UNIVERSITY OF TORINO

M.Sc. in Stochastics and Data Science

Final dissertation

A nonparametric empirical Bayes approach to disclosure risk assessment

Supervisor: Prof. Stefano FAVARO

Candidate: Francesca PANERO (763925)

ACADEMIC YEAR 2016/2017

Contents

Introduction

In the study of survey and census data, microdata are constituted by records of individual answers regarding a set of variables (*e.g.* age, home address, educational level, employment status...). This type of collection represents a precious source of information to perform data analysis, but to preserve their correctness and availability, the sensitive information that may be contained inside must be protected from intruders. If a sample of these data has to be published, it is necessary to ask ourselves if someone that is participating to the sample can be disclosed. This crucial issue is known as *disclosure risk*.

Historically, two approaches have been considered to tackle disclosure risk. A cryptographic and younger one, called *differential privacy*, that aims at distorting data with noise, still preserving the information that can be extracted from these datasets. The second - the one we will follow - is statistical and older, and aims at quantifying the risk by estimating some quantities related to the sample.
Among the quantities mostly used in literature we must mention the number of population uniques and the number of sample uniques that are also population uniques. In order to give an intuition of this last measure, we briefly need to describe how datasets are constructed. Microdata are constituted by rows, one for each individual, and columns, one for each variable. Since each person is characterized by his tuple of answers, a sample unique is defined as an individual whose combination of answers is different from any other tuple of responses in the sample. The underlying idea of this measure is that a person present in the sample could be disclosed only if he is unique in that sample and also unique in the population. In fact, if he was unique in the sample, but in the population there were two or more people with answers identical to his ones, the probability of picking the right person would decrease a lot. This measure permit to assess if a sample can be published or not, by fixing a threshold maximal value and observing if the estimated number of sample uniques that are also population uniques lays below that threshold. In that case, it is reasonably safe to publish those sample.

This problem of estimation is closely related to species sampling problems, a typical

issue in Bayesian nonparametric statistics. In fact, each tuple of answers can be identified as belonging to a given species, and our statistic aims at finding how many species that appeared only once in the sample won't appear anymore in the rest of the population. One of the most famous examples of species sampling problems dealt with Bayesian statistics arose in 1953 and 1956, when Good and Toulming [5], [6] provided an estimator of the missing mass (that is, how many species we did not observe in the sample but exist in the population), using for the first time, even if not formalizing it, a nonparametric empirical Bayesian approach. In 1955, Robbins [9] formalized this approach in a theoretical framework, and only after twenty years, in 1976, Efron and Thisted [4] highlighted that the estimator produced by Good and Toulmin was an anticipatory application of the nonparametric empirical Bayesian statistics in the sense of Robbins (as it is now known).

In these last decades, instead, another empirical Bayesian approach has become famous: the parametrical one of Efron and Morris [3]. In line with this, one of the first estimator of sample uniques that are also population uniques was proposed by Bethlehem et al. in 1990 [1] and improved by Skinner et al. in 1994 [12].
Even if very popular, these estimators lack of robustness. In fact, the parameters of the prior of the proposed model need to be estimated somehow with the data (for example, with the method of moments or maximum likelihood estimation). Skinner in 2002 [13] highlighted this problem '[...] In summary, we suggest that no inference procedure is currently available which robustly estimates the probability of population uniques $\mathbb{P}(\mathrm{PU})$ or predicts population uniques that are also sample uniques $\mathbb{P}(\mathrm{PU}|\mathrm{SU})$ across the wide range of possible population structures that may exist in surveys and for small sampling fractions". And from 2002 there was no improvement in this sense.

In this work, we propose a robust nonparametric empirical Bayesian estimator of sample uniques that are also population uniques.
In the first chapter, we will present Bethlehem's and Skinner's approaches, to which we will compare our estimator. In the second one, we will explain the nonparametric empirical Bayes approach, that constitutes the theoretical basis of our proposal. Finally, in the third chapter, we will provide the derivation of our estimator and present our results regarding the decay of the normalizing mean square error of the estimator, that will permit to offer the limit of predictability of it. In the conclusions, we will test our estimators on five different samples taken from real census data and discuss the open problems linked to our approach.

Chapter 1

Disclosure control of microdata

1.1 Two similar approaches

One of the first approaches to tackle the problem of disclosure risk with statistical methods arose in March 1990, when J. G. Bethlehem, W. J. Keller and J. Pannekoek published in the Journal of American Statistical Association the paper "Disclosure Control of Microdata".
They considered a population of N individuals divided in K species. To each species x is assigned a superpopulation parameter $p_x > 0$ and a random variable F_x that represents the frequency of that species in the population, and they assumed $F_x \sim Poisson(Np_x)$. f_x, $x = 1, \dots, K$ are their corresponding values in the sample. Then, the expected number of population uniques is simply equal to

$$U_p = \mathbb{E}\left[\sum_{x=1}^{K} \mathbb{1}_{F_x=1}\right] = \sum_{x=1}^{K} e^{-Np_x} Np_x.$$

Their approach to estimate U_p exploited the classical *Poisson-Gamma* model, where data are distributed as Poisson random variables and their means are modelled with Gamma prior distributions. More formally:

$$P_x \overset{iid}{\sim} gamma(\alpha, \beta)$$
$$F_x \sim Poisson(Np_x)|P_x = p_x$$

Of course, we should have $\sum_{x=1}^{K} P_x = 1$, but for simplicity the authors simply assumed $\sum_{x=1}^{K} \mathbb{E}(P_x) = K\alpha\beta = 1$. In this way, $\alpha = 1/K\beta$ and we can reduce our estimation just to one parameter, β. When the exact K is not available, Bethlehem's suggestion is to estimate assuming a uniform distribution over the species:

$$\hat{K} = N/\overline{f} \tag{1.1}$$

where $\overline{f}$ is the sample mean of the f_x.
This model implies that the marginal distribution of each F_x is a negative binomial. In fact:

$$\begin{aligned}
\mathbb{P}(F_x = i_x) &= \int \mathbb{P}(F_x = i_x|p_x) gamma(dp_x) = \\
&= \int e^{-Np_x} \frac{(Np_x)^{i_x}}{i_x!} \frac{1}{\Gamma(\alpha)\beta^\alpha} e^{-\frac{p_x}{\beta}} {p_x}^{\alpha-1} dp_x = \\
&= \frac{N^{i_x}}{\Gamma(\alpha)\beta^\alpha i_x!} \int e^{-(N+1/\beta)p_x} p_x^{(\alpha+i_x)-1} dp_x = \\
&= \frac{\Gamma(\alpha+i_x)}{\Gamma(\alpha)\Gamma(i_x+1)} \frac{\beta^{i_x} N^{i_x}}{\cancel{\beta^\alpha \beta^{i_x}}} \cancel{\beta^{\alpha+i}} (N\beta+1)^{-(\alpha+i_x)} = \\
&= \frac{\Gamma(\alpha+i_x)}{\Gamma(\alpha)\Gamma(i_x+1)} \left(\frac{N\beta}{N\beta+1}\right)^{i_x} \left(1 - \frac{N\beta}{N\beta+1}\right)^{\alpha} = \\
&= \binom{\alpha+i_x-1}{i_x+1} \left(1 - \frac{N\beta}{N\beta+1}\right)^{\alpha}
\end{aligned}$$

i.e. F_x is distributed as a $Negative\ Binomial\left(\alpha, \frac{N\beta}{N\beta+1}\right)$, whose expected value and variance are

$$\begin{aligned}
\mathbb{E}[F_x] &= N\alpha\beta \\
Var(F_x) &= N\alpha\beta(1+N\beta)
\end{aligned}$$

Under this model, the expected value of population uniques can be computed as:

$$\begin{aligned}
U_p = \mathbb{E}\left[\sum_{x=1}^{K} \mathbb{1}_{F_x=1}\right] &= K\mathbb{P}(F_x = 1) \\
&= K\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\Gamma(2)} \frac{N\beta}{(1+N\beta)^{1+\alpha}} = \\
&= N\frac{K\alpha\beta}{(1+N\beta)^{1+\alpha}} = \\
&= N(1+N\beta)^{-(1+\alpha)}
\end{aligned}$$

To estimate the expected number of uniques in sample who are also uniques in the population, they simply computed the sample proportion of the estimated population uniques: $\hat{U}_{ps}^B = \hat{U}_p n/N$.

Since the idea of simply taking a proportion was not convincing, four years later Skinner et al. proposed in 1994 [12] an improvement of this model to ameliorate this estimation. In particular, they added the specification of the distribution of f_x: $f_x \sim$

$Poisson(np_x)|P_x = p_x$, where n is the sample size, from which it easily follows that $F_x - f_x \sim Poisson((N-n)p_x)|P_x = p_x$. The probability of being population unique is

$$\mathbb{P}(\text{population unique}) = \frac{1}{N}U_p = \frac{1}{N}K\mathbb{P}(F_x = 1) = (1+N\beta)^{-(1+\alpha)}$$

whereas the probability of observing a sample unique is

$$\begin{aligned}\mathbb{P}(\text{sample unique}) =& \frac{1}{n}U_s = \frac{1}{n}K\mathbb{P}(f_x = 1) = \frac{1}{\cancel{n}}K\cancel{n}\alpha\beta(1+n\beta)^{-(1+\alpha)} = \\ =& (1+n\beta)^{-(1+\alpha)}\end{aligned} \tag{1.2}$$

where, analogously, U_s is the expected number of sample uniques.
The probability for a single species to be a population and a sample unique is found exploiting the distribution of $F_x - f_x$:

$$\begin{aligned}\mathbb{P}(F_x = 1, f_x = 1) &= \int \mathbb{P}(F_x = 1, f_x = 1|p_x)Gamma(dp_x) = \\ &= \int \mathbb{P}(F_x - f_x = 0|p_x)\mathbb{P}(f_x = 1|p_x)Gamma(dp_x) = \\ &= \int e^{-(N-n)p_x}e^{-np_x}np_x\frac{1}{\Gamma(\alpha)\beta^\alpha}e^{-\frac{p_x}{\beta}}p_x^{\alpha-1}dp_x = \\ &= \frac{n}{\Gamma(\alpha)\beta^\alpha}\int e^{-(N+\frac{1}{\beta})}p_x^{(\alpha+1)-1}dp_x = \\ &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\frac{n\beta}{\cancel{\beta^\alpha}\cancel{\beta}}\left(\frac{\cancel{\beta}}{N\beta+1}\right)^{\alpha+1} = \\ &= n\alpha\beta(1+N\beta)^{-(1+\alpha)}\end{aligned}$$

so the probability that an individual is both a sample and a population unique can be computed

$$\begin{aligned}\mathbb{P}(\text{sample and population unique}) &= \frac{1}{n}\mathbb{E}[\text{population and sample uniques}] = \\ &= \frac{1}{n}K\mathbb{P}(F_x = 1, f_x = 1) = \\ &= (1+N\beta)^{-(1+\alpha)}\end{aligned} \tag{1.3}$$

and exploiting 1.2 and 1.3, we obtain

$$\mathbb{P}(\text{population unique} \mid \text{sample unique}) = \left(\frac{1+N\beta}{1+n\beta}\right)^{-(1+\alpha)} \tag{1.4}$$

The estimates $\hat{\alpha}, \hat{\beta}$ of α, β can be obtained by equating the observed proportion of sample uniques to the probability of being a sample unique 1.2 and solving with Newton's

iterative method. Plugging these estimates in 1.4, an estimate of the proportion of sample uniques that are also population uniques can be obtained with:

$$\begin{aligned} U_{ps}^{S} &= k\mathbb{P}(F_x = 1|f_x = 1) = \\ &= k\frac{n\alpha\beta(1+N\beta)^{-(1+\alpha)}}{n\alpha\beta(1+n\beta)^{-(1+\alpha)}} = \\ &= k\left(\frac{1+N\beta}{1+n\beta}\right)^{-(1+\alpha)} \end{aligned}$$

