Statistical and Stochastic Models in Community Ecology

Todd L. Parsons

Laboratoire de Probabilités, Statistique et Modélisation (LPSM, UMR 8001)

Microbial Communities : Current Approaches and Open Challenges
Cambridge, October 12th 2022

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS
DEPUIS 1257

LPSM
LABORATOIRE DE PROBABILITÉS
STATISTIQUE & MODÉLISATION

cnrs
dépasser les frontières

1. Introduction
2. Fisher's Log-series
3. Mechanistic Models I : Demographic Stochasticity
4. Mechanistic Models II : Environmental Stochasticity
5. Preston's Lognormal

Outline

1. Introduction
2. Fisher’s Log-series
3. Mechanistic Models I : Demographic Stochasticity
4. Mechanistic Models II : Environmental Stochasticity
5. Preston’s Lognormal

The Darwinian Orthodoxy

"When we look at the plants and bushes clothing an entangled bank, we are tempted to attribute their proportional numbers and kinds to what we call chance. But how false a view is this!"

– Charles Darwin, *The Origin of Species*

Probabilistic Heresy

Chance enters into the formation of ecological communities in many important ways. For example,

- **Demographic stochasticity :** variation in individual birth and death rates, independent between individuals.
- **Environmental stochasticity :** fluctuations in the environment, experienced by all individuals in a correlated manner.
- **Random dispersal :** organisms arrive in habitats by chance.
- **Random mutation :** novel variation is created at random.
- **Sampling effects :** we observe random samples from a population.

Your Speaker’s Creed

Attributed to John von Neumann by Enrico Fermi :

> *“With four parameters I can fit an elephant, and with five, I can make him wiggle his trunk”*
>
> *Dyson (2004) “A meeting with Enrico Fermi” Nature 427 p. 297*

In (stochastic) modelling, **simple is beautiful**.

In this tutorial, I hope to illustrate some of the tools that we can use to study simple models in the context of a classical question in community ecology.

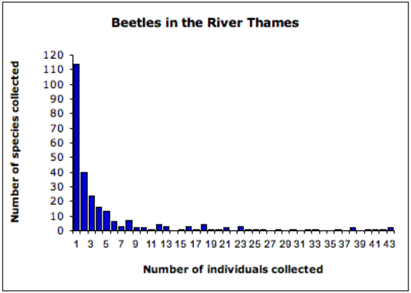
Relative Species Abundance

Perhaps no question in community ecology has commanded more theoretical and empirical attention than that *species diversity* and its two components :

- **Species Richness :** the total number of species in a community, and
- **Species Evenness :** their relative abundance.

Extensive effort goes into sampling even the rarest species, while theorists try to find descriptive statistics that can be fit to reveal qualitative features of data and mechanistic models that predict species abundance from biological principles that can be compared to data.

Relative Abundance Plot



Wikipedia, Magurran (2004) and Williams (1964)

We'll look at (and around) one of the very first efforts to predict and explain these curves.

Outline

1. Introduction
2. Fisher’s Log-series
3. Mechanistic Models I : Demographic Stochasticity
4. Mechanistic Models II : Environmental Stochasticity
5. Preston’s Lognormal

Fisher's Log-series

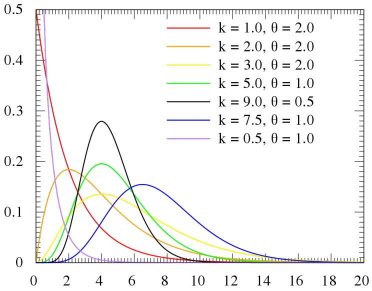
- Fisher, Corbet & Williams (1943) was an early and foundational semi-mechanistic model for species abundances.
- Corbet and Williams had extensively sampled butterflies in the Malay Islands and at the Rothamsted Experimental Station, observing a consistent trend of abundant rare species and relatively rare abundant species. They brought their data to esteemed statistician R. A. Fisher in search of biologically motivated model that would best fit their data.
- In the first mathematical model of species abundance, Fisher would already account for both population variability and sampling effects.

Fisher’s Population Variability

He accounted for population variability by proposing that population abundances would be random, proposing a Gamma distribution :

$$\mathbb{P}\{\text{abundance in } [x, x+dx)\} = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}\, dx$$

This gives a unimodal distribution on $[0, \infty)$.



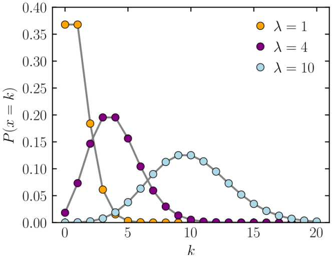
Wikipedia

Fisher’s Sampling Effects

He accounted for variability in sampling by supposing that the number collected would be Poisson distributed with rate proportional to species abundance :

$$\mathbb{P}\{n \text{ sampled}|\text{abundance } x\} = \frac{(\lambda x)^n e^{-\lambda x}}{n!}$$

Poisson samples arise when we observe any given example of a large number of individuals with low probability.



