as missing information. In our case, we assume X and Z are stochastically related through $Z\sim\mathrm{Bi}(X, p)$. Here, and elsewhere in the article, we take $Z\sim\mathrm{Bi}(X, p)$ to mean that given $X=x$, Z is a draw from the binomial distribution $\mathrm{Bi}(x, p)$.

We have noticed several examples for this in our own field of mathematical biology and below we list a number of these. We start out with an example that was the motivation for this paper. In the biological world and elsewhere it is often the case that data is well represented by a stochastic graph. For example, interactions between genes, proteins, or species in a foodweb can all be represented by a graph, where each gene, protein or species is a node in the network or graph, and interactions are connections between nodes. Maybe, the best-known and understood stochastic graph is an Erdös–Renyi random graph

[*] Author for correspondence (wiuf@birc.dk).

Received 18 May 2005
Accepted 23 November 2005

© 2006 The Royal Society

(e.g. Bollobás 2001). Each node, i, in the network has $X_i \sim \mathrm{Bi}(M-1, q)$ connections to other nodes, where M is the total number of nodes in the network. Other examples which have been studied extensively, are small-world networks and scale-free graphs.

However, it is often the case that our experimental devices, or observational strategies, only allow us to observe each connection in a network with a certain probability. Perhaps, the most parsimonious sampling scheme is one where each node is sampled with the same probability p, independently of its connections to other nodes or other features of the network. For an Erdös–Renyi random graph, where node i has $X_i \sim \mathrm{Bi}(M-1, q)$ true connections, only $Z_i \sim \mathrm{Bi}(M-1, pq)$ are observed. Interestingly, the sampled distribution has the same form as the true distribution. In other examples, such as small-world networks and scale-free graphs, the sampled distribution does not take the same form as the true distribution (Stumpf *et al.* 2005; Stumpf & Wiuf 2005). Nonetheless, data are analysed under the implicit—though generally unacknowledged—assumption that the distribution of Z_i has the same form as that of X_i (see Stumpf *et al.* 2005).

Other examples include the following. (i) Assume X_t is drawn according to a counting process, where t denotes time. If each event (X_t in total) only can be observed with probability p, then the number of observable events, Z_t, is $Z_t \sim \mathrm{Bi}(X_t, p)$. For example, let X_t be the number of mutations in a DNA sequence over time and Z_t the number of mutations that change the DNA sequence. If mutations arrive at rate λ, then $X_t \sim \mathrm{Po}(\lambda t)$ and $X_t \sim \mathrm{Po}(p\lambda t)$, where p is the probability of a DNA changing mutation and $1-p$ the probability of a mutation that does not alter the DNA sequence (e.g. Felsenstein 2004). (ii) Another example comes from cancer research (Koed *et al.* 2005). Cancer genomes are instable and undergo frequent chromosomal alterations. Here we focus on loss of chromosomal regions. A normal genome has two copies of all chromosomes, whereas a cancer genome might have lost one of these copies, or lost regions of a chromosome (theoretically, loss of both copies is possible, but this is very rare). Loss can be inferred experimentally from SNP (single nucleotide polymorphism) or bi-allelic DNA markers typed in a normal genome and a cancer genome from the same patient: if two different alleles are observed in the normal genome (i.e. the marker is heterozygous), but only one allele is observed in the cancer genome, then one allele must have been lost. If both are observed, then nothing has been lost. Conversely, if only one allele is observed in the normal cell (i.e. the marker is homozygous), and also one is observed in the cancer cell, it is impossible to tell whether an allele has been lost or not, because the cancer genome could still have one or two copies of that allele. If the number of losses in a region with n markers is X_n and the marker is heterozygous with probability p, then the number of observable losses Z_n is $Z_t \sim \mathrm{Bi}(X_n, p)$. X_n might be modelled by a k-order Markov chain in n.

(iii) Lastly, in field ecology capture–release experiments can be used in order to measure species abundance (Southwood & Henderson 2000). Quite generally, there is considerable interest in ecological time series (e.g. Powell & Steele 1994). However, at each time-point, it is impossible to sample all individuals of all species present in a given area. When averaged over the duration of the observation each individual is observed with probability p. Note, however, that some individuals may be sampled more than once by chance. If the true species abundance is given by a process X_t, then the observed number of individuals at each time-point, Z_t, is given by $Z_t \sim \mathrm{Bi}(X_t, p)$. There are many other related biological processes which can be modelled in this way. These include, for example, all processes where a source generates particles (e.g. hepatitis C virus (HCV) produced in the liver), which leave the source vicinity with probability p, and where the measurements of the number of particles occur at some distance from the source (e.g. measuring HCV abundance in the blood).

In some of these examples, the distribution of Z will have the same functional form as the distribution of X. This property, of course, relies on the distribution of X. Our own interest was mainly motivated by analysis of biological network data, where we noticed that many models other than Erdös–Renyi random graphs, e.g. scale-free graphs, did not have the property. In the analysis of real data, it is often assumed that data confirms to the distribution of the entire network, and that sampling does not change this distribution (maybe apart from a change in parameters), see Stumpf *et al.* (2005) for a discussion. This may be erroneous and consequently lead to inferential mistakes. In other examples, e.g. modelling the evolution of a DNA sequence, the process of DNA changing mutations has the same functional form as the process of all mutations, and the same model can be used for both (with a change in parameter to $p\lambda t$ instead of λt). However, this has the cost that p and λ cannot be separated easily.

In this paper, we discuss families with the property of being *closed under binomial subsampling*, i.e. families that possess the property that the distribution of $Z \sim \mathrm{Bi}(X, p)$ is in the same family of probability distributions as X. We characterize such families in terms of probability generating functions (pgfs) and in terms of moments (if these exist). We also determine the power series families with this property. The results are illustrated by a number of examples. Although some of the results are relatively straightforward to derive, they and their implications have, to our knowledge, not been discussed together in the literature. Finally, we make a few comments in relation to statistical inference.

2. Closure under binomial subsampling

Let $\mathcal{P}$ be a family of probability distributions on $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. It is convenient to think of $\mathcal{P}$ as a family of distributions of random variables X given by $P(x) = \mathbb{P}(X = x)$ for $P \in \mathcal{P}$, and we write $X \sim P$, if X has distribution P. The degenerate distribution $P(0) = \mathbb{P}(X = 0) = 1$ is denoted by P_0.

