Measuring and predicting anonymity

Koot, M.R.

Citation for published version (APA):
Analysis of singletons in generalized birthday problems

5.1 Introduction

Consider again a population of $k$ people, each of them independently assigned a certain ‘feature’ (for instance: gender, birthday, age, ...) which is element of $\{1, \ldots, N\}$; in case of gender $N = 2$ (we simplify reality for clarity of exposition), in case of birthday $N = 365$ (neglecting leap years), in case of age $N$ can be taken, say, 95. We assume the distribution of the feature over $\{1, \ldots, N\}$ is given, which is not a priori assumed to be uniform (birthday and gender will be roughly uniform, whereas age will not). In the literature this setting is often referred to as that of the generalized birthday problem (see Chapter 4), or, alternatively, the birthday problem with unequal probabilities. There is a vast literature on this topic, e.g. [26, 31, 40, 41, 53, 62].

Some of the outcomes will be assigned to just one of the $k$ people in the population; we call these singletons. The objective of this Chapter\(^1\) is the analysis of the distribution of the number of singletons $S$. We subsequently address its mean and variance, as well as a computational scheme for evaluating the distribution of $S$. It is noted that existing literature, and also Chapter 4, primarily focus on the probability that all $k$ samples are unique (where it was

\(^1\)This Chapter is based on M. Koot and M. Mandjes, The analysis of singletons in generalized birthday problems, Probability in the Engineering and Information Sciences, April 2012) [42].
obviously assumed that \( k \leq N \).

Similar to Chapter 3, this Chapter will assume that the adversary knows, beforehand, the set of identities of those whose data is present within the de-identified data set.

The contributions of this Chapter are as follows. Our results cover both the homogeneous setting (that is, all outcomes \( 1, \ldots, N \) being equally likely, that is, have probability \( 1/N \)) and the heterogeneous setting. In the latter, we assume there are \( F_i \) ‘bins’ that have probability \( \alpha_i/N \); obviously we require that \( F_1 + \cdots + F_d = N \) and \( \alpha_1 F_1 + \cdots + \alpha_d F_d = N \).

- In Section 5.2 we first derive an explicit expression for the mean number of singletons \( \mathbb{E}S \). We then scale the number of samples and number of outcomes per group by \( N \), that is, \( k \equiv aN \) and \( F_i \equiv f_i N \). We then show that the mean number of singletons in the scaled model, that is, \( \mathbb{E}S(N) \) can be accurately approximated by

\[
\mathbb{E}S(N) \approx aN e^{-a} \left( 1 + \frac{a}{2} (a - 2) \kappa \right),
\]

where \( \kappa \) is the Kullback-Leibler distance \([23, 47]\) between our heterogeneous distribution and the homogeneous one. This approximation nicely reflects the impact of heterogeneity on the number of singletons. As we will argue, this effect is both quantitatively and qualitatively different for different values of \( a \): for low values of \( a \), \( \mathbb{E}S(N) \) is increasing in \( \kappa \), whereas for high values of \( a \), \( \mathbb{E}S(N) \) is decreasing in \( \kappa \). We illustrate the theory by an example.

- In Section 5.3 we perform a similar analysis for the variance of \( S \). Again we first derive an exact expression for \( \text{Var}S \), and then consider approximations in the scaled model.

- Section 5.4 first develops a recursive algorithm that identifies the full distribution of \( S \) for the homogeneous case. A crucial role is played by a technique to find the probability of no singletons, i.e., \( \mathbb{P}(S = 0) \). Then it is demonstrated how to extend the analysis to the heterogeneous case, for which also a more explicit approximation is presented.

- Section 5.5, finally, is devoted to numerical experiments. Based on demographic data of all Dutch municipalities, we estimate the heterogeneity \( \kappa \), and then assess the accuracy of the approximations for \( \mathbb{E}S \) and \( \text{Var}S \).

