Species sampling models

Tommaso Rigon

University of Milano-Bicocca

Lifeplan weekly meeting

The Finnish fungal dataset (Abrego et al., 2020)

- We consider $S = 174$ **samples** out of 180, i.e. excluding technically failed sequencing.
- We observed a total of $K = 79,155$ distinct species (OTUs), i.e. the overall **richness**.
- Within each sample, we **dichotomize the OTU abundances**, so that

$$Z_{js} = 1 \qquad \Longrightarrow \qquad \text{The } j\text{th species is present in the } s\text{th sample,}$$

  and $Z_{js} = 0$ otherwise, for $j = 1, \ldots, K$ and $s = 1, \ldots, S$.
- Samples can be of type `Air`, `Soil`, `Urban` and `Natural`, depending on the geographical location of the sample.
- We will take into account these differences at the end of the presentation.

Species frequencies (a.k.a. abundance)

- We consider a further summary of the data, i.e. we **count how many times a species has been observed across samples**.
- More formally, we have that for $j = 1, \dots, K$

$$n_j = \sum_{s=1}^{S} Z_{js} = \sum_{s=1}^{S} \mathbb{1}(\text{"The } j\text{th species is present in the } s\text{th sample"}).$$

- The frequencies $n_1, \dots, n_K$ will represent a **sufficient statistics** for our modeling.
- Obviously, each of these frequencies are such that

$$1 \leq n_j \leq S, \qquad j = 1, \dots, K.$$

- The integer $n = \sum_{j=1}^{K} n_j = 196,619$ is the global count of (non distinct) species.

Frequencies of frequencies

- The data can be summarized (without loss of information!) even more, using **frequencies of frequencies**.

- We define the integers

$$m_k = \sum_{j=1}^{K} \mathbb{1}(n_j = k) = \text{"How many times species occured } k \text{ times across samples"},$$

  for $k = 1, \ldots, K$.

- In our case, we have that for example $m_1 = 53,431$, $m_2 = 10,762$, $m_3 = 4,553$, $m_4 = 2,469$ and so on until the last terms $m_{135} = 1$ and $m_{146} = 1$.

- In this representation, by construction we have that

$$\sum_{k=1}^{K} m_k = K, \qquad \sum_{k=1}^{K} k\, m_k = n.$$

Species sampling models

- Let $X_1, \ldots, X_n$ be some collection of "species" with frequencies $n_1, \ldots, n_K$ such that for $i = 1, \ldots, n$,

$$(X_i \mid \tilde{p}) \overset{\text{iid}}{\sim} \tilde{p}, \qquad \tilde{p} = \sum_{h=1}^{H} \pi_h \delta_{\theta_h},$$

  with $\tilde{p}$ being an unknown **discrete sampling distribution**.

- The weights $\sum_{h=1}^{H} \pi_h = 1$ are the **species proportions**, with H being the total number of species in the **population**.

- The values $\theta_1, \ldots, \theta_H$ are instead the **distinct species** in the population.

- This is essentially a **multinomial model** and therefore the frequencies $n_1, \ldots, n_K$ are a sufficient statistics.

- Crucially, we have that $K \leq H$, i.e. the number of discovered species is smaller than the true number.

Goal I: Sample coverage

Sample coverage

- The **sample coverage** is the sum of the proportions of species that has been observed.
- More precisely, it is defined as

$$C_n = \sum_{h \in \mathcal{H}} \pi_h, \qquad \mathcal{H} = \{\text{"Indexes of the observed species among } n \text{ data"}\}.$$

- A very old method by Turing (Good 1953) for "estimating" C_n is:

$$\hat{C}_n = 1 - \frac{m_1}{n},$$

  which is based on the **number of species observed only once**.

- In the Lifeplan global data one has that $\hat{C}_n = 1 - 53,431/196,619 = 0.728$, a fairly high (?) number.

Goal II: Accumulation curves and rarefaction

Accumulation curve

Suppose the species $X_1, \ldots, X_n$ are observed sequentially. An accumulation curve is the number of distinct species K_n observed as the sample size n increases.

