to Jesus

Contents

Chapter 1

Introduction

Statistical institutions are continually facing an increasing demand for releasing census data, also called micro-data files. A challenging aspect to consider for these institutions is to decide whether to release these files to protect individuals against potential information disclosure. Census micro-data are composed of individual records containing information collected on persons and households that can be private and sensitive.

In the last ten years, the technological revolution and the impact of big data technologies have challenged the concept of privacy. Businesses understood the importance of personal data, and by collecting a massive amount of information from individuals, they can personalize their offerings, profiling their consumers and increase their revenues. Because of this trend and the prompt reaction of the public opinion regarding the priceless value of privacy, releasing census micro-data started to be seen as an opaque practice, and census participants lost their trust in data privacy, due also to public scandals such as Cambridge Analytica in 2016.

According to a report conducted by the Pew research center (February 2020) in the United States, roughly 34% of the people that participated in this study believed that census asks for too much personal information. Moreover, respondents do not trust governments and how public entities can use their valuable, sensitive data.

This thesis aims to show why disclosure is a problematic topic, and to propose a predictive approach for the estimation of the disclosure risk, based on a Bayesian parametric model introduced by Bethlehem *et al.* (1990) and Skinner *et al.* (2002) using

the Dirichlet-Multinomial model. Before entering the details and the derivation of the statistical model, the simulations, and the results obtained on real data, an introduction to the basic concept of statistical disclosure control is provided to the reader.

The fundamental concept of disclosure risk is described in Section 1.1. Closely related to it, the concept of identification is introduced in Section 1.2. In Section 1.3, the concept of uniqueness is defined. The role of uniqueness is vital as it is the central point of the computation of the disclosure risk. Then, different types of disclosure risk measures available in the literature are described in Section 1.4. Finally, Section 1.5 discusses various standard practices for statistical disclosure control.

1.1 The Disclosure Problem

Before analyzing the disclosure problem, it is essential to define the difference between personal data collected for administrative purposes and those for statistical ones.

On the one hand, administrative data are collected for non-statistical purposes by governments and agencies. These datasets are created for record-keeping and administrative purposes, such as birth and death records, pensions, and taxations. Governments produce this type of data to provide a historical memorandum that is not invasive to the population. On the other hand, statistical data are confidential data differentiated by sensitive and non-sensitive information such as income, social class, age, race, etc. The disclosure of this private information can put at risk the security and the privacy of those individuals who participated in a survey. For this reason, statistical confidentiality, which will be further discussed in the next section, plays a fundamental role in keeping these data secure while still permitting statistical studies.

According to the recitals (51) to (56) of the GDPR (General Data Protection Regulation), personal data are classified as sensitive when their processing could create significant risks to the fundamental rights and freedoms of an individual. For example, sensitive data are personal data revealing racial or ethnic origin, political opinions, health-related data, or data concerning an individuals' sex life.

In the context of statistical disclosure control, the term disclosure risk is defined

as the risk deriving from a third actor that can infer or directly identify confidential information about a specific individual from a released statistical dataset, and reveal those to third parties.

Information disclosure (and disclosure risk) can be differentiated into two types:

1. Attribute Disclosure
2. Identity Disclosure

Attribute Disclosure is the practice of inferring the value of an attribute (such as the ethnicity) that was not directly present in the fields of the dataset. An example of attribute disclosure could be related to determining voting behavior. By sharing a large amount of personal information on social media such as Facebook, Twitter, LinkedIn, an agency could use algorithms of sentiment analysis that can match and predict the political preference of individuals.

However, the estimation of the disclosure risk in this paper is related to the computation of the identity disclosure risk. Identity disclosure instead is the practice of inferring the identity of an individual given a set of specific attributes. That is the identification of an individual by looking at some personal dimensions or fields (such as income, job, race, and sex). It is clear to the reader that identification is the first step leading to the subsequent concept of information disclosure (both attribute and identity).

Identification means finding a one-to-one correspondence between a set of combined attributes from a released dataset and an individual. An intruder who knows the identifiers of an individual can pinpoint its record in the published dataset.

Nevertheless, why is information disclosure considered a violation of personal rights and freedoms? Disclosure is a problem from a legal perspective. At the EU level, according to the Regulation 223/2009, individual data collected by statistical offices for statistical compilation, whether they refer to legal or natural persons, can be processed exclusively for statistical purposes. Moreover, the researchers have to prevent the disclosure of information concerning an individual by applying protective measures on micro-data.

In the United States, the disclosure problem is faced similarly. In fact, before publishing any statistic, agencies apply safeguards, such as disclosure avoidance, to prevent an intruder to trace that statistic back to a specific respondent. Section 1.5 will describe these statistical disclosure control practices in a more detailed way.

In addition to the legal perspective, there is a very practical reason why information disclosure is a problem. If the respondents do not trust statistical institutions and governments, they will probably avoid participating in surveys.

This paper aims to estimate the identity disclosure risk of a released public dataset and propose a statistical model to quantify the disclosure problem.

1.2 The concept of Identification

As previously said, the disclosure problem is closely related to the concept of identification. According to Eurostat, the European Statistical System, identification can be both direct and indirect. Direct identification means identifying the respondent, a statistical unit, from its formal identifiers, or identifying information.
Indirect identification instead refers to the inference of the identity from the combination of dimensions or characteristics (such as age, income, sex, etc.) that can also be sensitive information.

It is essential to define also the difference between identifying and sensitive information. Identifying information refers to those variables in the dataset (also called key variables) that let a user identify a respondent. A basic example of key variables are names or addresses, but also age, race, and sex can help identify an individual.
Sensitive information instead refers to those variables that describe the private sphere of an individual and must be protected by law. In the EU, the GDPR explicitly describes various example of variables that constitutes sensitive information and are subject to specific processing conditions. Examples are race, religious beliefs, trade-union memberships, genetic data, health-related data, and sexual orientation.

Having defined the basic concept of information disclosure and identification, the fundamental rule for tabular disclosure control is proposed to the reader.

A micro-data set should be released in a way that it is impossible for a reader to identify a respondent by using the key variables present in the dataset and his preliminary knowledge about the individual. Prior knowledge is an important prerequisite in identification, and therefore in information disclosure. In fact, if a potential intruder does not have any prior knowledge about a specific individual, identification is impossible and therefore the risk of information disclosure becomes null. In our study, we assume that every possible record in the dataset is at risk and the intruders have prior knowledge of whom they are looking for. Therefore, the risk of disclosure does not depend on prior knowledge of the intruders, but on the composition of the dataset.

1.3 The concept of Uniqueness

A fundamental concept for the disclosure risk computation is the notion of uniqueness. To give a proper definition of uniqueness in the statistical disclosure control area, we need to define a cell in a microdata file. Suppose that a dataset is composed of M categorical variables (that can be both key or sensitive) such as age, sex, income, and education. Each of these variables is divided internally into several subsets (or categories) by the provider of the released dataset (also to protect the anonymity of the respondents). For example, sex has two categories: male and female or education may have four different levels, according to the degree of specificity chosen.
The cross-classification of all the categories of each variable represents the number of possible cells, called K. For example, if sex has two categories, age has nine, and education four, the total possible number of cells is $2 * 9 * 4 = 72$.

It is easy to notice that K depends exclusively on the nature of the dataset and the chosen categories for each variable. Every respondent has some combination of characteristics; thus, it will belong to one and only one cell. Besides, the number of cells whose participation is different from zero is always smaller or equal than K, as some combinations of categories may not be even feasible or real (e.g., imagine an individual within the 0-2 age category and with a high school diploma).

An individual is said to be unique in the population when it is the only one to

belong to a specific cell. Therefore, if the frequency of a cell is equal to 1, then that combination of categories represents a dangerous condition for the dataset. It would mean that there is only a person with those unique sets of values. For example, in certain regions, some professions are unique. Imagine the doctor or the dentist in a small city. They are unique with just two-dimension (profession and region). This could facilitate the identification process for an intruder looking for them.

