The use of saturated count models for synthesis of large confidential administrative databases

James Edward Jackson, BSc (Hons), MSc

Department of Mathematics and Statistics

Lancaster University

A thesis submitted for the degree of

Doctor of Philosophy

December, 2022

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. A rough estimate of the word count is 25,000.

The use of saturated count models for user-friendly synthesis of large confidential administrative databases

James Edward Jackson, BSc (Hons), MSc.

Department of Mathematics and Statistics, Lancaster University

A thesis submitted for the degree of *Doctor of Philosophy*. December, 2022

Abstract

Synthetic data sets are being increasingly used to protect data confidentiality. In the three decades since they were first introduced, methods for synthetic data generation have evolved, but mainly within the domain of survey data sets. As greater interest is being taken in utilising administrative data for statistical purposes, there is inevitably greater interest in creating synthetic administrative databases. Yet there are characteristics of these databases that require special attention from a synthesis perspective, such as their size and the presence of structural zeros. This thesis, through the fitting of saturated models in conjunction with overdispersed count distributions, presents a mechanism that allows large administrative databases to be synthesized e□ciently. This thesis also proposes a concept of satisfying risk and utility metrics *a priori* - that is, prior to synthetic data generation - using the synthesis mechanism's tuning parameters, allowing a more formalized approach to synthesis. The methods are demonstrated empirically throughout, primarily through synthesizing a database that can be viewed as a close substitute to the English School Census.

Publications

The thesis was based on - with many passages taken directly from - the following publications:

Jackson, J. E., Mitra, R., Francis, B. J. & Dove, I. (2022), Using saturated count models for user-friendly synthesis of categorical data', Journal of the Royal Statistical Society: Series A (Statistics in Society). Early view at https://rss.onlinelibrary.wiley.com/doi/10.1111/rssa.12876.

Jackson, J. E., Mitra, R., Francis, B. J. & Dove, I. (2022), On integrating the number of synthetic data sets m into the risk-utility trade-o. In *Privacy in Statistical Databases 2022, Proceedings*, (eds. J. Domingo-Ferrer and M. Laurent), 205–219. Cham: Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-031-13945-1_15.

Jackson, J. E., Mitra, R., Francis, B. J. & Dove, I. (2022), On the e ectiveness of the discretized gamma family distribution for data synthesis. (Currently under review)

Acknowledgements

Just a brief thank you to all who have helped me in the last three years, especially my family and my supervisors, Robin and Brian.

A thank you also to the ESRC for funding the project and to Iain Dove and the ONS for their support throughout.

James Jackson

30th September 2022

Contents

Chapter 1

Introduction

The General Data Protection Regulation (GDPR), implemented by both the European Union (EU) and the UK in May 2018, requires that businesses and organisations adhere to certain standards when processing personal data, that is, information that relates to identifiable individuals (Information Commissioner's O□ce, 2020). A notable requirement is that organisations now have to demonstrate *how* they are complying with the regulations. Even data with direct identifiers removed - such as names and identification numbers - are regarded as personal data with respect to GDPR; only *truly anonymous* data are exempt.

However, in general, anonymisation is a contentious issue and there is some ambiguity as to what anonymised data actually constitute.[1] The need to comply with GDPR is coupled with the ever-growing volume of data available through the internet, thus allowing intruders (attackers) - malicious users of the data - to link records

[1] In November 2019, The University of Manchester held a debate entitled $Debate\ on\ `Anonymisation'$ where speakers gave their opinions on the notion of anonymisation. For instance, one speaker argued that "data can either be useful or anonymised but never both", whereas another argued that anonymisation should be considered with respect to the "data environment", that is, the conditions under which the data are released.

from di□erent databases. The well-repeated, infamous example of this occurring is the supposedly anonymised Netflix data set, where re-identifications were possible by linking records to publicly available information on the Internet Movie Database (IMDb) website (see Schneier 2007).

There is a drive - by the UK Government, for example (HM Government, 2018) - to fully utilize administrative data. These databases, collected by organisations such as government departments for administrative - non-statistical - purposes, can sometimes hold records for an entire population. Owing to their population-like qualities, large administrative databases are increasingly being primed to supplement - or even replace - future censuses.

But there are challenges to address before using administrative data for statistical analysis. Protecting privacy is the overriding consideration. Often individuals would be unaware that their information is even being held in these databases, and it would not be feasible to obtain formal consent for their data to be used for analysis, as they would do in a survey, for example. Also, unlike census data, these databases may be held in a relatively insecure environment where, for example, any members of sta□ can access versions of the original database. Obstacles from a statistical perspective have been considered later in the thesis; the overriding obstacle, though, is the sheer size of administrative databases, which presents a challenge when protecting privacy.