### 5.2 Mean number of identifiable objects

In this Section we analyze the mean number of singletons. We find an exact expression, as well as approximations that show how the heterogeneity affects this quantity.
5.2. MEAN NUMBER OF IDENTIFIABLE OBJECTS

5.2.1 Explicit expressions

We first consider the homogeneous case: suppose one throws \( k \) balls into \( N \) bins, uniformly at random. Then, the probability that a given bin contains exactly one ball (a ‘singleton’) is

\[
k \cdot \frac{1}{N} \left(1 - \frac{1}{N}\right)^{k-1};
\]

here we make use of the fact that the number of balls in that bin obeys a binomial distribution with parameter \( k \) and \( 1/N \). As there are \( N \) bins, it follows that the mean number of singletons is

\[
\mathbb{E}[S] = k \left(1 - \frac{1}{N}\right)^{k-1}.
\]

The result for the homogeneous case is standard, but interestingly it can be extended to the heterogeneous setting in a straightforward manner. Let there be \( F_i \) bins that have probability \( \alpha_i/N \); obviously \( F_1 + \cdots F_d = N \) and \( \alpha_1 F_1 + \cdots \alpha_d F_d = N \). Let \( N_i \) the number of balls that end up in bins of type \( i \), and let \( S_i \) be the number of singletons among them; observe that \( N_i \) has a binomial distribution with parameters \( k \) and \( \alpha_i F_i/N \). It is clear that

\[
\mathbb{E}[S_i] = k F_i \left(1 - \frac{\alpha_i}{N}\right)^{k-1} \frac{\alpha_i}{N}.
\]

We arrive at the following statement.

**Proposition 5.1** In the heterogeneous model defined above, the mean number of singletons equals

\[
\mathbb{E}[S] = k \sum_{i=1}^{d} F_i \left(1 - \frac{\alpha_i}{N}\right)^{k-1} \frac{\alpha_i}{N}.
\]

We now consider the number of singletons \( S(N) \) in the asymptotic regime in which there are \( aN \) balls, and \( F_i \) is scaled by \( N \) (that is, \( F_i \equiv f_i N \)). After straightforward calculus we find the following result.

**Proposition 5.2** In the scaled heterogeneous model defined above, the mean number of singletons satisfies, as \( N \to \infty \),

\[
\frac{\mathbb{E}[S(N)]}{N} \to a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a}.
\]

This result essentially states that the number of singletons equals roughly \( Na \) (the number of balls), but thinned by a factor \( \sum_{i=1}^{d} \alpha_i f_i \exp(-\alpha_i a) \); from the requirement that \( \alpha_1 f_1 + \cdots \alpha_d f_d = 1 \) it is immediately seen that this factor is smaller than 1.
5.2.2 Impact of heterogeneity: an approximation

Similar to Chapter 4 and [53], we can assess the impact of heterogeneity by parameterizing \( \alpha_i = 1 + \beta_i \varepsilon \), for \( \varepsilon \) typically small; evidently, \( \beta_1 f_1 + \cdots + \beta_d f_d = 0 \). Relying on the Taylor series \( e^x = 1 + x + x^2/2 + O(x^3) \), it is now immediate that

\[
ad X_i = 1 + \beta_i \varepsilon \quad \text{for } \beta_i \text{ typically small; evidently, } \left( \frac{d}{df} \right) f(1 + \beta_i \varepsilon) = 0.
\]

The Kullback-Leibler distance \( \kappa \) of the non-uniform probabilities \( (1 + \beta_i \varepsilon) / N \) with respect to the uniform probabilities \( 1 / N \) reads, as described in Chapter 4,

\[
\kappa := \sum_{i=1}^{d} f_i N \left( \frac{1 + \beta_i \varepsilon}{N} \right) \log \left( \frac{\left( \frac{1 + \beta_i \varepsilon}{N} \right)}{\left( \frac{1}{N} \right)} \right) = \frac{1}{2} \sum_{i=1}^{d} f_i (\beta_i \varepsilon)^2 + O(\varepsilon^3).
\]

This suggests the approximation

\[
\mathbb{E} S \approx ke^{-k/N} \left( 1 + \frac{k}{N} \left( \frac{k}{N} - 2 \right) \cdot \kappa \right) .
\]

It can even be computed what the fraction \( \gamma_j \) of bins is that is covered by \( j \) balls, when sampling \( aN \) balls. In the homogeneous case this leads to the known result that

\[
\gamma_j = \lim_{N \to \infty} \left( \begin{array}{c} aN \\ N \end{array} \right) \left( \frac{1}{N} \right)^j \left( 1 - \frac{1}{N} \right)^{aN-j} = e^{-a \frac{a^j}{j!}},
\]

we recognize the Poisson distribution. In the heterogeneous model described above,

\[
\gamma_j = \sum_{i=1}^{d} f_i e^{-\alpha_i \varepsilon} \frac{\alpha_i \varepsilon}{j!}.
\]