Further, let $G_X(s) = \mathbb{E}(s^X)$ denote the pgf of X, and similarly let $G_P(s)$ denote the pgf of P. If $X \sim P$, then $G_X(s) = G_P(s)$. Any open interval I, where $G_X(s)$, $s \in I$, is finite, defines the distribution of X uniquely; i.e. if $G_X(s) = G_Y(s)$ for $s \in I$, then X and Y have the same distribution. Therefore, for convenience, $G_X(s)$, $s \in I$, is also called the pgf of X.

We use the following notation througout the paper. Let $x_{[k]}$ denote the kth descending factorial of x, $x_{[k]} = x(x-1)\ldots(x-k+1)$, and $x_{(k)}$ the kth ascending factorial of x, $x_{(k)} = x(x+1)\ldots(x+k-1)$.

Definition 2.1. Let $\mathcal{P}$ be a family of distributions on $\mathbb{N}_0$, X an $\mathbb{N}_0$-valued random variable, and define Z by $Z \sim \mathrm{Bi}(X, p)$ for $p \in [0,1]$. If Z has distribution in $\mathcal{P}$, whenever X has distribution in $\mathcal{P}$, then $\mathcal{P}$ is said to be closed under binomial subsampling.

In definition 2.1, the dependence of p on Z is suppressed. The distribution of Z is given by

$$\mathbb{P}(Z = z) = \sum_{x \geq z} \binom{x}{z} p^z (1-p)^{x-z} \mathbb{P}(X = x). \tag{2.1}$$

Definition 2.1 is equivalent to the requirement that

$$Z = Z_1 + \cdots + Z_X, \tag{2.2}$$

has distribution in $\mathcal{P}$, whenever X has distribution in $\mathcal{P}$ and Z_j is a Bernoulli variable with parameter p.

Further, $p=0$ implies that P_0 always belongs to a family closed under binomial subsampling. Definition 2.1 also implies that if $\mathcal{P}_i$, $i \in I$, are families closed under binomial subsampling, then so are $\cup_{i \in I} \mathcal{P}_i$ and $\cap_{i \in I} \mathcal{P}_i$.

Definition 2.2. Let $\mathcal{P}$ be closed under binomial subsampling, and let Q, $P \in \mathcal{P}$. Define a relation $\diamond$ on $\mathcal{P}$ in the following way: $Q \diamond P$ if and only if there exists a random vector (Z, X) with $Z \sim Q$ and $X \sim P$, such that either $Z \sim \mathrm{Bi}(X, p)$ or $X \sim \mathrm{Bi}(Z, p)$ for some $p \in (0,1]$. Note that P_0 is only related to itself.

$Q \diamond P$ is a relation between univariate distributions, but holds if a suitable bivariate distribution, which links Q and P together, exists.

Theorem 2.3. *The relation $\diamond$ defined in definition 2.2 is an equivalence relation on $\mathcal{P}$.*

The equivalence class consisting of only P_0 is called the *trivial* class.

Theorem 2.4. *Let $\mathcal{P}$ be closed under binomial subsampling and $\mathcal{C}$ a non-trivial class of $\mathcal{P}$. If $P \in \mathcal{C}$, then there is an interval J_P with left endpoint zero but not containing it, such that $Q \in \mathcal{C}$ if and only if*

$$G_Q(s) = G_P(1-r+rs), \tag{2.3}$$

for some $r \in J_P$.

In theorem 2.4, P is called a *class representative.* Theorem 2.4 provides an alternative characterization of the relation $\diamond$ in terms of generating functions. This characterization is perhaps more natural than the one given in definition 2.2; in that it does not involve bivariate distributions. However, definition 2.2 is closer to how we originally perceived the problem.

Example 2.5. Let $\mathcal{P}$ be closed under binomial subsampling and let $X_i \sim P \in \mathcal{P}$ for $i=1, 2, \ldots$. Further let N be a random variable with distribution on $\mathbb{N}$, and assume all X_i and N are mutually independent. Then the family of distributions defined by $X = \sum_{n=1}^{N} X_i$ is closed under binomial subsampling. The pgf of X is given by

$$G_X(s) = \sum_{n=1}^{\infty} G_{X_1}(s)^n \mathbb{P}(N = n). \tag{2.4}$$

If $Z\sim\mathrm{Bi}(X, p)$, then $Z=\sum_{n=1}^{N} Z_i$, with $Z_i\sim\mathrm{Bi}(X_i, p)$; hence

$$\begin{aligned} G_Z(s) &= \sum_{n=1}^{\infty} G_{Z_1}(s)^n \mathbb{P}(N=n) = \sum_{n=1}^{\infty} G_{X_1}(1-p+ps)^n \mathbb{P}(N=n) \\ &= G_X(1-p+ps), \end{aligned} \tag{2.5}$$

and the results follow from theorem 2.4.

Theorem 2.6. *Let X be an $\mathbb{N}_0$-valued random variable and define a family of functions by*

$$G_r(s) = G_X(1-r+rs), \tag{2.6}$$

for $r\in\mathbb{R}_0$ and $s\geq\max(0, 1-1/r)$. Then there exists $1\leq\rho\leq\infty$, such that $G_r(s)$ is the pgf of a random variable with values in $\mathbb{N}_0$ if and only if $r\leq\rho$, $r\neq\infty$. The family of distributions defined by $G_r(s)$, $r\leq\rho$, $r\neq\infty$, is closed under binomial subsampling.

Example 2.7. Let X_1 and X_2 be independent random variables. Further, let $\mathcal{P}_i$, $i=1, 2$, be the family of distributions generated by X_i, $i=1, 2$. Let Z_i be defined by $G_{Z_i}(s)=G_{X_i}(1-r+rs)$ for $r\leq\min(\rho_1, \rho_2)$ and ρ_i given as in theorem 2.6. Then also the family $\mathcal{P}$ of distributions defined by Z_1+Z_2 (assuming Z_1 and Z_2 are independent) is closed under binomial subsampling.