This thesis will explore the use of synthetic data methods to protect privacy in administrative databases. Synthetic data sets are generated by fitting and simulating from a model (synthesis model) fit to the original data. The idea is that they can preserve the statistical properties of the original data, hence providing analysts with valid inferences (albeit with more uncertainty associated with estimates). As values

are not real, disclosure risk should be much lower than in the original data. Synthetic data methods have continued to develop since 1993, when the *Journal of O cial Statistics* published a special issue on data confidentiality. Rubin (1993) and Little (1993) proposed, in separate articles, similar ideas which later led to the genesis of synthetic data sets. Rubin (1987) proposed to adapt the methods he himself had developed for the multiple imputation of missing data (Rubin, 1987). He e ectively suggested treating those individuals from the population who were not included in the original data as missing and multiply imputing values for these individuals. Little (1993) proposed replacing the original data's sensitive values by simulating from models fit to the original data. Rubin (1993)'s idea led to the branch of synthetic data sets now known as fully synthetic (Raghunathan et al., 2003); whereas Little (1993)'s idea led to partially synthetic data sets (Reiter, 2003).

The aim of synthesis models is to capture the underlying distribution governing the original data. Many models have been proposed to achieve this: from generalised linear models (see, for example, Drechsler 2011), to tree-based non-parametric methods such as classification and regression trees (CART) (Reiter, 2005d) and random forests (Caiola and Reiter, 2010), to Bayesian non-parametric latent class models (Manrique-Vallier and Hu, 2018), to machine learning techniques such as generative adversarial networks (GANs) (Kaloskampis, 2019). The quality of synthetic data depends entirely on the quality of these models.

Synthetic data are subject to the same risk-utility trade-o (see Duncan et al. 2001) inherent in all statistical disclosure control (SDC) methods: that data with high utility comes at the expense of high risk of disclosure. A compromise needs to be struck: that is, it needs to be ascertained whether a given level of utility is worth

the cost in terms of risk. As part of a pilot study into synthetic data conducted by the UK's O□ce for National Statistics (ONS) (Bates et al., 2019), a spectrum was devised as a way of classifying types of synthetic data with respect to this trade-o□. These range from "structural" synthetic data sets where only the basic structure of the original data are preserved, such that the utility and risk are essentially null; these can be useful as test data sets, for example, for highlighting any issues with code. At the other end of the spectrum are "replica" synthetic data sets, which closely resemble the original and, consequently, have high risk and utility. A similar spectrum was proposed by Kokosi et al. (2022), who use the term univariate synthetic data for the case when risk and utility are minimal and complex modality synthetic data for when risk and utility are at their highest. An advantage of the methods developed in this thesis is that they can be tuned - through the adjustment of parameters - to produce synthetic data anywhere on this spectrum.

When any data set - synthetic or otherwise - relating to individuals is released, the considerations for the various parties involved - the individuals included, the data-holder (for example, a government department), the analysts (researchers using the data) and the intruders - are di□erent and need to be considered separately. The individuals involved demand that their privacy be protected; analysts expect the data to facilitate the obtaining of valid inferences; the data holder's concern is two-fold: whether the data protects confidentiality while also yielding valid inferences; and intruders seek to glean sensitive information about the respondents. When first proposing the use of synthetic data sets for SDC, Rubin (1993) points out that they neatly satisfy these di□ering concerns: individuals' privacy should be protected since no real values are released (thereby deterring intruders, too); while the data-

holder is responsible for the imputation aspect - the heavy lifting - and can employ a trained statistical modeller to do so. This contrasts favourably with some other SDC techniques - such as the addition of random noise - that place the burden on analysts to undertake relatively complex statistical techniques.

The overriding advantage, then, of synthetic data is that values are artificial - not real. Fully synthetic data sets contain no real values whatsoever; and partially synthetic data replace sensitive values (why bother replacing more values than is necessary?). Either way, the disclosure risk is low and it is hard to imagine another SDC technique being able to lower this risk further. This is the reason why synthetic data are being considered as a way of making administrative data - a particularly "high-risk" source of data - available to analysts.

In their raw state, these data sets tend to follow a microdata format - an $n \quad p$ matrix of values - where rows relate to individuals (also known as records, subjects or units) and the columns relate to variables. A key feature of administrative databases is that variables tend to be categorical in nature. This leads to an alternative - but equivalent - way of representing them. A microdata set comprising entirely of categorical variables can be aggregated into a contingency table, such that cell counts give the frequencies with which the various combination of categories across variables (cells) are observed. In general, if there are p categorical variables with $m_1, \ldots, m_p$ categories, respectively, then the data can be expressed as a multi-dimensional contingency table with $m_1 \quad \ldots \quad m_p$ cells. There is no loss of information when categorical microdata sets are tabulated in this way; though it usually allows the data to be presented more compactly.

This thesis focuses on the synthesis of categorical data at the tabular level; this

naturally produces synthetic data sets also at the tabular level, which can then, of course, either be released in tabular format or expanded back to microdata.

Chapter 2

Literature Review

2.1 Disclosure risk in statistical data sets

From a disclosure risk perspective, variables in a data set can be broadly classified into two kinds: *sensitive variables* and *key variables* (Bethlehem et al., 1990). *Sensitive variables*, such as income, are confidential and individuals have a right for these to be kept private; gaining access to these variables can be viewed as the motivation behind an intruder attack. To ensure confidentiality is protected, all variables can be considered sensitive, which is the stance taken by Statistics Netherlands (Willenborg and De Waal, 1996). *Key variables* (also known as *quasi-identifiers*) are those that are not inherently sensitive but which can serve to identify an individual, thus are intrinsic to disclosure risk in much the same way as sensitive variables are; examples of key variables include sex, age, ethnicity, etc. A data set is also almost certain to include *formal identifiers* (or *direct identifiers*), which are variables that explicitly identify individuals, such as individuals' names, social security numbers, national insurance numbers, etc. As these variables add little - if any - value from a utility

perspective, they are often just removed entirely from a data set. An alternative is to use pseudonymization, which is when false names (or numbers) replace the real names. *Formal identifiers* have been ignored henceforth.