Notice that the \( \gamma_j \) sum to 1, as desired (recall they represent fractions). Again parameterizing \( \alpha_i = 1 + \beta_i \varepsilon \), for \( \varepsilon \) small, we obtain

\[
\gamma_j = \sum_{i=1}^{d} f_i e^{-\beta_i \varepsilon a + \beta_i \varepsilon a^2 / 2} \left( 1 + j \beta_i \varepsilon + \frac{1}{2} j(j-1) (\beta_i \varepsilon)^2 \right) + O(\varepsilon^3)
\]

\[
= e^{-a \frac{a^j}{j!}} \left( 1 + a^2 + j(j-1) - 2ja \right) \frac{1}{2} \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2 + O(\varepsilon^3).
\]
5.2. MEAN NUMBER OF IDENTIFIABLE OBJECTS

Replacing \( a \) by \( k/N \) and \( \frac{1}{2} \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2 \) by \( \kappa \), this gives an approximation for the fraction of bins covered by \( j \) balls, as a function of \( k, N \), and the ‘non-homogeneity’ \( \kappa \):

\[
\gamma_j \approx e^{-k/N} \frac{(k/N)^j}{j!} \left( 1 + \left( \frac{k^2}{N^2} + j(j-1) - 2j \frac{k}{N} \right) \kappa \right). \tag{5.5}
\]

5.2.3 Remarks, example

Remark 5.3 The above findings also enable us to compute the fraction \( \phi_j \) of people being in a group of size \( j \); cf. the concept of \( k \)-anonymity in privacy \[1, 77\]. After an elementary computation, we obtain for \( j = 1, 2, \ldots \)

\[
\phi_j = \frac{\sum_{\ell=1}^{\infty} \ell \gamma_{\ell} j \gamma_j}{\sum_{\ell=1}^{\infty} \ell \gamma_{\ell}} = \sum_{i=1}^{d} f_i \alpha_i e^{-\alpha_i} \frac{(\alpha_i a)^{j-1}}{(j-1)!}.
\]

As before, this can be approximated by an expression in terms of \( k, N \), and \( \kappa \) only, under \( \alpha_i = 1 + \beta_i \varepsilon \):

\[
\phi_j \approx e^{-k/N} \frac{(k/N)^{j-1}}{(j-1)!} \left( 1 + \left( \frac{k^2}{N^2} + j(j-1) - 2j \frac{k}{N} \right) \kappa \right). \tag{5.6}
\]

It is a matter of elementary calculus to verify that both the approximation of \( \gamma_j \) (as given in Eqn. (5.5)) and the approximation of \( \phi_j \) (as given in Eqn. (5.6)) add up to 1 (summing over \( j = 1, 2, \ldots \), as it should.

Remark 5.4 We now study for which value of \( k \) the above approximation (5.4) is maximized. We do so by looking at the scaled version:

\[
\max_{a \geq 0} a e^{-a} \left( 1 + \frac{1}{2} a(a - 2) \kappa \right).
\]

This yields the first order condition

\[
e^{-a} \left( (1 - a) - \frac{a \kappa}{2} (a - 1)(a - 4) \right) = 0,
\]

yielding the optimizer \( a = 1 \) (which is easily seen to be a maximizer for \( \kappa < \frac{2}{3} \)).

We observe that (5.4) first increases in \( k \), reaches a maximum \( N/e \) at \( k^* = N \), and then decreases, with limiting value 0 as \( k \to \infty \). This qualitative behavior can be understood easily. For small \( k \) there are few singletons, as there are few samples; for large \( k \) quite likely all possible outcomes have been sampled more than once, also resulting in a low number of singletons.

For instance in case of birthdays, assuming they are equally spread over the 365 days, then sampling 365 individuals maximizes the number of identifiable objects, which is (on average) 134.
Remark 5.5 Expression (5.3) confirms the claim that (for \( a \leq 2 \), at least) the mean number of singletons is maximized by the uniform distribution (that is, \( \beta_i = 0 \) for all \( i = 1, \ldots, d \)) — this is due to the absence of a linear (in \( \varepsilon \)) term in the expression in (5.3).