Let $\mathcal{P}$ be a family closed under binomial subsampling and let $\psi:\Psi\to\mathcal{P}$ be a parametrization of the classes that maps ψ to a unique member P_ψ of the class $\mathcal{C}$. For convenience, ψ is called a *class parameter* (which may be vector-valued). Then $\mathcal{P}$ is parameterized by (ψ, r), where $\psi\in\Psi$, $r\in J_{P_\psi}$, and J_{P_ψ} is an interval with boundary zero (theorem 2.4). According to theorem 2.6, $J_{P_\psi}\subseteq(0, \rho_\psi]$ for some ρ_ψ. If equality holds, then the class is said to be *full*; and if not, the class can be extended to a full class accordingly. A full class is said to be generated by the class representative P_ψ. Note that the class is generated by any member of the class.

Theorem 2.8. *Let $\mathcal{P}$ be the family generated by X (with $r\leq\rho$ as in theorem 2.6). Assume $X=\sum_{i=1}^{n} X_i$, X_i iid $\mathbb{N}_0$-valued random variables and let $\mathcal{Q}$ be the family generated by X_1 (with $r\leq\rho_1$). Then $Z=\sum_{i-1}^{n} Zi$, Z_i iid, has distribution in $\mathcal{P}$, whenever Z_1 has distribution in $\mathcal{Q}$. It follows that $\rho_1\leq\rho$.*

3. Finite moments of all orders

We will now impose some further constraints on the family $\mathcal{P}$. Explicitly, we will assume that all moments of a random variable X with distribution P in $\mathcal{P}$ exist, and that P is determined by these moments.

Abusing notation slightly we say that X is in $\mathcal{P}$ whenever X has a distribution in $\mathcal{P}$. Further, if $\mathcal{P}$ is closed under binomial subsampling, then we let Z denote an element in the class of X, and assume that (Z, X) is constructed, such that definition 2.2 is fulfilled, i.e. either $Z\sim\mathrm{Bi}(X, p)$ or $X\sim\mathrm{Bi}(Z, p)$.

Note that for Z in the class of X,

$$\mathbb{E}(Z) = \frac{\mathrm{d}}{\mathrm{d}s} G_Z(s)_{|s=1} = \frac{\mathrm{d}}{\mathrm{d}s} G_X(1-r+rs)_{|s=1} = r\mathbb{E}(X), \tag{3.1}$$

and $\mathcal{P}$ is equivalently parameterized by (ψ, τ), where τ denotes expectation and ψ is a class parameter.

Henceforth, we assume a family $\mathcal{P}$ of distributions is parameterized by $\omega = (\phi, \tau) \in \Omega$, where τ denotes expectation of the distribution and ϕ is an additional parameter. The parameterization is assumed to be one-to-one. No topological constraints are put on the space Ω.

Theorem 3.1. *A parameterized family $\mathcal{P}$ is closed under binomial subsampling with ϕ as a class parameter if and only if*

$$\mathbb{E}(X_{[k]}) = a_k(\phi)\tau^k, \tag{3.2}$$

where $a_k(\phi)$ is a constant depending on k and ϕ only, $a_1(\phi)=1$, $(\phi,\tau) \in \Omega = \{(\phi,\tau) | \phi \in \Phi, \tau \in T_\phi\}$, and T_ϕ is an interval containing 0, either $T_\phi = [0, t_\phi]$, or $T_\phi = [0, t_\phi)$ (here t_ϕ is potentially infinity).

Corollary 3.2. *Any series of positive numbers a_k, $k=1, 2, \ldots$, such that $a_1 = 1$, defines a family of distributions closed under binomial subsampling, by the requirement*

$$\mathbb{E}(X_{[k]}) = a_k\tau^k, \tag{3.3}$$

for τ in some interval T containing 0.

The family defined in corollary 3.2 is said to be generated by $\{a_k\}_k$ and T is said to be the range of $\{a_k\}_k$. Lemma 3.3 establishes how $\mathbb{P}(X = x)$ can be found from $\{a_k\}_k$ and $\tau = \mathbb{E}(X)$ for X in the family generated by $\{a_k\}_k$.

Lemma 3.3. *Let X be an $\mathbb{N}_0$-valued random variable with finite moments. If*

$$\lim_{k\to\infty} \frac{\mathbb{E}(X_{[k]})k^i}{k!} = 0, \tag{3.4}$$

for all i, then

$$\mathbb{P}(X = j) = \frac{1}{j!}\sum_{k=j}^{\infty}(-1)^{k-j}\frac{\mathbb{E}(X_{[k]})}{(k-j)!}. \tag{3.5}$$

In particular, the condition in lemma 3.3 is fulfilled if $\sup_k a_k < +\infty$.

Example 3.4. Let $a_k = (k+1)2^{-k}$ for $k = 1, 2, \ldots$. Then

$$\mathbb{P}(X = x) = \frac{(\tau/2)^x}{x!}(x + 1 - \tau/2)\mathrm{e}^{-\tau/2}, \tag{3.6}$$

for $x = 0, 1, \ldots$ and $\tau \in [0,2]$, defines a family closed under binomial subsampling with $\mathbb{E}(X) = \tau$. For $\tau = 2$, $X-1$ is Poisson distributed with intensity 1.

Example 3.5. If the state space is $\{0,1\}$, then the family generated by $\mathbb{E}(X) = \tau$ is the binomial family with distributions $\mathrm{Bi}(1, \tau)$.

Example 3.6. If the state space is $\{0,1,2\}$, then the family generated by $\mathbb{E}(X) = \tau$ and $\mathbb{E}(X(X-1)) = a_2\tau^2$, $a_2 > 0$, has

$$\mathbb{P}(X = 2) = \frac{a_2}{2}\tau^2, \tag{3.7}$$

$$\mathbb{P}(X = 1) = \tau(1 - a_2\tau), \tag{3.8}$$

and

$$\mathbb{P}(X=0) = 1 - \tau + \frac{a_2}{2}\tau^2, \tag{3.9}$$

with $\tau \leq 1/a_2$ for $0.5 \leq a_2$ and $\tau \leq (1-\sqrt{1-2a_2})/a_2$ for $0 < a_2 < 0.5$. If $a_2 \neq 0.5$, then the family is not a binomial family and the family does not contain any binomial distributions, apart from the degenerated distribution P_0.

4. Power series families

We start with a number of well-known examples.