Skinner (1992) defined two types of disclosure that can occur: re-identification and prediction (attribute) disclosure. Re-identification is when an intruder can establish the true identity of an anonymized record. Once a re-identification takes place, an intruder obtains all information pertaining to that individual, including their sensitive values. Even if the intruder cannot extract any sensitive information from the re-identification, privacy has still been compromised, that is, re-identification in itself represents a form of disclosure. Re-identifications are often achieved by linking records in the data set to other data sources, such as publicly-available data. An infamous example of this was the supposedly anonymized 2006 Netflix data set, where re-identifications were possible by linking records to publicly available data on the Internet Movie Database (IMDb) website (see Schneier, 2007).

Prediction disclosure (Skinner, 1992) - now more commonly referred to as attribute disclosure - is when an intruder can precisely estimate the sensitive values of a target (an individual of interest to the intruder) without necessarily re-identifying them. For example, suppose a data set is released in tabular format and one of the variables is, say, average income; and suppose an intruder knows to which cell a target belongs and that they "dominate" the cell; that is, the target's contribution to the cell is much larger than any of the other individuals' contributions. Then an intruder may be able to estimate with a reasonable degree of accuracy the target's income.

Another type of disclosure discussed in the literature is inferential disclosure (Duncan and Lambert, 1989), the notion of which was first mentioned by Dalenius

(1977). It relates to the *additional* information that an intruder can gain about a target through re-identification; as Drechsler (2011) points out, this is a similar concept to differential privacy (Dwork et al., 2006). If the intruder gains no extra information through re-identification, then the corresponding inferential disclosure would be relatively small.

In general, increasing the number of individuals in a data set (increasing n) increases the disclosure risk because there is less sampling uncertainty. But this assumption is conditional on an intruder not knowing which individuals are present in the data. If an intruder knows, however, that the data includes a particular individual, smaller n actually heightens the risk as there are fewer records to search through (Bethlehem et al., 1990).

An important concept is perceived disclosure risk (see Couper et al., 2008), which relates to an individual's personal feeling that their data are sufficiently protected. This, though, can be difficult to quantify. Disclosure risk metrics - which seek to quantify the level of risk - tend to focus on either re-identification or attribute disclosure.

2.1.1 Uniqueness

The risk of re-identification is typically examined with respect to the key variables in a data set. As these variables tend to be categorical, individuals can be aggregated according to these key variables into a contingency table, where cells pertain to the various cross-classifications. Individuals who belong to a cell with a small cell count, particularly a cell count of one (sample uniques), are at greater risk of re-identification. In general, if an intruder knows that a target belongs to a cell of size k, they can be

re-identified with probability $1/k$.

Individuals who are unique in the population (population uniques) on a given set of key variables are at greatest risk. It follows that a population unique is always a sample unique (Bethlehem et al., 1990) (the converse does not follow). Much work has been devoted to estimating the probability that a sample unique is also unique in the population; for example, see Skinner et al. (1994) and Skinner and Shlomo (2008).

Elliot et al. (1998) extended the concept of a sample unique to develop the concept of a special unique. Special uniques are those that are unique on a subset of the data's p variables. In the extreme case, a special unique can be unique with respect to just one variable. Special uniques and the concept of di□erent levels of sample uniqueness allow a finer attribution of risk than simply unique or not unique.

The properties of k-anonymity (Sweeney, 2002) and L-diversity (Machanavajjhala et al., 2007) are linked to uniqueness. A data set has the property of k-anonymity if, for every individual in the data, there exists at least $k \quad 1$ other individuals with the same set of values; that is, when the data set is tabulated according to all its variables, k-anonymity is achieved if the minimum cell size is k (excluding zeros). Thus any data set that achieves 2-anonymity guarantees the absence of sample uniques.

The metric L-diversity is a similar concept. Suppose the variables in a data set comprise key variables and one sensitive variable. The property of L-diversity is achieved if, for each observed set of key variable values, the sensitive variable takes at least L distinct values.

In a broad sense, k-anonymity relates to re-identification and L-diversity relates to attribute disclosure. These concepts require special consideration in a synthetic data setting. There is only a potential risk of disclosure if neither the original data set nor

the synthetic data set satisfy k-anonymity and L-diversity - and if it is the same cells which causes the definitions to fail. If the original data satisfies k-anonymity and L-diversity, then it does not matter from a risk perspective if the synthetic data do not. The ability, however, to obtain risk guarantees in the manner that k-anonymity and L-diversity provide is an appealing feature when protecting administrative databases. It may not be possible to check the risk for every individual, so measures which provide guarantees for the entire data set are useful.

