

A Bayesian theory for estimation of biodiversity

SIS Bayes 2025

Tommaso Rigon

Università degli Studi di Milano-Bicocca

2025-09-05

[Home page](#)

Warm thanks



Ching-Lung Hsu
(*Duke University*)



Alessandro Zito
(*Harvard School of Public Health*)



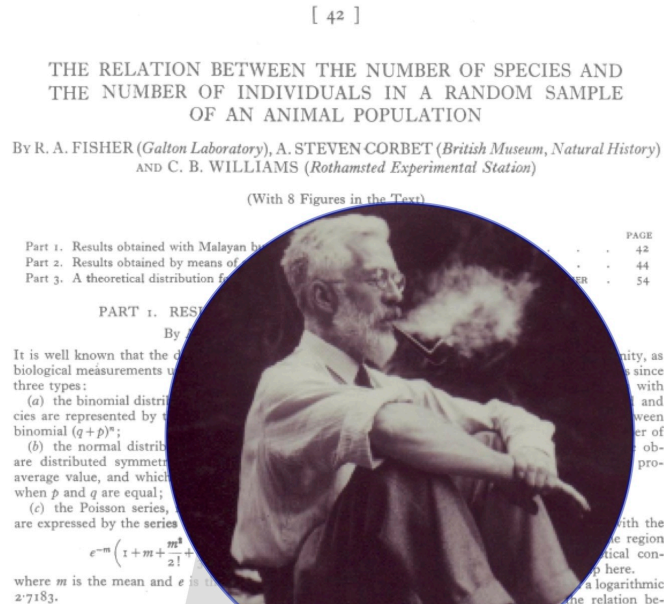
David Dunson
(*Duke University*)

A journey into statistical biodiversity

Biodiversity timeline



Approximately 80 years ago... Fisher et al. (1943)



Fisher, Corbet
and Williams
(1943)

Fisher's log-
series model

Expected number
of species appearing
 r times $=$ α $\times$ $\frac{x^r}{r}$

Constant derived
from the data

N. of species

Sample size

$$Y^{(n)} = \alpha \log \left(1 + \frac{n}{\alpha} \right)$$

Early ideas (Good 1953; Good and Toulmin 1956)

VOLUME 40, PARTS 3 AND 4

DECEMBER 1953

THE POPULATION FREQUENCIES OF SPECIES AND THE ESTIMATION OF POPULATION PARAMETERS

By I. J. GOOD

A random sample is drawn from a population of animals of various species. (The theory may also be applied to studies of literary vocabulary, for example.) If a particular species is represented r times in the sample of size N , then r/N is not a good estimate of the population frequency, p , which would be r/N if the sample were infinitely large. Methods are given for estimating p , assuming virtually nothing about the underlying population. The estimates are expressed in terms of smoothed values of the numbers n_r ($r=1, 2, \dots$), where n_r is the number of distinct species that are each represented r times in the sample. (The estimate of p as 'the frequency of the frequency r ') Turing is acknowledged for the most interesting part of the work. An estimate of the proportion of the population represented by species occurring r times in the sample is an immediate corollary. Estimates are made of measures of population, including Yule's 'characteristic' and Shannon's 'entropy'. Methods do depend on assumptions about the underlying population. It is here that the work of other writers. It is pointed out that a hypothesis can give a good fit to the data, but quite the wrong value for Yule's characteristic. An example of this is Fisher and Williams's on Macrolepidoptera.

1. *Introduction.* We imagine a random sample to be drawn from a population of animals of various species. Let the sample size be N and let n_r be the number of species represented exactly r times in the sample, so that

$$\sum_{r=1}^{\infty} r n_r = N.$$

The sample tells us the values of $n_1, n_2, \dots$, but not of n_0 . In fact it is not even clear that n_0 should be finite though we shall find it convenient to suppose that it is.



[45]

THE NUMBER OF NEW SPECIES, AND THE INCREASE IN POPULATION COVERAGE, WHEN A SAMPLE IS INCREASED

By I. J. GOOD AND G. H. TOULMIN

A sample of size N is drawn at random from a population of animals of various species. Methods are given for estimating, knowing only the contents of this sample, the number of species which will be represented r times in a second sample of size λN ; these also enable us to estimate the number of different species and the proportion of the whole population represented in the second sample. A formula is found for the variance of the estimate; when λ is small, this variance becomes in general very large, so that the estimate is useless without modification. This difficulty can be partly overcome, at least for $\lambda < 5$, by using Euler's method with a suitable parameter or the methods described by Shanks (1955) to hasten the convergence of the series by which the estimate is expressed. The methods are applied to the words from *Our Mutual Friend*, to an entomological sample, and to a sample of Macaulay's essay on Bacon.

1. INTRODUCTION

This is a further development of the theory expounded by Good (1953); that paper will be referred to, for brevity, by the letter G throughout.