Example 4.1. The members of $\mathcal{P}$ have distribution given by

$$\mathbb{P}(X=x) = \binom{n}{x} q^x (1-q)^{n-x}, \tag{4.1}$$

for $x=0, 1, \ldots, n$, $n \in \mathbb{N}_0$ and $q \in [0,1]$. Here $\tau = nq$, ψ is zero-dimensional, and $a_k = n_{[k]}/n^k$ for $k \geq 1$.

Example 4.2. The Poisson distribution, Po(λ). The members of $\mathcal{P}$ have distribution given by

$$\mathbb{P}(X=x) = \frac{\lambda^x}{x!}\exp(-\lambda), \tag{4.2}$$

for $x \in \mathbb{N}_0$ and $\lambda \geq 0$. Here $\tau = \lambda$, ψ is zero-dimensional and $a_k = 1$ for all k.

Example 4.3. The negative binomial distribution, NB(q, ψ). The members of $\mathcal{P}$ have distribution given by

$$\mathbb{P}(X=x) = \binom{\alpha+x-1}{x} q^x (1-q)^\alpha = \frac{\Gamma(\alpha+x)}{\Gamma(x+1)\Gamma(\alpha)} q^x (1-q)^\alpha, \tag{4.3}$$

for $x \in \mathbb{N}_0$, $q \in (0,1)$ and $\alpha > 0$. Here $\tau = \alpha q/(1-q)$, $\psi = \alpha$ and $a_k(\alpha) = \alpha_{(k)}/\alpha^k$ for $k \geq 1$.

Definition 4.4. A k-order power series family is a family $\mathcal{P}$ of the form

$$\mathbb{P}(X=x) = b(x)\lambda^x g(\lambda)^{-1}, \tag{4.4}$$

for $x \geq k \in \mathbb{N}_0$, $b(k) > 0$, and

$$\mathbb{P}(X=x) = c(x;\lambda), \tag{4.5}$$

for $0 \leq x < k$, $\lambda \in \Lambda \subseteq \mathbb{R}$. Here $g(\lambda)$ is a normalizing constant, such that

$$c(\lambda) = \sum_{x=0}^{k-1} c(x;\lambda) = 1 - \sum_{x=k}^{\infty} b(x)\lambda^x g(\lambda)^{-1}. \tag{4.6}$$

In particular, a k-order power series is an m-order power series for all $m \geq k$.

Examples of 0-order power series families are given in examples 4.1–4.3, whereas example 3.4 is not a power series family for any k. However, according to definition 4.4, the family in example 3.6 is a 2-order power series family.

Theorem 4.5. *A k-order power series family closed under binomial subsampling fulfills, for some suitable parameterization $\lambda \in \Lambda$ and choice of $g(\lambda)$, one of the following conditions: for $x \geq k$, either (1),*

$$b(x) = \binom{n+k}{x}, \tag{4.7}$$

for fixed $n \in \mathbb{N}_0$; or (2),

$$b(x) = \frac{\alpha_{(x-k)}}{x!}, \tag{4.8}$$

for fixed $\alpha > 0$; or (3),

$$b(x) = \frac{1}{x!}. \tag{4.9}$$

Corollary 4.6. *The only 0-order power series families closed under binomial subsampling are the binomial family, the Poisson family and the negative binomial family for fixed α.*

Example 4.7. The (modified) logarithmic distributions with $c \in (0,1]$ and $\psi \in (0, \infty)$,

$$\mathbb{P}(X = x) = \frac{c}{\psi}\left(\frac{\tau\psi}{1+\tau\psi}\right)^x \frac{1}{x}, \tag{4.10}$$

for $x \in \mathbb{N}_0 \backslash \{0\}$ and

$$\mathbb{P}(X = 0) = 1 - \frac{c \log(1+\tau\psi)}{\psi}, \tag{4.11}$$

form a 1-order power series family closed under binomial subsampling for fixed c and ψ. Here

$$\tau \in T_\psi = \left[0, \frac{1}{\psi}(\mathrm{e}^\psi - 1)\right], \tag{4.12}$$

thus the range of $\tau = \mathbb{E}(X)$ depends on ψ.

5. Mixing

The construction of $Z \sim \mathrm{Bi}(X, p)$ can naturally be regarded as a mixture of binomial distributions $\mathrm{Bi}(n, p)$ over n with prior distribution $\mathbb{P}(X = n)$. The resulting family of distributions for $p \in [0,1]$ is closed under binomial subsampling, simply because the binomial families $\mathrm{Bi}(n, p)$, $p \in [0,1]$, are closed under binomial subsampling.

We will give some further examples of mixing. For simplicity, in examples 5.1 and 5.3, assume all moments exist and that the parameter space of a family $\mathcal{P}$ closed under binomial subsampling is a product space, $\Omega = \Psi \times T$. This assumption can easily be relaxed at the cost of a more complex notation.

Example 5.1. Mixing over Ψ. Let $g(x; \phi)$ be a prior on $\psi \in \Psi$, depending on the parameter ϕ. Then

$$\mathbb{E}(X_{[k]}) = \tau^k \int_\Psi a_k(x) g(x; \phi) \mathrm{d}x = \beta_k(\phi)\tau^k, \tag{5.1}$$

assuming the integrals exist. Let $\upsilon = \beta_1(\phi)\tau$, then

$$\mathbb{E}(X_{[k]}) = \beta_k(\phi)\tau^k = \frac{\beta_k(\phi)}{\beta_1(\phi)^k}\upsilon^k, \tag{5.2}$$

defines a family $\mathcal{P}_{\text{mix}}$ closed under binomial subsampling with class parameter ϕ. A subcase is finite mixtures

$$\beta_k(q) = \sum_i q_i a_k(\psi_i), \tag{5.3}$$

where $q = (q_1, \ldots, q_m)$ and $\sum_i q_i = 1$.

Example 5.2. Let $\mathcal{P}$ be a k-order power series family closed under binomial subsampling, and let $P_{\lambda_0} \in \mathcal{P}$. Define $\delta = 1 - c(\lambda_0)$ and Q_{λ_0} by $Q_{\lambda_0}(x) = P_{\lambda_0}(x)/\delta$ for $x \geq k$ and $Q_{\lambda_0}(x) = 0$ otherwise. Let $\mathcal{Q}$ be the family generated by Q_{λ_0}. It follows that $\mathcal{Q} = \{Q_\lambda | \lambda \leq \lambda_0\}$ is a k-order power series family, such that $\delta Q_\lambda(x) = P_\lambda$ for all $x \geq k$ and $\lambda \leq \lambda_0$.