1.2 Estimation of the parameters

In both models, in order to find U_{ps} we must rely on some methods to estimate α and β. In both cases, we can exploit the method of moments and the maximum likelihood estimation, keeping in mind also the condition $\alpha = 1/(K\beta)$.
In the former case, using the observed sample moments (i.e. of the observed values f instead of F), the estimators are:

$$\begin{aligned} N\hat{\alpha}^{\text{MM}}\hat{\beta}^{\text{MM}} &= \overline{f} \\ N\hat{\alpha}^{\text{MM}}\hat{\beta}^{\text{MM}}(1+N\hat{\beta}^{\text{MM}}) + (N\hat{\alpha}^{\text{MM}}\hat{\beta}^{\text{MM}})^2 &= \overline{f^2} \end{aligned}$$

Where $\bar{\cdot}$ represents the sample mean. Substituting the first equation into the second:

$$\begin{aligned} &\overline{f}(1+N\hat{\beta}^{\text{MM}}) + \overline{f}^2 = \overline{f^2} \\ &\hat{\beta}^{\text{MM}} = \frac{\overline{f^2} - \overline{f}^2}{N\overline{f}} - \frac{1}{N} \rightarrow \hat{\beta}^{\text{MM}} = \frac{1}{k-1}\sum_{x=1}^{k}\left(\frac{(f_x - \overline{f})^2}{N\overline{f}}\right) - \frac{1}{N} \end{aligned}$$

where $s^2 = \frac{1}{k-1}\sum_{x=1}^{k}\left(f_x - \overline{f}\right)^2$ is the corrected sample variance.

We can also rely on maximum likelihood estimation of $\hat{\alpha}^{\text{MLE}}$ and $\hat{\beta}^{\text{MLE}}$. In fact, being $F_x \overset{iid}{\sim} Negative\ Binomial\left(\alpha, \frac{N\beta}{N\beta+1}\right)$, defining $p = \frac{N\beta}{N\beta+1}$ we can compute the likelihood and the log-likelihood functions:

$$\begin{aligned} \mathcal{L}(\alpha, p) &= \prod_{x=1}^{k}\frac{\Gamma(\alpha+i_x)}{\Gamma(\alpha)\Gamma(i_x+1)}p^{i_x}(1-p)^{\alpha} \\ \ell(\alpha, p) = \ln(\mathcal{L}(\alpha, p)) &= \sum_{x=1}^{k} ln\left(\Gamma(\alpha+i_x)\right) - \sum_{x=1}^{k}\ln(i_x!) - k\ln(\Gamma(\alpha)) + \\ &+ \sum_{x=1}^{k} i_x \ln(p) + k\alpha\ln(1-p) \end{aligned}$$

from which we can take derivatives:

$$(\bullet)\ \frac{\partial \ell}{\partial p}(\alpha, p) = \frac{n}{p} - \frac{k\alpha}{1-p}$$
$$(\star)\ \frac{\partial \ell}{\partial \alpha}(\alpha, p) = \sum_{x=1}^{k} \psi(i_x + \alpha) - k\psi(\alpha) + k\ln(1-p)$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. Equating $(\bullet)$ to 0 we get

$$p = \frac{n}{n + k\alpha}$$

Substituting in $(\star)$ and equating to 0 we get

$$\frac{\partial \ell}{\partial \alpha} = \sum_{x=1}^{k} \psi(i_x + \alpha) - k\psi(\alpha) + k\ln\left(1 - \frac{n}{n+k\alpha}\right) = 0$$

Since there does not exist a closed solution for this equation, relying on numerical methods we can find a approximate solutions $\hat{\alpha}^{\text{MLE}}$ and $\hat{\beta}^{\text{MLE}}$.

Hence, exploiting these two methods, we can provide and compare four different estimates of U_{ps}, two for Bethlehem's $\hat{U}^B_{ps} = \hat{U}_p n/N$ and two for Skinners $\hat{U}^S_{ps} = k\left((1+N\hat{\beta})/(1+n\hat{\beta})\right)^{-(1+\hat{\alpha})}$.

1.3 The problem of constant estimates

Unfortunately, Bethlehem's approach is not sensitive to changes in population's size.
Let us first consider $\hat{U}^{\text{MM}}_{ps}$ computed with the estimates $\hat{\alpha}^{\text{MM}}$ and $\hat{\beta}^{\text{MM}}$ obtained with the method of moments.

$$\begin{aligned}
\hat{U}^{MM}_{ps} &= \frac{n}{N}\hat{U}^{\text{MM}}_{p} = \\
&= \frac{n}{N} N(1+N\hat{\beta}^{\text{MM}})^{-(1+\hat{\alpha}^{\text{MM}})} \overset{\bullet}{=} \\
&= n\left(1 + N\left(\frac{1}{k-1}\sum_{x=1}^{k}\left(\frac{(f_x-\overline{f})^2}{N\overline{f}}\right) - \frac{1}{N}\right)\right)^{-1-\overline{f}/(N\hat{\beta})} = \\
&= n\left(1 + \left(\frac{1}{k-1}\sum_{x=1}^{k}\left(\frac{(f_x-\overline{f})^2}{\overline{f}}\right) - 1\right)\right)^{-1-\overline{f}/(\frac{1}{k-1}\sum_{x=1}^{k}\frac{(f_x-\overline{f})^2}{\overline{f}}-1)}
\end{aligned}$$

where in $\bullet$ we used the estimate of K 1.1. As we can see, $\hat{U}^{\text{MM}}_{ps}$ does not depend on N, and this seems puzzling from a physical point of view. In fact, as N grows, we

expect a non-increasing behavior for this estimator, since it becomes less likely for the sample uniques of being also population uniques, as described by the original $\hat{U}_{ps}^{\text{MM}} = n(1+N\hat{\beta}^{\text{MM}})^{-(1+\hat{\alpha}^{\text{MM}})}$ that is a decreasing function of N. But imposing $\alpha = 1/(K\beta)$ and $\hat{K} = N/\overline{f}$, we loose this dependence. This way, the estimator is useless, since it cannot understand the changes in the population's size.

The same happens with $\hat{U}_{ps}^{\text{MLE}}$, computed with the estimates $\hat{\alpha}^{\text{MLE}}$ and $\hat{\beta}^{\text{MLE}} = 1/\hat{\alpha}^{\text{MLE}}$ and $\hat{K} = N/\overline{f}$.

$$\begin{aligned}\hat{U}_{ps}^{\text{MLE}} &= \frac{n}{N}N(1+N\beta^{\text{MLE}})^{-(1+\alpha^{\text{MLE}})} = \\ &= n\left(1+\frac{N\overline{f}}{N\alpha^{\text{MLE}}}\right)^{-(1+\alpha^{\text{MLE}})} = \\ &= n\left(1+\frac{\overline{f}}{\alpha^{\text{MLE}}}\right)^{-(1+\alpha^{\text{MLE}})}\end{aligned}$$

that again does not depend on N.

Chapter 2

Empirical Bayes approach

Differently from the previous parametric approaches, we want to rely on nonparametric statistics in order to find a robust estimator. In particular, we will exploit Herbert Robbins' empirical version of Bayesian statistics, born in 1955 when Robbins introduced it during the Third Berkeley Symposium on Mathematical Statistics and Probability.

In a classical Bayesian model, we assume to have data represented by a random variable X - that we assume for simplicity to be discrete - whose distribution function depends on a parameter θ:

$$p(x|\bar{\theta}) = \mathbb{P}(X = x|\theta = \bar{\theta})$$

We know that one of the biggest differences between the Bayesian and the frequentist approach lies in the nature of the parameter θ: frequentistis believe θ to be a fixed value, while Bayesians think of it as a random variable.
Hence, θ has its own distribution function G over its support Θ:

$$G(\bar{\theta}) = \mathbb{P}(\theta \leq \bar{\theta})$$

Then, the unconditional distribution of the data is obtained by integrating out the parameter:

$$p_G(x) = \mathbb{P}(X = x) = \int_\Theta p(x|\theta)G(\theta)d\theta \tag{2.1}$$

The estimator $\phi(X)$ of θ that minimizes $\mathbb{E}[\phi(X) - \theta]^2$, the expected square deviation from the parameter itself, is the expected value of the *a posteriori* distribution of θ given $X = x$:

$$\phi_G(X) = \frac{\int_\Theta p(x|\theta)G(d\theta)}{\int_\Theta p(x|\theta)G(d\theta)} \tag{2.2}$$

Robbins did not want to impose a prior distribution for the parameter. He believed in the information that data carried with them, and wished to exploit it, instead of choosing a

prior distribution, maybe in situations in which the experimenter did not know anything about this function.
Suppose we collect a sample of size n of data and the corresponding generating parameters:

$$(X_1, \theta_1), (X_2, \theta_2), \ldots, (X_n, \theta_n)$$

We know that the empirical distribution function of the parameter is

$$G_{n-1}(\theta) = \frac{\text{number of terms } \theta_1, \ldots, \theta_{n-1} \text{ that are } \leq \theta}{n-1} \tag{2.3}$$

Taking G_{n-1} as prior distribution for the parameter, the estimator of θ becomes:

$$\psi_n(x) = \frac{\int_\Theta p(x|\theta)\theta G_{n-1}(d\theta)}{\int_\Theta p(x|\theta) G_{n-1}(d\theta)}$$

Since $G_{n-1}(\lambda) \overset{a.s.}{\to} G(\lambda)$, $\psi_n(x)$ will tend to ϕ_G defined in 2.2 for any fixed x under suitable regularity conditions on the kernel $p(x|\lambda)$. Actually, since we cannot observe the true values of $\lambda_1, \ldots, \lambda_{n-1}$, 2.3 cannot be computed.
Robbins' idea was to exploit the observable values of $X_1, \ldots, X_{n-1}$ to compute the estimator 2.2. We observe that the empirical frequency of data

$$p_n(x) = \frac{\text{number of terms } X_1, \ldots, X_n \text{ which are equal to } x}{n}$$

tends almost surely to $p_G(x)$ 2.1, regardless of the *a priori* distribution of the parameter. The question now aims to understand if from an approximate value of 2.1, where $p(x|\theta)$ is known, it is possible to find the prior G, or at leat the value of the estimator 2.2 (and we will restrict our consideration to this case). This possibility will depend on the nature of $p(x|\theta)$ and on the set $\mathcal{G}$ of functions to which G belongs. In order to understand this idea better, we will report a few examples.