A random sample of size N , the *basic sample*, to be drawn from an infinite population of animals of various species, and suppose that n_r distinct species are each represented exactly r times in the sample, so that

$$\sum_{r=1}^{\infty} r n_r = N. \quad (1)$$

$$d = \sum_{r=1}^{\infty} n_r,$$

Notion of **sample coverage**

+

Prediction of species on a further sample

Fisher, Corbet
and Williams
(1943)

Fisher's log-
series model

Good
(1953)

Prediction of N.
of species in a
further sample

Good and
Toulmin
(1956)

The Ewens sampling formula Ewens (1972)

Distribution of **allelic partition** from neutral selection

m_1 alleles represented ones, m_2 represented twice etc...

$$\mathbb{P}(m_1, \dots, m_n) = \frac{\alpha^n}{\alpha(\alpha+1)\cdots(\alpha+n-1)} \prod_{j=1}^n \frac{n!}{j^{m_j} m_j!}$$

with $\sum_{j=1}^n j \cdot m_j = n$

THEORETICAL POPULATION BIOLOGY 3, 87-112 (1972)

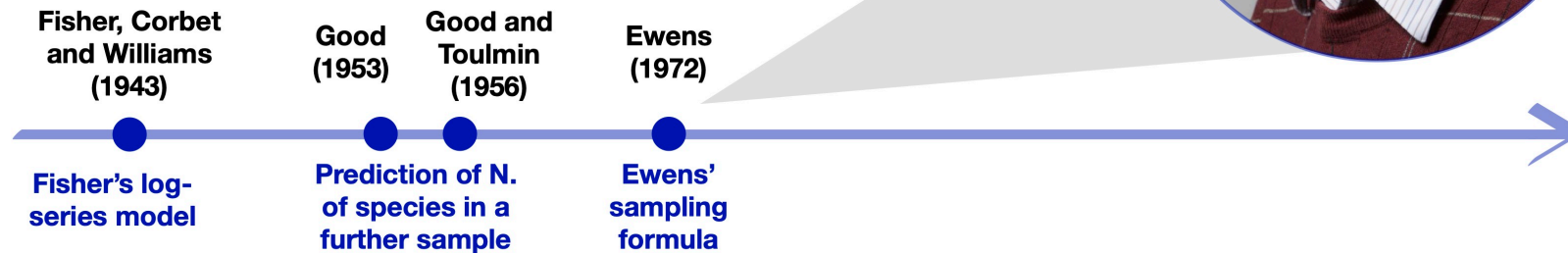
The Sampling Theory of Selectively Neutral Alleles*

W. J. EWENS†

Department of Zoology, University of Texas at Austin, Austin, Texas, 78712

Received August 17, 1971

DEDICATED TO THE MEMORY OF KEN KOJIMA



The Dirichlet process (Ferguson 1973)

The Annals of Statistics
1973, Vol. 1, No. 2, 209-230

A BAYESIAN ANALYSIS OF SOME NONPARAMETRIC PROBLEMS¹

BY THOMAS S. FERGUSON

University of California, Los Angeles

1. Introduction and summary. The Bayesian approach to statistical problems, though fruitful in many ways, has been rather unsuccessful in treating nonparametric problems. This is due primarily to the difficulty in finding workable prior distributions on the parameter space, which in nonparametric problems is taken to be a set of probability distributions on the sample space. There are two desirable properties of a prior distribution for nonparametric problems.

(I) The support of the prior distribution should be the set of all probability distributions on the sample space.

(II) Posterior distributions should be proper.



The Annals of Statistics
1984, Vol. 12, No. 6, 1159-1174

MIXTURES OF DIRICHLET PROCESSES WITH APPLICATIONS TO BAYESIAN NONPARAMETRIC STATISTICS

BY CHARLES E. ANTONIAK

University of California, Berkeley

A random process called the Dirichlet process was introduced by Ferguson (1973). It is almost surely a probability measure and has been proposed as a prior distribution in an approach to analyzing nonparametric problems from a Bayesian point of view.

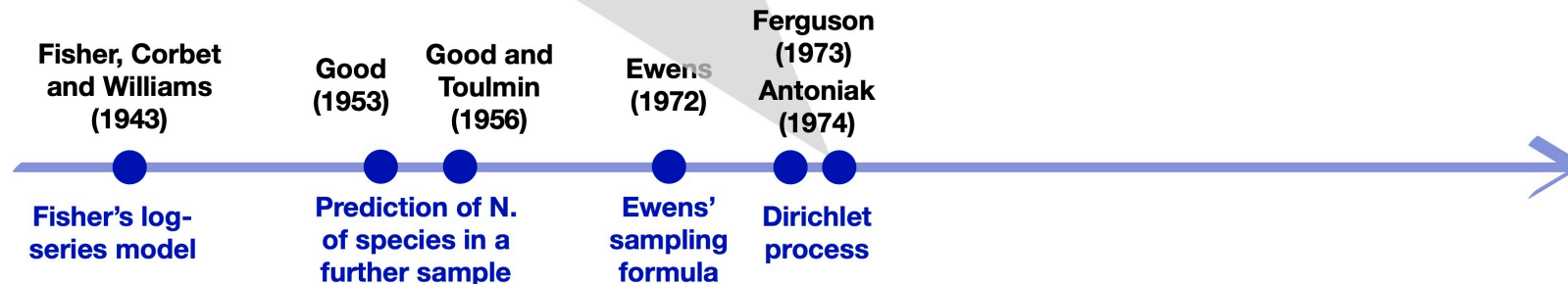
An important result obtained by Ferguson in this approach is that if observations are made on a random variable whose distribution is a random sample function of a Dirichlet process, then the conditional distribution

The partition from a DP is just a re-write of the Ewens formula!

$$\mathbb{P}(n_1, \dots, n_y) = \frac{\alpha^y}{(\alpha)_n} \prod_{j=1}^y (n_j - 1)!$$