Define $R_\lambda(x)$ by $R_\lambda(x) = [P_\lambda(x) - \delta Q_\lambda(x)]/(1-\delta)$, $\lambda \leq \lambda_0$. Note that $R_\lambda(x) = 0$ for $x \geq k$ and $R_\lambda(x) \geq 0$ for $0 \leq x < k$. Also R_λ is a probability measure for all $\lambda \leq \lambda_0$, and further the family $\mathcal{R} = \{R_\lambda | \lambda \leq \lambda_0\}$ is closed under binomial subsampling.

In consequence, any $P_\lambda \in \mathcal{P}$, with $\lambda \leq \lambda_0$ can be written as a mixture of two measures $Q_\lambda \in \mathcal{Q}$ and $R_\lambda \in \mathcal{R}$,

$$P_\lambda(x) = \delta Q_\lambda(x) + (1-\delta)R_\lambda(x), \tag{5.4}$$

such that $\mathcal{Q}$ and $\mathcal{R}$ are closed under binomial subsampling, and such that $\mathcal{Q}$ is generated by a member of a 0-order power series family with $b(x) = 0$ for $x \leq k$, and $\mathcal{R}$ has support in $\{0, 1, \ldots, k\}$. The two families cannot be extended beyond λ_0, because $Q_{\lambda_0}(x) = 0$ for $x \leq k$.

Example 4.7 provides one example: $Q_{\lambda_0}(x)$ can here be given as the probability measure with $c = 1$ and $\lambda_0 = (\mathrm{e}^\psi - 1)/\psi$.

Example 5.3. Mixing over T. Let a prior distribution, $f(x; \nu, \phi)$, be given on T, depending on the parameter $(\nu, \phi) \in \mathbb{R}_0 \times \Phi$. If the moments of the prior fulfill

$$\int_0^\infty x^k f(x; \nu, \phi)\mathrm{d}x = c_k(\phi)\nu^k, \tag{5.5}$$

for $k \geq 1$, then the mixture of $\mathcal{P}$ with $f(x; \nu, \phi)$ is also closed under binomial subsampling, because

$$\mathbb{E}(X_{[k]}) = \int_0^\infty a_k(\psi)x^k f(x; \nu, \phi)\mathrm{d}x = a_k(\psi)c_k(\phi)\nu^k. \tag{5.6}$$

Let $\mu = c_1(\phi)\nu$, hence

$$\mathbb{E}(X_{[k]}) = \frac{a_k(\psi)c_k(\phi)}{c_1(\phi)^k}\mu^k = \alpha_k(\psi, \phi)\mu^k, \tag{5.7}$$

for $k \geq 1$, $\mu \in \mathbb{R}_0$ and $\alpha_k(\psi, \phi) = a_k(\psi)c_k(\phi)/c_1(\phi)^k$ defines a family $\mathcal{P}_{\text{mix}}$ closed under binomial subsampling with class parameter (ψ, ϕ). Prior distributions fulfilling equation (5.5) include the Gamma distributions, $\Gamma(\nu, \phi)$, with moments

$$\int_0^\infty x^k \frac{\nu^{-\phi}}{\Gamma(\phi)} x^{\phi-1}\,\mathrm{e}^{-x/\nu}\,\mathrm{d}x = \frac{\Gamma(\phi + k)}{\Gamma(\phi)}\nu^k, \tag{5.8}$$

and the folded normal distribution (i.e. the absolute value of a normal) with moments

$$\int_0^\infty x^k \frac{\sqrt{2}}{\sqrt{\pi}\nu} \mathrm{e}^{-x^2/(2\nu^2)} = \frac{2^{k/2}}{\sqrt{\pi}} \Gamma\left(\frac{k+1}{2}\right) \nu^k. \tag{5.9}$$

If $\mathcal{P}$ is the family of Poisson distributions with $a_k = 1$, then mixing $\mathcal{P}$ over λ with the folded normal yields

$$\alpha_k = \pi^{(k-1)/2} \Gamma\left(\frac{k+1}{2}\right), \tag{5.10}$$

and from lemma 3.3,

$$\begin{aligned} \mathbb{P}(X=j) &= \frac{1}{j!} \sum_{k=j}^{\infty} (-1)^{k-j} \int_0^\infty x^k \frac{\sqrt{2}}{\sqrt{\pi}\nu} \mathrm{e}^{-x^2/(2\nu^2)}\, \mathrm{d}x \\ &= \frac{1}{j!} \frac{\sqrt{2}}{\sqrt{\pi}\nu} \int_0^\infty x^j\, \mathrm{e}^{-x-x^2/(2\nu^2)}\, \mathrm{d}x. \end{aligned} \tag{5.11}$$

The expectation of X is $\mathbb{E}(X) = \nu\sqrt{2/\pi}$.

6. Inference

Assuming $Z \sim \mathrm{Bi}(X, p)$, then the conditional distribution of Z given $X = x$, $Z|X = x$, is S-ancillary about inference on p, and X is S-sufficient for inference on parameters, $\omega \in \Omega$, describing the distribution of X. The distribution of Z is generally not sufficient for inference on either of the parameters p and ω.

Let us now consider a one-dimensional family of distributions closed under binomial subsampling. Denote the compound parameter by $\theta = p\tau$. Then the distribution of Z does only depend on θ and the conditional distribution $X|Z = z$ can be found from that of X, Z and $Z|X = x$. For some of the families discussed in this paper, the distribution of $X|Z = z$ is in fact L-nonformative about θ (Barndorff-Nielsen 1999). Hence, it can be argued that Z contains all the information available about θ and Z is said to be L-sufficient for inference on θ. Families for which Z is L-sufficient include the 0-order and 1-order power series families. However, this is not a general feature of families closed under binomial subsampling. For instance, $X|Z = z$ is not L-nonformative about θ in examples 3.4 and 3.6; despite the family in example 3.6 being a 2-order power series family.