Finally, as Fienberg and Makov (1998) point out, even cells with a count of zero have a disclosure risk attached, because an intruder can use these to make deductions about the non-zero counts; for example, by looking at the pattern of zeros, an intruder may be able to deduce, say, that a target's income lies between £60k-£80k .

2.2 Statistical disclosure control (SDC) for categorical data

Statistical Disclosure Control (SDC), also known as Statistical Disclosure Limitation (SDL), is the process of protecting the privacy of individuals included in a data set. There is an inherent trade-o when applying SDC methods (see Duncan et al., 2001), which is analogous to many other trade-os, such as the work-life balance, risk and reward in finance and the bias-variance trade-o. The literature on SDC is extensive and includes many books, manuals and articles; see, for example, Duncan et al. (2011), Hundepool et al. (2012), Templ (2017) and SDAP Working Group (2019).

SDC methods for categorical data can either be applied pre- or post-tabulation, and can either be applied at a local level to specific entries or cells, or at a global

level to the entire data set. Methods can also be roughly divided into two kinds: coarsening methods, which reduce or suppress the amount of information released; and perturbative methods, which alter the released data.

2.2.1 Data coarsening methods

Global recoding, also called *table redesign*, is an umbrella term for a variety of alterations that can be made to the structure of a data set. Firstly, one example is to drop a key variable, which reduces the number of uniques, thus lessening the risk of re-identification. Alternatively, categories can be collapsed to reduce granularity, for example, using county- rather than town-level geographical regions. Similarly, continuous variables can be re-coded as categorical variables, for example, ages can be expressed as integers rather than exact date of births; the loss of information can be clearly seen here, since, for example, a child recently turned three cannot be distinguished from a child aged nearly four. A related concept is *top and bottom coding*, which is where a maximum or minimum value is specified, and no values can surpass the threshold. The classic example in the literature is setting a maximum income value at, say, $£100,000$. The motivation behind top and bottom coding is that it protects the values at the tail of the distribution - the extremes - which tend to be relatively high risk. It is worth noting here that SDC applied data will support some analyses, but not others. For example, a top coded data set would have low-utility for an analyst wishing to learn about extremes.

Another data coarsening technique, suited to microdata or tabular data, is *local suppression*, which is when certain values - or cell counts - are omitted. Suppression effectively results in data with missing values, which can still present a disclosure

risk, especially when an intruder can understand *why* a value is suppressed. Primary suppressions may also require secondary suppressions, particularly in contingency tables which have marginal counts (row and column sums): when certain cell counts are suppressed but marginal counts are not, an intruder can deduce - or at least estimate - the underlying suppressed counts.

2.2.2 Data perturbation methods

The addition of *random noise* is a well-known technique that usually assumes that a variable in the data is normally distributed. The sample mean and variance of this normal distribution are estimated, before replacement values are generated by drawing random variates. A downside of noise addition is that it complicates the subsequent analyses performed by researchers, who have to use the measurement error models - such as those given by Fuller (1987) - to obtain valid inferences.

Other perturbative techniques include - but are not limited to - the following: *microaggregation* (see Mateo Sanz and Domingo Ferrer, 1998), which groups similar records together and replaces original values with average values from these groups; *data swapping* (see Dalenius and Reiss, 1982), which involves switching values between records in the data set; and *post-randomisation method (PRAM)* (see Gouweleeuw et al., 1998), whereby values in the original data are changed according to a Markov matrix. PRAM is a general method that includes noise addition, suppression and global recoding (Hundepool et al., 2010).

Data perturbation methods, however, alter the original values, which can distort relationships between variables. This is in contrast to coarsening methods, which seek to hide the original values. Purdam and Elliot (2007) discovered that applying

these traditional SDC methods on samples of anonymised records (SARs) from the UK census had a significant impact on the data's utility.

2.3 Introduction to synthetic data

The generation of synthetic data sets can be viewed as a unique type of SDC method. They are generated by simulating from a model fit to the original data. The notion is that synthetic data sets can be released to analysts instead of the original data. As the synthetic values are inherently artificial, individuals' privacy should be protected, while, as synthetic values are based on the original values, analysts' ability to obtain valid inferences should remain undiminished. The method relies on the synthesizer – he or she responsible for generating the synthetic data – accurately modelling the data's underlying distribution; such models are known as synthesis models.

The theory of synthetic data evolved from the multiple imputation of missing data (Rubin, 1987), and as with multiple imputation, it is typical to generate and release multiple data sets to allow analysts – through combining rules; see Drechsler (2011) - to formally account for the uncertainty from simulation.

Traditionally, synthetic data sets are either fully or partially synthetic. The distinction between the two is confusing, owing to the way in which the synthetic data literature has developed over time. The generation of fully synthetic data, as first proposed by Rubin (1993) and later developed by Raghunathan et al. (2003), involves creating a synthetic population by imputing values for those individuals who belong to the population but who *were not* included in the original data. If the original data is not a simple random sample, values must be imputed using the Bayesian posterior predictive distribution, to fully account for the underlying uncertainty. On the other

hand, partially synthetic data, as first proposed by Little (1993) and later formalized by Reiter (2003), involves generating synthetic values for some - or even all - of the individuals who *were* included in the original data. As values are simply replaced, simulating from the Bayesian posterior predictive distribution is never necessary; it suffices to simulate values directly from the fitted model (see Reiter and Kinney, 2012). This approach - referred to as the "plug-in" approach, as the model parameters' MLEs are "plugged" directly into the synthesis model - involves fewer steps, thus has advantages in terms of computational time, particularly as it can be difficult to obtain parameters' standard errors, hence posterior distributions, in large data sets. Raab et al. (2016) showed that this approach is equally valid for fully synthetic data, when the original data constitute a sample random sample.