Statistical institutions do not release the whole population dataset, but they release a sample, appropriately modified to avoid information disclosure. It is crucial to define a principle that will walk the reader along with this paper. Uniqueness in the population implies uniqueness in the sample, but the reverse is false. This means that if an individual is unique in the population and he has been selected in the sample (in which he is still unique), then this individual is at serious risk of information disclosure. However, if an individual is unique in the sample, then it is not always true that it will be unique in the population.

Different measures of disclosure risk available in the literature are defined and described in the next section, together with the one adopted in this paper.

1.4 Measures of Disclosure Risk

As proposed by Skinner *et al.* (2002) and Bethlehem *et al.* (1990), there are different measures for the estimation of the disclosure risk.
The first measure reported is the proportion of unique records in the population. This measure is denoted as PR(PU), that is the probability of population uniqueness from unit randomly drawn from the population, defined as:

$$\Pr(PU) = \frac{N_1}{N} = \frac{\sum_{j=1}^{N} \mathbb{1}_{F_j=1}}{N}.$$

The second measure of disclosure risk relies on the consideration that only sample unique can be population unique. It represents the proportion of sample unique that

are also population unique, which is indicated as:

$$\Pr(PU \mid SU) = \frac{\sum_{j=1}^{N} \mathbb{1}_{f_j=1,F_j=1}}{\sum_{j=1}^{N} \mathbb{1}_{f_j=1}}.$$

The third measure of disclosure risk instead extends this concept, by taking into consideration not only the sample unique but also the risk deriving from those records that are not population unique, denoted by:

$$\theta = \frac{\sum_{j=1}^{N} \mathbb{1}_{f_j=1}}{\sum_{j=1}^{N} F_j \mathbb{1}_{f_j=1}}.$$

The measure of disclosure risk that is adopted in this paper represents the sum of cells that are unique in the sample and are also unique in the population, and it is defined as:

$$\boldsymbol{\tau} = \sum_{i=1}^{K} \mathbb{1}_{f_i=1,F_i=1}.$$

1.5 Statistical Disclosure Control

Statistical disclosure control can be achieved in different ways. By law, it is compulsory to make the confidential dataset anonymous before releasing it. Both in the US and EU, statistical confidentiality (or disclosure avoidance) is adopted through different methods such as:

- Physical Protection
- Statistical Disclosure Control

Physical protection refers to physically securing confidential data and giving access only to specific individuals or entities, with explicit authorization. In this paper, three different types of Statistical Disclosure Control (SDC) methods are described, differentiated by the data format to which they are applied:

1. Tabular SDC methods
2. Queryable Database Protection

3. SDC methods for micro-data

Firstly, tabular methods for disclosure control can be classified as perturbative and non-perturbative. Perturbative methods aim to identify sensitive cells and modify their values, making the released information different than the true one. Non-perturbative methods, instead, do not modify the values inside sensitive cells, but they entirely suppress them. A well-known non-perturbative method is called cell suppression (CS) that consists of suppressing sensitive cells from the database and those secondary cells that could help infer the primary ones.

Secondly, SDC for queryable database protection can be of two different types: query perturbation and query restriction. The practice of adding noise to micro-data records or the result of input queries is called query perturbation. Query restriction instead refers to limiting the possibility of asking a precise question (queries) to the database. Finally, there are two different methods of SDC for micro-data:

- Data Masking
- Data Synthesis

Data masking refers to the practice of generating a mask dataset D_1 from the original micro-data set D, on which tabular SDC methods (both perturbative and non-perturbative) are applied. Data Synthesis instead refers to the practice of generating an artificial dataset with similar characteristics of the original one.

Chapter 2

Theoretical Background

In this chapter, the mathematical background and the derivation of the estimator are provided to the reader. Section 2.1 introduces the disclosure problem through mathematical formalism and notation. In Section 2.2, the Dirichlet and Multinomial distributions are analyzed. Section 2.3 describes the primary derivation of this paper, the estimator of the disclosure risk. Section 2.4 discusses the maximum likelihood estimation procedure to choose the parameters of the Dirichlet-Multinomial model.

2.1 Framework and Notation

Let the tabular micro-data file consist of K possible cells that correspond to the cross-classification of all the categorical variables: K is the number of possible cells. Let $X_1, \ldots, X_n$ be a collection of independent variables with $X_i \in \{1, \ldots, K\}$ where each X_i corresponds to an individual.

In particular, if $X_i = k$, then it means that the i-th individual belongs to the k-th cell. That is, it has the k-th configuration of categorical variables.
Let f_i be the frequency of each cell i in the sample. Then

$$f_1 = \sum_{i=1}^{n} \mathbb{1}_{x_i=1}, \ldots, f_K = \sum_{i=1}^{n} \mathbb{1}_{x_i=K},$$

that is f_k measures the number of individuals from the sample that belong to the k-th

cell. Let N be the population size, that is, N is the number of individuals present in the dataset. Let $X_1, \ldots, X_N$ be the entire population, then a superpopulation model is proposed. Let $X_1, \ldots, X_n$ be the sample and $X_{n+1}, \ldots, X_N$ be the unobserved part of the population. Then, let $F_1, \ldots, F_K$ be the frequencies on the entire population.

Where

$$F_1 = \sum_{i=1}^{N} \mathbb{1}_{x_i=1},$$

$$\vdots$$

$$F_K = \sum_{i=1}^{N} \mathbb{1}_{x_i=K}.$$

That is, F_k is the number of individuals from the entire population that belong to the k-th cell.

Then the disclosure risk estimation problem is defined as follows: by observing the variables $X_1, \ldots, X_n$, the frequencies $(f_1, \ldots, f_K)$ are computed, and the following disclosure risk index is estimated.

Let $\boldsymbol{\tau}$ be the index of disclosure risk defined as

$$\boldsymbol{\tau} = \sum_{i=1}^{K} \mathbb{1}_{f_i=1, F_i=1},$$

which represents the sum of all the cells that are unique in the sample and the population. This statistic was first introduced by Bethlehem *et al.*(1990) and further commented by Takemura (1997) and Skinner *et al.*(2002). This measure of disclosure gives a better performance than PR(PU), described in the previous chapter, as it considers the risk deriving from those records that are population unique and sample unique. However, it treats all the sample unique records in the same way, ignoring the fact that some records that are sample unique are more likely to be population unique.

2.2 Dirichlet-Multinomial Model

This section introduces the Dirichlet-Multinomial model. First, the Dirichlet and the Multinomial distribution are proposed separately to the reader.

In probability theory and statistics, the Dirichlet distribution is a well-known family of continuous multivariate probability distributions characterized by a vector parameter $(\alpha_1, \ldots, \alpha_K)$, where $\alpha_i > 0$ for all $i \in \{1, \ldots, K\}$. It is a multivariate generalization of the Beta distribution.

The density function takes the form:

$$f(p_1, p_2, \ldots, p_K \mid \alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{\Gamma(\boldsymbol{\alpha})}{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \ldots p_K^{\alpha_K - 1},$$

where $\{p_K\}_{i=1}^{i=K}$ belongs to the $K-1$ simplex or

$$\sum_{i=1}^{K} p_i = 1 \text{ and } p_i \geq 0 \text{ for all } i \in \{1, \ldots, K\}$$

and

$$\sum_{i=1}^{K} \alpha_i = \boldsymbol{\alpha};$$

$$\boldsymbol{\alpha} \geq 0;$$

The parameter vector $(\alpha_1, \ldots, \alpha_K)$ governs the shapes of the distribution. The Dirichlet distribution is sometimes called a distribution over distribution, and it can be thought of as a distribution of probabilities themselves. Dirichlet distributions are used in Bayesian statistics as a prior distribution. It has been proved that it is a conjugate prior to the multinomial distribution and the categorical distribution.