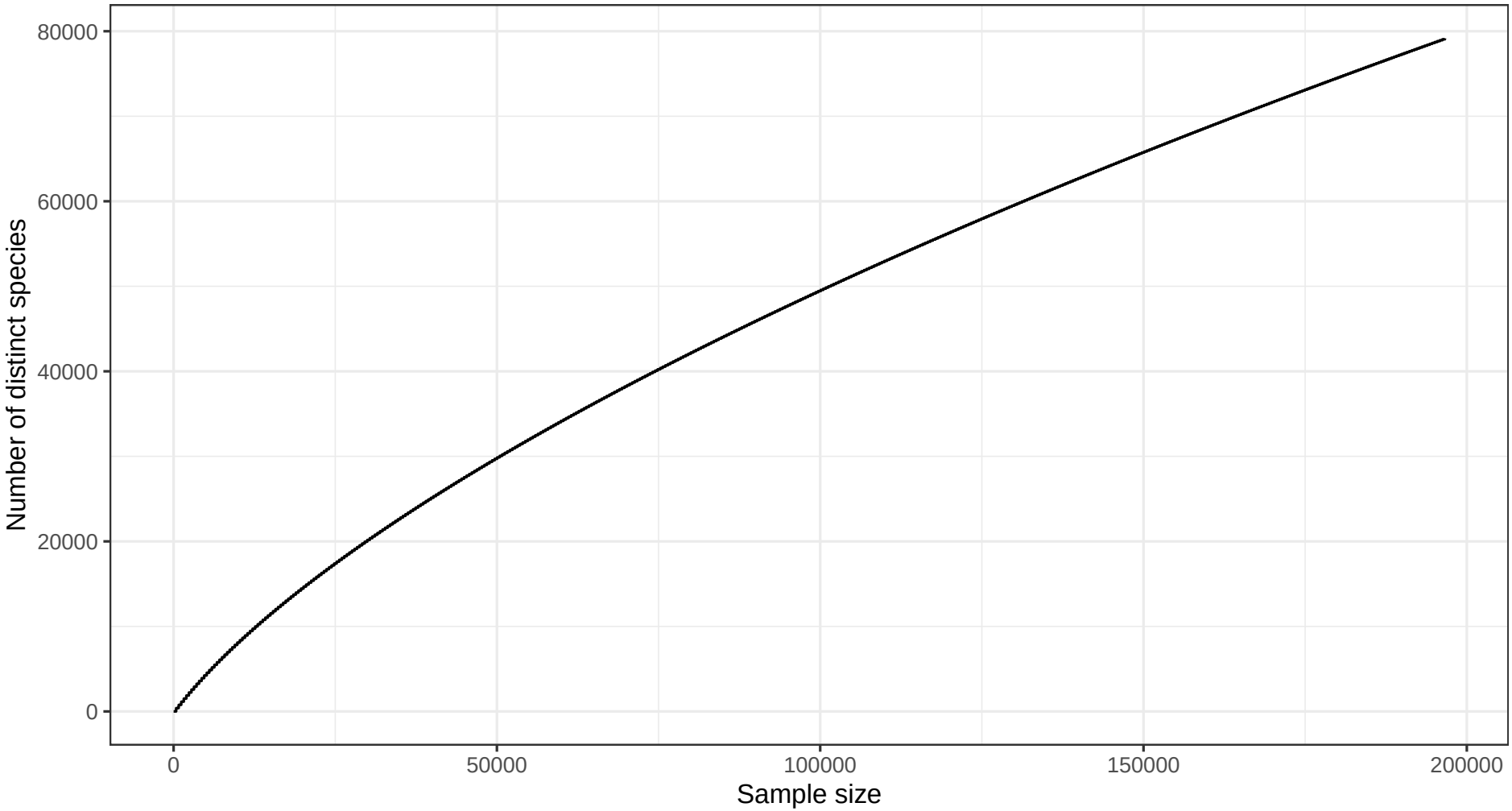
Rarefaction

- Several times data are not observed sequentially. The **rarefaction** is the average accumulation curve over the space of permutations.
- Combinatorial calculus leads to the following formula

$$\mathbb{E}(K_i) = K - \binom{n}{i}^{-1} \sum_{j=1}^{K} \binom{n - n_j}{i}.$$

- Note that such a formula only depends on the frequencies $n_1, \ldots, n_K$.