Confusion arises when all original values are replaced with synthetic values using the partially synthetic framework of Reiter (2003). In this instance, the synthetic data contains synthetic values only, but, by definition, are not classified as fully synthetic. Instead of referring to fully and partially synthetic data sets, Raab et al. (2016) introduced the more intuitive definitions of *completely* and *incompletely synthesized data sets*. The former includes any synthetic data set with no original values; whereas the latter includes any synthetic data set with some original values. The newer term of completely synthesized data sets, therefore, encompasses a wider range of data sets than the older term of fully synthetic data sets; that is, the latter is a subset of the former. Completely synthesized data sets include all data sets that would previously have been described as fully synthetic, plus some data sets that would previously have been described as partially synthetic.

2.3.1 The generation of synthetic data sets

Suppose a synthesizer wishes to synthesize a data set with p variables $Y = (Y_1, Y_2, \ldots, Y_p)$. Generating synthetic data involves modelling the underlying joint distribution of Y, which for categorical data sets can be done by modelling Y by a model such as the Poisson log-linear model. In practice, though, because a data set is likely to comprise of a mixture of continuous and categorical variables, it is more common to decompose the joint model $Y = (Y_1, Y_2, \ldots, Y_p)$ into a product of conditional models, that is

$$P(Y \mid X) = P(Y_1 \mid X) \prod_{j=2}^{p} P(Y_j \mid Y_{j-1} \ldots, Y_2, Y_1, X), \tag{2.1}$$

where X are any additional variables not requiring synthesis. Separate models to be specified and fit to each of these variables. The ordering of the variables can dramatically affect the utility of the resulting synthetic data. This first variable to be modelled, Y_1, subsequently becomes a predictor when modelling the other variables. Thus a biased synthesis model for Y_1 propagates bias throughout the synthesis. Ideally, the variable chosen as Y_1 is the one which is the strongest predictor on the other variables, yet this is, of course, subjective.

Each model is fitted to the original data to obtain, depending on whether or not the Bayesian posterior predictive distribution is used, either the model parameters' joint posterior distribution or the model parameters' maximum likelihood estimates (MLEs). In some cases, the *exact* joint posterior distribution can be derived, but more often an *approximate* posterior distribution is derived, by estimating the variance-covariance matrix numerically, for example, via the R function vcov from the

stats package. Synthetic values are then obtained by simulating from each variable sequentially; this is repeated m times to obtain m synthetic data sets.

2.3.2 Obtaining inferences from synthetic data sets

When carrying out a full synthesis, simple random samples of size n_{syn} are taken from each of the m synthetic populations to release to analysts. Reiter and Raghunathan (2007) emphasize that, in practice, entire synthetic populations do not need to be generated; it su□ces to take a random sample of individuals from the population first, and then synthesize values for these individuals only. For a partial synthesis, n_{syn} is usually set equal to the size of the original sample n. Metadata may be released alongside these samples, such as information about the synthesis models, even model parameters. However, such information would heighten the risk.

The synthesizer has control over m and n_{syn} and can tweak these to find an optimal balance between risk and utility; see, for example, Reiter (2005b). The motivation for keeping m and n_{syn} small is two-fold: firstly, fewer records (smaller n_{syn}) typically implies lower risk; and secondly, fewer data sets (smaller m) means the synthesis is less computationally burdensome, which is especially relevant for large, administrative databases. This optimal combination of m and n_{syn} is likely to depend on a range of factors, such as the synthesis model, the number of variables being synthesised and the estimands of interest to an analyst.

When estimating a univariate population parameter Q from $m \quad 1$ synthetic data sets, the analyst treats each synthetic data set as if it is the original data set, and uses complete-data methods for inference. A point estimate $q^{(l)}$ and its corresponding variance estimate $v^{(l)}$ is obtained from each synthetic data set, $l = 1, \ldots, m$. The next

step is to calculate the following three quantities Drechsler (2011):

$$\bar{q}_m = \frac{1}{m}\sum_{l=1}^{m} q^{(l)}, \qquad b_m = \frac{1}{(m \quad 1)}\sum_{l=1}^{m}(q^{(l)} \quad \bar{q}_m)^2, \qquad \bar{v}_m = \frac{1}{m}\sum_{l=1}^{m} v^{(l)}, \tag{2.2}$$

where $\bar{q}_m$ is the mean estimate, b_m is the 'between-synthesis variance', that is, the sample variance of the $m > 1$ estimates, and $\bar{v}_m$ is the mean 'within-synthesis variance', the mean of the estimates' variance estimates. When $m = 1$, it is not possible to calculate b_m; the combining rules for $m = 1$ derived by Raab et al. (2016) do not use this quantity (see Table 2.1). Note also that the variability in v_l is assumed to be ignorable, therefore it is su□cient to use any v_l instead of $\bar{v}_m$.