Wikipedia

Fisher's Log-series

- This Poisson-Gamma *mixture* model gives us a negative binomial distribution :

$$\begin{aligned}\mathbb{P}\{n \text{ sampled}\} &= \int_0^\infty \mathbb{P}\{n \text{ sampled}|\text{abundance } x\}\,\mathbb{P}\{\text{abundance in } [x, x+dx)\} \\ &= \frac{1}{\Gamma(k)\theta^k}\int_0^\infty \frac{(\lambda x)^n e^{-\lambda x}}{n!} x^{k-1} e^{-\frac{x}{\theta}}\,dx \\ &= \frac{\Gamma(n+k)}{n!\Gamma(k)}\left(1-\frac{\lambda\theta}{\lambda\theta+1}\right)^k\left(\frac{\lambda\theta}{\lambda\theta+1}\right)^n\end{aligned}$$

- When k is a positive integer, we can understand this as the probability we sample n individuals of a species, assuming we succeed with probability $p = \frac{\lambda\theta}{\lambda\theta+1}$, and stop sampling once we have sampled k individuals not of the target species, but it is a well defined for any $k > 0$.
- The negative binomial is often used to model positive integer valued random variables more dispersed (*i.e.* having a greater variance to mean ratio) than the Poisson distribution.

Fisher's Log-series

- Fisher then reasoned that we could only speak of species that were actually observed, and thus restricted his attention to the probability of observing $n > 0$ individuals.
- Moreover, if the number of species, S, is sufficiently large, then the law of large numbers tells us that number that are observed n times is asymptotically

$$S\mathbb{P}\{n \text{ sampled}\} = S\frac{\Gamma(n+k)}{n!\Gamma(k)}(1-p)^k p^n$$

Fisher's Log-series

- Finally, Fisher reasoned that S was very large, but unobserved, so he took the limit $S \to \infty$.
- But, absent other assumptions, this would lead to an infinite number of individuals sampled, so he further posited that most species had very low abundance, and thus would not be observed.
- Thus, he simultaneously took $k \to 0$, making $x = 0$ the maximal abundance in his gamma model.
- Seeking a nontrivial limit for the number of species observed, Fisher recalled that $\Gamma(k) \sim \frac{1}{k}$, so for $S \gg 1$ and $k \sim 0$, the number of species observed was approximately

$$\frac{\Gamma(n+k)}{n!\Gamma(k)}(1-p)^k p^n \sim Sk\frac{\Gamma(n-1)}{n!}p^n = Sk\frac{p^n}{n}.$$

- He thus assumed that Sk tended to a constant in the simultaneous limit : $\lim_{S\to\infty,k\to 0} Sk = \alpha$

Fisher's Log-series

- The resulting distribution is now known as Fisher's log-series, since

$$-\log(1-p) = \sum_{n=1}^{\infty} \frac{p^n}{n},$$

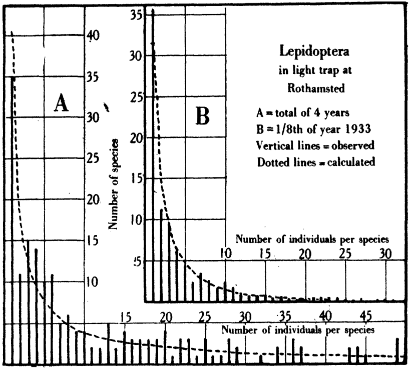
  so the number of species observed at least once is $S' = -\alpha \log(1-p)$.
- If our sampling effort is sufficient, we would expect that $S \propto S' \propto \alpha$.
- The number of individuals observed is

$$N = \alpha \sum_{n=1}^{\infty} n \frac{p^n}{n} = \frac{\alpha p}{1-p}.$$

  These two equations can be (numerically) solved to give us estimates for α and p.

Fisher’s Log-series

Fisher’s model showed excellent fits to Corbet and Williams’ data :



Fisher, Corbet & Williams (1943).

Fisher’s α has become a commonly used measure of species richness.

Fisher’s Log-series

However. . . Fisher’s modelling is based upon a number of reasonable, but not *a priori* justified assumptions :

- A large number of species,
- most of which are vanishingly rare,
- **whose abundance is gamma distributed.**

Outline

1. Introduction
2. Fisher's Log-series
3. Mechanistic Models I : Demographic Stochasticity
4. Mechanistic Models II : Environmental Stochasticity
5. Preston's Lognormal

Two Paths to Fisher’s Gamma (and One Direct to his Negative Binomial)

- Fisher chose a gamma distribution with *shape* $k > 0$ and *scale* $\theta > 0$ as his population model.
- Why ? Because for him, it had desired properties (only taking positive real values, having a mean and mode that he could adjust).
- Is there a mechanistic species model, based on biological processes like birth, death, and migration, that can give us Fisher’s gamma ?
- What are the biological interpretations of θ and k or α and p ? What is the **biological** interpretation of the limit $k \to 0$?

A Stochastic Model for Species **Number**

- About the simplest possible stochastic model is a *Markovian* birth and death process.
- Formulated in continuous time, we assume that individuals independently give birth and die at per-capita rates b and d :

$$\mathbb{P}\{\text{a given individual gives birth in } [t, t+\Delta t\} = b\Delta t + o(\Delta t)$$

- Let $N(t)$ denote the the number of individuals alive at time t.
- When $N(t) = n$, the population increases by 1 at rate $q_{n,n+1} = bn$ and decreases by 1 at rate $q_{n,n-1} = dn$. These are the *transition rates*.

Simulating (and Constructing) the Markov Model

- How would we simulate this model ?
- In practice, we can't deal with infinitesimals dt, so we could discretize time into small segments Δt.
- Suppose $N(t) = n$. Provided $\Delta t < \min\left\{\frac{1}{bn}, \frac{1}{dn}\right\}$, then we have a birth in $[t, t + \Delta t)$ with **probability** $bn\Delta t$, a death with probability $dn\Delta t$, and nothing happens with probability $1 - (b + d)n\Delta t$.
- Obvious shortcomings : Δt has to be very small if n is large, and if $\Delta t \ll \frac{1}{n}$, then one has to wait a long time for events.