It is observed the mean number of singletons decreases in \( \kappa \) for small \( a \) (that is, \( a < 2 \)), but increases for large \( a \) (that is, \( a > 2 \)). This can be intuitively understood.

- For small \( a \), most bins will be empty or occupied by just one or two balls. Then heterogeneity leads to a smaller number of singletons, as it increases the probability that two balls end up in the same bin.

- For large \( a \), most bins will be occupied by multiple balls. The more heterogeneity, the larger the probability that it is actually just one ball, thus leading to more singletons.

Example 5.6 Consider the following (somewhat atypical) example. Suppose one has data of a set of individuals, consisting of (a) postal code, and (b) age. Assume that ages range from 0 to 94, and (for the moment) that all these ages are equally likely — below we indicate how to deal with heterogeneity. Now suppose that \( k \) people share a postal code, and that \( k \) needs to be chosen so as to optimize the number of uniquely identifiable individuals.

If there is no penalty imposed on the number of postal codes introduced, it is evident that it is optimal to give any individual her or his own postal code. It is more realistic to assume that there are costs, say \( C \), for every postal code introduced. If the set of people has size \( M \), then about \((M/k) \cdot k \exp(-k/N)\) individuals can be uniquely identified. We are therefore faced with the optimization problem

\[
\max_k M e^{-k/N} - C \frac{M}{k};
\]

observe that the value of \( M \) is irrelevant when determining the optimum group size \( k^* \).

It is a matter of elementary computation to conclude that for \( C = 1 \) one should have 10 individuals per postal code; for \( C = 10 \) we obtain 38. Adaptation to the heterogeneous case is straightforward: then

\[
M e^{-k/N} \left( 1 + \frac{k}{N} \left( \frac{k}{N} - 2 \right) \cdot \kappa \right) - C \frac{M}{k}
\]

should be maximized.

5.2.4 Continuous model

The result of Proposition 5.2 can be further refined. We now present its continuous counterpart. Let \( \varphi(\cdot) \) be a continuous density on \([0, 1]\), and define the
probability that an arbitrary ball is put in bin $i$ by

$$\Phi_{i,N} := \int_{(i-1)/N}^{i/N} \varphi(x) dx.$$ 

Then, in the scaled model, due to Proposition 5.1,

$$\lim_{N \to \infty} \frac{\mathbb{E} S(N)}{N} = a \lim_{N \to \infty} \sum_{i=1}^{N} \left(1 - \Phi_{i,N}\right)^{a N - 1} \Phi_{i,N}$$

$$= a \lim_{N \to \infty} \sum_{i=1}^{N} \left(1 - \frac{1}{N} \varphi\left(\frac{i}{N}\right)\right)^{a N - 1} \frac{1}{N} \varphi\left(\frac{i}{N}\right).$$

Now it is a matter of straightforward analysis to derive the following result.

**Proposition 5.7** In the scaled heterogeneous model defined above, the mean number of singletons satisfies, as $N \to \infty$,

$$\mathbb{E}S(N) \to a \int_{0}^{1} \varphi(x) e^{-\varphi(x) a} dx.$$ 

**Example 5.8** Consider the density $\varphi_{\gamma}(x) = \gamma(x - \frac{1}{2}) + 1$, with $\gamma \in [-2, 2]$. The substitution $y := a(\gamma(x - \frac{1}{2}) + 1)$ substitution yields

$$a \int_{0}^{\infty} \varphi_{\gamma}(x) e^{-\varphi_{\gamma}(x) a} dx = \frac{1}{\gamma a} \int_{a(1-\gamma/2)}^{a(1+\gamma/2)} y e^{-y} dy.$$ 

After some calculus, this expression can be rewritten as

$$e^{-a} \left(\frac{a+1}{\gamma a}\right) \left(e^{a\gamma/2} - e^{-a\gamma/2}\right) \left(1 - \frac{2}{e^{a\gamma/2} + e^{-a\gamma/2}}\right).$$

For instance for $\gamma = 2$, we thus find

$$\frac{\mathbb{E} S(N)}{N} \to \frac{1 - e^{-2a} - 2ae^{-2a}}{2a},$$

which is maximized for $a \approx 0.90$.