The way in which these quantities are combined depends on several factors, such as whether fully or partially synthetic data sets are generated, whether the original data are a simple random sample and whether an analyst is using the synthetic data to estimate a population parameter Q or an observed data estimate $\hat{Q}$. The later arises, for example, when synthetic data are acting as test data, where the aim is to obtain inferences comparable to those that would have been obtained had the analyses been run on the original data. In this instance, estimating variances is simpler because the combining rules need only take account of the uncertainty from synthesis; whereas when estimating a population parameter directly, the sampling uncertainty inherent in the observed data also needs to be taken into account, that is, there is an extra source of variability. This, incidentally, is the purpose of synthetic data in the synthpop package in R (Nowok et al., 2016).

Regardless of whether the synthetic data are being used to estimate Q or $\hat{Q}$, the average (mean) of the point estimates $\bar{q}_m$ is an unbiased estimate of Q. An unbiased estimate for Var$(\bar{q}_m)$ can be obtained by applying the relevant combining rules. Table

2.1 lists valid combining rules in different scenarios. Under certain conditions, there are different sets of combining rules that yield valid inferences, for example, Raab et al. (2016) give two sets of valid combining rules for fully synthetic data sets generated for an original data set that is a simple random sample. Drechsler (2018) also clarifies when each set of combining rules should be used.

Estimating variance-covariance matrices for multivariate population estimands requires more work on the part of the analyst and there is a notable increase in the complexity of combining rules; see Reiter (2005c) and Reiter and Raghunathan (2007).

These inferential frameworks were originally developed for small survey data sets. They were later extended for the generation of synthetic census data (Reiter and Drechsler, 2010), which, for instance, involved consideration of the finite population correction factor. There are further complications - factors to consider - when generating synthetic administrative data; these are discussed further in Section 2.8.

2.4 Synthesis models for categorical data

As mentioned above, the original data are often modelled through a product of conditional models because the data may comprise a range of variable types - continuous, categorical, count variables, *et cetera*. The ability to convert categorical data into tabular format, increases the options available to the synthesizer because the synthesis can either be undertaken at the microdata or tabular level.

Combining rules	Description
$T_s = \left(1 + \frac{1}{m}\right) b_m \quad \bar{v}_m$	Fully synthetic data sets generated via the Bayesian posterior predictive distribution; derived in Raghunathan et al. (2003).
$T_s^{\text{PPD}} = \bar{v}_m \left[\frac{n_{\text{syn}}}{n} + \frac{1}{m}\left(1 + \frac{n_{\text{syn}}}{n}\right)\right]$	Fully synthetic data sets generated via the Bayesian posterior predictive distribution when the original data constitute a simple random sample; derived in Raab et al. (2016).
$T_s^{\text{plug}} = \bar{v}_m \left(\frac{n_{\text{syn}}}{n} + \frac{1}{m}\right)$	Completely synthesized data when the "plug-in" approach is used and the original data constitute a simple random sample; derived in Raab et al. (2016).
$T_p = \bar{v}_m + \frac{b_m}{m}$	Any kind of partially synthetic data sets; derived in Reiter (2003).
$T_p = \frac{v_m n_{\text{syn}}}{n} + \frac{b_m}{m}$	Partially synthetic data when all values are replaced and a sample of size $n_{\text{syn}} < n$ is released; derived in Raab et al. (2016).
$T_d^1 = \frac{b_m}{m}$	When the original data is the entire population and small synthetic samples are randomly drawn; derived in Drechsler and Reiter (2010).
$T_d^2 = \bar{v}_m + \left(\frac{1}{m} \quad 1\right) b_m$	When the original data is the entire population and observations for the same set of individuals are released in each synthetic sample; derived in Drechsler and Reiter (2010).
$T_d^3 = \frac{1 \quad n_{\text{syn}}}{n} + \frac{b_m}{m}$	As above, but when the synthetic sample size n_{syn} is not small enough to ignore the finite population correction factor; derived in Drechsler and Reiter (2010).

Table 2.1: The combining rules for variance estimates when estimating a univariate population estimand from multiple synthetic data sets in di□erent scenarios. For clarity, n_{syn} is the size of the synthetic data sets, n is the size of the original data set, m is the number of synthetic data sets generated, b_m is the between-synthesis variance and $\bar{v}_m$ is the within-synthesis variance.

2.4.1 Multinomial logistic regression

If following the conditional approach, a series of multinomial logistic regression models can be fitted. Although not strictly a generalised linear model (GLM) owing to the multivariate response, the multinomial logistic regression model can be viewed as an extension to the logistic regression model to the case where the response has three or more levels. The standard parameterisation sets one category of the response variable - often, by default, the first - as the reference category and then has a distinct set of regression coefficients for every other level of the response relative to the reference.

When a multinomial logistic regression model involves variables with many categories, however, there are many parameters to estimate, and model-fitting algorithms can take too long to converge, thus rendering R functions such as `multinom` from the nnet package (the commonly used function in R to fit multinomial lgoistic regressions) infeasible. Note, to estimate parameter estimates, `multinom`, which builds single-layer neural networks, calls the function `optim`, which in turn uses the BFGS algorithm (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970).