Simulating (and Constructing) the Markov Model

- Consider the waiting time, T, until the next event if $N(t) = n$.
- Starting with discretized time,

$$\mathbb{P}\{T > t\} = (1 - (b+d)n\Delta t)^{\frac{t}{\Delta t}} .$$

so, as $\Delta t \to 0$ (*i.e.*, passing to the infinitesimal limit),

$$\mathbb{P}\{T > t\} \to e^{-(b+d)nt},$$

so the waiting time is *exponentially distributed* with rate $(b+d)n$ and mean $\frac{1}{(b+d)n}$.

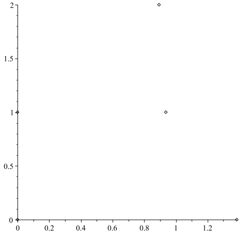
Simulating (and Constructing) the Markov Model

- Next, we know that nothing happens until time $t+T$.
- At this time, a birth happens with probability **proportional** $bn\Delta t$, a death happens with probability **proportional** $bn\Delta t$, and one or the other must happen, so the probability a birth or death occurs is

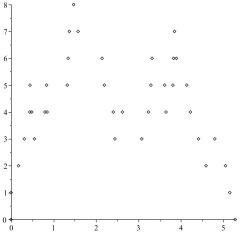
$$\frac{bn\Delta t}{bn\Delta t + dn\Delta t} = \frac{b}{b+d} \quad \text{and} \quad \frac{dn\Delta t}{bn\Delta t + dn\Delta t} = \frac{d}{b+d}$$

  respectively.
- The discrete time Markov chain that keeps the transitions, but ignores the waiting time between events, is called the *skeleton*.

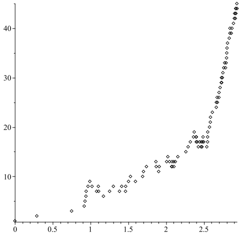
Simulating (and Constructing) the Markov Model



Subcritical : $b = \frac{1}{2}, d = 1$



Critical : $b = 1, d = 1$



Supercritical : $b = 2, d = 1$

Gillespie's Algorithm

Although the construction considerably predates Gillespie (1976), this is commonly called Gillespie's algorithm :

- Assume a continuous-time Markov chain on an arbitrary state space $\mathscr{S}$ with transition rates $q_{x,y}$ from state $x \in \mathscr{S}$ to state $y \in \mathscr{S}$.
- Pick an arbitrary initial state x_0 at time $t = 0$.
- Suppose that the chain is in state x at time t. The waiting time T to the next event is exponentially distributed with rate $\sum_z q_{x,z}$.
- At time $t + T$ the process jumps to state y with probability $\frac{q_{x,y}}{\sum_z q_{x,z}}$.
- Repeat.
- Many, many refinements for speed, large populations, *e,g,* τ-leaping, adaptive τ-leaping,…

Shortcomings of the Birth and Death Process (For Our Purposes)

- The three cases simulated are characteristic : if $b < d$, the population quickly goes extinct ; if $b = d$, it goes extinct, but slowly ; if $b > d$, it can grow (exponentially) indefinitely.
- None of these are conducive to having a distribution of population sizes (0 or ∞ are not terribly useful).
- We’ll can fix this by assuming source-sink dynamics :
  - Assume $d > b$, so the population isn’t self-sustaining
  - Assume that the population is sustained by immigration from a larger source population.
- These are conjectured to be the population dynamics sustaining many island communities.
- But first, let’s examine my claim above.

An Aside on Exponential Growth

- Consider the expected size of the population, or more precisely, it’s change in a small time step Δt :

$$\begin{aligned}\mathbb{E}\left[N(t+\Delta t)\right] &= \mathbb{E}\left[\mathbb{E}\left[N(t+\Delta t)|N(t)\right]\right] \\ &= \mathbb{E}\left[(N(t)+1)bN(t)\Delta t + (N(t)-1)dN(t) + (1-(b+d)\Delta t)N(t)\right] \\ &= ((b-d)\Delta t + 1)\,\mathbb{E}\left[N(t)\right],\end{aligned}$$

- Rearranging gives the Newton quotient :

$$\frac{\mathbb{E}\left[N(t+\Delta t)\right]-\mathbb{E}\left[N(t)\right]}{\Delta t} = (b-d)\mathbb{E}\left[N(t)\right],$$

- Taking $\Delta t \to 0$ gives us an ODE : $\frac{d}{dt}\mathbb{E}\left[N(t)\right] = (b-d)\mathbb{E}\left[N(t)\right]$.
- Solving gives $\mathbb{E}\left[N(t)\right] = N(0)e^{(b-d)t}$, so exponential growth or decay as $b > d$ or $b < d$, whereas $\mathbb{E}\left[N(t)\right] = N(0)$ for all $t \geq 0$.

An Aside on Extinction

- In the simulated populations, those with $b \leq d$ went extinct
- This is easily shown via the skeleton. Let q be the probability of extinction starting from a single individual.
- Consider a population with $N(0) = 1$, and consider the first event that happens, birth or death :

$$q = \mathbb{P}\{\text{death}\} + \mathbb{P}\{\text{birth}\}q^2 = \frac{d}{b+d} + \frac{b}{b+d}q^2.$$

- Either the ancestor dies, or it and it's offspring both give rise to independent birth and death processes starting from a single individual. Extinction occurs if both populations go extinct.
- Solving this quadratic, we find that $q = 1$ is always a solution, and if $b > d$ $q = \frac{d}{b}$.
- Thus, extinction is certain if $b \leq d$ (despite $\mathbb{E}[N(t)] = 1$ for all $t \geq 0$ is $b = d$).