The Multinomial distribution is a multivariate generalization of the binomial distribution. The binomial distribution models the number of occurrences X of success in n independent repetitions of a basic experiment, with two possible outcomes: success and failure. Therefore, during a binomial experiment, it is necessary to keep track of the number of successes alone, which is X, as the number of failures will be $n - X$. We can imagine it as drawing with replacement red and blue balls from an urn, with probability p and $1-p$. Instead, in the case of a multinomial distribution, we are drawing with replacement n balls appearing in K colors where each color can be drawn with probabilities $p_1, \ldots, p_K$. The multinomial distribution extends the binomial concept

to K categories instead of 2. That is, the basic experiment is now with K possible outcomes $O_1, \ldots, O_K$. Each experiment is repeated n times independently, and each category has the same probability of $P(O_i)$ in each trial.

Such that:

$$\sum_{i=1}^{K} P(O_i) = 1.$$

Then each X_i measures the number of occurrences of the i-th outcome in the n trials of the basic experiment. If there are n experiments, or n units to allocate into K categories then

$$\sum_{i=1}^{K} X_i = n.$$

Then, let $X_1, \ldots, X_K$ be distributed as

$$(X_1, ..., X_K) \sim \text{Multinomial}\,(n, p_1, ..., p_K)\,,$$

where the probability mass function (pmf) is

$$\begin{aligned} f(x_1, \ldots, x_K; n, p_1, \ldots, p_K) &= \Pr(X_1 = x_1 \text{ and } \ldots \text{ and } X_K = x_K) \\ &= \begin{cases} \dfrac{n!}{x_1! \cdots x_K!} p_1^{x_1} \times \cdots \times p_K^{x_K}, & \text{when } \sum_{i=1}^{K} x_i = n \\ \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

and can be rewritten using Gamma functions as:

$$f(x_1, \ldots, x_K; p_1, \ldots, p_K) = \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{K} p_i^{x_i},$$

whose support is:

$$x_i \in \{0, \ldots, n\}, \quad i \in \{1, \ldots, K\};$$

$$\sum_{i=1}^{K} x_i = n.$$

Given the theoretical explanation and formalisation of the model selected, consider a Multinomial model where $(f_1, ..., f_K) \sim \text{Multinomial}\,(n, \pi_1, ..., \pi_K)$ with a Dirichlet prior on the probabilities, that is $(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$.
In Bayesian statistics evaluating a posterior distribution could be difficult, but if a prior is a conjugate, then computing the posterior is immediate, avoiding the need for numerical integration.

Definition 1. *Let $X_1, ..., X_n \mid \theta \sim f(x \mid \theta)$, where θ lies in Θ. A family F of distributions over Θ is said to be a class of conjugate priors w.r.t. the likelihood at the hand if: $p(\theta) \in F$ implies that the posterior $p(\theta \mid x_1, ..., x_n) \in F$, for every $n \geq 1$ and for every $x_1, \ldots, x_n$.*

This definition says that if the prior is in F, then the posterior belongs to F as well, and the only thing required is to update the parameters.

As previously introduced, the Dirichlet distribution is a conjugate prior for the multinomial distribution, that means it is a conjugate prior for the probabilities $\pi_1, \ldots, \pi_K$. Then if

$$(\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K),$$

prior to observing the data, then given the frequencies $(f_1, \ldots, f_K)$ for the different cells,

$$(\pi_1, \ldots, \pi_K) \mid (f_1, \ldots, f_K) \sim \text{Dirichlet}(\alpha_1 + f_1, \ldots, \alpha_K + f_K).$$

Moreover, the higher the value of α_i, the greater mass assigned to p_i (recall that p_i must sum up to 1). If the $\alpha_i < 1$, it can be thought as the distribution is pushing the mass toward the extremes, whereas when $\alpha_i > 1$ it attracts p_i toward some central value. When $\alpha_1 = \cdots = \alpha_K = \alpha$ then the Dirichlet distribution is symmetric and when $\alpha_1 = \cdots = \alpha_K = 1$ then the points are uniformly distributed.

2.3 Derivation of the Estimator

The derivation of the estimator for the disclosure risk, indicated with $\boldsymbol{\tau}_1$, finds its root in the Dirichlet-Multinomial model. Assume the following Bayesian model where

$$X_j | \, \widetilde{\pi}_1 \cdots \widetilde{\pi}_K \overset{iid}{\sim} \sum_{i=1}^{K} \widetilde{\pi}_i \cdot \delta_{\{i\}}(\,\cdot\,) = \widetilde{p}\,(\,\cdot\,),$$

and

$$(\widetilde{\pi}_i \cdots \widetilde{\pi}_K) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K)$$

are drawn from a Symmetric Dirichlet distribution with parameter vector α. Then

$$\widetilde{p}\,(\,\cdot\,) = \sum_{i=1}^{K} \widetilde{\pi}_i \cdot \delta_{\{i\}}(\,\cdot\,)$$

is a discrete random probability measure with masses $(\widetilde{\pi}_i \cdots \widetilde{\pi}_K)$ where

$$\Pr(X_i = k \,|\widetilde{p}) = \widetilde{\pi}_k = \Pr(X_i = k \mid \pi_1, \ldots, \pi_K).$$

In this case, the random variables $X_i \geq 1$ are exchangeable, that means the order is not relevant but only the way they are collected. For this reason, the joint distribution of $(X_1, \ldots, X_n)$ depends only on the number of observations that fall in the $1, \ldots, K$ cells but not on any particular value of X_i.

The law governing $X_1, \ldots, X_n$ is a function only of $f_1, \ldots, f_K$ because it does not depend on the order through which these observations are collected but only on how many fall in every cell. Therefore, the multinomial distribution conditioned on the Dirichlet distribution is computed as follows.

$$\Pr(f_1, \ldots, f_K) = \Pr(f_1 \text{ observations in cell } 1, \ldots, f_K \text{ observations in cell K}),$$

$$= \binom{n}{f_1 \cdots f_K} * \pi_1^{f_1} \cdots \pi_K^{f_K}$$

where

$$\pi_K = (1 - \pi_1 \cdots - \pi_{K-1}),$$

moreover

$$f(\pi_1, \pi_2, \ldots, \pi_K) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} * \pi_1^{\alpha-1}\pi_2^{\alpha-1} \ldots \pi_K^{\alpha-1}.$$

Therefore we get

$$f(\pi_1, \pi_2, \ldots, \pi_K \mid f_1, \ldots, f_K) \sim \text{Dirichlet}(\alpha + f_1, \ldots, \alpha + f_K).$$

We want to find the expected value of $\boldsymbol{\tau}$ given $X_1 \ldots X_n$, that is its estimator called $\boldsymbol{\tau}_1$. In formulae:

$$\mathrm{E}[\boldsymbol{\tau} \mid X_1, \ldots, X_n]$$

$\boldsymbol{\tau}$ is not a simple statistic to study because it also depends on the observations.

Theorem 1. *By observing each realizations of the X_i and given a prior Dirichlet distribution over the cells, the estimator of $\boldsymbol{\tau}$ under a squared loss function is:*

$$\mathrm{E}[\boldsymbol{\tau} \mid X_1, \ldots, X_n] = m_1 * \frac{\Gamma(N + K\alpha - \alpha - 1) * \Gamma(n + K\alpha)}{\Gamma(n + K\alpha - \alpha - 1) * \Gamma(K\alpha + N)}. \tag{2.1}$$

Proof.