To provide an example, consider the Poisson family. The conditional distribution of $X|Z = z$ is given by

$$\binom{x}{z} p^z (1-p)^{x-z} \frac{z!\theta^x}{x! p^x \theta^z} \mathrm{e}^{-\theta(1/p-1)}, \tag{6.1}$$

where $X \sim \mathrm{Po}(\tau) = \mathrm{Po}(\theta/p)$. Let $\hat{p}_\theta$ be the profile likelihood estimate of p for fixed θ. Then

$$\frac{1}{\hat{p}_\theta} = \frac{x-z}{\theta} + 1, \tag{6.2}$$

and the relative conditional profile likelihood becomes

$$\frac{L(\theta_1, \hat{p}_{\theta_1})}{L(\theta_2, \hat{p}_{\theta_2})} = \left(\frac{\theta_1}{\theta_2}\right)^z \mathrm{e}^{-\theta_1+\theta_2}, \tag{6.3}$$

for two values of θ, θ_1 and θ_2. Since it only depends on θ_1, θ_2 and z, *not* x, it is L-nonformative about θ (Barndorff-Nielsen 1999) and Z is L-sufficient for inference on θ, the compound parameter.

Let us return to the general setting. The likelihood function of z is

$$\mathbb{P}(Z = z) = \sum_{x \geq z} \binom{x}{z} p^z (1-p)^{x-z} \mathbb{P}(X = x). \tag{6.4}$$

If the family is closed under binomial subsampling, simulation from the distribution of X (for arbitrary parameter ω) and simulation from the distribution of Z are computationally the same, as Z and X belong to the same family of distributions. Thus, the likelihood function for $Z = z$ has the same computational tractability as the likelihood function for $X = x$. Conversely, if the family is not closed under binomial subsampling, then simulation from the distribution of Z might be a considerably harder task than simulation from the distribution of X, because the obvious, or straightforward, way to proceed is to simulate X, then sample Z from X. The upside here is, of course, that (given sufficient data) we can separate p and ω in the likelihood of $Z = z$. This is not possible if we consider Z with distribution in a family closed under binomial subsampling.

C.W. is supported by a grant from the Danish Cancer Society. M.P.H.S. is a Wellcome Trust Research Fellow. Financial support from the Carlsberg Foundation and the Royal Society is gratefully acknowledged. Oskar Hagberg is thanked for several useful suggestions that improved the presentation.

Appendix A

The following lemma is required.

Lemma A 1. *Let X be a random variable with values in $\mathbb{N}_0$. If $Z \sim \mathrm{Bi}(X, p)$ for some $p \in [0,1]$, then $G_Z(s) = G_X(1-p+ps)$.*

Proof of lemma A 1. If $Z \sim \mathrm{Bi}(X, p)$, then

$$\begin{aligned} G_Z(s) &= \sum_z \sum_{x \geq z} s^z \binom{x}{z} p^z (1-p)^{x-z} \mathbb{P}(X = x) \\ &= \sum_x \mathbb{P}(X = x) \sum_{0 \leq z \leq x} \binom{x}{z} (sp)^z (1-p)^{x-z} \\ &= \sum_x (1-p+ps)^x \mathbb{P}(X = x) = G_X(1-p+ps), \end{aligned} \tag{A 1}$$

and the lemma is proved. ■

Proof of theorem 2.3. Clearly, $\diamond$ is reflexive, and symmetric by definition. To prove transitivity, i.e. if $Q \diamond P$ and $P \diamond R$, then $Q \diamond R$, we distinguish between three possible cases:

(1) $Z\sim\text{Bi}(X_1, p)$ and $X_2\sim\text{Bi}(Y, q)$
(2) $Z\sim\text{Bi}(X_1, p)$ and $Y\sim\text{Bi}(X_2, q)$
(3) $X_1\sim\text{Bi}(Z, p)$ and $X_2\sim\text{Bi}(Y, q)$,

where X_1, $X_2\sim P$, $Y\sim R$ and $Z\sim Q$. (The case $X_1\sim\text{Bi}(Z, p)$ and $Y\sim\text{Bi}(X_2, q)$ is identical to case (1) by symmetry.)

If (1), then we can choose $X_1=X_2$ and Z, such that the conditional distribution of Z given (X_1, Y) does not depend on Y. In consequence, $Z\sim\text{Bi}(Y,pq)$ and $Q\diamond R$.

To prove (2) and (3), assume $q\leq p$ (the proof for $q>p$ is similar).

(2) Define a random variable by $Y'\sim\text{Bi}(Z,q/p)$. Now $G_{X_1}(s)=G_{X_2}(s)$, and hence lemma A 1 gives $G_{Y'}(s)=G_Z(1-q/p+qs/p)=G_{X_1}(1-q+qs)=G_Y(s)$, since $1-q/p+qs/p\geq 0$ for $s\geq 0$; thus, $Y'\sim R$ and in consequence $Q\diamond R$.

(3) Similar calculations as above yield that Z' or Y' defined as in case (2) have $Z'\sim Q$ or $Y'\sim R$, and in consequence $Q\diamond R$. ■

Proof of theorem 2.4. Theorem 2.3 and definition 2.2 define the possible relationships between Q and P. Since $\diamond$ is an equivalence relation all $Q\in\mathcal{C}$ are related to P. Assume as in definition 2.2, such that $Z\sim Q$ and $X\sim P$. If $Z\sim\text{Bi}(X, p)$, then equation (2.3) follows from lemma A 1 with $r=p$. If $X\sim\text{Bi}(Z,p)$, then $G_X(s)=G_Z(1-p+ps)$, or $G_Z(s)=G_X(1-1/p+s/p)$ for $s\geq 1-p$. It has the form of equation (2.3) for $r=1/p$. It remains to be proven that r lies in an interval of the desired form. Assume equation (2.3) is fulfilled for all r in some set J_P. By definitions 2.1 and 2.2, the distribution R of $Y\sim\text{Bi}(Z,p)$ is in $\mathcal{C}$, because $R\diamond Q$ and $Q\in\mathcal{C}$. If $G_Z(s)=G_X(1-r+rs)$, then from lemma A 1,

$$G_Y(s) = G_Z(1-p+ps) = G_X(1-r+r(1-p+ps)) = G_X(1-rp+rps), \quad \text{(A 2)}$$

for $s\geq\max(0,1-1/(pr))$, i.e. $r\in J_P$ implies $r'\in J_P$ for all $r'<r$. Hence, J_P has the desired form and the theorem is proven. ■