We could use an approximation in the spirit of (5.3) to approximate $\mathbb{E}S$. For this model, it takes a straightforward computation to obtain that the Kullback-Leibler distance, as a function of the ‘asymmetry parameter’ $\gamma$ equals

$$\kappa = \frac{1}{2\gamma} \left(\left(1 + \frac{\gamma}{2}\right)^2 \log \left(1 + \frac{\gamma}{2}\right) - \left(1 - \frac{\gamma}{2}\right)^2 \log \left(1 - \frac{\gamma}{2}\right) - 1\right);$$

Observe that $\kappa$ is minimal for $\gamma = 0$ (corresponding with the uniform distribution), and symmetric around 0, as could be expected.
5.3 Variance of the number of identifiable objects

This Section considers the variance of the number of singletons. Again, after giving exact expressions and approximations, we study the impact of heterogeneity.

5.3.1 Explicit expressions

As usual, we start with the homogeneous case. Let \( I_j \) be the indicator function of the event that there is exactly one ball in bin \( j \) (where \( j = 1, \ldots, N \)). It was observed before that

\[
P(I_j = 1) = k \cdot \frac{1}{N} \left(1 - \frac{1}{N}\right)^{k-1},
\]

but it is easily verified that for \( j_1 \neq j_2 \),

\[
P(I_{j_1} = 1, I_{j_2} = 1) = k(k-1) \cdot \frac{1}{N^2} \left(1 - \frac{2}{N}\right)^{k-2}.
\]

Observe that \( S = I_1 + \cdots + I_N \). From

\[
\text{Var}S = \mathbb{E}S^2 - (\mathbb{E}S)^2 = \sum_{i=1}^{N} \mathbb{E}I_i + \sum_{i \neq j} \mathbb{E}I_i I_j - (\mathbb{E}S)^2,
\]

we find (noting that there are \( N(N-1) \) terms for which \( i \neq j \))

\[
\text{Var}S = k \left(1 - \frac{1}{N}\right)^{k-1} + k(k-1) \cdot \frac{N-1}{N} \left(1 - \frac{2}{N}\right)^{k-2} - k^2 \left(1 - \frac{1}{N}\right)^{2k-4}.
\]

Again we can consider the random variable \( S(N) \), after scaling \( k \equiv aN \). Directly from the previous formula, we obtain

\[
\lim_{N \to \infty} \frac{\text{Var}S(N)}{N} = ae^{-a} + a^2 \lim_{N \to \infty} N \left( \left(1 - \frac{2}{N}\right)^{aN} - \left(1 - \frac{1}{N}\right)^{2aN} \right).
\]

It is clear that

\[
\lim_{N \to \infty} N \left( \left(1 - \frac{2}{N}\right)^{aN} - \left(1 - \frac{1}{N}\right)^{2aN} \right) = f'(0),
\]

with \( f(x) = (1 - x)^{2a/x} \). Straightforward calculus yields that \( f'(0) = -ae^{-2a} \). In other words, in the homogeneous model,

\[
\lim_{N \to \infty} \frac{\text{Var}S(N)}{N} = ae^{-a}(1 - a^2e^{-a}).
\]
5.3. VARIANCE OF THE NUMBER OF IDENTIFIABLE OBJECTS

We now consider the heterogeneous case. Recall the standard relation

\[
\text{Var} S = \sum_{i=1}^{d} \text{Var} S_i + \sum_{i \neq j} \text{Cov}(S_i, S_j).
\]

Let us first compute \(\text{Var} S_i = \mathbb{E} S_i^2 - (\mathbb{E} S_i)^2\). Observing that we already found \(\mathbb{E} S_i\) in (5.2), we now focus on \(\mathbb{E} S_i^2\). Conditioning on the number of objects that ends up in group \(i\) (which we assumed to have \(F_i\) elements, each with probability \(\alpha_i/N\)) yields

\[
\mathbb{E} S_i^2 = \sum_{j=0}^{k} \binom{k}{j} \left(\frac{\alpha_i F_i}{N}\right)^j \left(1 - \frac{\alpha_i F_i}{N}\right)^{k-j} \cdot \mathbb{E} (S_i^2 | N_i = j).
\]

As earlier,

\[
\mathbb{E} (S_i^2 | N_i = j) = j \left(1 - \frac{1}{F_i}\right)^{j-1} + j(j-1) \cdot \frac{F_i - 1}{F_i} \left(1 - \frac{2}{F_i}\right)^{j-2},
\]

so that

\[
\mathbb{E} S_i^2 = kF_i \left(1 - \frac{\alpha_i}{N}\right)^{k-1} + k(k-1)F_i^2 \left(1 - \frac{2\alpha_i}{N}\right)^{k-2} \left(\frac{\alpha_i}{N}\right)^2. \tag{5.7}
\]