$$\mathrm{E}[\boldsymbol{\tau} \mid X_1, \ldots, X_n] = \mathrm{E}\left[\sum_{i=1}^{K} \mathbb{1}_{(f_i=1, F_i=1)} \mid X_1, \ldots, X_n\right],$$

$$= \sum_{i=1}^{K} \mathrm{E}[\mathbb{1}_{(f_i=1, F_i - f_i = 0)} \mid X_1, \ldots, X_n]$$

$$= \sum_{i=1}^{K} \mathrm{E}[\mathbb{1}_{(f_i=1)}, \mathbb{1}_{(F_i - f_i = 0)} \mid f_1, \ldots, f_K]$$

$$= \sum_{i=1}^{K} \mathbb{1}_{(f_i=1)} * \Pr(F_i - f_i = 0 \mid f_1, \ldots, f_K)$$

$$= \sum_{\{i:f_i=1\}} \Pr(F_i - f_i = 0 \mid f_1, \ldots, f_K)$$

$$= \sum_{\{i:f_i=1\}} \mathrm{E}[\Pr(F_i - f_i = 0 \mid f_1, \ldots, f_K, \pi_1, \ldots, \pi_k) \mid f_1, \ldots, f_K]$$

$$= \sum_{\{i:f_i=1\}} \mathrm{E}[\Pr(F_i - f_i = 0 \mid \pi_1, \ldots, \pi_k) \mid f_1, \ldots, f_K]$$

$$= \sum_{\{i:f_i=1\}} \mathrm{E}[\binom{N-n}{0} * \pi_i^0 * (1-\pi_i)^{N-n} \mid f_1, \ldots, f_K]$$

$$= \sum_{\{i:f_i=1\}} \mathrm{E}[(1-\pi_i)^{N-n} \mid f_1, \ldots, f_K]$$

$$= \sum_{\{i:f_i=1\}} \mathrm{E}[(1-\pi_i)^{N-n} \mid f_1, \ldots, f_K],$$

notice that $\pi_i \mid f_1 \ldots f_K \sim \mathrm{Beta}\,(\alpha + f_i, n + K\alpha - \alpha - f_i)$, and as we are summing over the indexes i with $f_i = 1$ then $\pi_i \mid f_1 \ldots f_K \sim \mathrm{Beta}\,(\alpha + 1, n + K\alpha - \alpha - 1)$

$$= \sum_{\{i:f_i=1\}} \int_0^1 \frac{\pi^{(\alpha+f_i-1)} * (1-\pi)^{(n+K\alpha-\alpha-f_i+N-n-1)}}{\mathrm{B}(\alpha + f_i, n + K\alpha - \alpha - f_i)}\, d\pi$$

$$= \sum_{\{i:f_i=1\}} \int_0^1 \frac{\pi^{(\alpha+1-1)} * (1-\pi)^{(n+K\alpha-\alpha-1+N-n-1)}}{\mathrm{B}(\alpha + 1, n + K\alpha - \alpha - 1)}\, d\pi.$$

As the integral does not depend on i, and calling with m_1 the number of those cells that have a frequency equal to 1 in the sample, we have:

$$m_1 = \sum_{i=1}^{K} \mathbf{1}_{f_i=1}.$$

We then obtain:

$$= m_1 * \int_0^1 \frac{\pi^{(\alpha+1-1)} * (1-\pi)^{(K\alpha-\alpha+N-1-1)}}{\mathrm{B}(\alpha+1, n+K\alpha-\alpha-1)}\, d\pi$$

$$= m_1 * \frac{\Gamma(n+K\alpha) * \Gamma(\alpha+1) * \Gamma(K\alpha-\alpha+N-1)}{\Gamma(\alpha+1) * \Gamma(n+K\alpha-\alpha-1) * \Gamma(K\alpha+N)}$$

$$= m_1 * \frac{\Gamma(N+K\alpha-\alpha-1) * \Gamma(n+K\alpha)}{\Gamma(n+K\alpha-\alpha-1) * \Gamma(K\alpha+N)}$$

by rotation of gamma functions and as for $\lim_{n\to\infty} \frac{\Gamma(n+\alpha)}{\Gamma(n+\beta)} \sim n^{\alpha-\beta}$

$$\approx m_1 * N^{K\alpha-\alpha-1-K\alpha} * n^{K\alpha-K\alpha+\alpha+1}.$$

Therefore:

$$\boldsymbol{\tau}_1 = \mathrm{E}[\,\boldsymbol{\tau} \mid X_1, \ldots, X_n\,] \approx m_1 * N^{-\alpha-1} * n^{\alpha+1} \approx m_1 * (\frac{n}{N})^{\alpha+1}.$$

□

We observe that the parameter α can be fixed or estimated via Maximum Likelihood Estimation (MLE), as proved in the next section, we obtain:

$$\begin{cases} \text{for } \alpha \to 0 & m_1 * \frac{n}{N} \\ \\ \text{for } \alpha \to \infty & 0 \end{cases}$$

2.4 Estimation of α trough MLE

The vector parameter $(\alpha_1, \ldots, \alpha_K)$ that governs the Dirichlet distribution can be estimated from the data, as shown in this section.

Observation 1. *By integrating out the Dirichlet distribution from the multinomial*

distribution, we can compute the likelihood of our model for a set of individual outcomes:

$$\Pr(\mathbf{f} \mid \boldsymbol{\alpha}) = \int_{\mathbf{p}} \Pr(\mathbf{f} \mid \mathbf{p}) \Pr(\mathbf{p} \mid \boldsymbol{\alpha}) d\mathbf{p}$$

results as

$$\Pr(\mathbf{f} \mid \boldsymbol{\alpha}) = \frac{\Gamma\left(\sum \alpha_k\right)}{\Gamma\left(\sum f_k + \sum \alpha_k\right)} \prod_{k=1}^{K} \frac{\Gamma(f_k + \alpha_k)}{\Gamma(\alpha_k)}.$$

We know that maximizing the likelihood is the same as maximizing the log-likelihood. For this reason, we rewrite the problem as:

$$\log \Pr(\mathbf{f} \mid \boldsymbol{\alpha}) = \log \Gamma\left(\sum \alpha_k\right) - \log \Gamma\left(\sum f_k + \sum \alpha_k\right) + \sum_{k=1}^{K} (\log \Gamma(f_k + \alpha_k) - \log \Gamma(\alpha_k)).$$

We assume that $\alpha_1 = \ldots = \alpha_k = \alpha$, then we want to find the parameter α that maximize the log-likelihood of the data:

$$\arg\max_{\alpha} \log \Gamma(K * \alpha) - \log \Gamma(\sum f_K + K * \alpha) + \sum_{k=1}^{K} (\log \Gamma(f_k + \alpha) - \log \Gamma(\alpha)),$$

that is the same as:

$$\arg\min_{\alpha} - \log \Pr(\mathbf{f} \mid \boldsymbol{\alpha})$$

For the estimation of α, the L-BFGS-B method provided by the SciPy.optimize library was used, and an example of maximization is given in the next figure. In this case, the frequencies are created from a multinomial distribution with $K = 20.000$ cells and $N = 1.000.000$ individuals, with a prior Dirichlet distribution over the probabilities drawn with fixed parameter $\alpha = 1$.

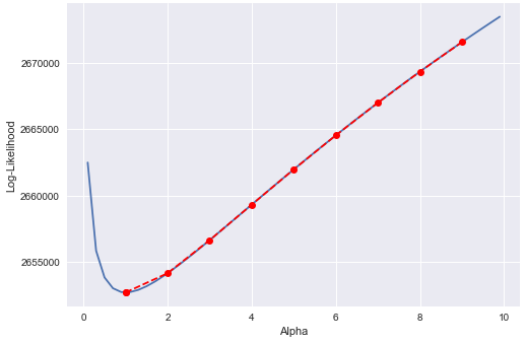

Figure 2.1: Maximization of Log-Likelihood of Dirichlet-Multinomial Model

Chapter 3

Methods

This section introduces the reader to the dataset created for the analysis of the information disclosure risk problem, what data have been collected and which operations had been performed on them. Section 3.1 deals with the creation of the micro-data set from IPSUM USA, a repository of public use surveys. In the last two sections, the datasets are created through statistical simulations. Section 3.2 illustrates the simulations from a Zipfian distribution through the VGAM R statistical package. Finally, Section 3.3 describes the creation of a dataset through the simulation directly from the Dirichlet-Multinomial, the model used to derive the estimator.

3.1 Creation of the micro-data set

We created a micro-data set from a subset of $N = 3.000.000$ households and persons provided by the 2013 to 2018 American Community Survey public use of micro-data. These surveys are constructed in such a way that there is a 1-in-100 national random sample of the population; the data include persons in group quarters, which are weighted samples. The number of categorical variables (or fields) selected is seven, and their subcategories are illustrated in Table 3.1. In Table 3.2, instead, an explanation of each variable is given. By computing the cross-classification of these variables and their subcategories, we obtain a number of potential cells equal to 8640. However, some combinations of categories are unreal and infeasible. Thus the 3.000.000 individuals

occupy 2050 of these cells.