Chinese Restaurant Process
(Blackwell and MacQueen, 1973)

$$X_{n+1} = \begin{cases} X_j^* & \text{w.p. } n_j/(\alpha + n) \\ \text{new} & \text{w.p. } \alpha/(\alpha + n) \end{cases}$$



... and species sampling models (Pitman 1996)

Statistics, Probability and Game Theory
IMS Lecture Notes – Monograph Series (1996) Volume 30

Species sampling models: exchangeable
generalization of the Dirichlet process

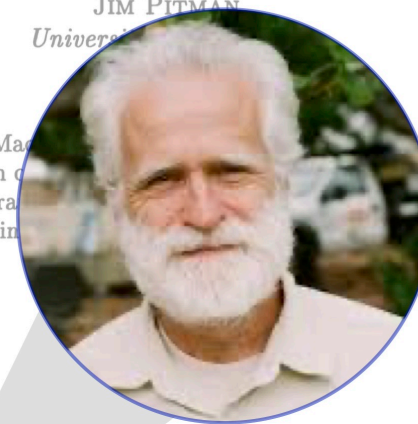
Notion of **σ -diversity**: asymptotic behavior of
the number of clusters adjusted for the growth

$$\frac{Y^{(n)}}{c_{\sigma}(n)} \rightarrow S_{\sigma} = \alpha$$

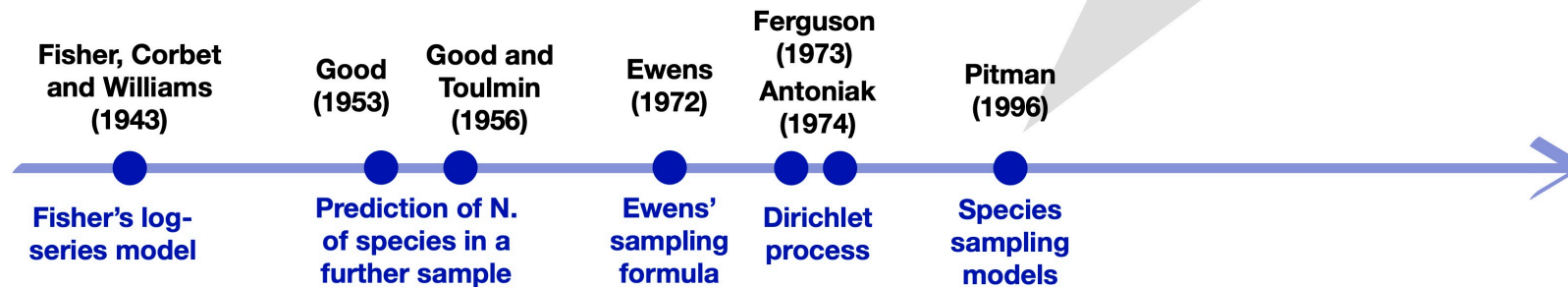
$\log n$ in the Dirichlet process

SOME DEVELOPMENTS OF THE BLACKWELL-MACQUEEN URN SCHEME

JIM PITMAN
University of California, San Diego



The Blackwell-MacQueen random distribution of a general family of random distributions obtained by application of Kingman's

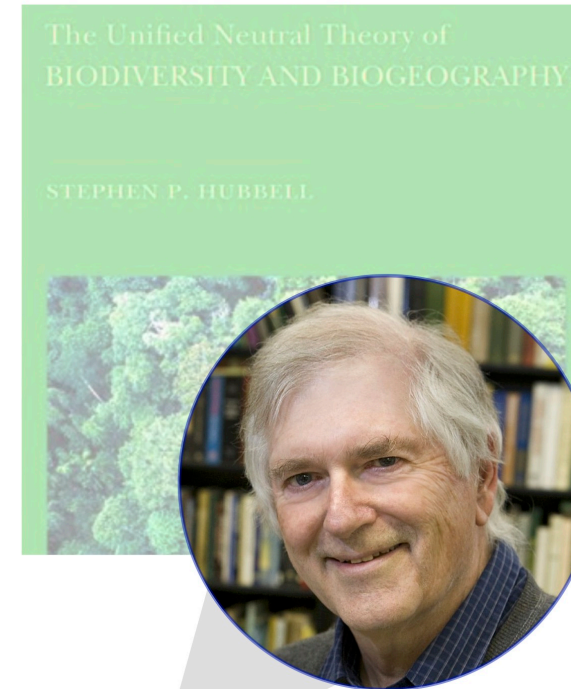
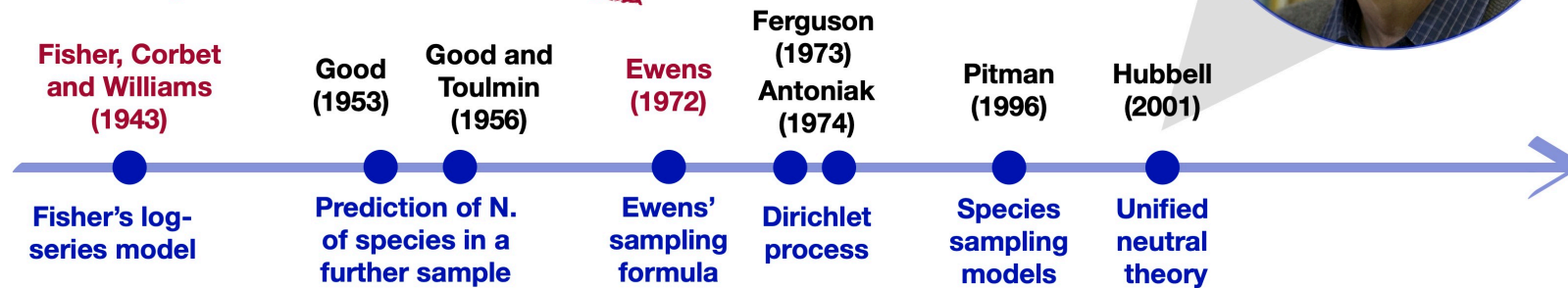


Ewens formula governs the neutral theory! (Hubbell 2001)

Unified neutral theory of biodiversity and existence of a **fundamental biodiversity number α**

$$\alpha = \text{Twice the metacommunity size times speciation rate}$$

This is asymptotically equivalent to the α in Fisher log-series!