A (Last) Aside on Indefinite Growth

- What about $b > d$? Look at the maximum M of the birth and death process.
- Fix m and let T_m be the first time that $N(t) = m$. $M > m$ if and only if $T_m < T_0$.
- Let $p_{n,m}$ be the probability that, starting from n, $T_m < T_0$.
- Clearly, $p_{0,m} = 0$ and $p_{m,m} = 1$.
- In between, using the skeleton : $p_{n,m} = \frac{b}{b+d} p_{n+1,m} + \frac{d}{b+d} p_{n-1,m}$.
- Rearranging, $p_{n+1,m} - p_{n,m} = \frac{d}{b}(p_{n,m} - p_{n-1,m})$.
- Iterating, $p_{n,m} - p_{n-1,m} = \cdots = \left(\frac{d}{b}\right)^{n-1} (p_{1,m} - p_{0,m}) = \left(\frac{d}{b}\right)^{n-1} p_{1,m}$.
- Summing, $p_{n,m} - p_{0,m} = \sum_{i=1}^{n} p_{i,m} - p_{i-1,m} = p_{1,m} \sum_{i=1}^{n} \left(\frac{d}{b}\right)^{n-1} = p_{1,m} \frac{1-\left(\frac{d}{b}\right)^n}{1-\frac{d}{b}}$.
- Lastly, $1 = p_{m,m} = p_{1,m} \frac{1-\left(\frac{d}{b}\right)^m}{1-\frac{d}{b}}$ gives us $p_{1,m}$ and thus

$$p_{n,m} = \frac{1-\left(\frac{d}{b}\right)^n}{1-\left(\frac{d}{b}\right)^m} \to 1 - \left(\frac{d}{b}\right)^n.$$

as $m \to \infty$.

A Birth, Death, and Migration Process

- Suppose that in addition to births and deaths as previously, we allow migrants to arrive at rate ν.
- Our transition rates are now $q_{n,n+1} = bn + \nu$ and $q_{n,n-1} = dn$.
- Migration prevents extinction.
- Proceeding as above, we find $\frac{d}{dt}\mathbb{E}[N(t)] = (b-d)\mathbb{E}[N(t)] + \nu$, with solution

$$\mathbb{E}[N(t)] = e^{(b-d)t}N(0) + \nu \frac{e^{(b-d)t} - 1}{b-d}.$$

- If $d > b$, this converges to $\frac{1}{d-b}$ as $t \to \infty$, so no indefinite growth.
- What is the behaviour of the stochastic model ?

The Master Equation

- If we knew the probability distribution $p_n(t) = \mathbb{P}\{N(t) = n\}$, we would have a complete description of the process.
- We'll attack it again, by looking at it's change over a time step Δt :

$$p_n(t+\Delta t) = (b(n-1)+\nu)\Delta t p_{n-1}(t) + d(n+1)\Delta t p_{n+1}(t) + (1-(bn+\nu+dn)\Delta t)p_n(t)$$

- Just as before, we can rearrange this to get a Newton quotient on the right hand side, and take $\Delta t \to 0$ to get a system of ODEs :

$$\frac{dp_n}{dt} = (b(n-1)+\nu)p_{n-1}(t) + d(n+1)p_{n+1}(t) - (bn+\nu+dn)p_n(t).$$

- We could solve this, but in general, it gives only limited insight. Still, it's very useful.

The Stationary Distribution

- To posit, as Fisher did, time-independent distribution of species abundances, we need to ignore transient behaviour and consider

  $$\pi(n) = \lim_{t\to\infty} p_n(t)$$

  (if it exists).
- The existence of this *stationary distribution* does not imply that the process stops changing.
- Rather, as we observe the process at different times, it will be in different states, but the probability of being in those states (and thus the fraction of times we will observe any given state) remain constant.
- With time dependence gone, the derivative in the Master Equation disappears, leaving a system of linear equations for the $\pi(n)$:

  $$(b(n-1)+\nu)\pi(n-1)+d(n+1)\pi(n+1) = (bn+\nu+dn)\pi(n)$$

Detailed Balance

- Rather than try and solve this infinite system of linear equations, we can make a simplifying observation :

$$\begin{aligned}&(b(n-1)+\nu)\pi(n-1)+d(n+1)\pi(n+1)-(bn+\nu+dn)\pi(n)\\ =&((b(n-1)+\nu)\pi(n-1)-dn\pi(n))-((bn+\nu)\pi(n)-d(n+1)\pi(n+1)).\end{aligned}$$

- This will equal zero provided for each $n \geq 1$,

$$(b(n-1)+\nu)\pi(n-1) = dn\pi(n)$$

- These are known as *detailed balance relations* : the rate of flow from state x to state y balances the flow from y to x.
- A stationary distribution can still exist when detailed balance fails, but it holds in almost every example where we can easily compute the stationary distribution.