First, consider the Poisson kernel:

$$\begin{aligned} p(x|\lambda) &= e^{-\lambda}\frac{\lambda^x}{x!} \qquad x = 0, 1, \ldots; \lambda > 0 \\ p_G(x) &= \int p(x|\lambda) dG(\lambda) = \int_o^\infty e^{-\lambda}\lambda^x dG(\lambda)/x! \\ \phi_G(x) &= \frac{\int_0^\infty e^{-\lambda}\lambda^{x+1} dG(\lambda)}{\int_0^\infty e^{-\lambda}\lambda^x dG(\lambda)} \end{aligned}$$

Observe that

$$\phi_G(x) = (x+1)\frac{p_G(x+1)}{p_G(x)}$$

and construct the function

$$\phi_n(x) = (x+1)\frac{p_n(x+1)}{p_n(x)}$$

Then, regardless of the prior distribution G, we have that

$$\phi_n(x) \overset{a.s.}{\to} \phi_G(x) \text{ for } n \to +\infty$$

For a binomial kernel, instead:

$$p_r(x|\lambda) = \binom{r}{x}\lambda^x(1-\lambda)^{r-x}\ x = 0, \dots, r; 0 \leq \lambda \leq 1$$
$$p_{G,r}(x) = \binom{r}{x}\int_0^1 \lambda^x(1-\lambda)^{r-x}dG(\lambda)$$
$$\phi_{G,r}(x) = \frac{\int_0^1 \lambda^{x+1}(1-\lambda)^{r-x}dG(\lambda)}{\int_0^1 \lambda^x(1-\lambda)^{r-x}dG(\lambda)}$$

Hence,

$$\phi_{G,r}(x) = \frac{x+1}{r+1}\frac{p_{G,r+1}(x+1)}{p_{G,r}(x)}$$

Given

$$p_{n,r}(x) = \frac{\text{number of terms in } X_1, \dots, X_n = x}{n}$$

we know that $p_{n,r} \overset{a.s.}{\to} p_{G,r}(x)$ as $n \to +\infty$.
Now consider a new sequence of random variables $X_1^{'}, \dots, X_n^{'}, \dots$ that represent the number of successes in the first $r-1$ out of r trials which produced the original sequence and let $p_{n,r-1}$ be the empirical distribution function of these new variables. Then, $p_{n,r-1} \overset{a.s.}{\to} p_{G,r-1}(x)$ as $n \to +\infty$. Thus if we consider

$$\phi_{n,r}(x) = \frac{x+1}{r}\frac{p_{n,r}(x+1)}{p_{n,r-1}(x)}$$

we obtain the almost sure convergence to the function

$$\phi_{G,r-1}(x) = \frac{x+1}{r}\frac{p_{G,r}(x+1)}{p_{G,r-1}(x)}$$

If we take as our estimate of Λ_n the value $\phi_{G,r}(X_n^{'})$ we will do about as well as we knew the *a priori* G but limited to the first $r-1$ trials, and for large r this won't be a problem.

Chapter 3

The empirical Bayes estimator

3.1 The setting

Consider a population of individuals $(X_i)_{i\geq 1}$ divided into species and a sample $(X_1, \dots, X_N)$ of them. To identify the membership of each individual of the sample to a group, we assign to each X_i a label X_i^* that indicates his species. Each species is associated with an unknow proportion $(p_x)_{x\geq 1} \overset{iid}{\sim} G$, where G is a distribution function over the simplex $\Delta^\infty = \left\{x = (x_1, x_2, \dots) : \sum_{i\geq 1} x_i = 1\right\}$.
We count how many individuals belong to the species observed and record this information in a frequency vector of size $K_N \leq N$: $(N_{1,N}, \dots, N_{K_N,N})$, where $N_{x,N} = \sum_{i=1}^N \mathbb{1}_{(X_i = X_x^*)}$. Of course, it must be $\sum_{x=1}^{K_N} N_{x,N} = N$.
We record also the frequencies of the frequencies, *i.e.* the vector $(M_{1,N}, \dots, M_{N,N})$ that represents how many time we observed a species with given cardinality, where $M_{i,N} = \sum_{x\geq 1} \mathbb{1}_{(N_{x,N}=i)}$ represents the number of species whose observed cardinality was i. Then, $\sum_{i=1}^N M_{i,N} = K_N$ and $\sum_{i=1}^N iM_{i,N} = N$. Define m_i as the observed quantity in the sample of $M_{i,N}$.

Among several species sampling models to which the "extrapolation approach" has been applied, the most common are the Multinomial and the Poisson abundance model. Both assume that the initial observable samples are independent and identically distributed according to the species proportion $(p_x)_{x\geq 1}$. In the Multinomial model the sample size is a fixed number $n \geq 0$, whereas in the Poisson model it is assumed to be a Poisson random variable with parameter n. Hence, under the Poisson model, $N \sim Pois(n)$ and the frequencies $N_{x,N} \overset{indep}{\sim} Pois(np_x)$. In this work we will follow this model, but similar results should be expected with the Multinomial model, even if with a more tedious derivation.

Since we want to make inference on the individuals that we do not observe, we define the size of the rest of the population λN and consider the additional sample $(X_{N+1}, \dots, X_{\lambda N+N})$. Again, $(N_{x,\lambda N})_{x\geq 1}$ represents the vector of frequencies of each species x in the sample of size λN.

Sample uniques that are also population uniques are the species x that appear in the first sample just one time, *i.e.* $N_{x,N} = 1$, an do not appear in the "second sample" (the rest of the population), *i.e.* $N_{x,\lambda N} = 0$. This means that the estimator we are looking for is

$$T_1 = \sum_{x\geq 1} \mathbb{1}_{(N_{x,N}=1)} \mathbb{1}_{(N_{x,\lambda N}=0)}$$

3.2 Empirical Bayes estimator for $\lambda < 1$

Since T_1 is a random quantity and its estimation is far more difficult, we switch to the estimation of its expected value. In a setting where $N_{x,N} \sim Pois(p_x n)$ and $N_{x,\lambda N} \sim Pois(\lambda p_x n)$, the statistics we care about becomes:

$$\mathbb{E}(T_1) = \sum_{x\geq 1} \mathbb{P}(N_{x,N} = 1)\mathbb{P}(N_{x,\lambda N} = 0) = \sum_{x\geq 1} e^{-p_x n} \frac{np_x}{1!} e^{-\lambda p_x n} = \sum_{x\geq 1} e^{-(\lambda+1)p_x n} np_x$$

Robbins approach suggests us to consider the posterior Bayes estimator

$$\begin{aligned}
\hat{T}_1^i &= \frac{\int \mathbb{E}(T_1) Pois(np) G(dp)}{\int Pois(np) G(dp)} = \\
&= \frac{\int e^{-(\lambda+1)np} \frac{(np)^1}{1!} e^{-np} \frac{(np)^i}{i!} G(dp)}{\int e^{-np} \frac{(np)^i}{i!} G(dp)} = \\
&\overset{\star}{=} \frac{\sum_{j=0}^{\infty} \int \frac{(np)^1}{1!} \frac{(-(\lambda+1)np)^j}{j!} e^{-np} \frac{(np)^i}{i!} G(dp)}{\int e^{-np} \frac{(np)^i}{i!} G(dp)} = \\
&= \frac{\sum_{j=0}^{\infty} \frac{(-(\lambda+1)^j)}{j!} \frac{(i+j+1)!}{1!i!} \overbrace{\int e^{-np} \frac{(np)^{i+j+1}}{(i+j+1)!} G(dp)}^{\mathbb{E}[M_{i+j+1}]}}{\underbrace{\int e^{-np} \frac{(np)^i}{i!} G(dp)}_{\mathbb{E}[M_i]}}
\end{aligned}$$

Where in $\star$ we exploited the series of the exponential function $e^t = \sum_{j=0}^{\infty} \frac{t^j}{j!}$. We substitute the expected value of $M_{i,N}$ and $M_{i+j-1,N}$ with their observed values m_i, m_{i+j+1}. We can then write:

$$\hat{T}_1^i = \sum_{j=0}^{\infty} (-(\lambda+1))^j \frac{(i+j+1)!}{1!i!j!} \frac{m_{i+j+1}}{m_i}$$

Since this expression represents the estimator for a fixed value i of the species' occurrence, we must sum over all the frequencies i:

$$\begin{aligned}
\hat{T}_1 &= \sum_{i=0}^{\infty} m_i \sum_{j=0}^{\infty} (-(\lambda+1))^j \frac{(i+j+1)!}{1!i!j!} \frac{m_{i+j+1}}{m_i} = \\
&= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} (-(\lambda+1))^j \frac{(i+j+1)!}{1!i!j!} m_{i+j+1} = \\
&= \sum_{i=0}^{\infty} \sum_{j=i}^{\infty} (-(\lambda+1))^{j-i} \frac{(j+1)!}{1!i!(j-i)!} m_{j+1} = \\
&= \sum_{j=0}^{\infty} \sum_{i=0}^{j} (-(\lambda+1))^{j-i} \frac{(j+1)!}{1!i!(j-i)!} m_{j+1} = \\
&= \sum_{j=0}^{\infty} (-(\lambda+1))^j \frac{(j+1)!}{1!} m_{j+1} \underbrace{\sum_{i=0}^{j} (-(\lambda+1))^{-i} \frac{1}{i!(j-i)!}}_{\star} = \\
&= \sum_{j=0}^{\infty} (-(\lambda+1))^j \frac{(j+1)!}{1!} m_{j+1} \left(\frac{\lambda}{\lambda+1}\right)^j = \\
&= \sum_{j=0}^{\infty} (-1)^j (j+1) m_{j+1} \lambda^j
\end{aligned}$$

where in $\star$ we used the binomial theorem.

This series is a geometrical one, and it does converge if and only if his ratio λ is less or equal to 1. Calling N^* the size of the population, since $\lambda = \frac{N^*-N}{N}$, we reach the following condition for the convergence:

$$\lambda = \frac{N^* - N}{N} \leq 1 \iff N^* \leq 2N$$

that is, if and only if our sample size represents at least half of the population. Obviously, in most of the cases this sample size is not achieved in surveys. Hence, in most of real cases, this estimator would diverge and hence be useless.

Our estimator belongs to the more general class of linear estimators

$$T^h = \sum_{i \geq 1} m_i h_i$$

where the coefficients h_i are identified by a power series $h(y) = \sum_{i \geq 1} \frac{h_i y^i}{i!}$. In fact, our

estimator:

$$\hat{T_1} = \sum_{i\geq 1} \underbrace{(-\lambda)^{i-1} i}_{h_i} m_i$$

$$h(y) = \sum_{i\geq 1} \frac{h_i y^i}{i!} = \sum_{i\geq 1} \frac{(-\lambda)^{i-1} i y^i}{i!} = y \sum_{i\geq 1} \frac{(-\lambda y)^{i-1}}{(i-1)!} = y \sum_{i\geq 0} \frac{(-\lambda y)^i}{i!} = y e^{-\lambda y}$$

From now on, for the sake of a lighter notation, we will define $N_x := N_{x,N}$ and $N_x^{'} := N_{x,\lambda N}$. Also, all the logarithms, unless otherwise specified, are considered with natural basis.

Theorem 3.2.1. *$\hat{T_1}$ is an unbiased estimator for T_1*

Proof. The bias of the estimator is

$$\begin{aligned}
\mathbb{E}[\hat{T_1} - T_1] &= \mathbb{E}[\sum_{i\geq 1} m_i h_i - \sum_x \mathbb{1}_{N_x=1} \mathbb{1}_{N_x^{'}=0}] = \\
&= \sum_x \mathbb{E}[\sum_{i\geq 1} \mathbb{1}_{N_x=i} h_i - \mathbb{1}_{N_x=1} \mathbb{1}_{N_x^{'}=0}] = \\
&= \sum_x \sum_{i\geq 1} \Big(\mathbb{P}(N_x = i) h_i - \mathbb{P}(N_x = 1) \mathbb{P}(N_x^{'} = 0) \Big) = \\
&= \sum_x \left(\sum_{i\geq 1} e^{-np_x} \frac{np_x^i}{i!} h_i - e^{-np_x} np_x e^{-\lambda np_x} \right) = \\
&= \sum_x e^{-np_x} \sum_{i\geq 1} \left(\frac{(np_x)^i h_i}{i!} - e^{-\lambda np_x} np_x \right) = \\
&= \sum_x e^{-np_x} (h(np_x) - e^{-\lambda np_x} np_x) = 0
\end{aligned}$$

□

Theorem 3.2.2.

$$Var(\hat{T_1} - T_1) = \mathbb{E}[(\hat{T_1} - T_1)^2] \leq \mathbb{E}[M_+] \Psi^2(\lambda) - \frac{1}{\lambda + 1} \mathbb{E}[M_{1,N+\lambda N}]$$

where $M_+ = \sum_{i\geq 1} M_{i,N}$ and $\Psi(\lambda) := (i^* + 1)\lambda^{i^*}$ with $i^* = argmax_{i\geq 1} h_i^2 = \left\lfloor \frac{2\lambda - 1}{1 - \lambda} \right\rfloor \vee 0$.