All ties together (McCullagh 2016)



Statistical Science
2016, Vol. 31, No. 1, 1–19
DOI: 10.1214/15-875529
© Institute of Mathematical Statistics, 2016

The Ubiquitous Ewens Sampling

Harry Crane

Abstract. Ewens's sampling formula exemplifies the harmony of mathematical theory, statistical application, and scientific discovery. It not only contributes to the foundations of evolutionary models, the neutral theory of biodiversity, Bayesian nonparametric stochastic processes, and inductive inference but also embeds fundamental concepts in probability theory, algebra, and number theory. We highlight these and many other consequences of Ewens's formula.

Key words and phrases: Ewens's sampling formula, Poisson process, random partition, coalescent process, inductive inference, logarithmic combinatorial structures, Chinese restaurant process, Ewens–Pitman distribution, Dirichlet process, stick breaking, permanent partition model, cyclic product distribution, clustering, Bayesian nonparametrics, α -permanent.

“It is fair to say that Fisher almost discovered the Ewens sampling formula. He had it in his grasp. After all, he had only to compute the conditional distribution given the sample size, a task that was both statistically natural and, for him, mathematically trivial.”



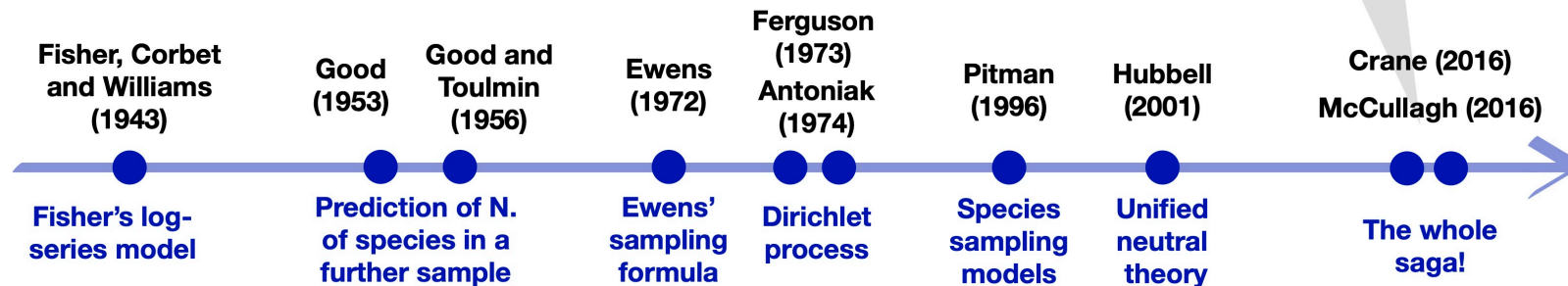
Statistical Science
2016, Vol. 31, No. 1, 21–31
DOI: 10.1214/15-875530
Manuscript received 10/11/15; accepted 1/16/16
© Institute of Mathematical Statistics, 2016

Two Early Contributions to the Ewens Saga

Peter McCullagh

Abstract. The mixture model devised by Fisher, Corbet and Williams [*Journal of Animal Ecology* 12 (1943) 42–58] for species sampling and the sequential prediction approach pioneered by Good [*Biometrika* 40 (1953) 237–264] and Good and Toulmin [*Biometrika* 43 (1956) 45–63] are both closely related to the Ewens sampling formula. Fisher's two-parameter joint distribution for the species counts includes the Ewens distribution as the conditional distribution given the sample size. The log-series model, as it is known in the ecological literature, is closely related to a Poisson process model devised by Arratia, Barbour and Tavaré [*Ann. Appl. Probab.* 2 (1992) 519–535]. Oddly, despite its advantages for statistical inference, Fisher does not mention the conditional distribution. Likewise, although Good (1953) pioneered the sequential prediction approach, neither he nor Toulmin discovered the Ewens process in a form equivalent to the modern-day Chinese restaurant process.

Key words and phrases: Chinese restaurant process, Poisson process, species richness, species sampling.



Species sampling models

- Let $X_1, \dots, X_n$ be some collection of **species** or **taxa**. Suppose X_n are conditionally iid samples from a **species sampling model**:

$$(X_n \mid \tilde{p}) \stackrel{\text{iid}}{\sim} \tilde{p}, \quad \tilde{p}(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{Z_h}(\cdot), \quad n \geq 1,$$

where $(\pi_h)_{h \geq 1}$ is a set of probabilities (**species proportions**) and Z_h represent distinct species.