Proof of theorem 2.6. Assume $G_r(s)$ defines a pgf of a random variable Z with values in $\mathbb{N}_0$ for some r. Then Y defined by $Y\sim\text{Bi}(Z,p)$ has pgf given by (lemma A 1)

$$\begin{aligned} G_Y(s) &= G_Z(1-p+ps) = G_r(1-p+ps) = G_X(1-r+r(1-p+ps)) \\ &= G_X(1-rp+rps), \end{aligned} \quad \text{(A 3)}$$

for $s\geq\max(0,1-1/(pr))$. Hence, $G_{r'}(s)$ is the pgf of an $\mathbb{N}_0$-valued random variable for all $r'\leq r$. In consequence, there exists a $\rho\geq 0$, such that $G_r(s)$, $s\geq\max(0,1-1/r)$, defines a pgf for all $r<\rho$. If $\rho<\infty$, then

(1) $G_{r_n}(s)\to G_\rho(s)$ for $r_n\to\rho$, $r_n<\rho$, and $s\geq\max(0, 1-1/\rho)$ and
(2) $G_\rho(s)\to G_\rho(1)=1$ for $s\to 1$, $s<1$.

Conditions (1) and (2) are sufficient to prove that $G_\rho(s)$, $s\geq\max(0,1-1/\rho)$ also defines a pgf (e.g. Hoffmann-Jørgensen 1994). It is achieved as the weak limit of $\mathbb{N}_0$-valued random variables; hence, the limit variable is also $\mathbb{N}_0$-valued. If $r\leq 1$, then $Z\sim\text{Bi}(X,r)$ has pgf $G_Z(s)=G_r(s)$. Hence, $\rho\geq 1$. Equation (A 3) also proves that the family of pgfs is closed under binomial subsampling. The proof is completed. ■

Proof of theorem 2.8. A random variable Z with distribution in $\mathcal{P}$ fulfills

$$G_Z(s) = G_X(1-r+rs) = G_{X_1}(1-r+rs)^n. \tag{A 4}$$

If $r \le \rho_1$, then $G_{Z_1}(s) = G_{X_1}(1-r+rs)$ defines a distribution in $\mathcal{Q}$, hence $Z = \sum_{i=1}^{n} Z_i$ has distribution in $\mathcal{P}$ and $\rho_1 \le \rho$. ■

Proof of theorem 3.1. Assume $\mathcal{P}$ is closed under binomial subsampling. Then Z in the class of X has factorial moments given by

$$\mathbb{E}(Z_{[k]}) = \frac{\mathrm{d}^k}{\mathrm{d}s^k} G_Z(s)_{|s=1} = \frac{\mathrm{d}^k}{\mathrm{d}s^k} G_X(1-r+rs)_{|s=1} = r^k \mathbb{E}(X_{[k]}). \tag{A 5}$$

In particular, $\mathbb{E}(Z) = r\mathbb{E}(X)$, and as a consequence T_ϕ must have one of the specified forms, whenever $\phi \in \Phi$ is a parameterization of the orbits of $\mathcal{P}$.

Let $\tau_r = r\tau$ and $fk(\tau, \phi) = \mathbb{E}(X_{[k]})$, then

$$f_k(\tau_r, \phi) = r^k f_k(\tau, \phi), \tag{A 6}$$

and

$$\frac{f_k(\tau_r, \phi)}{\tau_r^k} = \frac{f_k(\tau, \phi)}{\tau^k}. \tag{A 7}$$

For fixed ϕ, vary τ and r, such that $\tau_r = r\tau$ remains constant. This is possible for all $\tau_r \in T_\phi$, because T_ϕ is an interval. It follows that

$$f_k(\tau, \phi) = a_k(\phi)\tau^k, \tag{A 8}$$

for some constant $a_k(\phi)$ depending on k and ϕ only, and that $a_1(\phi) = 1$, as required.

Assume now that the descending moments fulfill the relations in the theorem for $(\tau, \phi) \in \Omega$, and that T_ϕ has one of the specified forms. Then Z, given by $Z \sim \mathrm{Bi}(X, p)$ for X in $\mathcal{P}$, has descending moments

$$\mathbb{E}(Z_{[k]}) = a_k(\phi)(p\tau)^k, \tag{A 9}$$

(similar calculation as above). Z has the same descending moments as the distribution $P_{(p\tau,\, \phi)}$ in $\mathcal{P}$ ($(p\tau, \phi) \in \Omega$, because T_ϕ is an interval by assumption); hence, Z has distribution $P_{(\tau_p, \phi)} = P_{(p\tau, \phi)}$, because the descending moments determine the distribution uniquely (by assumption). It follows that $\mathcal{P}$ is closed under binomial subsampling. ■

Proof of corollary 3.2. Clearly, $\mathbb{E}(X_{[k]}) = 0$ for $k > 0$ defines the degenerated distribution P_0. Hence, $0 \in T$ and T is non-empty. Assume $\tau \in T$ for some $\tau \ge 0$, and that τ determines the distribution of X by $\mathbb{E}(X_{[k]}) = a_k \tau^k$. Then τ^* fulfilling $0 \le \tau^* \le \tau$ determines the distribution of $Z \sim \mathrm{Bi}(X, \tau^*/\tau)$ by $\mathbb{E}(Z_{[k]}) = a_k(\tau^*)^k$. Hence, T is an interval, and the family defined by T is closed under binomial subsampling. ■

Proof of lemma 3.3. See Bollobás (2001). ■

Proof of theorem 4.5. Let a k-order power series family be given. Let $Z\sim\mathrm{Bi}(X,p)$ and $z\geq k$, then

$$\begin{aligned}\mathbb{P}(Z=z) &= \sum_{x\geq z}\binom{x}{z}p^z(1-p)^{x-z}b(x)\lambda^x g(\lambda)^{-1}\\ &= \frac{1}{z!}(p\lambda)^z g(\lambda)^{-1}\frac{\mathrm{d}^z}{\mathrm{d}u^z}\sum_{x\geq k}b(x)u^x_{|u=\lambda(1-p)}\\ &= \frac{1}{z!}(p\lambda)^z g(\lambda)^{-1}\frac{\mathrm{d}^z}{\mathrm{d}u^z}(1-c(u))g(u)_{|u=\lambda(1-p)}. \end{aligned} \tag{A 10}$$