Goal II: Accumulation curves and rarefaction



Goal III: Species diversity

Gini heterogeneity index

- The Gini heterogeneity index is a measure for **quantifying biodiversity**.
- The Gini index G is the probability that two randomly taken species from the population are different, namely

$$G = 1 - \sum_{h=1}^{H} \pi_h^2.$$

- A simple way for estimating G is setting $\hat{G} = 1 - 1/n^2 \sum_{j=1}^{K} n_j^2 = 0.9999369$.

- Other heterogeneity indexes might be considered, e.g. the Shannon entropy.
- **Note**: the Simpson index is simply $S = 1 - G = \sum_{h=1}^{H} \pi_h^2$.

Issues with these approaches

- The properties of these estimators are often based on asymptotic considerations $\Longrightarrow$ Bayesian inference could be helpful.
- If **prior information** is available, there is not a simple way to incorporate it into the modeling.
- It is even more problematic to incorporate these estimators into more complex models accounting for covariates and to borrow strength across locations.
- Unclear how to perform e.g. testing and uncertainty quantification in such complex settings.
- **Solution**: use Bayesian statistics to obtain model-based "estimators" within a unified setting.

Bayesian nonparametric priors

- The sampling distribution $\tilde{p}$ encodes all the **relevant information** but it is unknown, so we are interested in learning it from the data $X_1, \ldots, X_n$.
- In the Bayesian framework, this amounts to the choice a **nonparametric prior** for the sampling distribution $\tilde{p}$.
- Then, one can study the following posterior law

$$\tilde{p} \mid X_1, \ldots, X_n.$$

- **Note**: all the previous quantities of interest are functions of $\tilde{p}$ $\Longrightarrow$ this leads to natural Bayesian estimators for coverage, diversity, etc.
- Common nonparametric priors are the Dirichlet process (DP) and the Pitman–Yor (PY) process.

The Pitman–Yor process

Stick-breaking of the PY

$$\tilde{p} = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \qquad \pi_h = \nu_h \prod_{\ell=1}^{h} (1 - \nu_\ell), \quad \nu_h \overset{\text{ind}}{\sim} \text{BETA}(1 - \sigma, \alpha + \sigma h),$$

with $\sigma \in [0, 1)$ and $\theta > -\sigma$.

Urn-scheme

$$X_{n+1} \mid X_1, \ldots, X_n \sim \frac{\alpha + \sigma K}{\alpha + n} (\text{"new species"}) + \frac{1}{\alpha + n} \sum_{j=1}^{K} (n_j - \sigma) \delta_{X_j^*}.$$

Posterior distribution

$$(\tilde{p} \mid X_1, \ldots, X_n) = \sum_{j=1}^{K} W_j \delta_{X_j^*} + W_{k+1} \tilde{q},$$

with $(W_1, \ldots, W_{k+1}) \sim \text{DIR}(n_1 - \sigma, \ldots, n_k - \sigma, \alpha + \sigma K)$ and $\tilde{q}$ is a $\text{PY}(\alpha + \sigma K, \sigma)$.

Estimation of the parameters

- Inference on the hyperparameters (α, σ) can be conducted trough MCMC.
- However, it is common to replace them with their maximum likelihood estimate, i.e. an empirical Bayes procedure.
- In the PY model, the **likelihood** is the following quantity

$$\mathscr{L}(\alpha, \sigma \mid X_1, \ldots, X_n) = \frac{\prod_{j=1}^{K-1}(\alpha + j\sigma)}{(\alpha+1)_{n-1}} \prod_{j=1}^{K} (1-\sigma)_{n_j - 1}.$$

- The maximizer of this likelihood can be easily obtained using **off-the-shelf numerical routines** (i.e. `optim` R command) in fraction of seconds.

Goal I: Sample coverage

Lemma 1

In a PY model, the posterior distribution of the **sample coverage** is

$$(C_n \mid X_1 \ldots, X_n) \sim \text{BETA}(n - \sigma K, \alpha + \sigma K).$$

Moreover, the posterior mean coincides with

$$\mathbb{E}(C_n \mid X_1, \ldots, X_n) = \mathbb{P}(X_{n+1} = \text{"old species"} \mid X_1, \ldots, X_n) = \frac{n - \sigma K}{\alpha + n}.$$

- An empirical Bayes procedure applied to the Finnish Fungal data leads $\hat{\alpha} = 7080.164$, $\hat{\sigma} = 0.6138$.
- Recalling that $n = 196,619$ and $K = 79,155$, we get a **Bayesian estimate** for the sample coverage $\mathbb{E}(C_n \mid X_1, \ldots, X_n) = 0.7267$.
- This is remarkably similar to Good & Turing estimators, but uncertainty quantification can be conducted in a very simple manner.