- The discreteness of $\tilde{p}$ identifies $Y^{(n)} = y$ distinct taxa $X_1^*, \dots, X_y^*$ with frequencies $n_1, \dots, n_y$, called **abundances** in ecology.
- Gibbs-type priors** have emerged as the **most natural extension** of the DP (De Blasi et al. 2015).
- The **predictive distribution** of a Gibbs-type prior is given by:

$$\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \dots, X_n) = \frac{V_{n+1, y+1}}{V_{n, y}} P(\cdot) + \frac{V_{n+1, y}}{V_{n, y}} \sum_{j=1}^y (n_j - \sigma) \delta_{X_j^*}(\cdot).$$

where $(a)_n$ denotes a rising factorial, and $\sigma < 1$ is the discount parameter. The $V_{n, y}$'s are non-negative weights satisfying a forward recursive equation.

Three notable examples

- **Dirichlet multinomial** ($\sigma < 0$). For $\sigma < 0$ and $H \in \mathbb{N}$, a valid set of Gibbs coefficients is given by:

$$V_{n,y}(\sigma, H) := \frac{|\sigma|^{y-1} \prod_{j=1}^{y-1} (H - j)}{(H|\sigma| + 1)_{n-1}} \mathbb{I}(y \leq H).$$

- **Dirichlet process** ($\sigma = 0$). For $\sigma = 0$ and $\alpha > 0$, a valid set of Gibbs coefficients is given by:

$$V_{n,y}(\alpha) := \frac{\alpha^y}{(\alpha)_n}.$$

- **σ -stable Poisson–Kingman process** ($0 < \sigma < 1$). For $\sigma \in (0, 1)$ and gamma $\gamma > 0$, a valid set of Gibbs coefficients is defined as:

$$V_{n,y}(\sigma, \gamma) := \frac{\sigma^y \gamma^y}{\Gamma(n - y\sigma) f_\sigma(\gamma^{-1/\sigma})} \int_0^1 s^{n-1-y\sigma} f_\sigma((1-s)\gamma^{-1/\sigma}) \, ds,$$

where $f_\sigma(t) = (\pi)^{-1} \sum_{h=1}^{\infty} (-1)^{h+1} \sin(h\pi\sigma) \Gamma(h\sigma + 1) / t^{h\sigma+1}$.

A characterization theorem

Theorem (Gnedin and Pitman 2005)

The Gibbs coefficients $V_{n,y}$ satisfy the recursive equation in the following three cases:

- If $\sigma < 0$, whenever $V_{n,y} = \sum_{h=1}^{\infty} V_{n,y}(\sigma, h)p(h)$, for some discrete random variable $H \in \mathbb{N}$ with pdf $p(h)$, where the $V_{n,y}(\sigma, h)$'s are those of the **Dirichlet multinomial**.
- If $\sigma = 0$, whenever $V_{n,y} = \int_{\mathbb{R}^+} V_{n,y}(\alpha)p(d\alpha)$, for some positive random variable α with probability measure $p(d\alpha)$, where the $V_{n,y}(\alpha)$'s are those of the **Dirichlet process**.
- If $\sigma \in (0, 1)$, whenever $V_{n,y} = \int_{\mathbb{R}^+} V_{n,y}(\sigma, \gamma)p(d\gamma)$, for some positive random variable γ with probability measure $p(d\gamma)$, where the $V_{n,y}(\sigma, \gamma)$'s are those of the **σ -stable PK**.

- The **Dirichlet multinomial**, **Dirichlet process**, and **σ -stable PK** form the foundation of any Gibbs-type prior.
- In fact, any Gibbs-type process can be represented **hierarchically**, involving a suitable **prior distribution** for the key parameters H , α , and γ .

The quantification of biodiversity (Lijoi et al. 2007a)

- The simplest measure of biodiversity is arguably the taxon **richness** $Y^{(n)} = y$.
- A priori, the distribution of $Y^{(n)}$ induced by a Gibbs-type prior has a simple form:

$$\mathbb{P}(Y^{(n)} = y) = V_{n,y} \frac{\mathcal{C}(n, y; \sigma)}{\sigma^y},$$

where $\mathcal{C}(n, y; \sigma)$ denotes a generalized factorial coefficient.