Solving Detailed Balance

- Rearranging the detailed balance condition gives us $\frac{\pi(n)}{\pi(n-1)} = \frac{b(n-1)+\nu}{dn}$. Mutliplying,

$$\frac{\pi(n)}{\pi(0)} = \frac{\pi(n)}{\pi(n-1)} \frac{\pi(n-1)}{\pi(n-2)} \cdots \frac{\pi(2)}{\pi(1)} \frac{\pi(1)}{\pi(0)} = \prod_{m=1}^{n} \frac{b(m-1)+\nu}{dm}.$$

- Still must determine $\pi(0)$: $\pi(n)$ is a probability distribution, so

$$1 = \sum_{n=0}^{\infty} \pi(n) = \pi(0) \sum_{n=0}^{\infty} \prod_{m=1}^{n} \frac{b(m-1)+\nu}{dm}$$

- And we can recognize the sum as a binomial series :

$$\sum_{n=0}^{\infty} \prod_{m=1}^{n} \frac{b(m-1)+\nu}{dm} = \sum_{n=0}^{\infty} \frac{\prod_{m=1}^{n} \frac{\nu}{b} + m - 1}{n!} \left(\frac{b}{d}\right)^{n}$$

$$= \sum_{n=0}^{\infty} \frac{\Gamma\left(\frac{\nu}{b}+n\right)}{n!\Gamma\left(\frac{\nu}{b}\right)} \left(\frac{b}{d}\right)^{n} = \left(1 - \frac{b}{d}\right)^{-\frac{\nu}{b}}$$

A Negative Binomial Stationary Distribution

- Putting all of the above together, we see that the stationary distribution is negative binomially distributed with $k = \frac{\nu}{b}$ and $p = \frac{b}{d}$:

$$\pi(n) = \frac{\Gamma\left(\frac{\nu}{b} + n\right)}{n!\Gamma\left(\frac{\nu}{b}\right)} \left(1 - \frac{b}{d}\right)^{\frac{\nu}{b}} \left(\frac{b}{d}\right)^{n}.$$

- Parameters now have biological meaning : taking $k \to 0$ means considering a very low migration rate. The success rate p is

$$b \times \frac{1}{d} = \text{birth rate} \times \mathbb{E}[\text{lifespan}] = \mathbb{E}[\text{total reproductive output}] = \text{fitness}$$

- Unlike Fisher's, however, this doesn't account for sampling effects.
- We can, however, account for these by assuming that each extant species is sampled with probability q.
- The resulting distribution is most easily characterized by using its probability generating function (pgf).

Probability Generating Functions

- The probability generating function of a random variable, say N, taking non-negative integer values is

$$G(z) = \mathbb{E}[z^N] = \sum_{n=0}^{\infty} \mathbb{P}\{N = n\} z^n,$$

  where z is a dummy variable.
- It's also a handy way to calculate moments :

$$\mathbb{E}\left[N^m\right] = \left(z \frac{d}{dz}\right)^m G(z),$$

- When it exists, it uniquely characterizes the probability distribution, as

$$\left.\frac{d^n}{dz^n}\right|_{z=0} G(z) = n! \mathbb{P}\{N = n\}.$$

(Motivated) Examples of Probability Generating Functions

- Using the binomial series as above, one finds that the negative binomial with rates r and p has pgf

$$G_{k,p}(z) = \left(\frac{1-p}{1-pz} \right)^k .$$

- Now, let B be a Bernoulli random variable with success probability q, *i.e.* $\mathbb{P}\{B=1\} = q$, $\mathbb{P}\{B=0\} = 1-q$. This has pgf

$$G_q(z) = (1-q) + qz.$$

The Sampled Negative Binomial Distribution

- Now, suppose a species has N individuals (N is a NB(k,p) random variable), and suppose that we labelled them $i = 1, \ldots, N$ and let B_i be a Bernoulli random variable which is 1 if we observe i (so with probability q).
- Then, the number observed is $N' = \sum_{i=1}^{N} B_i$.
- Now, assume N is known. N' has pgf

$$\mathbb{E}\left[z^{N'}\right] = \mathbb{E}\left[z^{\sum_{i=1}^{N} B_i}\right] = \prod_{i=1}^{N} \mathbb{E}\left[z^{B_i}\right] = ((1-q)+qz)^N.$$

- Now, if N is unknown, but negatively binomially distributed. We have to take the expectation over N as well :

$$\mathbb{E}\left[((1-q)+qz)^N\right] = G_{k,p}((1-q)+qz) = \left(\frac{1-\frac{pq}{1-p+pq}}{1-\frac{pq}{1-p+pq}z}\right)^k,$$

so still negative binomial, but with p replaced by $\frac{pq}{1-p+pq}$

FIsher's Gamma

- But what about the gamma distribution for the abundance ?
- We can recover it in the limit : recall

$$\pi(n) = \frac{\Gamma(k+n)}{n!\Gamma(r)} (1-p)^k (p)^n .$$

- $r = \frac{\nu}{b}, p = \frac{b}{d} = 1 - s < 1$ is individual fitness. Suppose that $s \ll 1$.

$$\pi(n) = \frac{\Gamma(k+n)}{n!\Gamma(k)} s^k e^{n \ln(1-s)} \sim \frac{\Gamma(k+n)}{n!\Gamma(k)} s^k e^{-ns}$$

- Simultaneously assume that $n \gg 1$. Then, using Stirling's approximation, $\Gamma(k+n) \sim \Gamma(k) n^k$,

$$\pi(n) \sim \frac{1}{\Gamma(k)} (ns)^k e^{-(ns)}$$

- So, if we let x measure n in units of $\frac{1}{s}$, $x = ns$, we get Fisher's gamma distribution.

Outline

1. Introduction
2. Fisher's Log-series
3. Mechanistic Models I : Demographic Stochasticity
4. Mechanistic Models II : Environmental Stochasticity
5. Preston's Lognormal

Density Dependence

- We saw that keeping the population finite, but non-zero, was necessary to speak of an abundance distribution.
- If we assumed source-sink dynamics sustained by migration, we could satisfy that demand.
- Another way to keep a population finite is to impose density dependence via a carrying capacity, say K.
- We can express this as a birth and death process $N(t)$ with state dependent birth and/or death dates, *e.g.*

$$q_{n,n+1} = bn \quad \text{and} \quad q_{n,n-1} = dn\left(1 + \frac{n}{K}\right)$$

- Unfortunately, like our birth and death processes with $b \leq d$, this eventually goes extinct with probability 1 (see *e.g.* Parsons (2018) for a relatively elementary proof).
- But, if $b > d$, the process takes a **long** time to go extinct, on the order of e^{cK} for $c > 0$, so if $K \gg 1$ it has a *quasi-stationary distribution*.