Table 3.1: Variables and chosen categories

Variable	Categories
SEX	1 = Male, 2 = Female
AGE	0 = 0-10, 1 = 10-20, 2 = 20-30, 3= 30-40, 4 = 40-50, 5 = 50-60, 6 = 60-70, 7 = 70-80, 8 = 80-90, 9 = 90+
MORTGAGE	0 = N/A, 1 = No, owned free and clear, 3 = Yes, mortgaged or similar debt, 4 = Yes, contract to purchase
MARST	1 = Married, spouse present, 2 = Married, spouse absent, 3 = Separated, 4 = Divorced, 5 = Widowed, 6 = Never married/single
HCOVANY	0 = No health insurance coverage, 1 = With health insurance coverage
SCHOOL	0 = N/A , 1 = No, not in school, 2 = Yes, in school
SSMC	0 = Households without same-sex married couple, 1 = Same-sex married couple household where not all relevant data shown as reported, 2 = All other same-sex married couple households

Table 3.2: Selected variables and explanation

Variable	Explanation
SEX	Sex of the individual
AGE	Age of the person
MORTGAGE	Indicates whether an owner-occupied housing unit was owned or was encumbered by a mortgage, a loan or any other type of debt
MARST	Gives each person's current marital status
HCOVANY	Indicates whether the person had any health insurance coverage at the time of the interview
SCHOOL	Indicates whether the individual attended school during a specific period
SSMC	Indicates whether the head of household and spouse are same-sex married couple

After we created the dataset, the common practice is to apply some data cleaning techniques. For this reason, all the rows with missing data have been checked and eliminated. Moreover, the column age has been constructed through a function that maps every age to its pre-selected interval. If the age of an individual were 15, then this function would map it to category 1.

Then the dataset has been translated from a record of categorical variables into a record of identification numbers. That is, every individual belongs to specific categories with a number representing an integer between 0 and 9. Therefore, as the dataset is consistent, every individual belongs to a category for every variable, an identification number is assigned to each record. An example is given in Table 3.3:

Table 3.3: ID creation

MORTG	SSMC	SEX	MARST	HCOV	SCHOOL	AGE	ID
3	0	2	1	2	1	2	3021212
3	0	1	1	2	1	3	3011213

Each record corresponds to an individual, and its ID is given by a string that encodes the information. It is easy to notice that each individual has a string of seven characters, and as the order through which this ID is generated is the same for all the individuals, some IDs will be equal. Besides, some IDs will be unique, as individuals have a singular combination of categorical variables. For simplicity, each ID is mapped into a number between 0 and the number of generated cells called K_1. That is, each ID represents a cell, and identical IDs correspond to individuals that belong to the same cell. It is possible to imagine every individual as a random variable whose realization is a value between $\{0, \ldots, K_1\}$.

Table 3.4: Mapping of ID

INDIVIDUAL	ID	CELL
1	3021212	0
2	3011213	1
3	3016200	2
...	...	...
3000000	3011213	1

Consequently, the frequency for each cell is computed by grouping the individuals by their ID and counting them, as Table 3.5 shows.

Table 3.5: Frequency for each cell

CELL	FREQUENCY
23	81510
5	80615
4	79963
...	...
...	...
1690	1
2047	1

Table 3.5 shows that some cells have the majority of individuals, while others exist only because of the presence of an individual with unique characteristics. As explained in Appendix A, this typology of data distribution resembles a power law, where the majority of individuals belong to an end of the distribution, and the minority is concentrated in the tails. What is interesting for this thesis is the presence of unique cells that will define the magnitude of the disclosure risk.

3.2 Simulation from the Zipfian Distribution

Regardless of the use of a micro-data set created by public sources, the model has been applied to simulations from data distributed according to the Zipfian distribution, whose formal explanation can be found in Appendix A. Our analysis could extend to data that behave accordingly to the power law. The behavior of the cells resembles a power-law distribution, as the majority of individuals are concentrated in specific cells, while some are associated with one individual only. Consequently, studying the model on data simulated by a Zipfian distribution is a close approximation to the study of the model with real data.

The Zipf's law is governed by a parameter s that describes the shape of the distribution. In our experiments, the simulations are obtained with a different parameter of s, ranging from 0.5 to 1.5. The resulting datasets and the performances of the model are commented in chapters 4 and 5, respectively.

For the creation of the dataset, the approach taken is based on the Zipf function provided by the package VGAM, in the R programming language. The function is implemented in such a way that we can choose the population size (our N), the number of cells (K), and the parameter s. A difference between the simulated data and real data is that with simulated data, $K_1 = K$, whereas, as described before, with real data $K_1 <= K$, that is some combination of variables are not feasible. Therefore, we simulate N random variables from a Zipfian distribution, and each of them takes a value in the interval $(0, \ldots, K)$. Through this, we construct the population. From this dataset, we want to extract a sample on which we will estimate the (population) disclosure risk,

according to our model and compare it with the real disclosure risk. The technique adopted for the sampling is random sampling: that is, a sampling method in which each sample has an equal probability of being chosen. The selection of random sampling is justified as the sample constructed is meant to be an unbiased representation of the total population, by its ease of use and effectiveness. The dimension of the sample chosen corresponds to 30% of the population size, therefore $n = 0.3 * N$.

In order to prepare the data for the application of the model, the frequency for each cell is computed by grouping each random variable according to its realization. As these data come from a power-law distribution, the behavior is similar to the real data previously analyzed, that is only some cells are populated with 80% of the population size, and the rest of the cells are either unique or with small participation. We, therefore, compute the frequencies of frequencies (FoF) in order to see this behavior. The FoF is computed on data generated from a Zipf distribution with $N = 100.000$, $K = 10.000$ and $s = 0.9$ and showed in table 3.6.

Table 3.6: Frequency of frequencies (FoF) from Zipfian with $s = 0.9$

ELEMENTS	**CELL**
2	1728
1	1469
3	1440
...	...
...	...
1255	1

Table 3.6 and the bar chart in figure 3.1 (a) show that there are 1728 cells with two elements, 1469 cells with one element, 1440 cells with three elements, and one cell with 1255 elements. The reader should know that this behavior illustrates the similarity between frequencies in cells created from micro-data and frequencies in cells from simulated data.

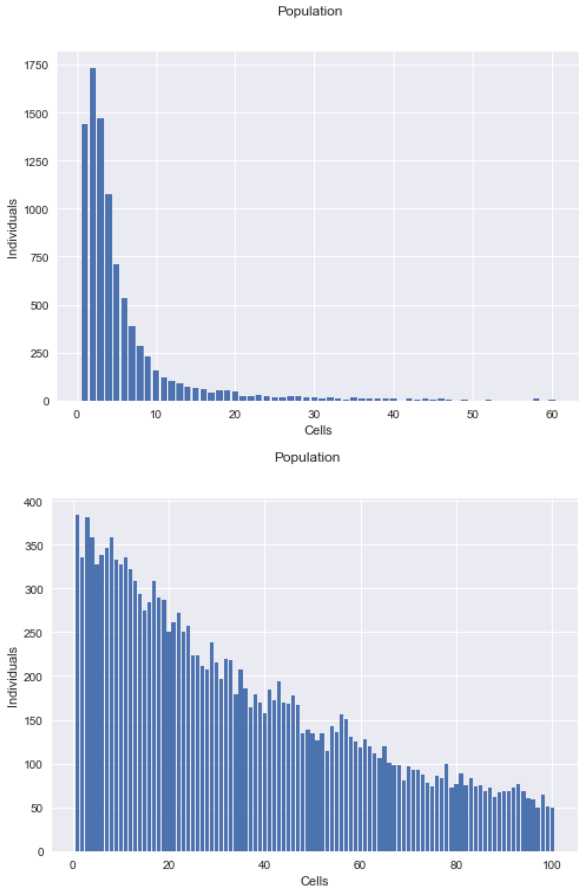

Figure 3.1: (a) FoF Zipf simulated data with $s = 0.9$
(b) FoF Dirichlet-Multinomial simulated data with $\alpha = 1$

3.3 Simulation from the Dirichlet-Multinomial Model

As the disclosure risk estimator is derived from the Dirichlet-Multinomial distribution, this thesis proposes the creation of a dataset from the model itself. The simulations were made through the Jupyter Environment using the Python programming language, and thanks to the function numpy.random.Dirichlet and numpy.random.multinomial implemented in the NumPy package.