Goal II: Accumulation curves and rarefaction

Lemma 2

In a PY model the rarefaction curve is a Markov process such that $K_1 = 1$ and

$$K_{n+1} \mid K_n = K_n + D_n, \qquad (D_n \mid K_n) \sim \text{BER}\left(\frac{\alpha + \sigma K_n}{\alpha + n}\right).$$

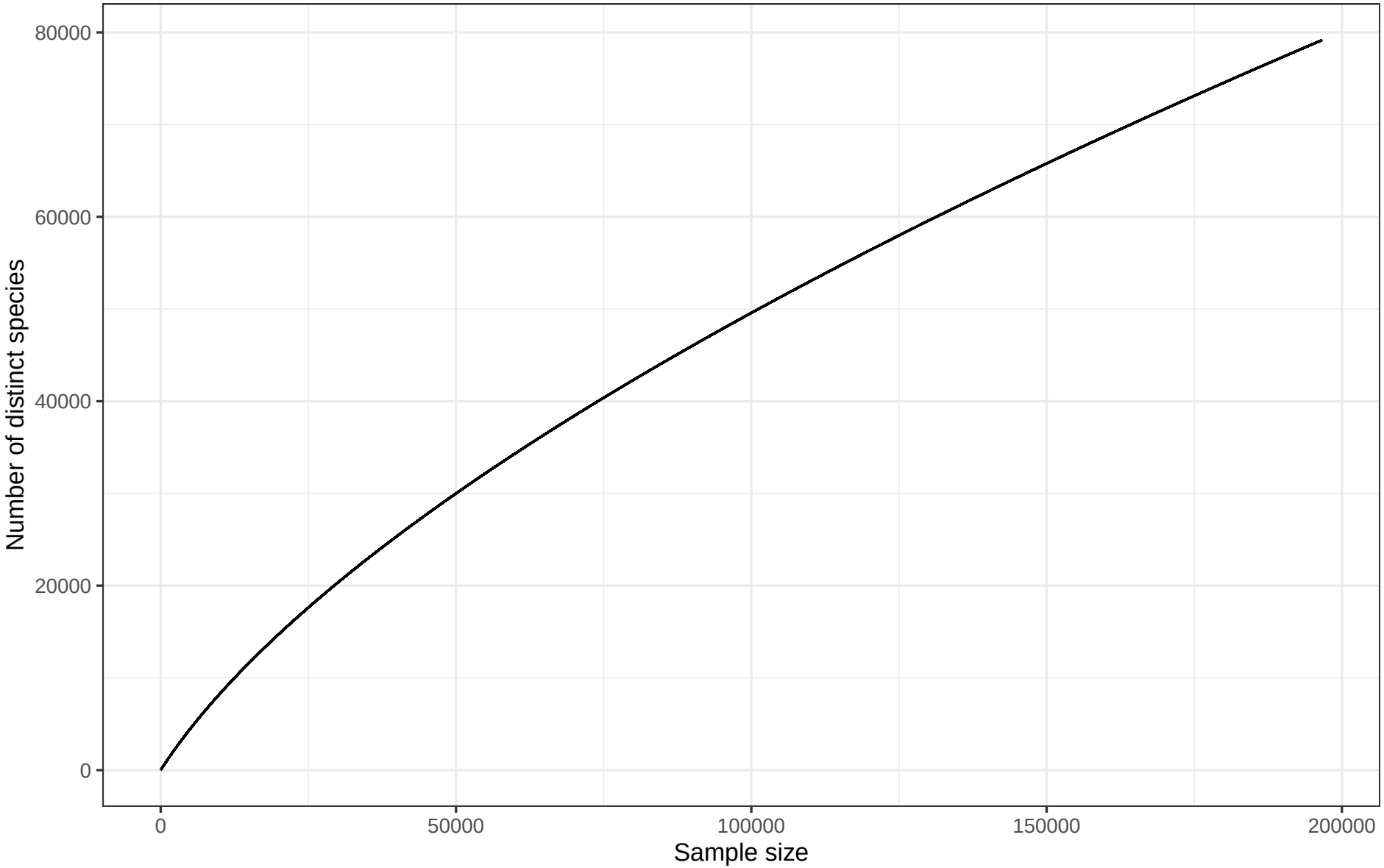
Moreover, a Bayesian estimate of the rarefaction curve is

$$\mathbb{E}(K_n) = \frac{\alpha}{\sigma}\left\{\frac{(\alpha+\sigma)_n}{(\alpha)_n} - 1\right\}.$$