We are now left with computing \(\text{Cov}(S_i, S_j) = \mathbb{E} S_i S_j - \mathbb{E} S_i \mathbb{E} S_j\) for \(i \neq j\). As we already know \(\mathbb{E} S_i\), we focus on \(\mathbb{E} S_i S_j\). It holds that

\[
\mathbb{E} S_i S_j = \sum_{\ell_i=0}^{k} \sum_{\ell_j=0}^{k} \binom{k}{\ell_i, \ell_j} \left(\frac{\alpha_i F_i}{N}\right)^{\ell_i} \left(\frac{\alpha_j F_j}{N}\right)^{\ell_j} \left(1 - \frac{\alpha_i F_i}{N} - \frac{\alpha_j F_j}{N}\right)^{k-\ell_i-\ell_j} \cdot \mathbb{E} (S_i S_j | N_i = \ell_i, N_j = \ell_j),
\]

and in addition a conditional independence argument yields that

\[
\mathbb{E} (S_i S_j | N_i = \ell_i, N_j = \ell_j) = \ell_i \left(1 - \frac{1}{F_i}\right)^{\ell_i-1} \ell_j \left(1 - \frac{1}{F_j}\right)^{\ell_j-1}.
\]

Standard computations now yield that

\[
\mathbb{E} S_i S_j = k(k-1)F_i F_j \left(1 - \frac{\alpha_i}{N} - \frac{\alpha_j}{N}\right)^{k-2} \frac{\alpha_i \alpha_j}{N^2}. \tag{5.8}
\]

Now all the above findings can be collected.
CHAPTER 5. ANALYSIS OF SINGLETONS

Proposition 5.9 In the heterogeneous model defined above, the variance of the number of singletons equals

\[ \text{Var} S = \sum_{i=1}^{d} (\mathbb{E}S_i^2 - (\mathbb{E}S_i)^2) + \sum_{i \neq j} (\mathbb{E}S_i S_j - \mathbb{E}S_i \mathbb{E}S_j), \]

with \( \mathbb{E}S_i \) given by (5.2), \( \mathbb{E}S_i^2 \) by (5.7), and \( \mathbb{E}S_i S_j \) by (5.8).

We now again look at the scaled variant. As before,

\[ \text{Var} S(N) = \frac{\sum_{i=1}^{d} a_i f_i e^{-\alpha_i a} (1 - a^2 \alpha_i f_i e^{-\alpha_i a})}{N^2} \]

Also, due to Lemma 4.1,

\[ \text{Cov}(S_i(N), S_j(N)) = -a^3 \alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i + \alpha_j)a}. \]

We arrive at the following statement.

Proposition 5.10 In the scaled heterogeneous model defined above, the variance of the number of singletons satisfies, as \( N \to \infty \),

\[ \frac{\text{Var} S(N)}{N} = \sum_{i=1}^{d} a_i^2 f_i e^{-\alpha_i a} (1 - a^2 \alpha_i f_i e^{-\alpha_i a}) - \sum_{i \neq j} a^3 \alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i + \alpha_j)a} \]

\[ = a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} - a^3 \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i^2 f_i \alpha_j^2 f_j e^{-(\alpha_i + \alpha_j)a} \]

\[ = a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} - a^3 \left( \sum_{i=1}^{d} \alpha_i^2 f_i e^{-\alpha_i a} \right)^2. \]

5.3.2 Impact of heterogeneity; an approximation

We again parameterize \( \alpha_i = 1 + \beta_i \varepsilon \). We already observed that

\[ a \sum_{i=1}^{d} \alpha_i f_i e^{-\alpha_i a} = a e^{-a} \left( 1 + \frac{a}{2} (a - 2) \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2 \right) + O(\varepsilon^3), \]

whereas it turns out that

\[ a^3 \left( \sum_{i=1}^{d} \alpha_i^2 f_i e^{-\alpha_i a} \right)^2 = a^3 e^{-2a} \left( 1 + (2 - 3a) \sum_{i=1}^{d} f_i \beta_i^2 \varepsilon^2 \right) + O(\varepsilon^3). \]