$K \gg 1$ Approximation

- Unlike previously, there's no stationary distribution to be found by detailed balance.
- The Master Equation still proves useful to analyze the problem. Recall, $p_n(t) = \mathbb{P}\{N(t) = n\}$

$$\frac{dp_n}{dt} = b(n-1)p_{n-1}(t) + d(n+1)\left(1 + \frac{n+1}{K}\right)p_{n+1}(t) - \left(bn + dn\left(1 + \frac{n}{K}\right)\right)p_n(t).$$

- Make a change of variable $x = \frac{n}{K}$ (*i.e.* replace $N(t)$ by $\frac{N(t)}{K}$), and set $p(x,t) = p_{Kx}(t) = \mathbb{P}\{X(t) = x\}$ and Taylor expand the result :

$$\begin{aligned}\frac{\partial p(x,t)}{\partial t} = & \, Kb\left(x - \tfrac{1}{K}\right)p\left(x - \tfrac{1}{K}, t\right) + Kd\left(x + \tfrac{1}{K}\right)\left(1 + x + \tfrac{1}{K}\right)p\left(x + \tfrac{1}{K}, t\right) \\ & - \left(b + d(1+x)\right)Kxp(x,t) \\ = & -\frac{\partial}{\partial x}\left[x\left(b - d(1+x)\right)p(x,t)\right] + \frac{1}{2K}\frac{\partial^2}{\partial x^2}\left[x\left(b + d(1+x)\right)p(x,t)\right] + \mathcal{O}\left(\tfrac{1}{K^2}\right).\end{aligned}$$

A Transport Equation

- If we take $K \to \infty$, we get a *transport equation* :

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[x\left(b-d(1+x)\right)p(x,t)\right].$$

- This tells us that if $p(x,0) = \delta_{x_0}(x)$ (*i.e.* $\mathbb{P}\{X(0) = x_0\} = 1$), then $X(t)$ is just the solution of the logistic ODE :

$$\frac{dX(t)}{dt} = X(t)(b-d-dX(t)) = rX(t)\left(1 - \frac{X(t)}{\kappa}\right)$$

  for $r = b-d$ and $\kappa = 1 - \frac{b}{d}$.

- More precisely, $p(x,t) = \delta_{X(t)}(x)$.
- **To think about :** $X(t) = \frac{N(t)}{K}$, so if we take $K \to \infty$, $X(t) = 0$ unless we assume that $N(t) \propto K$

Adding Environmental Stochasticity

- In this $K \gg 1$ limit, we have a finite, persistent population, but no variability.
- (If we kept terms of order $\frac{1}{K}$, we would have had demographic stochasticity, but that's a story for another day).
- But, what if populations are subject to environmental variability ? This is commonly approached by replacing the logistic ODE with a logistic SDE with multiplicative noise :

$$dX(t) = rX(t)\left(1 - \frac{X(t)}{\kappa}\right) dt + \sigma X(t)\, dB(t)$$

- This is a **phenomenological** approach to noise, and not without it's issues (forthcoming paper in *Oikos* proposing a more mechanistic alternative approach).
- Why is this called *multiplicative* ? The factor $X(t)$ multiplying $dB(t)$. It's role is to ensure $X(t) \geq 0$.

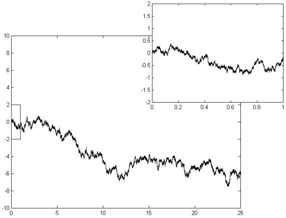
Brownian Motion aka Wiener Process

$B(t)$ is Brownian motion, it’s defining characteristics are that :

- $B(t)$ is a continuous, random function, and
- $\Delta B(t) = B(t+\Delta t) - B(t)$ is a normally distributed random variable with mean 0 and variance Δt :

$$\mathbb{P}\{\Delta B(t) \in [x, x+dx)\} = \frac{1}{\sqrt{2\pi\Delta t}} e^{-\frac{x^2}{2\Delta t}}.$$

- Non-overlapping increments $\Delta B(t_1)$, $\Delta B(t_2)$ are independent.
- *Standard Brownian motion* has initial condition $B(0) = 0$.



Wikipedia

Stochastic Differential Equations : Understanding and Simulating Them

- One can say a lot about SDEs, but like our Markov processes, we can get a handle on them for practical purposes by discretizing time.
- We numerically integrate ODEs via Euler's algorithm : if $\frac{dX}{dt} = b(X(t))$ and $X(0) = x_0$, then we iteratively evaluate the ODE :

$$X(t+\Delta t) = X(t) + b(X(t))\Delta t.$$

- We numerically integrate a *sample path* of the SDE $dX(t) = b(X(t))\,dt + \sigma(X(t))\,dB(t)$ with $X(0) = x_0$, via the *Euler-Maruyama* algorithm : we iteratively evaluate

$$X(t+\Delta t) = X(t) + b(X(t))\Delta t + \sigma(X(t))\Delta B(t),$$

  independently drawing a new $N(0,\Delta t)$ random variable $\Delta B(t)$ at each time step.
- **Important :** I'm only talking about Itô SDEs.