Finally, the growth rate of the curve is $K_n \sim c_\sigma n^\sigma$ a.s.

- The marginal distribution of K_n is available in closed form and can be easily simulated.
- The rarefaction curve does not depend on the ordering.
- This 2-parameter curve is virtually indistinguishable from the usual rarefaction curve.

Goal II: Accumulation curves and rarefaction



Goal II: extrapolation

Lemma III (Favaro et al., JRSS-B, 2009)

Let $K_m^{(n)}$ be the number of new species we observe in an additional sample of size m. Then

$$\mathbb{E}(K_m^{(n)} \mid K_n = K) = \left(K + \frac{\alpha}{\sigma}\right) \left\{ \frac{(\alpha + n + \sigma)_m}{(\alpha + n)_m} - 1 \right\}.$$

- Suppose we re-conduct the Finnish fungal experiment. How many new-species will we get, assuming other conditions are mostly unchanged (locations, pre-processing, etc.)?
- A Bayesian estimate is $\mathbb{E}(K_n^{(n)} \mid K_n = 79,155) = 46,610$.
- **Caution zone**. Extrapolating the data is always a risky practice. The results are reliable only if the model is correctly specified — which is hard to check.

Goal III: species diversity

Lemma IV

An a priori Bayesian estimate of the Gini index is

$$\mathbb{E}(G) = \mathbb{P}(X_i \neq X_j) = \frac{\alpha + \sigma}{\alpha + 1}.$$

An a posteriori Bayesian estimate of the Gini index is

$$\mathbb{E}(G \mid X_1, \ldots, X_n) = 1 - \frac{1}{(\alpha + n)_2} \left\{ (1 - \sigma)(\alpha + K\sigma) + \sum_{j=1}^{K} (n_j - \sigma)_2 \right\}.$$