This leads to the approximation (for the unscaled model)

\[ \text{Var} S \approx k e^{-k/N} \left( 1 - \frac{k}{N} (\frac{k}{N} - 2) \kappa \right) - \frac{k^3}{N^2} e^{-2k/N} \left( 1 + \left( 4 - \frac{K}{N} \right) \kappa \right). \]
5.3.3 Continuous model

We now consider the continuous model, analogously to Section 5.2.4; the probability of a ball being put in bin $i$ is $\Phi_{i,N}$, equalling the integral over the density $\varphi(\cdot)$ between $(i-1)/N$ and $i/N$, for $i = 1, \ldots, N$. The proof of following result is similar to the proof of Proposition 5.7.

**Proposition 5.11** In the scaled heterogeneous model defined above, the variance of the number of singletons satisfies, as $N \to \infty$,

$$\frac{\text{Var} S(N)}{N} \to \int_0^1 a \varphi(x)(1 - a^2 \varphi(x)) e^{-\varphi(x)a} dx.$$ 

We conjecture that $(S(N) - \mathbb{E}S(N))/\sqrt{\text{Var} S(N)/N}$ converges to a standard Normal random variable.

5.4 Probability of at least one singleton

Let $\xi(k, N)$ be the probability of at least one identifiable object, that is, the probability $\mathbb{P}(S > 0)$ of at least one singleton. Particularly if $k$ is large relative to $N$, this is an interesting anonymity metric. (An example could be: suppose one receives data about the ages of 300 people; is there anyone among these 300 people whose age is unique within that group?). In this Section we develop a recursive scheme to evaluate $\xi(k, N)$.

5.4.1 Recursive scheme

We analyze this probability by computing the probability $\zeta(k, N)$ of its complement (that is, no singletons); we start with the homogeneous case. Consider an arbitrary ball that ends up in an arbitrary bin. As there should not be singletons, it means that at least one more ball (out of the remaining $k - 1$) should be in that bin as well; realize that the number of balls that are in that bin (apart from the tagged one) follows a binomial distribution with parameters $k - 1$ and $1/N$. We thus find

$$\zeta(k, N) = \sum_{j=1}^{k-1} \binom{k-1}{j} \left( \frac{1}{N} \right)^j \left( 1 - \frac{1}{N} \right)^{k-1-j} \zeta(k - 1 - j, N - 1).$$

The initialization of this recursion is $\zeta(k, 1) = 1$ and $1 - \zeta(0, N) = \zeta(1, N) = 0$ for any $k = 2, 3, \ldots$ and $N = 1, 2, \ldots$. The first steps can be done easily:

$$\zeta(2, N) = \frac{1}{N}, \quad \zeta(3, N) = \frac{1}{N^2}, \quad \zeta(4, N) = \frac{3N - 2}{N^3},$$
and, with a bit more effort,

\[ \zeta(5, N) = \frac{10N - 9}{N^4}, \quad \zeta(6, N) = \frac{15N^2 - 20N + 6}{N^5}, \]

\[ \zeta(7, N) = \frac{105N^2 - 259N + 155}{N^6}. \]

Table A.1 in Appendix A presents the values of \( \zeta(k, N) \) for \( k = 1, \ldots, 50 \) and \( N = 1, \ldots, 20 \). It shows that \( \zeta(k, N) \) goes to 1 for \( k \) large. In addition, for fixed \( k \), \( \zeta(k, N) \) decreases with \( N \). A nice sanity check for formulae for \( \zeta(k, N) \) is the relation \( (k - 3) \zeta(k, 2) = 1 \).

### 5.4.2 Full distribution of number of singletons

The above results immediately lead to the full distribution of the number of singletons \( S \); for ease we restrict ourselves to the uniform case. It is seen that

\[ \mathbb{P}(S = j) = \binom{N}{j} k(k - 1) \cdots (k - j + 1) \left( \frac{1}{N} \right)^j \left( 1 - \frac{j}{N} \right)^{k-j} \zeta(k - j, N - j); \]

evidently \( \mathbb{P}(S = 0) = \zeta(k, N) \). Evidently, for \( k \leq N \), we already knew from the standard birthday problem that

\[ \mathbb{P}(S = k) = \frac{N!/(N - k)!}{N^k}. \]

### 5.4.3 Heterogeneous case

Once we have computed the numbers \( \zeta(i, j) \) (that correspond to the homogeneous case), it is fairly easy to deal with the heterogeneous case:

\[ \zeta(k, N) = \sum_j \binom{k}{j_1, \ldots, j_d} \prod_{i=1}^d \left( \frac{\alpha_i F_i}{N} \right)^{j_i} \zeta_u(j_i, F_i), \]

where \( \zeta_u(\cdot, \cdot) \) refers to the probability of no singletons in the uniform case, and the summation is over vectors \( j \in \{0, 1, \ldots\}^d \) such that \( j_1 + \cdots j_d = k \).