Proof.

$$Var\left(\sum_{i\geq 1} \mathbb{1}_{N_x=i} h_i - \mathbb{1}_{N_x=1} \mathbb{1}_{N_x^{'}=0} \right) \leq \mathbb{E}[(\sum_{i\geq 1} \mathbb{1}_{N_x=i} h_i - \mathbb{1}_{N_x=1} \mathbb{1}_{N_x^{'}=0})^2] =$$

$$= \mathbb{E}[(\sum_{i\geq 1} \mathbb{1}_{N_x=i} h_i)^2] + \mathbb{E}[\mathbb{1}_{N_x=1} \mathbb{1}_{N_x^{'}=0}] - 2\mathbb{E}[\underbrace{(\sum_{i\geq 1} \mathbb{1}_{N_x=i} h_i) \mathbb{1}_{N_x=1}}_{\mathbb{1}_{N_x=1} h_1} \mathbb{1}_{N_x^{'}=0}] =$$

$$= \mathbb{E}[\sum_{i\geq 1} \mathbb{1}_{N_x=i} h_i^2 + 2\sum_{i<j} \underbrace{\mathbb{1}_{N_x=i}\mathbb{1}_{N_x=j}}_{0}] + \mathbb{E}[\mathbb{1}_{N_x=1}]\mathbb{E}[\mathbb{1}_{N'_x=0}] - 2\underbrace{h_1}_{1}\mathbb{E}[\mathbb{1}_{N_x=1}]\mathbb{E}[\mathbb{1}_{N'_x=0}] =$$

$$= \sum_{i\geq 1} \mathbb{E}[\mathbb{1}_{N_x=i}]h_i^2 - \mathbb{E}[\mathbb{1}_{N_x=1}]\mathbb{E}[\mathbb{1}_{N'_x=0}]$$

Since the variance of the sum of independent random variables is the sum of their variances:

$$\begin{aligned} Var(\hat{T}_1 - T_1) &\leq \sum_x \sum_{i\geq 1} \mathbb{E}[\mathbb{1}_{N_x=i}]h_i^2 - \sum_x \mathbb{E}[\mathbb{1}_{N_x=1}]\mathbb{E}[\mathbb{1}_{N'_x=0}] = \\ &= \sum_{i\geq 1} \mathbb{E}[M_{i,N}]h_i^2 - \sum_x e^{-n(\lambda+1)p_x} np_x \leq \\ &\leq \mathbb{E}[M_+] max_{i\geq 1} h_i^2 - \frac{1}{\lambda+1}\mathbb{E}[M_{1,N+\lambda N}] \end{aligned}$$

where $M_+ = \sum_{i\geq 1} M_{i,N}$. Now, we can compute the value that maximizes h_i^2. Observe that:

$$\begin{aligned} h_{i+1}^2 = (i+2)^2\lambda^{2i+2} &\geq (i+1)^2\lambda^{2i} = h_i^2 \\ (i+2)\lambda &\geq i+1 \\ i &\leq \frac{2\lambda - 1}{1-\lambda} \end{aligned}$$

Then, $i^* = argmax_{i\geq 1} h_i^2 = \left\lfloor \frac{2\lambda-1}{1-\lambda} \right\rfloor \vee 0$. Defining $\Psi(\lambda) := (i^*+1)\lambda^{i^*}$, we obtain the thesis. □

Since $\hat{T}_1$ is an unbiased estimator for $\lambda < 1$ and we have a bound for $Var(\hat{T}_1 - T_1)$, we may obtain confidence intervals via Bernstein inequalities with variance factor as in 3.2.2. See Boucheron et al. (2013) [2] for details.

We start by bounding the log-Laplace random variable to obtain a concentration inequality for the difference $\hat{T}_1 - T_1$:

$$\begin{aligned} log\mathbb{E}\left[exp\{\mu(\hat{T}_1 - T_1)\}\right] &= log\mathbb{E}\left[exp\left\{\mu\sum_{i\geq 1}\sum_x \mathbb{1}_{N_x=i}h_i - \sum_x \mathbb{1}_{N_x=1}\mathbb{1}_{N'_x=0}\right\}\right] = \\ &= log\mathbb{E}\left[\left\{\mu\sum_x\left(\sum_{i\geq 1}\sum_x \mathbb{1}_{N_x=i}h_i - \mathbb{1}_{N_x=1}\mathbb{1}_{N'_x=0}\right)\right\}\right] = \\ &\overset{indep}{=} \sum_x log\mathbb{E}\left[exp\left\{\mu\left(\sum_{i\geq 1}\mathbb{1}_{N_x=i}h_i - \mathbb{1}_{N_x=1}\mathbb{1}_{N'_x=0}\right)\right\}\right] \end{aligned}$$

To limit the argument of the exponential function, we exploit the fact that the coefficients h_i admit a unique maximum, as proved in 3.2.2.

$$\left|\sum_{i\geq 1}\mathbb{1}_{N_x=i}h_i - \mathbb{1}_{N_x=1}\mathbb{1}_{N'_x=0}\right| \leq |h_i^*| = (-\lambda)^{i^*-1}i^* = \Psi(\lambda)$$

with $i^* = \left\lfloor \frac{2\lambda-1}{1-\lambda} \right\rfloor \vee 0$.
Now we can exploit Bennett's inequality:

$$log\mathbb{E}\left[e^{\mu(\hat{T}_1 - T_1)}\right] \leq \sum_x Var\left(\sum_{i\geq 1} \mathbb{1}_{N_x=i} h_i - \mathbb{1}_{N_x=1}\mathbb{1}_{N'_x=0}\right) \frac{\phi(\Psi(\lambda)\mu)}{\Psi^2(\lambda)} =$$
$$= Var(\hat{T}_1 - T_1)\frac{\phi(\Psi(\lambda)\mu)}{\Psi^2(\lambda)}$$

where $\phi(\lambda) := e^\lambda - 1 - \lambda$.
Fixing $y > 0$, thanks to this bound on the log-Laplace, the following Bernstein inequality holds true:

$$\mathbb{P}\left(|\hat{T}_1 - T_1)| \geq y\right) \leq 2exp\left\{-\frac{y^2}{2\left(Var(\hat{T}_1 - T_1) + \Psi(\lambda)\frac{y}{3}\right)}\right\}$$

To obtain the confidence interval, we write

$$\frac{y^2}{2\left(Var(\hat{T}_1 - T_1) + \Psi(\lambda)\frac{y}{3}\right)} = s$$
$$y^2 - 2sVar(\hat{T}_1 - T_1) - 2s\Psi(\lambda)\frac{y}{3} = 0$$
$$y = s\Psi(\lambda)\frac{1}{3} + \sqrt{s^2\Psi^2(\lambda)\frac{1}{9} + 2sVar(\hat{T}_1 - T_1)}$$

that yields the desired confidence interval for $\hat{T}_1$:

$$\mathbb{P}\left(\left|\hat{T}_1 - T_1\right| \geq s\Psi(\lambda)\frac{1}{3} + \sqrt{s^2\Psi^2(\lambda)\frac{1}{9} + 2sVar(\hat{T}_1 - T_1)}\right) \leq 2e^{-s}$$

3.3 Empirical Bayes estimator for $\lambda \geq 1$

The problem of this estimator arises when $\lambda > 1$, since the estimator is biased and $sup_{i\geq 1}|h_i| = sup_{i\geq 1}\lambda^{i-1}i = +\infty$, leading to a large variance. To avoid the large variance (but not the problem of the bias), we could think of truncating the estimator at a fixed threshold l, using the function

$$h^l(y) = y\sum_{i=0}^{l}\frac{(-\lambda y)^i}{i!}$$

for which $sup_{i\geq 1}|h_i| = \lambda^{l-1}l$. But this decision leads to a very large bias, since there exists always a distribution that exhibits most of the species l times.

Instead, in order to force the convergence of the coefficients of the series when $\lambda \geq 1$, one common method exploited is the Euler transformation.
Let $\lambda = \frac{u}{2-u}$ and substitute it in the original estimator $\hat{T}_1$, obtaining

$$\begin{aligned}
\hat{T}_1^E &= \sum_{i\geq 0}(-1)^i(i+1)\left(\frac{u}{2-u}\right)^i m_{i+1} = \\
&= \sum_{i\geq 0}(-1)^i \underbrace{(i+1)m_{i+1}}_{\eta_{i+1}} \left(\frac{u}{2}\right)^i \left(1-\frac{u}{2}\right)^{-i} = \\
&\overset{\star}{=} \sum_{i\geq 0}(-1)^i \eta_{i+1} \left(\frac{u}{2}\right)^i \sum_{y\geq 0}\binom{-i}{y}\left(-\frac{u}{2}\right)^y = \\
&= \sum_{i\geq 0}\sum_{y\geq 0}(-1)^{i+y}\eta_{i+1}\binom{-i}{y}\left(\frac{u}{2}\right)^{i+y} = \\
&= \sum_{i\geq 0}\sum_{z\geq i}(-1)^z\eta_{i+1}\binom{-i}{z-i}\left(\frac{u}{2}\right)^z = \\
&\overset{\bullet}{=} \sum_{z\geq 0}\sum_{i=0}^{z}(-1)^i\eta_{i+1}\binom{z-1}{i-1}\left(\frac{u}{2}\right)^z
\end{aligned}$$

where in $\star$ we used the Taylor expansion of $\left(1-\frac{u}{2}\right)^{-i}$ and in $\bullet$ the identity $(-1)^z\binom{-i}{z-i} = (-1)^i\binom{z-1}{i-1}$. Now, define

$$\xi_z = \sum_{i=0}^{z}(-1)^i\frac{1}{2^z}\eta_{i+1}\binom{z-1}{i-1}$$

and consider the Euler transformated estimator truncated at k and the original estimator truncated at k

$$\begin{aligned}
\Delta_k(u) &= \sum_{z=0}^{k}\xi_z u^z \\
\Delta_k(\lambda) &= \sum_{i=0}^{k}(-1)^i(i+1)\lambda^i m_{i+1}
\end{aligned}$$

These two truncated sums, for $k \to +\infty$ both converge to $\hat{T}_1$, but the Euler's truncated one is quicker than the original one.
We can exploit a different formula for $\Delta_k(u)$, trying to recover the probability of the tails of a binomial random variable

$$\Delta_k(u) = \sum_{z=0}^{k}\xi_z u^z = \sum_{z=0}^{k}\sum_{i=0}^{z}(-1)^i\eta_{i+1}\binom{z-1}{i-1}\left(\frac{1}{2^z}\right)u^z =$$

$$
\begin{aligned}
&= \sum_{i=0}^{k}(-1)^i\eta_{i+1}\sum_{z=i}^{k}\binom{z-1}{i-1}\left(\frac{1}{2^z}\right)u^z \\
&= \sum_{i=0}^{k}(-1)^i\eta_{i+1}\sum_{z=i}^{k}\binom{z-1}{i-1}\left(\frac{1}{2^z}\right)\left(\frac{2\lambda}{\lambda+1}\right)^z = \\
&= \sum_{i=0}^{k}(-1)^i\eta_{i+1}\sum_{z=i}^{k}\binom{z-1}{i-1}\left(\frac{\lambda}{\lambda+1}\right)^z = \\
&= \sum_{i=0}^{k}(-1)^i\eta_{i+1}\lambda^i\sum_{z=i}^{k}\binom{z-1}{i-1}\left(\frac{\lambda^{z-i}}{(\lambda+1)^z}\right) = \\
&= \sum_{i=0}^{k}(-1)^i\eta_{i+1}\lambda^i\sum_{z=i}^{k}\binom{z-1}{i-1}\left(\frac{1}{\lambda+1}\right)^i\left(1-\frac{1}{\lambda+1}\right)^{z-i} = \\
&= \sum_{i=0}^{k}(-1)^i\eta_{i+1}\lambda^i\mathbb{P}(L\geq i)
\end{aligned}
$$

where $L \sim Binom(k, \frac{1}{\lambda+1})$. However, regardless of the choice of k, there exists a distribution such that the $k-th$ term will dominates causing a large bias.