2.4.1.1 Poisson-multinomial equivalence

When all covariates are categorical, a multinomial logistic regression model's MLEs can be obtained through fitting the corresponding Poisson log-linear model, as described in Lang (1996). Unlike the multinomial logistic regression model, which has a clear categorical response variable, the log-linear model treats the *counts* in the corresponding contingency table as the response.

As an example, parameter estimates and standard errors for a multinomial logistic regression of A on B, C and D are identical to those from a hierarchical Poisson log-

linear model with all two-way interactions plus the three-way interaction between B, C and D. That is, the "slope" coefficients for the multinomial logistic regression are equal to the two-way interactions involving A in the Poisson log-linear model (see Asmussen and Edwards, 1983). Thus the latter model can be fit instead of the former, which has the benefit, as discussed in subsection 2.4.2, of allowing the IPF algorithm (see subsection 2.4.2.1) to be utilised.

2.4.2 The Poisson log-linear model

When a data set is made up entirely of categorical variables, such that it can be expressed as a contingency table, it is more efficient alternative is to fit a single model to the entire data set, by modelling the counts in the contingency table.

The Poisson log-linear model assumes that cell counts follow a multiplicative structure (or, equivalently, follow a linear structure on a logarithmic scale). For example, in the independence model, the probability of observing, say, outcome a for variable A and outcome b for variable B is equal to the marginal probability of observing a *multiplied* by the marginal probability of observing b.

The counts in the contingency table are assumed to be independent and Poisson distributed. It has a representation as a Poisson GLM under a log link function, in which it is parameterised by an intercept term, main effects and interactions. Including interactions relates to preserving marginal counts: for example, the all two-way interaction model ensures all expected two-way marginal counts are equal to the observed two-way marginal counts. The sufficient statistics, that is, the data required to fit the model, are the observed marginal tables for the margins which are to be preserved. For example, in the all two-way interaction model, the sufficient

statistics are the observed two-way marginal counts between every pair of variables. In a synthesis context the synthesizer does not require access to the full individual level microdata.

To generate the synthetic counts, the synthesizer takes draws from the fitted model. Rather than the Poisson, a multinomial approach could be taken here, too, by constraining the grand total - the sum of all synthetic counts - to a fixed n. This approach is comparable to the Poisson in terms of computational time, so it essentially comes down to ideology: how important is it to preserve known grand totals, given the data are synthetic?

2.4.2.1 The use of iterative proportional fitting (IPF)

The iterative proportional fitting (IPF) algorithm, first introduced by (Deming and Stephan, 1940) during the Second World War - one of the many developments accelerated due to the necessity of war - can be used to quickly obtain fitted counts (note these "counts" are not always integers) in a log-linear model. It works by adjusting the table's expected marginal counts to match - hence preserve - observed marginal counts (see Bishop et al., 1975). Its e□ciency makes it particularly useful for large, sparse contingency tables, when algorithms used in standard software are too slow. A downside, though, is that it is less straightforward to recover the log-linear model's parameter estimates and standard errors, which has implications for generating fully synthetic data in the sense of Raghunathan et al. (2003), where parameters' estimates and standard errors are intrinsic to estimating the Bayesian posterior predictive distribution.

There are several functions that can be used to carry out IPF in R. These include

the `ipf` function from the cat package and the function `loglm` from the MASS package; note `loglm` provides a user-friendly wrapper to the `loglin` function.

2.4.3 Multinomial-Dirichlet synthesizer

Abowd and Vilhuber (2008) proposed a multinomial-Dirichlet synthesis model to produce synthetic categorical data. As with the Poisson log-linear model, this approach is undertaken at the tabular level. The main advantage of this multinomial-Dirichlet method is that it can produce synthetic data that satisfies differential privacy (see Section 2.5.4).

Let the original data's contingency table's cell probabilities be denoted by $\boldsymbol{\pi}$. Then a Dirichlet prior with concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ is placed on $\boldsymbol{\pi}$, resulting in a posterior distribution of $\boldsymbol{\pi} \mid \mathbf{a}, \boldsymbol{\alpha}$ Dirichlet$(\mathbf{a} + \boldsymbol{\alpha})$. A set of synthetic counts $\mathbf{b}$ is then obtained by simulating from:

$$\mathbf{b} \mid \boldsymbol{\pi}, \mathbf{a}, \boldsymbol{\alpha} \quad \text{Multinomial}(n, \boldsymbol{\pi})$$

$$\boldsymbol{\pi} \mid \mathbf{a}, \boldsymbol{\alpha} \quad \text{Dirichlet}(\mathbf{a} + \boldsymbol{\alpha}).$$

$$\text{i.e.,} \quad p(\mathbf{b} \mid \boldsymbol{\pi}, \mathbf{a}, \boldsymbol{\alpha}) = \frac{n_{\text{syn}}!}{\prod_i b_i!} \prod_{i=1}^{K} \pi_i^{b_i} \quad \text{and} \quad p(\boldsymbol{\pi} \mid \mathbf{a}, \boldsymbol{\alpha}) = \frac{\Gamma(n + \sum_i \alpha_i)}{\prod_i \Gamma(a_i + \alpha_i)} \prod_{i=1}^{K} \pi_i^{a_i + \alpha_i - 1}.$$