For this simulation, the number of cells $K = 20.000$, the population size $N = 1.000.000$, the sample size $n = 0.3 * N$ and the vector $(\alpha_1, \ldots, \alpha_K)$ are selected such that $\alpha_1 = \ldots = \alpha_K = \alpha = 1$. Firstly, we generated from a symmetric Dirichlet distribution the vector $(p_1, \ldots, p_K)$ with fixed parameter α. Secondly, we generate N random variables from a multinomial distribution with parameters $(p_1, \ldots, p_K)$. In this case, we are already generating the frequencies as we are allocating N individuals into

K cells according to the multinomial distribution governed by the $p_1, \ldots, p_K$ vector. We obtain a dataset where every cell from $(0, \ldots, K)$ has several elements (individuals) inside it. An example is provided in table 3.7

Table 3.7: Simulated cell frequencies from Dirichlet-Multinomial distribution

CELL	FREQUENCY
1	218
2	44
3	110
...	...
19999	56
20000	176

We then compute the Frequencies of Frequencies and plot it through a bar chart plot, as shown in figure 3.1 (b). The two bar charts are not similar by choosing $\alpha = 1$, as previously explained, each cell has the same probability of being chosen. If we were going to choose $\alpha \to 0$, the FoF of such simulation would resemble a power law, as the masses are pushed toward the extremes, as shown in the next figure.

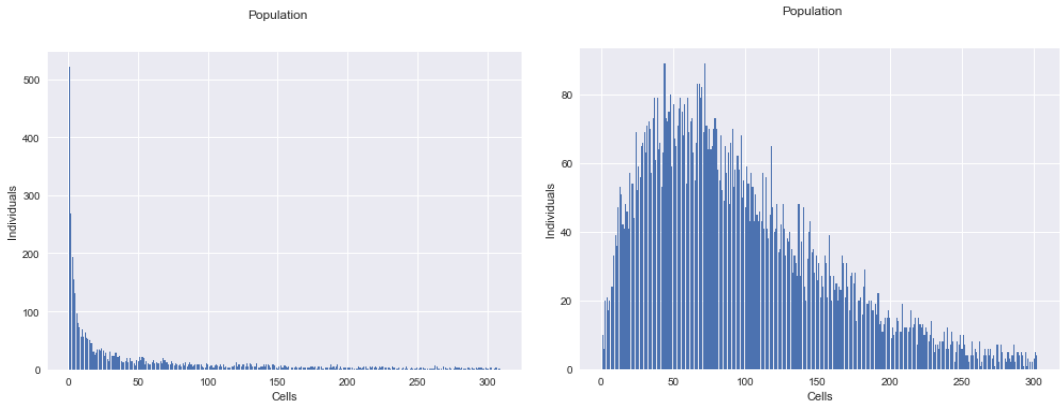

Figure 3.2: (a) FoF Dirichlet-Multinomial with $\alpha = 0.1$
(b) FoF Dirichlet-Multinomial with $\alpha = 1.99$

Chapter 4

Data

This section aims to summarize the empirical results obtained through our model's application on the datasets described in Section 3. The results of the estimation of the disclosure risk on the real micro-data set are discussed in Section 4.1. Then, in Section 4.2 and 4.3, the computation of $\boldsymbol{\tau}_1$ is provided, respectively, on the Zipfian and Dirichlet-Multinomial simulated datasets. The results are exposed "as they are" to allow the reader to develop its interpretation of the analysis performed. They will be further discussed and commented in Chapter 5.

4.1 Application on real micro-data set

We remind the reader of the micro-data set that we constructed and described in Section 3. The population is composed of 3.000.000 individuals, and the sample extracted through random sampling is 30%, therefore 900.000 individuals. The number of categorical variables is seven, and they are described in Tables 3.5 and 3.6. Moreover, the cross-classification of the fields yields a total number of possible cells equal to 8640, defined as K. Instead, by mapping each individual to an identification number (that is, a cell) and computing the frequencies, we obtain that the individuals present in the population occupy only 2050 out of the 8640 possible cells (23.72%). In contrast, the individuals present in the sample occupy 1659 cells (19.20%). With the frequency tables created both for the population and the sample, it is relatively easy to compute

the number of unique in the population and the sample, as we only need to count the number of cells that have frequency $f_i = 1$. We find that the number of unique in the population is 325, and the number of unique in the sample is 292. For the computation of the real disclosure risk of the population, we need to find the number of those individuals that are unique in the sample and remain unique in the population. We have to identify those cells whose frequency is equal to 1 in the sample and in the population. In order to do so, we create two sets that contain the "IDs" of the cells whose participation is equal to 1, respectively, for the sample and the population. Furthermore, we perform the simple mathematical operation of the intersection between sets. By doing so, we find that $\boldsymbol{\tau}$, the actual disclosure risk in the population, is equal to 84, that is, 84 individuals are simultaneously unique in the population and the sample, and therefore at serious risk of information disclosure. This is the real disclosure risk, and we now apply the model that we defined in Chapter 2, and to estimate it, we only observe the sample.

As computed in Section 2.3, the estimator of $\boldsymbol{\tau}$, $\boldsymbol{\tau}_1$ is equal to:

$$\boldsymbol{\tau}_1 \approx m_1 * (\frac{n}{N})^{\alpha+1}. \tag{4.1}$$

We estimated α by maximizing the log-likelihood as introduced in section 2.4, after observing the sample frequencies. By substituting then n, N, α and m_1 in equation 4.1 we find that $\boldsymbol{\tau}_1$ is equal to 87.58. Table 4.1 shows other simulations made on datasets with similar characteristics (that is, they have the same categorical variables, N, n, and K).

Table 4.1: Performance of the model on Real Data

Simulation	$\hat{\alpha}$	τ_1	τ	Difference
1	0.0254	87.58	84	+3.58
2	0.023	93.29	97	-3.71
3	0.034	88.29	95	-6.71
4	0.031	84.28	91	-6.72
5	0.019	92.38	94	-1.62

4.2 Application on Datasets from the Zipfian Distribution

The procedure for estimating the disclosure risk on the dataset generated from a Zipfian distribution is very similar to the one adopted in the previous section. However, we simulated data from Zipfian distribution governed by different parameter s, in order to understand which typology of census data our model would perform best and to identify its limitations. Table 4.2 describes the different behavior of a dataset generated through a Zipfian distribution and according to the choice of s.

Table 4.2: The Role of s

s	N	K	Pop. Unique	Sample Unique
0.2	100.000	10.000	14	1529
0.5	100.000	10.000	82	2230
0.7	100.000	10.000	456	2642
0.9	100.000	10.000	1456	2885
1.2	100.000	10.000	2647	1930
1.5	100.000	10.000	1235	664
2	100.000	10.000	185	126

We can notice two things from table 4.2. The first is that for values of $s \notin [0.5, 1.5]$,

the population or sample's uniqueness can be minimal. Second, for $s < 1$, the number of samples unique is more significant than the number of population unique. Finally, for $s > 1$, the population unique is always more substantial than the sample unique. This property strongly depends on the choice of the population size N, sample size n, and the number of categories K. Thus, we decided to study the performance of the estimator on data simulated from a Zipfian distribution with $s \in [0.5, 1.5]$ as we believe it resembles a more exciting and realistic scenario. For all the experiment we fixed $N = 100.000$, $K = 10.000$ and the sample randomly drawn is the 30%, therefore $n = 30.000$. In table 4.3, the computation of the real disclosure risk, together with the estimation of α from the frequency sample through MLE and the final estimation of $\boldsymbol{\tau}$, that we called $\boldsymbol{\tau}_1$ is provided.