Following Orlitsky et al. [8] we decided, instead of truncating at a fixed threshold l, to consider a random truncation at a value L, distributed as a given probability distribution function over the set of non-negative integers. If L is a random variable over $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$, define:

$$
\begin{aligned}
h^L(y) = \sum_{l\geq 0}\mathbb{P}(L=l)h^l(y) = \sum_{l\geq 0}\mathbb{P}(L=l)y\sum_{i=0}^{l}\frac{(-\lambda y)^i}{i!} = y\sum_{i\geq 0}\frac{(-\lambda y)^i}{i!}\mathbb{P}(L\geq i) = \\
= \sum_{i\geq 1}\frac{(-\lambda)^{i-1}iy^i}{i!}\mathbb{P}(L\geq i-1)
\end{aligned}
$$

This change in the function h leads to the estimator

$$\hat{T}_1^L = \sum_{i\geq 0}(-1)^i(i+1)\lambda^i\mathbb{P}(L\geq i)m_{n,i+1}$$

This smoothed estimator is linear with coefficients $h_i^L = (-\lambda)^{i-1}i\mathbb{P}(L\geq i-1)$. Note that for $L = +\infty$ we get the original estimator, and for $L = l$ the truncated estimator.

We now try to bound the bias and variance for this class of smoothed estimators.

Theorem 3.3.1. *For any $\lambda \geq 1$, $\hat{T}_1^L$ is a biased estimator of T_1. In particular,*

$$\mathbb{E}[\hat{T}_1^L] = \mathbb{E}[T_1] + \sum_x e^{-(\lambda+1)p_x n}p_x n\int_0^{\lambda np_x} e^s\mathbb{E}_L\left[\frac{(-s)^L}{L!}\right]ds$$

Proof. Using theorem 3.2.1, we know that

$$\mathbb{E}[\hat{T}_1^L - \hat{T}_1] = \sum_x \left(e^{-np_x}(h^L(np_x) - np_x e^{-\lambda np_x})\right)$$

Let us define the function

$$g(y) = \sum_{i\geq 0} \frac{(-y)^i}{i!}\mathbb{P}(L \geq i)$$

and note that $h^L(y) = yg(\lambda y)$. Compute

$$\begin{aligned}
g(y) - e^{-y} &= \sum_{i\geq 0} \frac{(-y)^i}{i!}\mathbb{P}(L \geq i) - \sum_{i\geq 0} \frac{(-y)^i}{i!} = \\
&= -\sum_{i\geq 0} \frac{(-y)^i}{i!}\mathbb{P}(L < i) = \\
&= -\sum_{i\geq 1}\sum_{j=0}^{i-1} \frac{(-y)^i}{i!}\mathbb{P}(L = j) = \\
&= -\sum_{j\geq 0}\sum_{i\geq j+1} \frac{(-y)^i}{i!}\mathbb{P}(L = j) \overset{\star}{=} \\
&= -\sum_{j\geq 0} \frac{e^{-y}}{j!}\int_o^{-y} \tau^j e^{-\tau} d\tau \mathbb{P}(L = j) = \\
&= -e^{-y}\int_0^{-y} e^{-\tau} d\tau \left(\sum_{j\geq 0} \frac{\tau^j}{j!}\mathbb{P}(L = j)\right) = \\
&= e^{-y}\int_0^{y} e^{s} ds \left(\sum_{j\geq 0} \frac{s^j}{j!}\mathbb{P}(L = j)\right) = \\
&= e^{-y}\int_0^{y} \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] e^s ds
\end{aligned}$$

where in $\star$ we used the definition of incomplete Gamma function:

$$\sum_{i\geq j+1} \frac{z^i}{i!} = \frac{e^z}{j!}\int_0^z \tau^j e^{-\tau} d\tau$$

and we swapped integral and sum with Fubini's theorem.
So,

$$\begin{aligned}
\mathbb{E}[\hat{T}_1^L - \hat{T}_1] &= \sum_x \left(e^{-np_x}(h^L(np_x) - np_x e^{-\lambda np_x})\right) \\
&= \sum_x e^{-np_x} np_x \left(g(\lambda np_x) - e^{-\lambda np_x}\right) =
\end{aligned}$$

$$= \sum_x e^{-np_x} np_x e^{-\lambda np_x} \int_0^{\lambda np_x} \mathbb{E}_L \left[\frac{(-s)^L}{L!} \right] e^s ds =$$
$$= \sum_x e^{-(\lambda+1)np_x} np_x \int_0^{\lambda np_x} \mathbb{E}_L \left[\frac{(-s)^L}{L!} \right] e^s ds$$

that proves the first part of our theorem.

□

Theorem 3.3.2.

$$Var(\hat{T}_1^L - T_1) \leq \mathbb{E}[M_+]\Psi_L(\lambda)^2 - \frac{1}{\lambda+1}\mathbb{E}[M_{1,\lambda N+N}]$$

with $\Psi_L(\lambda) = \mathbb{E}[\lambda^L(L+1)]$

Proof. For the variance, we exploit the bound on the variance of the original estimator: we need to find a bound for the coefficients h_i^L:

$$|h_i^L| \leq \lambda^{i-1} i \mathbb{P}(L \geq i-1) = \lambda^{i-1} i \sum_{j \geq i-1} \mathbb{P}(L=j) \leq \sum_{j \geq i-1} \mathbb{P}(L=j)\lambda^j(j+1) \leq$$
$$\leq \mathbb{E}[\lambda^L(L+1)]$$

Using this bound and the previous result we obtain:

$$Var(\hat{T}_1^L - T_1) \leq \mathbb{E}[M_+]\mathbb{E}^2[\lambda^L(L+1)] - \mathbb{E}[T_1] = \mathbb{E}[M_+]\mathbb{E}^2[\lambda^L(L+1)] - \frac{1}{\lambda+1}\mathbb{E}[M_{1,\lambda N+N}]$$

□

Now, our aim is finding a bound for the normalized mean square error of $\hat{T}_1^L$. The mean square error is defined as

$$MSE[\hat{T}_1^L] := \mathbb{E}[(\hat{T}_1^L - T_1)^2] = Var(\hat{T}_1^L - T_1) + Bias[\hat{T}_1^L]^2$$

Since $0 \leq T_{\lambda n}^L \leq n$, we can define the worst-case normalized mean square error as in [8].

Definition 3.3.1. For an estimator $\hat{T}$ for a statistic T such that $0 \leq T \leq n$, the worst-case **normalized mean square error** (NMSE) is

$$\mathcal{E}_{n,\lambda}(\hat{T}) = sup_L \mathbb{E}_L \left[\left(\frac{\hat{T} - T}{n} \right)^2 \right]$$

Combining theorems 3.3.1 and 3.3.2 we derive the following corollary.

Corollary 3.3.1. *For any $\lambda \geq 1$, the NMSE of $\hat{T}_1^L$ is upper bounded by*

$$\mathcal{E}_{n,\lambda}(\hat{T}_1^L) \leq \frac{1}{n^2}\mathbb{E}[M_+]\Psi_L(\lambda)^2 - \frac{1}{n^2(\lambda+1)}\mathbb{E}[M_{1,\lambda N+N}] + \frac{1}{n^2} + \\ + \frac{1}{n^2}\left(\sum_x e^{-(\lambda+1)p_x n} p_x n \int_0^{\lambda n p_x} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] ds\right)^2$$

We are looking for distributions L such that their associated NMSE vanishes as $n \to +\infty$.

3.3.1 Poisson smoothing

For $L \sim Pois(r)$, we first try to compute a bound for the bias of the estimator $\hat{T}_1^L$. We first need to bound $e^{-r}\sum_{l\geq 0}\frac{(r\lambda)^l}{l!}(l+1)$:

$$e^{-y}\int_0^y e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] ds = e^{-y}\int_0^y e^s \sum_{k=0}^{+\infty} e^{-r}\frac{r^k}{k!}\frac{(-s)^k}{k!} ds = \\ = e^{-(y+r)}\int_0^y e^s \sum_{k=0}^{+\infty}\frac{(rs)^k(-1)^k}{\Gamma(k+1)k!} ds = \\ \overset{\bullet}{=} e^{-(y+r)}\int_0^y e^s J_0(2\sqrt{sr}) ds$$

where in $\bullet$ we exploited the definition of the Bessel function:

$$J_0(z) := \sum_{k=0}^{+\infty}\frac{(-1)^k z^{2k}}{2^{2k}\Gamma(k+1)k!}$$

and since $J_0(z) \in [-1,1]$, $J_0(z) \leq 1$:

$$\left|e^{-y}\int_0^y e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] ds\right| \leq e^{-(r+y)}(e^y - 1) = e^{-r}(1 - e^{-y})$$

Therefore,

$$|Bias(\hat{T}_1^L)| = \left|\mathbb{E}[\hat{T}_1^L - T_1]\right| = \left|\sum_x e^{-(\lambda+1)np_x} np_x \int_0^{\lambda n p_x} e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] ds\right| \leq \\ \leq \sum_x e^{-np_x} np_x e^{-r}(1 - e^{-\lambda n p_x}) \leq \\ \leq e^{-r}\sum_x e^{-np_x} np_x = e^{-r}\mathbb{E}[M_{1,n}] \leq e^{-r} n$$

To bound the variance, we first need to compute

$$\mathbb{E}[\lambda^L(L+1)] = e^{-r}\sum_{l\geq 0}\frac{(r\lambda)^l}{l!}(l+1) =$$

$$= e^{-r} r\lambda \sum_{l\geq 1} \frac{(r\lambda)^{l-1}}{(l-1)!} + e^{r(\lambda-1)} =$$
$$= e^{r(\lambda-1)} r\lambda + e^{r(\lambda-1)} =$$
$$= e^{r(\lambda-1)}(r\lambda+1)$$

Then,

$$Var(\hat{T}_1^L - T_1) \leq e^{2r(\lambda-1)}(r\lambda+1)^2 \mathbb{E}[M_+] - \frac{1}{\lambda+1}\mathbb{E}[M_{1,\lambda N+N}] \leq n e^{2r(\lambda-1)}(r\lambda+1)^2$$

Theorem 3.3.3. *Let $\lambda \geq 1$ amd $L \sim Pois(r)$. If $r = \frac{1}{4\lambda} ln\left(\frac{n}{2\lambda-1}\right)$, then*

$$NMSE(\hat{T}_1^L) \leq \frac{A(\lambda)}{n^{1/(2\lambda)}}$$

where $A(\lambda) = \frac{2\lambda}{(2\lambda-1)^{1-\frac{1}{2\lambda}}}$ is continuous in $[1, +\infty)$ and $lim_{\lambda\to+\infty} = 1$.