It is observed that this expression is hard to evaluate, as one has to sum over all vectors \( j \) whose entries add up to \( k \), whose number grows explosively in \( k \). This explains the need for approximations. One such approximation relies on the idea of replacing the multinomial distribution by the corresponding Poisson distribution (where the individual components are assumed to be independent). Then one obtains

\[ \zeta(k, N) = \prod_{i=1}^d \left( \sum_{j=0}^{\infty} \exp \left( -\frac{k\alpha_i F_i}{N} \right) \left( \frac{k\alpha_i F_i}{N} \right)^j / j! \right) \cdot \zeta_u(j, F_i). \]
5.4. PROBABILITY OF AT LEAST ONE SINGLETON

Figure 5.1: Mean number of singletons, as a function of the Kullback-Leibler distance $\kappa$. Left panels: full population; right panels: ages 0–79 only. Top to bottom: $k = 60, 90, 120$. 
5.5 Numerical experiments

In this Section we report on numerical experiments that we ran with demo-
graphic data of all 428 Dutch municipalities that existed in 2010. For all of
them we first determined the distribution of age over the population. Ages
were truncated at 94 years, leaving us with 95 bins, viz. 0 up to and including
94). Then we computed on the basis of this data the Kullback-Leibler distance
\( \kappa \), and the mean and variance of the number of singletons. The mean \( \mathbb{E}S \), and
the mean plus/minus twice the standard error \( \mathbb{E}S \pm 2\sqrt{\text{Var}S} \) are depicted in
the left panels of Fig. 5.1, as a function of the \( \kappa \) — each dot represents one
municipality.

We also include Approximation (5.4), which is a linear function of \( \kappa \). As
argued in the derivation, it is supposed to perform well if the distance with
respect to the uniform distribution is relatively modest. From the left panels of
Fig. 5.1, it is seen that the approximation does not give an accurate prediction.
This is mainly due to the fact that the distribution is highly non-uniform for
the higher ages (ages above, say, 85 are hardly represented). In the right panels
we performed the same experiments, but just for the ages 0 up to and including
79, and there we indeed see an excellent fit.

Although the left panels indicate that Approximation (5.4) does not yield
an accurate estimate for the mean number of singletons \( \mathbb{E}S \) in case the non-
uniformity is too strong, the (nearly) linear shape of the scatter plot does show
that knowledge of the Kullback-Leibler distance accurately predicts \( \mathbb{E}S \). One
could for instance approximate \( \mathbb{E}S \) (as a function of \( \kappa \)) by the linear regression
\( \delta_0 + \delta_1 \kappa \), where \( \delta_0 \) and \( \delta_1 \) are estimated by a least squares procedure.

The left panels show that the mean number of singletons is highest for
\( k = 90 \), which could be expected from Remark 5.4 (recall that \( N = 95 \) here).
Additional experiments (not reported on here) show that when leaving out the
ages 80–94, the mean number of singletons is indeed highest around \( k = 80 \).

Fig. 5.2 shows a scatter plot of the Kullback-Leibler distance \( \kappa \) and the
variance \( \text{Var}S \). For the full population we observe three decreasing, more or
less linear lines; when leaving out the ages 79–94 there is hardly any sensitivity
in \( \kappa \).

5.6 Concluding remarks

This Chapter presented an analysis of the number of singletons in the setting of
the generalized birthday problem. Various metrics have been studied. Special
attention has been paid to obtaining insight into the impact of heterogeneity on
the number of singletons. In Chapter 7 we will discuss applications of the theory
developed here. Future research includes extensive testing with demographic
data.
Figure 5.2: Variance of the number of singletons, as a function of the Kullback-Leibler distance $\kappa$. Left panel: full population; right panel: ages 0–79 only.