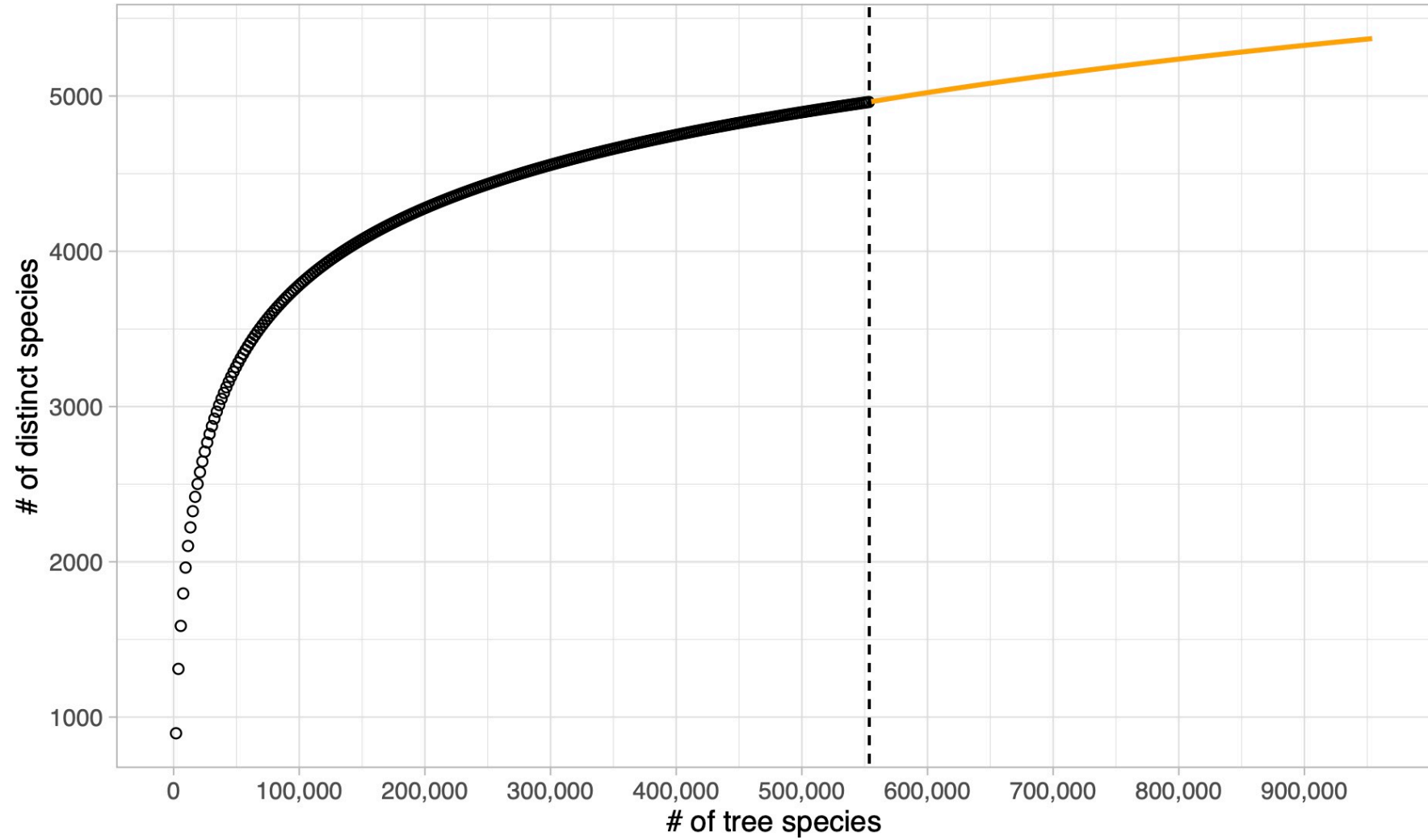
- The a priori expectations $\mathbb{E}(K_1), \dots, \mathbb{E}(K_n)$ define a model-based **rarefaction** curve.
- The **posterior distribution** of the number of previously unobserved taxa $Y_m^{(n)}$ is

$$\mathbb{P}(Y_m^{(n)} = j \mid X_1, \dots, X_n) = \frac{V_{n+m,y+j}}{V_{n,y}} \frac{\mathcal{C}(m, j; \sigma, -n + y\sigma)}{\sigma^j}, \quad j = 0, \dots, m,$$

where $\mathcal{C}(m, j; \sigma, -n + y\sigma)$ is the noncentral generalized factorial coefficient.

- The posterior expectations $\mathbb{E}(Y^{(n+1)} \mid Y^{(n)} = y), \dots, \mathbb{E}(Y^{(n+m)} \mid Y^{(n)} = y)$ represents a model-based **extrapolation** of the accumulation curve.

Rarefaction and extrapolation



The σ -diversity (Pitman 2003)

- Let $Y^{(n)}$ be the number of distinct values arising from a Gibbs-type prior:
 - i. Let $\sigma < 0$ and $V_{n,y}(\sigma, H)$ be the weights of a **Dirichlet multinomial**, then $Y^{(n)} \rightarrow H$ a.s.
 - ii. Let $\sigma = 0$ and $V_{n,y}(\alpha)$ be the weights of a **Dirichlet process**, then $Y^{(n)} / \log(n) \rightarrow \alpha$ a.s.
 - iii. Let $\sigma \in (0, 1)$ and $V_{n,y}(\sigma, \gamma)$ be the weights of a **σ -stable PK**, then $Y^{(n)} / n^\sigma \rightarrow \gamma$ a.s.
- Moreover, consider a generic set of weights $V_{n,y}$ and let

$$c_\sigma(n) = \begin{cases} 1, & \sigma < 0, \\ \log(n), & \sigma = 0, \\ n^\sigma, & \sigma \in (0, 1) \end{cases}$$

Then, as $n \rightarrow \infty$:

$$\frac{Y^{(n)}}{c_\sigma(n)} \xrightarrow{\text{a.s.}} S_\sigma.$$

The r.v. S_σ is the **σ -diversity** and its distribution coincides with the prior for H , α , and γ .

The σ -diversity (Pitman 2003)

- Broadly speaking, the σ -diversity can be seen as a **rescaled richness measure**.
- The σ -diversity is deterministic and assumed to be known in the Dirichlet multinomial, **Dirichlet process**, and σ -stable PK.
- However, the σ -diversities H , α , or γ are typically unknown, and they can be estimated employing a prior distribution, leading to a Gibbs-type prior.
- The **posterior law** of the σ -diversity is a key quantity for measuring **biodiversity**. The posterior law of S_σ has an elegant connection with accumulation curves.