Proof. Using corollary 3.3.1 with the two previously found bounds for the bias and the variance, we find a limitation for the NMSE in the Poisson case:

$$\mathcal{E}_{n,\lambda}(\hat{T}_1^L) \leq \frac{e^{2r(\lambda-1)}(r\lambda+1)^2}{n} + e^{-2r} \overset{\star}{\leq} \underbrace{e^{-2r} + \frac{e^{2r(2\lambda-1)}}{n}}_{\bullet}$$

where in $\star$ we used the fact that $1 + x \leq e^x$. Taking the derivative with respect to r of the NMSE we can observe that for $r = \frac{1}{4\lambda} log\left(\frac{n}{2\lambda-1}\right)$ the NMSE reaches its minimum value. In fact:

$$\frac{d}{dr}\bullet = -2e^{-2r} + 2(2\lambda-1)\frac{e^{4r\lambda}e^{-2r}}{n} = 2e^{-2r}\left(\frac{(2\lambda-1)}{n}e^{4r\lambda} - 1\right)$$

Imposing the derivative equal to 0 we get:

$$e^{4r\lambda} = \frac{n}{2\lambda-1} \iff r = \frac{1}{4\lambda} log\left(\frac{n}{2\lambda-1}\right)$$

Substituting in $\bullet$ we get

$$NMSE(\hat{T}_1^L) = \left(\frac{n}{2\lambda-1}\right)^{-\frac{1}{2\lambda}} + \frac{1}{n}\left(\frac{n}{2\lambda-1}\right)^{1-\frac{1}{2\lambda}} =$$
$$= \left(\frac{n}{2\lambda-1}\right)^{-\frac{1}{2\lambda}}\left(1 + \frac{1}{2\lambda-1}\right) =$$
$$= \frac{(2\lambda-1)^{1/(2\lambda)}}{n^{1/(2\lambda)}}\frac{2\lambda}{2\lambda-1} =$$

$$= \frac{1}{n^{1/(2\lambda)}} \frac{2\lambda}{(2\lambda-1)^{1-\frac{1}{2\lambda}}}$$

Regarding the limit, it is straightforward to see that

$$lim_{\lambda\to+\infty} \frac{2\lambda}{(2\lambda-1)^{1-\frac{1}{2\lambda}}} = lim_{\lambda\to+\infty} \frac{2\lambda}{(2\lambda)^{1-\frac{1}{2\lambda}}} = lim_{\lambda\to+\infty} \frac{2\lambda}{2\lambda} = 1$$

□

Corollary 3.3.2. *Let $\lambda \geq 1$ and $L \sim Pois(r)$ with $r = \frac{1}{4\lambda} ln\left(\frac{n}{2\lambda-1}\right)$, then for any $\delta \in (0,1)$ we have*

$$\lim_{n\to+\infty} \frac{max\{\lambda : NMSE(\hat{T}_1^L \leq \delta)\}}{log(n)} \geq \frac{1}{2log(A/\delta)}$$

where $A := max_{\lambda\geq 1} A(\lambda)$

Proof. We already know that

$$NMSE(\hat{T}_1^L) \leq \frac{A}{n^{1/(2\lambda)}}$$

Observe also that

$$\frac{A}{n^{1/(2\lambda)}} \leq \delta \iff \lambda \leq \frac{log(n)}{2log(A/\delta)} = \lambda^*$$

As a consequence, the maximum value of λ for which the inequality $NMSE(\hat{T}_1^L) \leq \delta$ is satisfied, is bigger or equal than λ^*:

$$max\{\lambda : NMSE(\hat{T}_1^L) \leq \delta\} \geq \frac{log(n)}{2log(A/\delta)}$$

The thesis follows by taking the limit of the previous inequality as $n \to +\infty$. □

3.3.2 Binomial smoothing

We can repeat the previous step in a binomial setting, that is when $L \sim Binom(x_0, q))$, where we assume $q = 1/(\lambda+1)$ as in Efron's paper.

$$\begin{aligned} e^{-y}\int_0^y e^s \mathbb{E}_L\left[\frac{(-s)^L}{L!}\right] ds &= e^{-y}\int_0^y e^s \sum_{k=0}^{x_0} \binom{x_0}{k} q^k (1-q)^{x_0-k} \frac{(-s)^k}{k!} ds = \\ &= e^{-y}\int_0^y e^s (1-q)^{x_0} \sum_{k=0}^{x_0} \frac{x_0!(-1)^k}{(x_0-k)!k!k!} \left(\frac{sq}{1-q}\right)^k ds = \\ &= e^{-y}\int_0^y e^s (1-q)^{x_0} \sum_{k=0}^{x_0} (-1)^k \frac{1}{k!} \binom{x_0}{x_0-k} \left(\frac{sq}{1-q}\right)^k ds \end{aligned}$$

here we need to recall Laguerre's polynomials, namely

$$L_n^0(z) := \sum_{m=0}^{n} (-1)^m \binom{n}{n-m} \frac{z^m}{m!}$$

Since $|L_n^0(z)| \leq e^{z/2}$, we can bound the previous integral by

$$\begin{aligned}
\left| e^{-y} \int_0^y e^s \mathbb{E}_L \left[\frac{(-s)^L}{L!} \right] ds \right| &\leq e^{-y} \int_0^y e^s (1-q)^{x_0} |L_{x_0}^0(sq/(1-q))| ds \leq \\
&\leq e^{-y} \int_0^y e^s (1-q)^{x_0} e^{sq/(2(1-q))} ds = \\
&\leq e^{-y} (1-q)^{x_0} \int_0^y exp\left\{ s\left(1 + \frac{q}{2(1-q)}\right)\right\} ds = \\
&= \frac{2(1-q)^{x_0+1}}{2-q} (e^{yq/(2(1-q))} - e^{-y})
\end{aligned}$$

Substituting y with $\lambda n p_x$, we obtain

$$\begin{aligned}
|Bias(\hat{T}_1^L)| &= \left| \sum_x e^{-np_x} np_x e^{-\lambda np_x} \int_0^{\lambda np_x} e^s \mathbb{E}\left[\frac{(-s)^L}{L!} \right] ds \right| \leq \\
&\leq \frac{2(1-q)^{x_0+1}}{2-q} \sum_x e^{-np_x} np_x \left(e^{\lambda np_x q/(2(1-q))} - e^{-\lambda np_x} \right) = \\
&= 2\left(\frac{\lambda}{\lambda+1}\right)^{x_0+1} \frac{\lambda+1}{2\lambda+1} \sum_x e^{-np_x} np_x \left(e^{\lambda np_x/(2\lambda)} - e^{\lambda np_x} \right) = \\
&= 2\left(\frac{\lambda}{\lambda+1}\right)^{x_0+1} \frac{\lambda+1}{2\lambda+1} \left(2\sum_x e^{-np_x/2} \frac{np_x}{2} - \sum_x e^{-(\lambda+1)np_x} np_x \right) = \\
&= 4\left(\frac{\lambda}{\lambda+1}\right)^{x_0+1} \frac{\lambda+1}{2\lambda+1} \mathbb{E}\left[M_{1,N/2}\right] - 2\left(\frac{\lambda}{\lambda+1}\right)^{x_0+1} \frac{\lambda+1}{2\lambda+1} \mathbb{E}\left[M_{1,\lambda N+N}\right] \leq \\
&\leq 4\left(\frac{\lambda}{\lambda+1}\right)^{x_0+1} \frac{\lambda+1}{2\lambda+1} \mathbb{E}\left[M_{1,N/2}\right] \leq \\
&\leq 2\left(\frac{\lambda}{\lambda+1}\right)^{x_0+1} \frac{\lambda+1}{2\lambda+1} n
\end{aligned}$$

Now we try to bound the variance of $\hat{T}_1^L$. Let us consider

$$\begin{aligned}
\Psi_L(\lambda) = \mathbb{E}_L\left[(L+1)\lambda^L\right] &= \sum_{k=0}^{x_0} (k+1)\lambda^k \binom{x_0}{k} q^k (1-q)^{x_0-k} = \\
&= \sum_{k=1}^{x_0} \frac{x_0!}{(k-1)!(x_0-k)!} (\lambda q)^k (1-q)^{x_0-k} + (1+q(\lambda-1))^{x_0} = \\
&= \sum_{k=0}^{x_0-1} \frac{x_0!}{k!(x_0-k-1)!} (\lambda q)^{k+1} (1-q)^{x_0-k-1} + (1+q(\lambda-1))^{x_0} =
\end{aligned}$$

$$= x_0\lambda q \sum_{k=0}^{x_0-1} \binom{x_0-1}{k} (\lambda q)^k (1-q)^{x_0-1-k} + (1+q(\lambda-1))^{x_0} =$$
$$= x_0\lambda q (1+q(\lambda-1))^{x_0-1} + (1+q(\lambda-1))^{x_0} =$$
$$= (1+q(\lambda-1))^{x_0-1}(x_0\lambda q + 1 + q(\lambda-1))$$

And substituting $q = \frac{1}{\lambda+1}$ we get

$$\Psi_L(\lambda) = \left(1+\frac{\lambda-1}{\lambda+1}\right)^{x_0-1}\left(x_0\frac{\lambda}{\lambda+1}+1+\frac{\lambda-1}{\lambda+1}\right) =$$
$$= \left(\frac{2\lambda}{\lambda+1}\right)^{x_0-1}\frac{(x_0+2)\lambda}{\lambda+1} =$$
$$= \left(\frac{\lambda}{\lambda+1}\right)^{x_0} 2^{x_0-1}(x_0+2)$$

Then, the bound for the variance becomes

$$Var(\hat{T}_1^L - T_1) \leq \mathbb{E}[M_+]\Psi_L(\lambda)^2 - \frac{1}{\lambda+1}\mathbb{E}[M_{1,\lambda N+N}] \leq$$
$$\leq \left(\frac{\lambda}{\lambda+1}\right)^{2x_0} 2^{2(x_0-1)}(x_0+2)^2\mathbb{E}[M_+] - \frac{1}{\lambda+1}\mathbb{E}[M_{1,\lambda N+N}] \leq$$
$$\leq \left(\frac{\lambda}{\lambda+1}\right)^{2x_0} 2^{2(x_0-1)}(x_0+2)^2 n$$

Theorem 3.3.4. *Let $\lambda \geq 1$ and $L \sim Binom(x_0, 1/(\lambda+1))$ with*

$$x_0 = \left\lfloor \frac{2}{7} log_2\left(\frac{n(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)\right\rfloor$$

then

$$NMSE(\hat{T}_1^L) \leq \frac{B(\lambda)}{n^{log_2(1+1/\lambda)4/7}}$$

where $B(\lambda) = \left(\frac{(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)^{4log_2(\lambda/(\lambda+1))/7}\left(\frac{(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)} + \left(\frac{2\lambda}{2\lambda+1}\right)^2\right)$ *is continuous in* $[1,+\infty)$ *and* $lim_{\lambda\to+\infty} = 1$.