Table 4.3: Performance of the Model on Data from a Zipfian Distribution

Simulation	**s**	$\hat{\boldsymbol{\alpha}}$	$\boldsymbol{\tau}_1$	$\boldsymbol{\tau}$	**Difference**
1	0.5	1.97	59.65	24	35.65
2	0.7	0.83	303.39	137	166.39
3	0.9	0.48	558.52	444	114.48
4	0.99	0.25	607.02	632	-24.98
5	1.2	0.09	484.44	782	-293.56
6	1.5	0.02	207.29	393	-185.71

4.3 Application on Dirichlet-Multinomial dataset

Finally, we evaluate the estimator on the dataset generated from the model itself. As described in Section 3.3, the dataset is created with $N = 1.000.000$ and $K = 20.000$, and the sample size corresponds again to 30% of the population size. As previously stated, we generated directly the frequencies of each cell according to a vector $(p_1, \ldots, p_K)$ of probabilities that comes from a symmetric Dirichlet distribution with fixed parameter $\alpha = 1$. Maximizing the log-likelihood of the sample frequencies with respect to α will lead us to an estimate close to 1. Therefore, the results obtained are described in

table 4.4, where we simulated multiple times from the Dirichlet-Multinomial model and record the value of $\boldsymbol{\tau}$, its estimate $\boldsymbol{\tau}_1$ and their difference.

Table 4.4: Performance of the Model on Data from a Dirichlet-Multinomial

Simulation	$\hat{\boldsymbol{\alpha}}$	$\boldsymbol{\tau}_1$	$\boldsymbol{\tau}$	**Difference**
1	0.988	106.74	130	-23.26
2	1.002	105.75	126	-20.25
3	0.997	105.21	112	-6.79
4	1.002	113.49	110	+3.49

Chapter 5

Discussion: Performance and Limitation

This section discusses and comments on the main result of this paper, providing the interpretation of the author. Section 5.1, 5.2, and 5.3 define the strengths, explain the performance, and the limitations of the model on the datasets created and described in the previous sections, both on simulated and real micro-data.

5.1 Real Micro-Data

The model's performance varies from dataset to dataset, and it depends entirely on its composition. As we can see from chapter 4, the model only behaves well on some typology of the dataset, giving a close estimate of the real disclosure risk. As expected, we can either overestimate or underestimate the real disclosure risk for some datasets after we estimated the parameter α from the sample frequencies. We noticed that our model performs well on those data where the FoF (Frequencies of Frequencies) for the population and the sample resemble the histogram showed in figures 5.1 and 5.2.

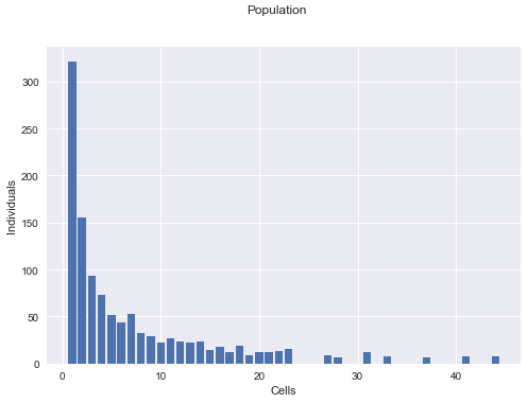

Figure 5.1: FoF Population Real Data

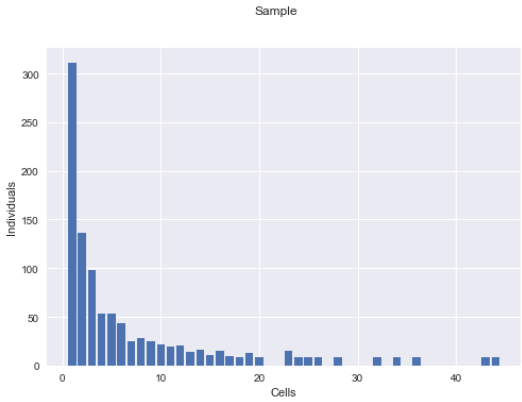

Figure 5.2: FoF Sample Real Data

We can see that the cells that contain only one element both in the sample and in the population correspond to more than 20% of all the cells. When applying our model to those datasets that follow this pattern, and after having estimated α from its sample frequencies, we found that our estimation is close to the true value of $\boldsymbol{\tau}$.

5.2 Simulated data: Zipfian Distribution

Moving to the simulations made on datasets created from the Zipfian distribution with different s parameter, problems start to emerge as we can see that our estimation is pretty far from its true value in some cases, as it is clear from Table 4.2. In particular, we see that the model does not perform well when the number of unique cells is more than 30% of the total cells or when the Frequencies of Frequencies have a shape similar to figure 5.3 and 5.4.

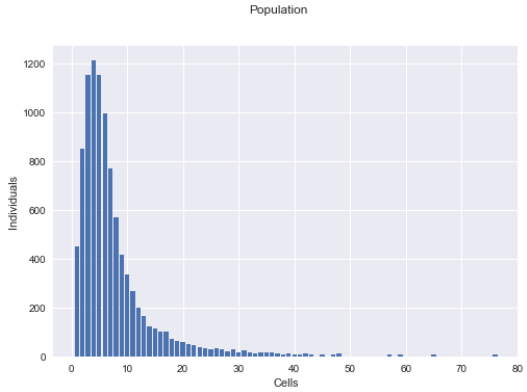

Figure 5.3: FoF Population dataset from Zipfian Distribution with s = 0.7

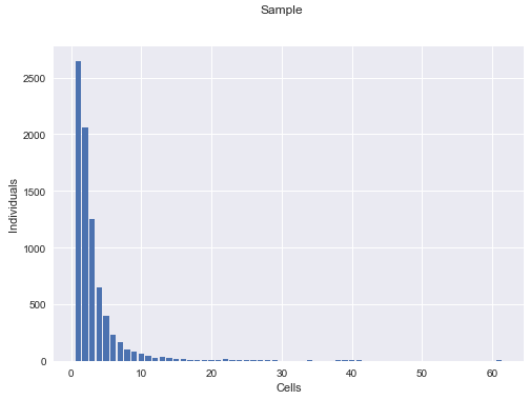

Figure 5.4: FoF Sample dataset from Zipfian Distribution with s = 0.7

However, when we simulated with a parameter $s \approx 1$, we can notice that the proposed model works better, and the estimation is closer to its true value. We expected this result for a different value of s, as it does not exist a true α for the Zipfian distribution since the model we are using is the Dirichlet-Multinomial.

5.3 Simulated Data: Dirichlet-Multinomial Model

Considering the case of the Dirichlet-Multinomial simulated dataset, here we notice that they have a different composition. The cells do not resemble a power-law distribution, as we can see its FoF in figure 5.5 and 5.6 both for the sample and the population.

For these typologies of datasets, by estimating α from the sample frequencies, we find that it is very close to 1, as we showed in Table 4.3, and our estimations are pretty close to the real disclosure risk, as we expected since we are simulating from the model itself.