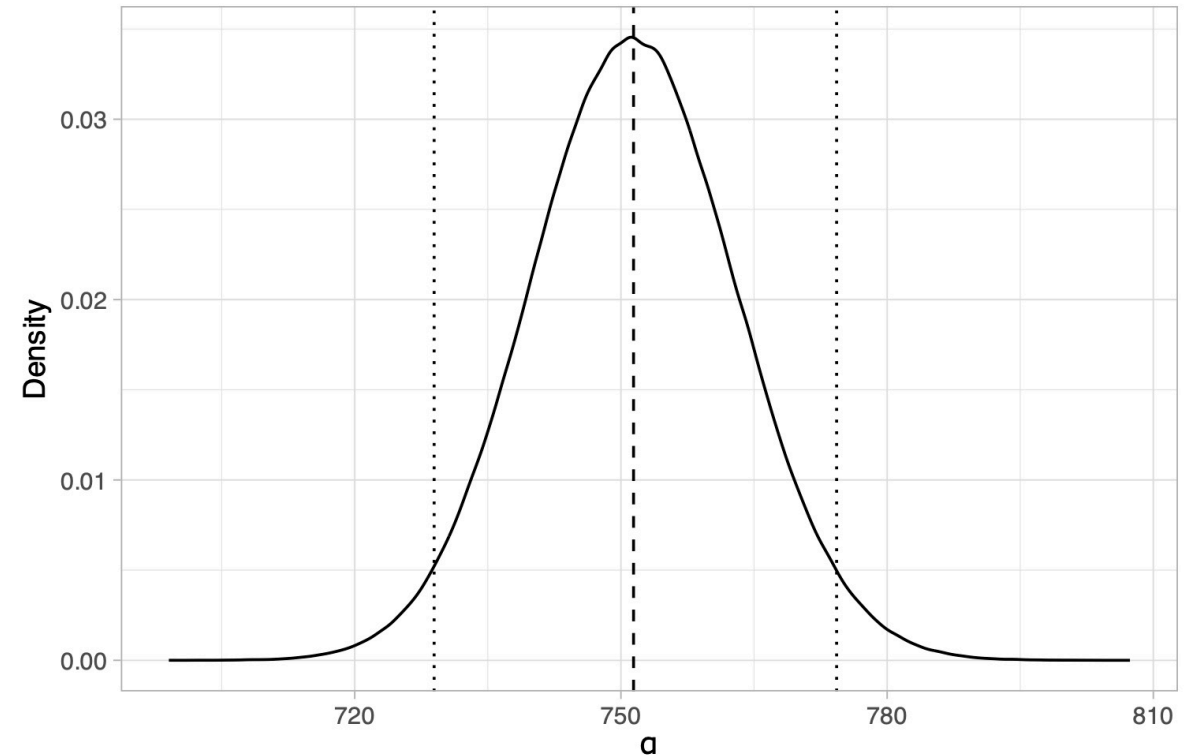
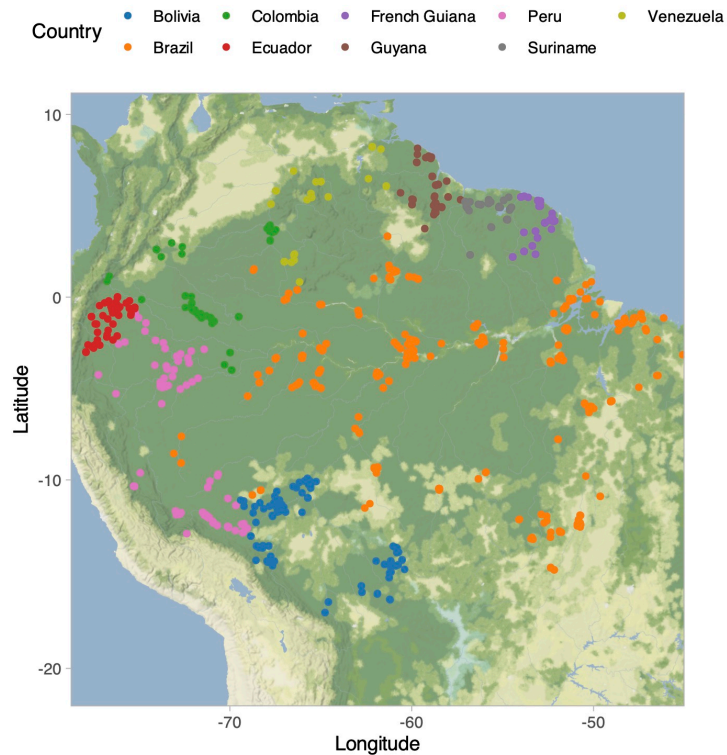
Theorem (Rigon et al. 2025b)

Let $X_1, \dots, X_{n+m}$ be a sample from a Gibbs-type prior with $Y^{(n+m)}$ distinct values. Then:

$$\left(\frac{Y^{(n+m)}}{c_\sigma(m)} \mid X_1, \dots, X_n \right) \xrightarrow{d} \tilde{S}_\sigma, \quad \tilde{S}_\sigma \stackrel{d}{=} (S_\sigma \mid X_1, \dots, X_n),$$

as $m \rightarrow \infty$, where S_σ is the σ -diversity.

Tree species in the Amazon Basin (Hubbell et al. 2008)



- How to get a **Bayesian** estimator for the biodiversity? We can place a **prior** on α and then compute its **posterior distribution** using Bayes theorem.
- **Posterior distribution** of the σ -diversity α , under a (conjugate) Stirling-gamma prior (Zito et al. 2024). The dotted lines represent 95% credible intervals. The dashed line is the posterior mean.