Also

$$\mathbb{P}(Z=z) = b(z)\lambda_p^z g(\lambda_p)^{-1}, \tag{A 11}$$

for some λ_p that depends on λ and p. Hence,

$$\frac{\mathrm{d}^z}{\mathrm{d}u^z}(1-c(u))g(u)_{|u=\lambda(1-p)} = z!b(z)\left(\frac{\lambda_p}{p\lambda}\right)^z\frac{g(\lambda)}{g(\lambda_p)}. \tag{A 12}$$

It follows that $p\lambda/\lambda_p$ and $g(\lambda)/g(\lambda_p)$ depend on λ and p only through $u=\lambda(1-p)$. Hence, we might consider them as functions of u. Put $\zeta(u)=(1-c(u))g(u)$, $f(u)=p\lambda/\lambda_p$, and $h(u)=f(u)^{-k}g(\lambda)/g(\lambda_p)$. Then the zth differential of $\zeta(u)$ can be written as

$$\frac{\mathrm{d}^z}{\mathrm{d}u^z}\zeta(u) = B(z)\frac{1}{f(u)^{z-k}}h(u), \tag{A 13}$$

where $B(z)=z!b(z)$. It follows that $f(0)=h(0)=1$, because $\lambda_p=\lambda$ for $p=1$ and $\lambda(1-p)=\lambda(1-1)=0$. Now

$$\begin{aligned}\frac{\mathrm{d}^{z+1}}{\mathrm{d}u^{z+1}}\zeta(u) &= \frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{\mathrm{d}^z}{\mathrm{d}u^z}\zeta(u)\right] = B(z)\left[-(z-k)\frac{1}{f(u)^{z+1-k}}f'(u)h(u)+\frac{1}{f(u)^{z-k}}h'(u)\right]\\ &= B(z+1)\frac{1}{f(u)^{z+1-k}}h(u),\end{aligned} \tag{A 14}$$

where f' and h' denote the differentials of f and h, respectively. Rewriting equation (A 14) yields

$$B(z)[-(z-k)f'(u)h(u)+f(u)h'(u)] = B(z+1)h(u). \tag{A 15}$$

In particular,

$$B(k)f(u)h'(u) = B(k+1)h(u), \tag{A 16}$$

and hence

$$f'(u) = \frac{-1}{z-k}\left(\frac{B(z+1)}{B(z)}-\frac{B(k+1)}{B(k)}\right), \tag{A 17}$$

which must be independent of z, i.e. $f'(u)=\beta$ for some constant $\beta\in\mathbb{R}$, and hence $f(u)=\beta u+1$, because $f(0)=1$. From equation (A 16)

$$\frac{h'(u)}{h(u)} = \frac{B(k+1)}{B(k)}\frac{1}{f(u)} = \frac{B(k+1)}{B(k)}\frac{1}{\beta u+1}, \tag{A 18}$$

and hence for $\beta = 0$

$$h(u) = \mathrm{e}^{uB(k+1)/B(k)}, \tag{A 19}$$

and $\beta \neq 0$

$$h(u) = (\beta u + 1)^{\beta B(k+1)/B(k)}, \tag{A 20}$$

where it has been used that $h(0) = 1$.

If $\beta = 0$, it follows from equation (A 17) that

$$b(x) = \frac{B(x)}{x!} = \frac{\gamma^{x-k} B(k)}{x!}, \tag{A 21}$$

with $\gamma = B(k+1)/B(k)$. Equation (A 21) has the form of equation (4.9) in the theorem by reparameterization of λ. If $\gamma = 0$, then the distribution can only be positive on $\{0, \ldots, k\}$ (this case is covered by equation (4.7) in the theorem).

If $\beta \neq 0$, then equation (A 17) yields

$$b(x) = \frac{B(x)}{x!} = \frac{\beta^{x-k} \alpha^{[x-k]} B(k)}{x!}, \tag{A 22}$$

$x \geq$k, where α is defined by $B(k+1)/B(k) = \beta\alpha$. If $\alpha > 0$, then also $\beta > 0$, and as a consequence α is an integer, $\alpha \in \mathbb{N}_0$. In this case, equation (A 22) takes the form of equation (4.7) in the theorem. If $\alpha < 0$, then also $\beta < 0$ and

$$b(x) = \frac{B(x)}{x!} = \frac{(-\beta)^{x-k} (-\alpha)_{(x-k)} B(k)}{x!}, \tag{A 23}$$

$x \geq k$. Equation (A 23) has the form of equation (4.8) in the theorem. If $\alpha = 0$, then the distribution is only positive on $\{0, \ldots, k\}$ (this case is covered by equation (4.7) in the theorem). The proof is completed. ■

References

Barndorff-Nielsen, O. E. 1999 *L*-nonformation, *L*-ancillarity and *L*-sufficiency. *Theor. Prob. Appl.* **44**, 225–229.

Bollobás, B. 2001 *Random graphs*. Cambridge: Cambridge University Press.

Felsenstein, J. 2004 *Inferring phylogenies*. Sunderland, MA: Sinauer Associates.

Hoffmann-Jørgensen, J. 1994 *Probability theory with a view towards statistics*. New York: Chapman & Hall.

Koed, K., Wiuf, C., Christensen, L.-L., Wikman, F. P., Zieger, K., Møller, K., Von Der Maase, H. & Ørntoft, T. F. 2005 High-density single nucleotide polymorphism array defines novel stage and location dependent allelic imbalances in human bladder tumors. *Cancer Res.* **65**, 34–45.

Powell, T. M. & Steele, H. J. (eds) 1994. *Ecological time series*. New York: Springer.

Southwood, R. & Henderson, P. 2000 *Ecological methods*. Oxford: Blackwell Science.

Stumpf, M. P. H. & Wiuf, C. 2005 Sampling properties of random graphs: the degree distribution. *Phys. Rev. E* **72**, 036 118. (doi:10.1103/PhysRevE.72.036118)

Stumpf, M. P. H., Wiuf, C. & May, R. M. 2005 Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl Acad. Sci. USA* **102**, 4221–4224. (doi:10.1073/pnas.0501179102)