Stochastic Differential Equations : Itô Integral

- Why the notation $dX(t) = b(X(t))\,dt + \sigma(X(t))dB(t)$?
- It is meant to evoke derivatives, without being derivatives : $\frac{dB(t)}{dt}$ **doesn't exist** (Paley, Wiener and Zygmund 1933) because $B(t)$ is too rough to have well defined tangents.
- The integral $\int_0^t f(X(t))dB(t)$ does exist, and is defined analogously to the Riemann-Stieltjes integral :

$$\int_0^t f(X(t))dB(t) = \lim_{\Delta t \to 0} \sum_{i=1}^{n} f(X(t_i))\Delta B(t_i).$$

- Just as we can interpret a differential equation $\frac{dX}{dt} = b(X(t))$ as an integral equation :

$$X(t) - X(0) = \int_0^t \frac{dX}{ds}\,ds = \int_0^t b(X(s))\,ds,$$

  the SDE $dX(t) = b(X(t))\,dt + \sigma(X(t))dB(t)$ is properly understood as an integral equation :

$$X(t) - X(0) = \int_0^t dX(t) = \int_0^t b(X(s))\,ds + \int_0^t b(X(s))\sigma(X(t))\,dB(s).$$

The Fokker-Planck Equation

- Just as the Master Equation characterizes the probability distribution function for a Markov chain, the Fokker-Planck equation characterizes the *probability density function* $p(x,t)$ for the solution of an SDE $dX(t) = b(X(t))\,dt + \sigma(X(t))\,dB(t)$:

$$\mathbb{P}\{X(t) \in [x, x+dx)\} = p(x,t)\,dx.$$

- $p(x,t)$ satisfies

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\left[b(x)p(x,t)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma(x)^2 p(x,t)\right]$$

- If $X(t)$ approaches a stationary distribution, $\lim_{t\to\infty} p(x,t) = \pi(x)$, we can similarly obtain it via the Fokker-Planck equation :

$$0 = -\frac{\partial}{\partial x}\left[b(x)\pi(x)\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma(x)^2 \pi(x)\right]$$

Zero-flux Solution

- We can write our equation for the stationary distribution as

$$0 = \frac{\partial}{\partial x}\left[-b(x)\pi(x) + \frac{1}{2}\frac{\partial}{\partial x}\left[\sigma(x)^2\pi(x)\right]\right]$$

- For finite t, $J(x,t) = -b(x)p(t,x) + \frac{1}{2}\frac{\partial}{\partial x}\left[\sigma(x)^2 p(t,x)\right]$ is called the flux.
- Analogous to detailed balance relations, we can find the stationary distribution as a *zero-flux* solution, $J(x,t) = 0$.
- Solving this gives

$$\pi(x) = \frac{C}{\sigma(x)^2} e^{-2\int \frac{b(x)}{\sigma(x)^2}\,dx},$$

and a stationary solution exists if and only if there exists a non-zero normalizing constant C that ensures that $\pi(x)$ integrates to 1.

Another Path to Fisher's Gamma

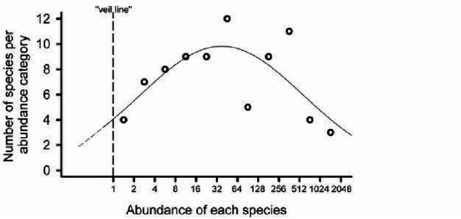
- For our logistic equation with environmental noise, $dX(t) = rX(t)\left(1 - \frac{X(t)}{\kappa}\right)dt + \sigma X(t)\,dB(t)$, we get
$$\pi(x) = \frac{\left(\frac{2r}{\sigma^2\kappa}\right)^{\frac{2r}{\sigma^2}-1}}{\Gamma\left(\frac{2r}{\sigma^2}-1\right)} x^{\frac{2r}{\sigma^2}-2} e^{-\frac{2r}{\sigma^2\kappa}x},$$
provided $r > \frac{\sigma^2}{2}$.
- Otherwise, the normalizing constant – and the stationary distribution – don't exist : noise can drive the population to extinction.
- This gives another biologically motivated derivation – based on very different mechanisms – for Fisher's Gamma.
- **An object lesson :** pattern doesn't imply process.

Outline

1. Introduction
2. Fisher’s Log-series
3. Mechanistic Models I : Demographic Stochasticity
4. Mechanistic Models II : Environmental Stochasticity
5. Preston’s Lognormal

Preston's Lognormal

- Unfortunately, Preston (1948) found that Fisher's log series was a poor fit to bird and moth data sets.
- Plotting numbers numbers of species against their abundance on a logarithmic scale, Preston found the bell-shaped curves :



Preston (1948).

- He argued that inadequate sampling concealed the decline in the number of exceedingly rare species.
- Such lognormal distributions have since been shown to be ubiquitous across diverse species.
- Let's make one last effort to find a lognormal model of species abundance.

Gompertz Growth

- In keeping with our last example, let's consider density dependent growth with environmental noise.
- Fits to data often improve on the logistic equation by making density dependence non-linear :

$$\frac{dX(t)}{dt} = rX(t)\left(1 - \left(\frac{X(t)}{\kappa}\right)^{\theta}\right).$$

- If we take $r = \frac{\rho}{\theta}$, then in the limit as $\theta \to 0$, we get Gompertz's growth equation :

$$\frac{dX(t)}{dt} = \rho X(t)\ln\left(\frac{\kappa}{X(t)}\right).$$

- Let's consider this with multiplicative environmental noise :

$$dX(t) = \rho X(t)\ln\left(\frac{\kappa}{X(t)}\right)dt + \sigma X(t)\,dB(t).$$

- We end up with one of the relatively few examples of an SDE that can be solved exactly (Engen and Lande, 1996). To do so, use Itôs Stochastic Calculus.