Historical summary and extensions

- The fundamental biodiversity number α is a **growth-adjusted richness**:

$$\frac{Y^{(n)}}{\log n} \xrightarrow{\text{a.s.}} \alpha, \quad n \rightarrow \infty$$

- The α from the neutral theory, DP and Fisher log-series is the same for large n :

$$\underbrace{y = \hat{\alpha}^{\text{Fisher}} \log \left(1 + \frac{n}{\hat{\alpha}^{\text{Fisher}}} \right)}_{\text{Solution from Fisher}} \underset{\text{Under large } n}{\approx} \underbrace{\sum_{j=1}^n \frac{\hat{\alpha}^{\text{DP}}}{\hat{\alpha}^{\text{DP}} + j - 1}}_{\text{MLE from Hubbell and DP}} = y$$

- There are a lot of possible extensions. This is a very **biased** list:
 - The choice $\sigma \in (0, 1)$ allows for polynomial growth (Lijoi et al. 2007b; Pitman and Yor 1997) while $\sigma < 0$ leads to a finite richness (Gnedin 2010; Lijoi et al. 2020).
 - Multiple samples, a.k.a. **partial exchangeability** (Camerlenghi et al. 2019; Franzolini et al. 2025)
 - Enriched processes, taxonomical data (Rigon et al. 2025a; Rigon et al. 2025b; Zito et al. 2023).
 - Incidence data and **feature models** (Ghilotti et al. 2025).
 - What about **covariates**?

References

- Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019), "Distribution theory for hierarchical processes," *Ann. Statist.*, 47, 67–92.
- Conley, T. G. (1999), "GMM estimation with cross sectional dependence," *Journal of Econometrics*, 92, 1–45.
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prunster, I., and Ruggiero, M. (2015), "Are Gibbs-type priors the most natural generalization of the Dirichlet process?" *IEEE Trans. Pattern Anal. Mach. Intell.*, IEEE, 37, 212–229.
- Ewens, W. (1972), "The sampling theory of selectively neutral alleles," *Theoretical Population Biology*, 3, 87–112.
- Ferguson, T. S. (1973), "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, 1, 209–230.
- Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943), "The relation between the number of species and the number of individuals in a random sample of an animal population," *J. Anim. Ecol.*, 12, 42–58.
- Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2025), "Multivariate species sampling models," *arXiv:2503.24004*.
- Ghilotti, L., Camerlenghi, F., and Rigon, T. (2025), "Bayesian analysis of product feature allocation models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Gnedin, A. (2010), "Species sampling model with finitely many types," *Electron. Comm. Probab.*, 15, 79–88.
- Gnedin, A., and Pitman, J. (2005), "Exchangeable Gibbs partitions and Stirling triangles," *Zapiski Nauchnykh Seminarov, POMI*, 325, 83–102.
- Good, I. J. (1953), "The population frequencies of species and the estimation of population parameters," *Biometrika*, 40, 237–264.
- Good, I. J., and Toulmin, G. H. (1956), "The number of new species, and the increase in population coverage, when a sample is increased," *Biometrika*, 43, 45–63.
- Hubbell, S. P. (2001), *The unified neutral theory of biodiversity and biogeography*, Princeton University Press.
- Hubbell, S. P., He, F., Condit, R., Borda-de-Água, L., Kellner, J., and Ter Steege, H. (2008), "How many tree species are there in the Amazon and how many of them will go extinct?" *Proceedings of the National Academy of Sciences of the United States of America*, 105, 11498–11504.
- Lijoi, A., Mena, R. H., and Prünster, I. (2007a), "Bayesian nonparametric estimation of the probability of discovering new species," *Biometrika*, Oxford University Press, 94, 769–786.
- Lijoi, A., Mena, R. H., and Prünster, I. (2007b), "Controlling the reinforcement in Bayesian non-parametric mixture models," *J. R. Statist. Soc. B*, Wiley Online Library, 69, 715–740.
- Lijoi, A., Prünster, I., and Rigon, T. (2020), "The Pitman–Yor multinomial process for mixture modeling," *Biometrika*, 107, 891–906.
- McCullagh, P. (2016), "Two Early Contributions to the Ewens Saga," *Statist. Sci.*, 31, 23–26.
- Pitman, J. (1996), "Some developments of the Blackwell–MacQueen urn scheme," in *Statistics, probability and game theory*, IMS lecture notes monogr. ser., Inst. Math. Statist., Hayward, CA, pp. 245–267.
- Pitman, J. (2003), "Poisson–Kingman partitions," *Lecture Notes–Monograph Series*, 40, 1–34.
- Pitman, J., and Yor, M. (1997), "The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator," *Ann. Probab.*, 25, 855–900.
- Rigon, T., Hsu, C., and B., D. D. (2025a), "A Bayesian theory for estimation of biodiversity," *arXiv:2502.01333*.
- Rigon, T., Petrone, S., and Scarpa, B. (2025b), "Enriched Pitman–Yor processes," *Scand. J. Statist.*
- Zito, A., Rigon, T., and Dunson, D. B. (2023), "Inferring taxonomic placement from DNA barcoding allowing discovery of new taxa," *Meth. Ecol. Evol.*, 14, 529–542.
- Zito, A., Rigon, T., and Dunson, D. B. (2024), "Bayesian nonparametric modeling of latent partitions via Stirling–gamma priors," *Bayesian Analysis*.