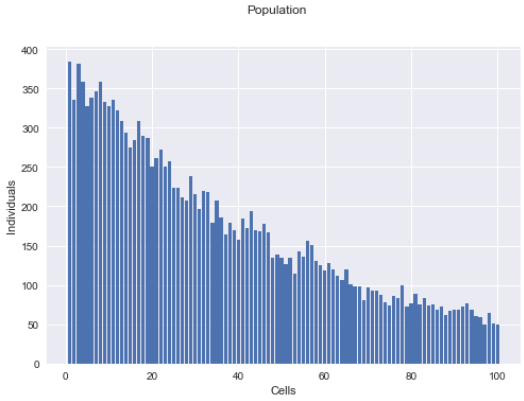

Figure 5.5: FoF Population dataset from Dirichlet-Multinomial with $\alpha = 1$

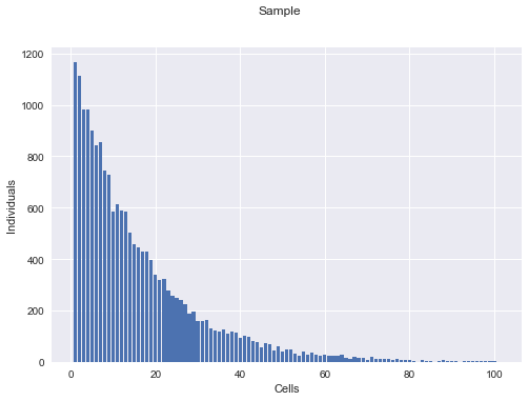

Figure 5.6: FoF Sample dataset from Dirichlet-Multinomial with $\alpha = 1$

Chapter 6

Future Applications

The study of the information disclosure risk and its estimation became an exciting topic as personal data and private information (appropriately anonymized) are released every day by institutions and corporations to third parties to exploit the hidden value of data. The analysis that we proposed was based on a parametric approach from the model introduced by Bethlehem *et al.* (1990), but with a predictive approach. The disclosure risk has been initially estimated through parametric models such as those described in Skinner and Holmes (1998), Carlson (2002), Elamir and Skinner (2006), Forster, and Webb (2007) and Skinner and Shlomo (2008). In the literature, many strategies have been implemented for the disclosure risk estimation, from maximum likelihood estimates to fully Bayesian ones. Moreover, the problem has been analyzed from a nonparametric and semi-parametric Bayesian framework, such as the nonparametric log-linear model adopted by Carota *et al.* (2014).

We believe that the future of statistical disclosure control is already affected by Big Data, that is, enormous and real-time data characterized by volume, velocity, and variety. Therefore, a future challenge could be the estimation of disclosure risk for NoSQL databases, which are unstructured datasets and not just tabular ones (called structured databases). Moreover, as data are considered to be the "new oil", the need for corporations to make data anonymized is urgent, as they are willing to analyze and extract useful information from their data lakes. Statistical disclosure control (SDC) measures described in Section 1.5 are already implemented through Machine Learning

algorithms for practices such as Data Synthesis and Data masking.

Different future studies could be done by analyzing the problem through a Bayesian nonparametric lens and finding a model that works with different typologies of datasets and under different conditions. Moreover, we only studied the problem by finding a point estimator, and we did not deal with the confidence interval estimation. For that purpose, we would need to find the Bayesian posterior distribution of $\boldsymbol{\tau}$ and derive the Bayesian confidence intervals.

Chapter 7

Conclusion

The collection and release of the micro-data file represent a fundamental challenge for businesses and institutions. Thanks to sophisticated algorithms and data-processing software, they are now able to understand better economic trends and tailor their offerings. Besides, this is a positive externality for our society as it generates an impact on products' quality and customer needs will be better satisfied. Moreover, governments and statistical agencies will be more effective in applying regulations and policy measures for our societies and make more informed decisions. Many times, statistical disclosure control practices suppress entirely some information that could be instead used for analysis. The aim of the information disclosure risk assessment is the first step of a more enthusiastic project: minimizing it.

In this paper, we took a predictive approach to estimate the information disclosure risk of a tabular dataset through an estimator based on a Bayesian conjugate model (Dirichlet-Multinomial) already introduced by Bethlehem *et al.* (1990). We tested the model on a real dataset published by IPUMS USA and on those created by us through simulations. By maximizing the log-likelihood of the sample frequencies with respect to the parameter α (that defines a symmetric Dirichlet distribution), we find that the model works well with real data and gives a precise estimate most of the times, even if it does not generalize in some cases.

Appendix A

The Zipf's law

IIn this section, we introduce the Zipf's law, a well-known power law, as it will be used to simulate the data for our analysis. The Zipf's law is an empirical law that was named after its inventor, the linguist George Kingsley Zipf, and was proposed between 1935-1949. It belongs to the family of power-law probability distributions. It is used to model different types of data coming from phenomenons in physical and social sciences, from the size of earthquakes to the frequency of use of words in languages.

In statistics, a power-law describes the relationship between two variables and states that a relative change in one variable results in a relatively proportional change in the other quantity. The power law can be used to describe phenomenons where a cluster of values dominates one end of the distribution. Empirical examples are the size of craters on the moon, frequencies of words in a text, frequency of family names, income distribution, and many others.

A distribution that follows a power law has the behavior shown in Figure A.1.

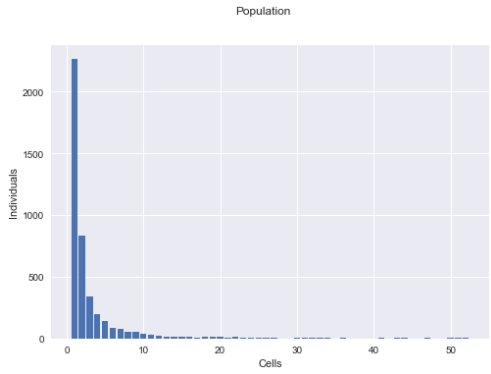

Figure A.1: Example of data following a power law distribution with $s = 1.3$

Identifying a power-law distribution can be difficult sometimes. The strategy to understand it is to plot the same histogram on logarithmic scales. The result of the log-log plot of the histogram of a quantity distributed according to a power-law should resemble a straight line, as shown in Figure A.2.

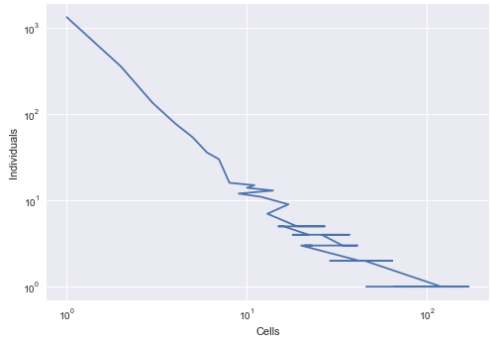

Figure A.2: Log-log plot of data following a power law distribution with $s = 1.3$

A distribution of discrete or continuous variables that follow the power law has the form:

$$P(x) = Cx^{-\alpha},$$

where the constant α is called the exponent of the power-law.
The Zipf's law was intended to measure the frequency of an event related to its rank in a frequency table. Applied to linguistic, the Zipf's law states that the frequency of any word is inversely proportional to its rank in the frequency table. The Zipf's law can be written as:

$$f(k, s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)},$$

where $s \geq 0, N \in \{1, 2, 3 \dots\}$ and whose support is $K \in \{1, 2, \dots, N\}$

Where N is the number of words (or elements), k is their rank and s is the value of the exponent that characterize the distribution.

List of Figures

List of Tables

Bibliography

[1] Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990). Disclosure control of microdata. *Journal of the American Statistical Association* **85**, 38-45.

[2] Carlson, M. (2002). Assessing microdata disclosure risk using the Poisson-inverse Gaussian distribution. *Statistics in Transition* **5** 901-925.

[3] Carota, Filippone, Leombruni, Polettini. (2015). Bayesian nonparametric disclosure risk estimation via mixed effects log-linear models. *The Annals of Applied Statistics* Vol. 9, No. 1, 525-546

[4] Elamir, E. A. H., Skinner C. J. (2006). Record level measures of disclosure risk for survey microdata. *Journal of Official Statistics* **22** 525-539.

[5] *EU General Data Protection Regulation (GDPR)*: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1. Article 4(13), (14), (15) and Article 9 and Recitals (51) to (56).

[6] Forster, J. J. and Webb, E. L. (2007). Bayesian disclosure risk assessment: Predicting small frequencies in contigency tables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **56** 551-570. MR2405419

[7] Newman, M. E. J. (2005). Power laws, pareto distributions, and zipf's law. *Complexity Digest 2005.02*, 1-27.

[8] Pew Research Center, February (2020). Most Adults Aware of 2020 Census and Ready to Respond, but Don't Know Key Details.

[9] Skinner C. J., Holmes D. J. (1998) Estimating the re-identification risk per record in microdata *Journal of Official Statistics* **14** 361-372.

[10] Skinner C.J., Shlomo N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103** 989-1001. MR2462887

[11] Takemura, A. (1999). Some superpopulation models for estimating the number of population uniques. *Proceedings of the Conference on Statistical Data Protection* 45-58. Eurostat, Luxembourg.