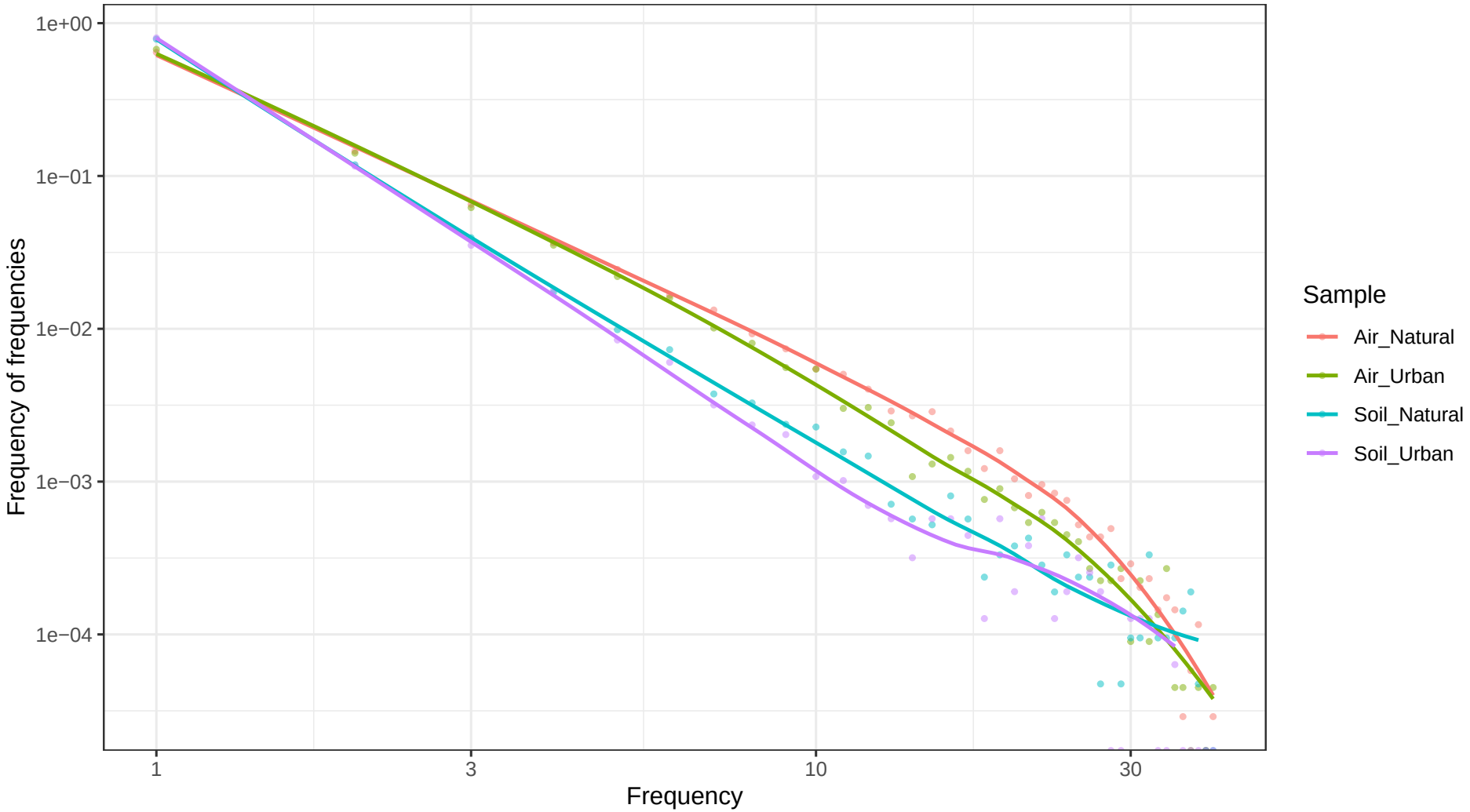
- **Note**. The prior and posterior distribution of G can be easily sampled thanks to the stick-breaking representation.
- Both prior and posterior moments (i.e. the variance) of G are also analytically available.
- Pluggin in the estimates for $\hat{\alpha}$ and $\hat{\sigma}$, we get in the Finnish fungal data

$$\mathbb{E}(G) = 0.9999455, \qquad \mathbb{E}(G \mid X_1, \ldots, X_n) = 0.9999422.$$