Proof. Using corollary 3.3.1 with the two previously found bounds for the bias and the variance, we can control the NMSE of $\hat{T}_1^L$ this way:

$$NMSE(\hat{T}_1^L) \leq \frac{1}{n}\left(\frac{\lambda}{\lambda+1}\right)^{2x_0} 2^{2(x_0-1)}(x_0+2)^2 + 4\left(\frac{\lambda}{\lambda+1}\right)^{2(x_0+1)}\left(\frac{\lambda+1}{2\lambda+1}\right)^2 =$$
$$= \left(\frac{\lambda}{\lambda+1}\right)^{2x_0}\left(\frac{2^{2x_0}}{n}\left(1+\frac{x_0}{2}\right)^2 + \left(\frac{2\lambda}{2\lambda+1}\right)^2\right) \leq$$
$$\leq \left(\frac{\lambda}{\lambda+1}\right)^{2x_0}\left(\frac{2^{7x_0/2}}{n} + \left(\frac{2\lambda}{2\lambda+1}\right)^2\right)$$

We can find the minimum of such function of x_0 studying the monotonicity of it.

$$\cancel{\left(\frac{\lambda}{\lambda+1}\right)^{2x_0}}\left(\frac{2^{7x_0/2}}{n}+\left(\frac{2\lambda}{2\lambda+1}\right)^2\right) \leq \cancel{\left(\frac{\lambda}{\lambda+1}\right)^{2x_0}}\left(\frac{\lambda}{\lambda+1}\right)^2\left(\frac{2^{7(x_0+1)/2}}{n}+\left(\frac{2\lambda}{2\lambda+1}\right)^2\right)$$

$$\frac{2^{7x_0/2}}{n}\left(1-2^{7/2}\left(\frac{\lambda}{\lambda+1}\right)^2\right) \leq \left(\frac{2\lambda}{2\lambda+1}\right)^2\left(\left(\frac{\lambda}{\lambda+1}\right)^2-1\right)$$

$$x_0 \leq \frac{2}{7}log_2\left(\frac{n\left(\frac{2\lambda}{2\lambda+1}\right)^2\left(\left(\frac{\lambda}{\lambda+1}\right)^2-1\right)}{\left(1-2^{7/2}\left(\frac{\lambda}{\lambda+1}\right)^2\right)}\right)$$

$$x_0 \leq \frac{2}{7}log_2\left(\frac{n(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)$$

This means that the minimum is reached for

$$x_0 = \left\lfloor \frac{2}{7}log_2\left(\frac{n(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)\right\rfloor$$

Substituting this value in the bound for the $NMSE(\hat{T}_1^L)$ we get

$$\begin{aligned}
NMSE(\hat{T}_1^L) \leq & \left(\frac{\lambda}{\lambda+1}\right)^{\frac{4}{7}log_2\left(\frac{n(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)} \\
& \left(\frac{(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}+\left(\frac{2\lambda}{2\lambda+1}\right)^2\right) \\
= & \left(\frac{n(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)^{\frac{4}{7}\frac{1}{log_{\lambda/(\lambda+1)^2}}} \\
& \left(\frac{(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}+\left(\frac{2\lambda}{2\lambda+1}\right)^2\right) = \\
= & \frac{1}{n^{4log_2(1+1/\lambda)/7}}\left(\frac{(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)^{4log_2(\lambda/(\lambda+1))/7} \\
& \left(\frac{(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}+\left(\frac{2\lambda}{2\lambda+1}\right)^2\right)
\end{aligned}$$

□

We can also discover the limit of predictability for λ.

Corollary 3.3.3. *Let $\lambda \geq 1$ and $L \sim Binom(x_0, 1/(\lambda+1))$ with*

$$x_0 = \left\lfloor \frac{2}{7}log_2\left(\frac{n(2\lambda)^2}{(2\lambda+1)((2^{7/2}-1)\lambda^2-2\lambda-1)}\right)\right\rfloor$$

then for any $\delta \in (0,1)$

$$\lim_{n\to+\infty} \frac{max\{\lambda : NMSE(\hat{T}_1^L) \leq \delta\}}{log(n)} \geq \frac{4}{7log(B/\delta)log(2)}$$

where $B := max_{\lambda\geq 1} B(\lambda)$.

Proof. Exploiting the previous theorem, we can write

$$NMSE(\hat{T}_1^L) \leq \frac{B}{n^{log_2(1+1/\lambda)4/7}}$$

and besides observe that the inequality

$$\frac{B}{n^{log_2(1+1/\lambda)4/7}} \leq \delta$$

holds if and only if

$$\lambda \leq \frac{1}{exp\{\frac{7log(B/\delta)log(2)}{4log(n)}\} - 1} := \lambda^*$$

Hence we have

$$max\{\lambda : \mathcal{E}_{n,\lambda}(\hat{T}_1^L) \leq \delta\}\frac{7log(B/\delta)log(2)}{4log(n)} \geq \frac{\frac{7log(B/\delta)log(2)}{4log(n)}}{exp\{\frac{7log(B/\delta)log(2)}{4log(n)}\} - 1}$$

The thesis follows by taking the limit of the previous inequality. □

Chapter 4

Conclusions

4.1 Estimation on real datasets

We tested the performance of our estimator $\hat{T}_1^L$ on some benchmark datasets from the 5% public use microdata sample of the U.S. 2000 census for the state of California. See Ruggles et al. (2010) [10] and references therein for details on the data. To compare it, we performed the estimation also with $\hat{U}_{ps}^B$ and $\hat{U}_{ps}^S$, with both method of moments and maximum likelihood estimation.
We treat the $N = 1150934$ individuals older than 21 years as the population, and consider a variety of contingency tables obtained from different sets of variables and independent random samples of different sizes. Specifically, we reconsider the variables (label in parentheses) specifed by Manrique-Vallier and Reiter (2012) [7]: number of children (A, 10 levels), age (B, 10 levels), sex (C, 2 levels), marital status (D, 6 levels), race (E, 5 levels), education (F, 5 levels), employment status (G, 3 levels), income (H, 10 levels), disability (I, 2 levels), and veteran status (J, 2 levels). Then, we consider three contingency tables spanned by the sets of variables (A, B, C, D, E, F, G), (A, B, C, D, E, F, G, I, J) and (A, B, C, D, E, F, G, H) respectively, and draw the random samples *California 1* ($n/N^* = 0.004$), *California 2* ($n/N^* = 0.05$), *California 3* ($n/N^* = 0.10$) from the first table, *California 4* ($n/N^* = 0.05$) from the second table and *California 5* ($n/N^* = 0.05$) from the third table.

In the table below, we compare the obtained results. For each dataset, we reported the sample size n, the number of species observed k, the number of sample uniques m_i, the true value of sample uniques that are also population uniques T_1^{true} and the six estimates. In the first three datasets, Skinner's estimates seem to be closer to the real value, in the fourth one Bethlehem's wins and in the last one $\hat{T}_1^{\text{bin}}$ performs better. In general, it seems

that Bethlehem's and Skinner's estimators vary a lot between method of moments and maximum likelihood and that the method of moments performs slightly better, whereas even with two different distributions, our estimates are more similar. Between the Poisson and the binomial distributions, the binomial is always closer to the true T_1.
The biggest lack of this table, of course, are confidence intervals. We will discuss why we couldn't provide them in the next paragraph.

Dataset	n	k	m_1	T_1^{true}	$\hat{U}_{ps}^{B\,mm}$	$\hat{U}_{ps}^{B\,mle}$	$\hat{U}_{ps}^{S\,mm}$	$\hat{U}_{ps}^{S\,mle}$	$\hat{T}_1^{\text{pois}}$	$\hat{T}_1^{\text{bin}}$
Cal 1	5000	1328	738	15	30	28	9	8	2	6
Cal 2	57547	4707	2072	211	202	429	214	129	119	181
Cal 3	115093	6406	2622	469	234	863	609	371	300	369
Cal 4	57547	7212	3575	498	272	419	312	180	202	288
Cal 5	57547	13546	7669	1169	429	358	371	245	462	707

In order to understand how these estimators behave according to a variation of the size of the population, N^*, we plotted the estimated values as functions of $\lambda = (N^* - n)/n$, letting λ vary between 1.1 and $(1150934 - n)/n$, since 1150934 is the real size of the population. To estimate the number of total species in the population, we used Bethlehem's suggestion $\hat{K} = N^*/\overline{f}$.
In figures from 4.1 to 4.5 is evident the decreasing behavior of Skinner's (blue and purple line) and our (red and yellow) estimators. The remaining two horizontal lines represent Bethlehem's estimates. We shouldn't be surprised, since in 1.3 we proved that $\hat{U}_{ps}^{B}$ is a constant function of N^*.