Itô Calculus

- We've already said that *stochastic differentials* $dX(t)$ are analogous to derivatives.
- An important similarity with derivatives is *Itô's chain rule* : if $dX(t) = b(X(t))\,dt + \sigma(X(t))dB(t)$ and $Y(t) = f(X(t))$, then

$$\begin{aligned} dY(t) &= f'(X(t))dX(t) + \frac{1}{2}f''(X(t))\sigma(X(t))^2\,dt \\ &= \left(f'(X(t))b(X(t)) + \frac{1}{2}f''(X(t))\sigma(X(t))^2\right)dt + f'(X(t))\sigma(X(t))dB(t). \end{aligned}$$

- Informally this arises from Taylor expansion if we assume that $dB(t)dB(s)`` ='' \delta(t-s)dt$ (recall, $\Delta B(t) = N(0,\Delta t)$, so $\mathbb{E}\left[(\Delta B(t))^2\right] = \Delta t)$:
- There's also an analogue to the product rule : if, for $i = 1,2$, $dX_i(t) = b_i(X_i(t))\,dt + \sigma_i(X_i(t))dB(t)$, then

$$d\left(X_1(t)X_2(t)\right) = X_1(t)dX_2(t) + X_2(t)dX_1(t) + \sigma_1(X_1(t))\sigma_2(X_2(t))\,dt$$

- We can (sometimes, not often) use these to solve SDEs.

Solving the Noisy Gompertz Equation

- Let $Y(t) = \ln X(t)$. Using Ito's chain rule,

$$\begin{aligned} dY(t) &= \frac{1}{X(t)}\left(\rho X(t)\ln\left(\frac{\kappa}{X(t)}\right) dt + \sigma X(t)\, dB(t)\right) - \frac{1}{2}\frac{1}{X(t)^2}\sigma^2 X(t)^2\, dt \\ &= \rho\left(\ln\kappa - \frac{\sigma^2}{2} - Y(t)\right) dt + \sigma\, dB(t). \end{aligned}$$

- This linear SDE is known as the Ornstein-Uhlenbeck process, and can be solved exactlty. Let $Z(t) = e^{\rho t}Y(t)$. Using Itô's product rule :

$$dZ(t) = e^{\rho t}\left(\rho Y(t)\, dt + dZ(t)\right) = \rho e^{\rho t}\left(\ln\kappa - \frac{\sigma^2}{2}\right) dt + e^{\rho t}\sigma\, dB(t)$$

- Solve by integrating :

$$Z(t) = Z(0) + \frac{\ln\kappa - \frac{\sigma^2}{2}}{\rho}\left(e^{\rho t} - 1\right) + \sigma\int_0^t e^{\rho s}\, dB(s)$$

$$Y(t) = e^{-\rho t}Y(0) + \left(\ln\kappa - \frac{\sigma^2}{2}\right)\left(1 - e^{\rho t}\right) + \sigma\int_0^t e^{-\rho(t-s)}\, dB(s)$$

Interpreting the Stochastic Integral

- Recall the meaning of the stochastic integral :

$$\int_0^t e^{-\rho(t-s)}\,dB(s) = \lim_{\Delta t\to 0}\sum_{i=1}^{n} e^{-\rho(t-t_i)}\Delta B(t_i).$$

- It is a sum of independent mean-zero normal random variables $\Delta B(t_i)$, and thus also a mean zero normal random variate.
- Let's compute it's variance :

$$\begin{aligned}\mathbb{E}\left[\left(\int_0^t e^{-\rho(t-s)}\,dB(s)\right)^2\right] &= \mathbb{E}\left[\left(\int_0^t e^{-\rho(t-s)}\,dB(s)\right)\left(\int_0^t e^{-\rho(t-u)}\,dB(u)\right)\right]\\ &= \mathbb{E}\left[\int_0^t\int_0^t e^{-\rho(t-s)}e^{-\rho(t-u)}\,dB(s)\,dB(u)\right]\\ &= \mathbb{E}\left[\int_0^t e^{-2\rho(t-s)}\,ds\right] = \frac{1}{2\rho}\left(1-e^{-2\rho t}\right).\end{aligned}$$

Interpreting $X(t)$ and $Y(t)$

- Recall $Y(t) = e^{-\rho t} Y(0) + \left(\ln\kappa - \frac{\sigma^2}{2}\right)\left(1 - e^{\rho t}\right) + \sigma \int_0^t e^{-\rho(t-s)}\, dB(s)$
- We just saw that the stochastic integral term is a normal distribution with mean 0 and variance $\frac{\sigma^2}{2\rho}\left(1 - e^{-2\rho t}\right)$.
- $Y(t)$ has mean $e^{-\rho t} Y(0) + \left(\ln\kappa - \frac{\sigma^2}{2}\right)\left(1 - e^{\rho t}\right)$
- As $t \to 0$, the mean and variance converge to $\ln\kappa - \frac{\sigma^2}{2}$ and $\frac{\sigma^2}{2\rho}$.
- Thus, $X(t)$ is lognormally distributed, with a stationary distribution with mean $e^{\ln\kappa - \left(1 - \frac{1}{\rho}\right)\frac{\sigma^2}{2}} = \kappa e^{-\left(1 - \frac{1}{\rho}\right)\frac{\sigma^2}{2}}$ and variance
$$\left(e^{\frac{\sigma^2}{2\rho}} - 1\right) e^{2\ln\kappa - \left(1 - \frac{1}{2\rho}\right)\sigma^2}.$$
- So, logistic growth plus environmental noise can also explain Preston's lognormal.