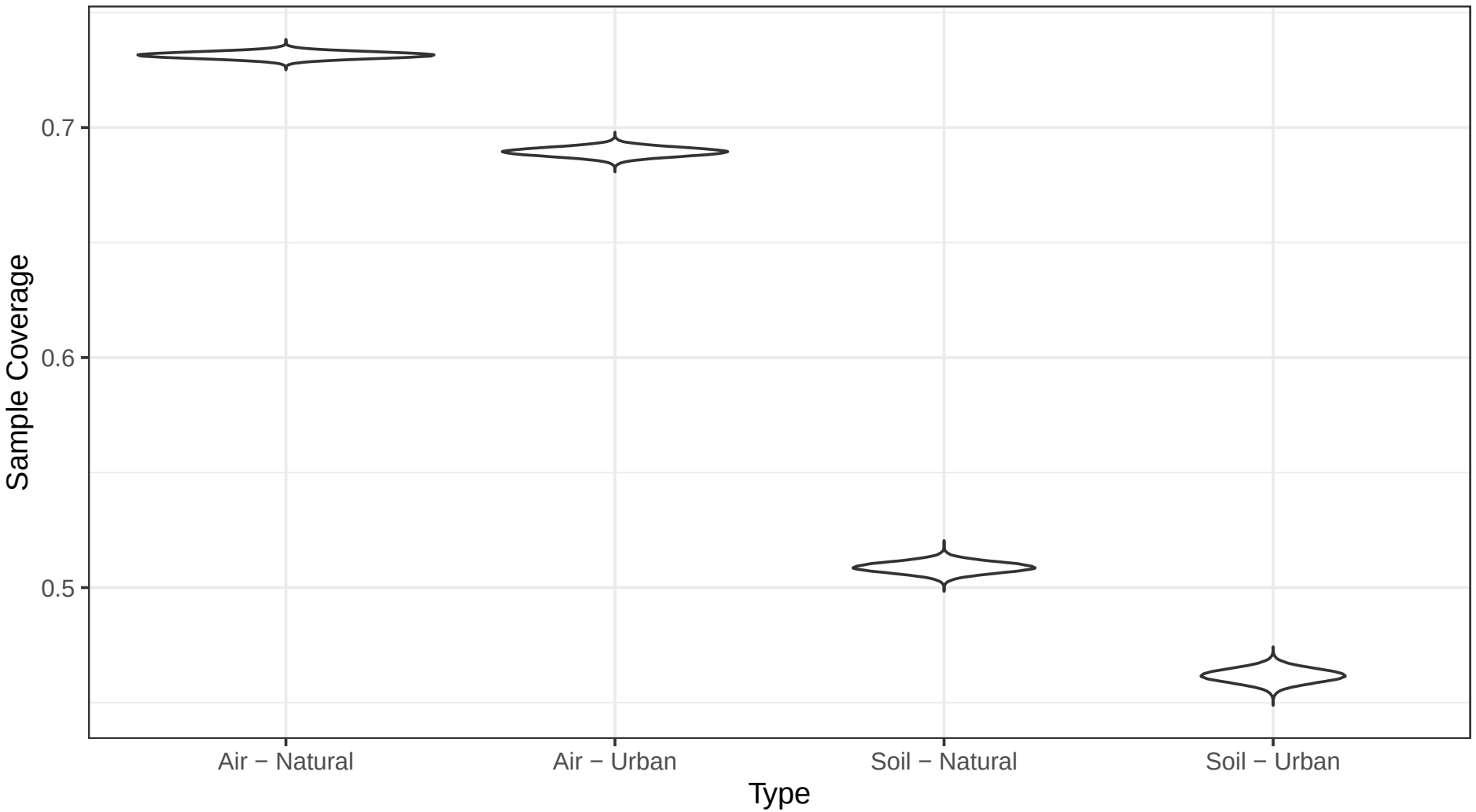
Air & Soil, Natural & Urban

- So far we applied the BNP modeling strategy on the whole Finnish fungal dataset.
- However, samples can be roughly divided into 4 groups: `Air`, `Soil`, `Natural` and `Urban`. We expected marked differences within these 4 groups.
- We re-conduct these analyses considering 4 groups of frequencies, one for each type of samples.
- We estimated 4 **independent** PY models.
- **Note**. To make the analyses comparable, we **randomly discarded** 14 samples out of 174, so that `Air`, `Soil`, `Natural` and `Urban` comprise 40 samples each.

Frequency of frequencies (m_k)



Posterior distribution of C_n (sample coverage)