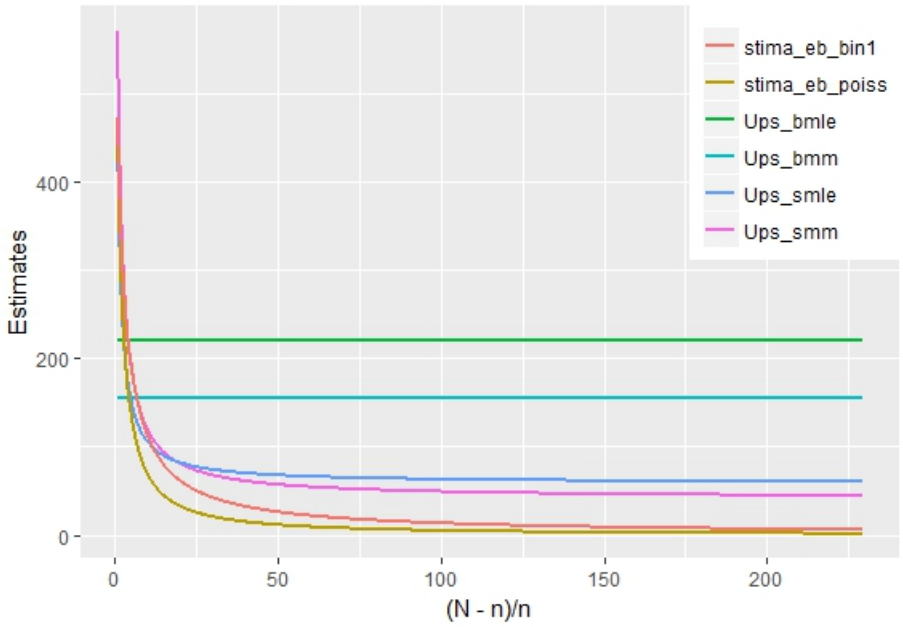


Figure 4.1: California 1

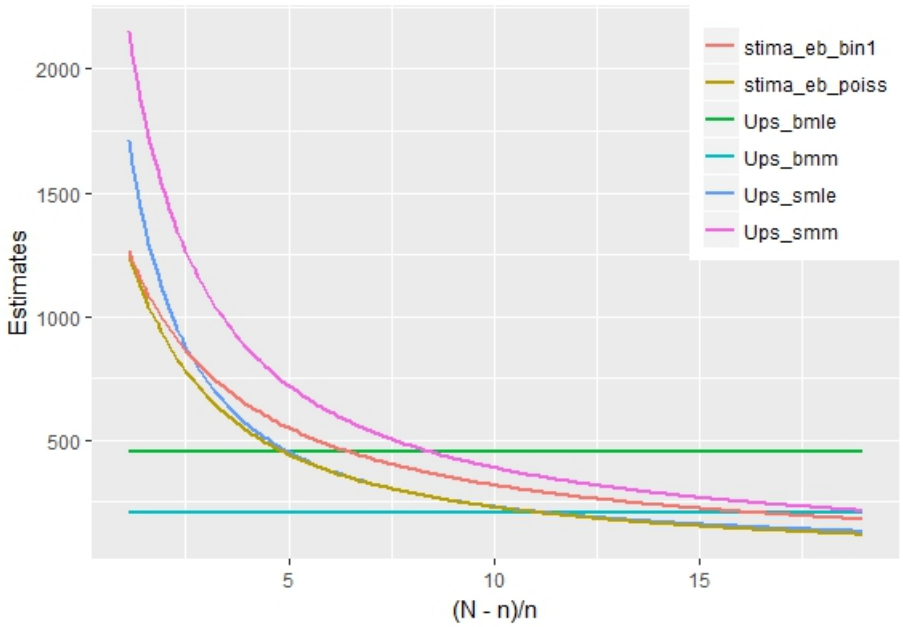


Figure 4.2: California 2

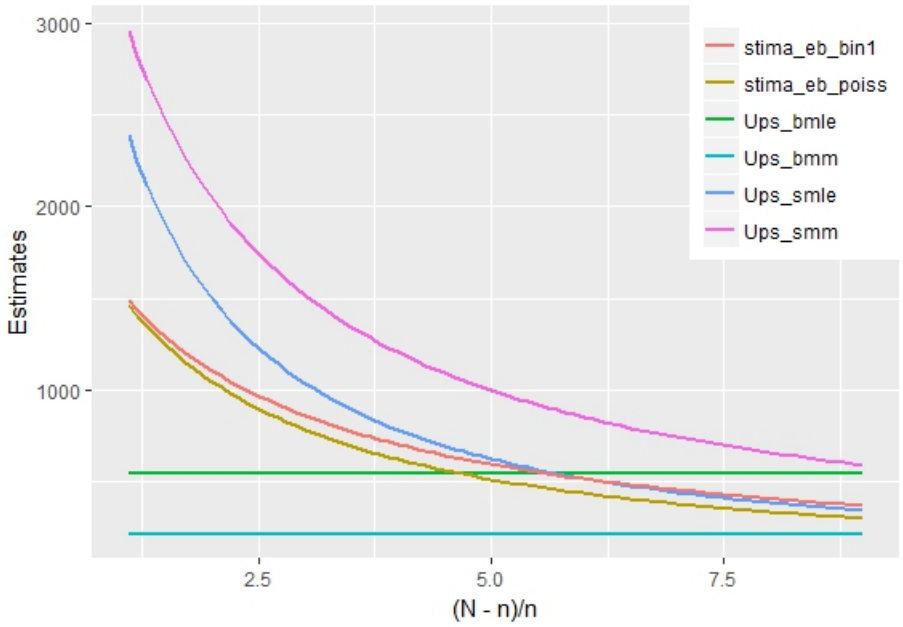


Figure 4.3: California 3

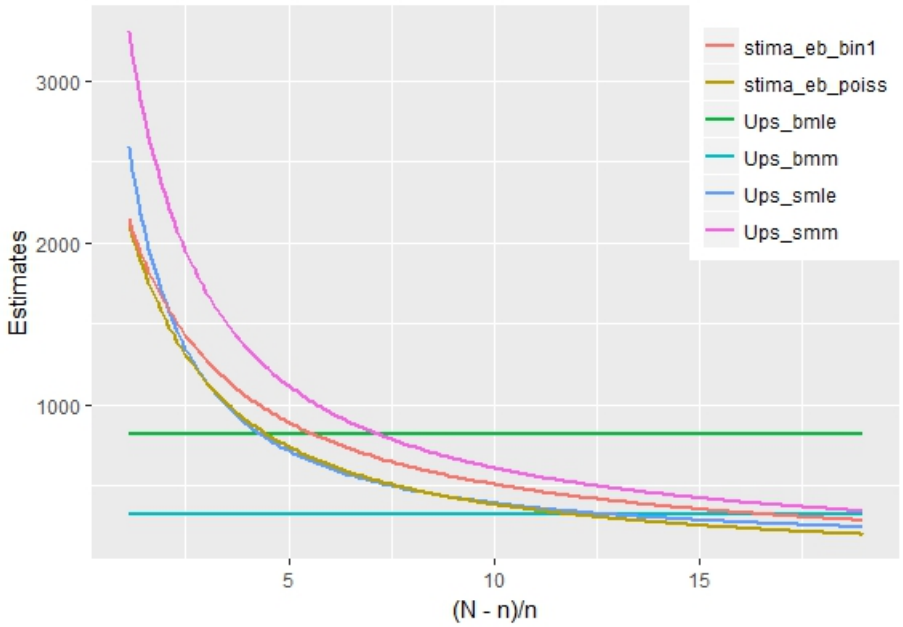


Figure 4.4: California 4

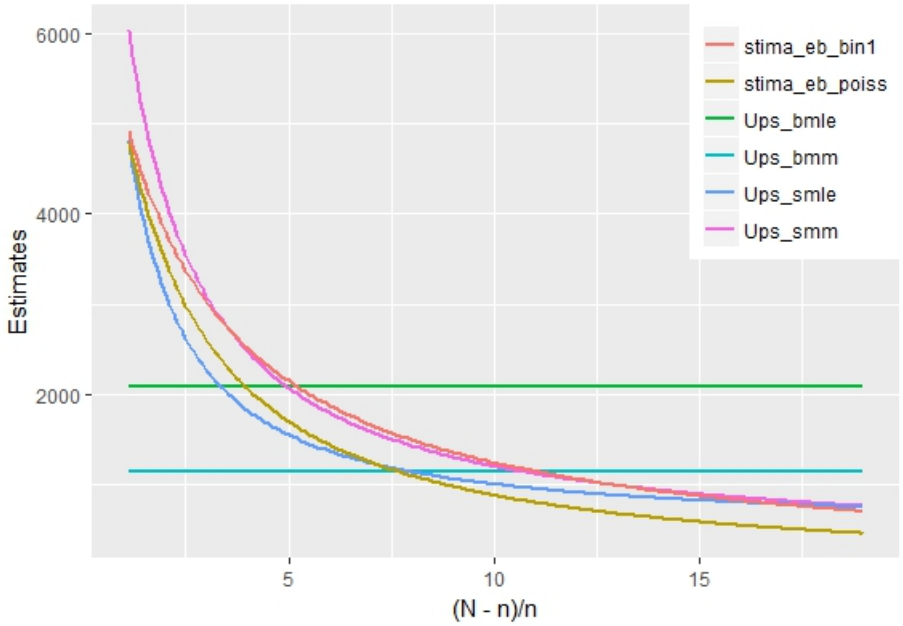
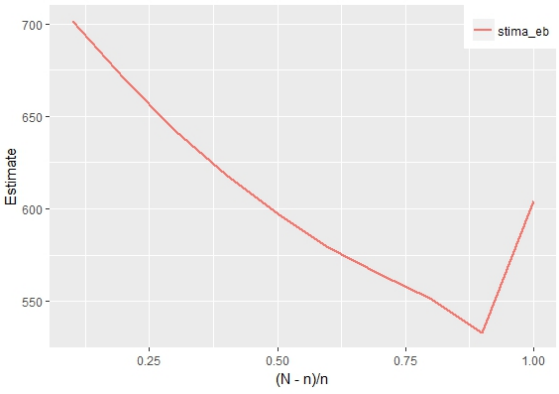
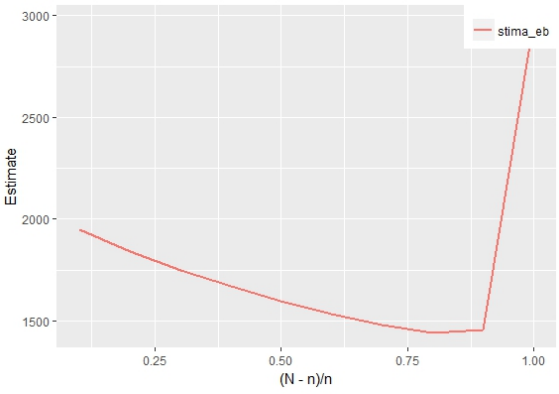


Figure 4.5: California 5

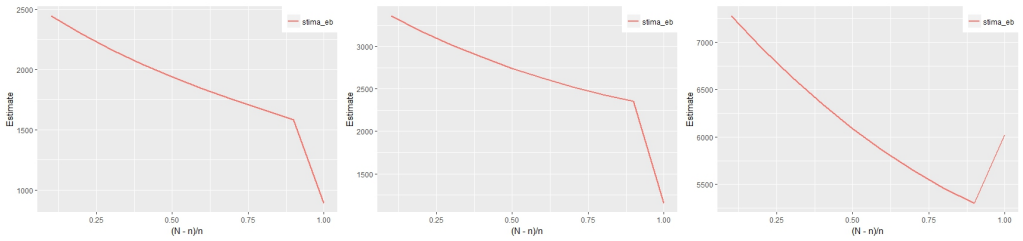
In the figures below, instead, we show the behavior of our first estimator $\hat{T}_1$ (before performing the smoothing correction) up to $\lambda = 1$, *i.e.* when the population size doubles the sample size. At $\lambda = 1$, the estimator clearly explodes.



(a) California 1



(b) California 2



(a) California 3 (b) California 4 (c) California 5

4.2 Concluding remarks and open problems

Our contribution with this work is the introduction of a new type of estimator in the setting of disclosure risk assessment, born from the nonparametric empirical Bayes statistical theory in the sense of Robbins. The importance of this estimator lies in its robust nature, derived by the nonparametrical assumptions, that makes it possible to overcome the problem of strong distributional assumptions and of estimation of parameters, necessarily required by the most common parametrical empirical Bayes approaches. This elegant approach is not sufficient alone to assure the convergence of the estimator, but exploiting Orlitsky et al. (2016) [8], we further smoothed it in order to obtain convergence and provided a limit of predictability for the estimator in relation to the sample and population sizes.

Nevertheless, we leave two open problems to which we will work in the future.
The first one is related to the big absent in this work: confidence intervals for our estimator when $\lambda \geq 1$. As we highlighted, for $\lambda \geq 1$ the estimator $\hat{T}_1^L$ is no longer unbiased, and not even asymptotically unbiased. In addition, the bound on the variance is not sharp enough to derive confidence intervals.
The second problem lies in the optimality of the estimator. We provided an upper bound for NMSE$(\hat{T}_1^L)$, but to obtain optimality we should propose also a result for the lower bound.

Bibliography

[1] Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (March 1990) *Disclosure Control of Microdata*, Journal of American Statistical Association.

[2] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration inequalities*. Oxford University Press.

[3] Efron, B. and Morris, C. (1973) *Stein's estimation rule and its competitors - an empirical Bayes approach*. Journal of American Statistical Association.

[4] Efron, B. and Thisted, R. (1976) *Estimating the number of unseen species: How many words did Shakespeare know?*. Biometrika.

[5] Good, I. (1953) *The population frequencies of species and the estimation of population parameters*. Biometrika.

[6] Good, I., Toulmin, G. (1956) *The number of new species, and the increase in population coverage, when a sample is increased.* Biometrika.

[7] Manrique-Vallier, D. and Reiter, J.P. (2012) *Estimating identification disclosure risk using mixed membership models.* J. Amer. Stat. Ass..

[8] Orlitsky, A., Suresh, A. T. and Wu, Y. (2016) *Optimal prediction of the number of unseen species* PNAS.

[9] Robbins, H. (1955) *An empirical Bayes approach to statistics* Third Berkeley Symposium on Mathematical Statistics and Probability.

[10] Ruggles, S., Alexander, J. T., Genadek, K., Goeken, R., Schroeder, M. B. and Sobek, M. (2010) *Integrated public use microdata series: Version 5.0* [Machine-readable database]. University of Minnesota, Minneapolis. Available at https://usa.ipums.org/usa/.

[11] Samuels, S. M. (1998) *A Bayesian, Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment* Journal of Official Statistics.

[12] Skinner, C. J. , Marsh, C., Openshaw, S. and Wymer, C. (1994) *Disclosure Control for Census Microdata* Journal of Official Statistics.

[13] Skinner, C. J. and Elliot, M. J. (2002) *A measure of disclosure risk for microdata* Journal of the Royal Statistical Society