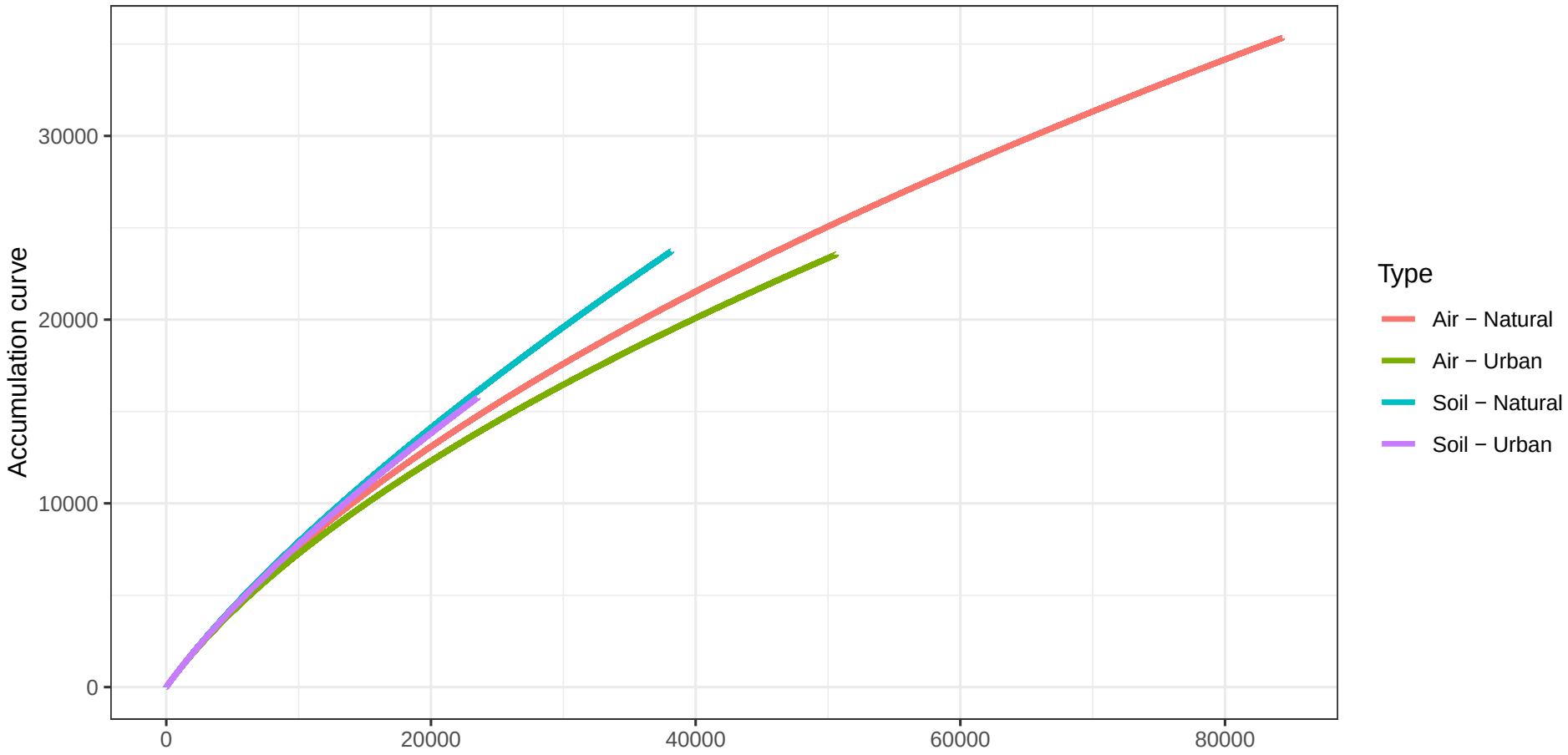


Summary statistics

Type	n	K	$\hat{\alpha}$	$\hat{\sigma}$
`Air - Natural`	84,323	35,332	5,788.10	0.52
`Air - Urban`	50,617	23,541	4,202.82	0.54
`Soil - Natural`	38,155	23,716	3,257.68	0.72
`Soil - Urban`	23,524	15,750	2,977.10	0.72

Type	$\mathbb{E}(C_n \mid \text{data})$	$\mathbb{E}(K_n^{(n)} \mid \text{data})$	$\mathbb{E}(G \mid \text{data})$	m_1/K
`Air - Natural`	0.731453	19,079	0.999923	0.641345
`Air - Urban`	0.689426	13,378	0.999893	0.667686
`Soil - Natural`	0.508769	16,963	0.999901	0.786726
`Soil - Urban`	0.461726	11,479	0.999889	0.800127

Accumulation curves



Comments

Ecological comments

- There are major differences in terms of all the indicators between the four groups. Differences are more marked between `Air` and `Soil`.
- `Air` samples have higher sample coverage and higher richness.
- Although in `Soil` we detect less species, the growth rate is higher than in `Air`.
- Growth rates of `Natural` vs `Urban` are very similar.
- Biodiversity measured by the Gini index is higher in `Natural` than in `Urban`

Statistical comments

- In order to confirm this differences, we could perform Bayesian testing e.g. through Bayes Factors.

Next directions

- With BNP modeling we can do much more than what we have shown here.
- Different biodiversity measures can be considered (Shannon entropy, Tsallis diversity, etc).
- The posterior distribution of the **proportions** π_h of **each species** is available in closed form. Hence, we can easily test for example whether a specific species / family / etc is more prevalent in `Air` vs `Soil` and quantify the associated uncertainty.
- We could consider more refined groups (`core` vs `edge`) or locations (`Helsinki` vs `Lahti` vs etc.). This calls for hierarchical specifications for the parameters α and σ, borrowing strength across samples.
- More sophisticate models (e.g. hierarchical PY) accounting for shared species can be also considered.