Integrating over $\boldsymbol{\pi}$ gives:

$$\begin{aligned} p(\mathbf{b} \mid \mathbf{a}, \boldsymbol{\alpha}) &= \frac{n_{\text{syn}}!\, \Gamma(n + \sum_i \alpha_i)}{\prod_i b_i!\, \Gamma(a_i + \alpha_i)} \int \prod_{i=1}^{K} \pi_i^{b_i + a_i + \alpha_i - 1} \, d\boldsymbol{\pi} \\ &= \frac{n_{\text{syn}}!\, \Gamma(n + \sum_i \alpha_i)}{\prod_i [b_i!\, \Gamma(a_i + \alpha_i)]} \quad \frac{\prod_i \Gamma(b_i + a_i + \alpha_i)}{\Gamma(n_{\text{syn}} + n + \sum_i \alpha_i)}. \end{aligned} \tag{2.3}$$

2.4.4 Latent class models

Vermunt et al. (2008) describe the use of latent class analysis (LCA) for the multiple imputation - and by extension, synthesis - of categorical data. This is a method that can be applied when log-linear analysis is unfeasible, so is especially useful for large, sparse data sets.

An individual i ($i = 1, \ldots, n$) in the data set is assumed to belong, independently, to one (and only one) of $K < 1$ latent classes. Using the notation of Si and Reiter (2013), let z_i denote i's latent class, and let π_k denote the probability of belonging to the kth latent class.

The individuals within a given latent class are assumed to follow a class-specific multinomial distribution; let ϕ_{kjl} denote the probability that an individual in latent class k observes outcome l for variable j for $l = 1, \ldots, L_j$, where L_j is the number of categories for variable j (note $\sum_{l=1}^{L_j} \phi_{kjl} = 1$). Let $\boldsymbol{\phi} = \{\phi_{kjl} : k = 1, \ldots, K, j = 1, \ldots, p, l = 1, \ldots, L_j\}$ denote the set of all cell probabilities and $\boldsymbol{\pi}$ denote the set of latent class inclusion probabilities. The finite mixture model (McLachlan and Peel, 2000) is then given as (Si and Reiter, 2013):

$$Y_{ij} \mid z_i = k, \phi \quad \text{Multinomial}(\phi_{kj1}, \phi_{kj2}, \ldots, \phi_{kjL_j}) \tag{2.4}$$

$$z_i \mid \pi \quad \text{Multinomial}(\pi_1, \pi_2, \ldots, \pi_K), \tag{2.5}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Vermunt et al. (2008) note, in this context, as the model is being used to estimate the joint distribution of the data, the parameters have little interpretative value. Hence, it does not really matter if the parameters are unidentifiable, nor if the

optimisation algorithm converges to a local - rather than a global - maximum.

Similarly, synthetic categorical data can be generated via Dirichlet process mixture of products of multinomial (DPMPM) distributions (Manrique-Vallier and Hu, 2018; Hu et al., 2014; Si and Reiter, 2013), which is the Bayesian analogue to the latent class model. As the model has a fully Bayesian specification, Markov-Chain Monte Carlo (MCMC) methods are required to obtain inferences. This can be carried out via the NPBayesImputeCat R package (Hu et al., 2021).

2.4.5 Non-parametric tree-based approaches

Classification and regression tree (CART) analysis was introduced by Breiman et al. (1984), a pioneering statistician who helped to bridge the gap between statistics and computer science (see Raper, 2020 for an interesting biography).

Reiter (2005d) proposed the use of CART to generate *partially* synthetic data - note, only partially synthetic data can be generated this way because CART replaces individuals' values - which has since become a popular method; undoubtedly this is because it is the default method in the excellent R package synthpop. Regression trees are "grown" for each variable - CART uses the approach of modelling the data variable-by-variable (see Section 2.3.1) - by repeatedly dividing the individuals in the data according to the predictor variables - usually with binary splits - until groups of similar individuals are formed; these groups are known as "leaves". For example, individuals may first be split by gender; then the males by age and the females by ethnicity, and so on, until some pre-specified "pruning" criterion is met, such as a minimum number of individuals in each leaf. As with the parametric approach, a separate tree is grown for each variable; though unlike the approach given in 2.3.1,

the tree is grown fully conditional on *all* other variables.

A natural extension to generating synthetic data sets through CART is to generate them through random forests (Breiman, 2001), as proposed by Caiola and Reiter (2010).

Drechsler and Reiter (2011) compared the utility and risk of synthetic data generated via these tree-based methods. They looked at CART, random forests and a similar method, bagging (Breiman, 1996) and demonstrated that these methods - CART, in particular - compare favourably with parametric approaches.

2.4.6 Advanced Machine Learning Techniques

Advanced machine learning techniques are becoming an increasingly popular area of research in relation to synthetic data generation. Drechsler (2010) investigated the plausibility of support vector machines (Boser et al., 1992) and, more recently, Kaloskampis (2019) looked at techniques such as generative adversarial networks (GANs) (Goodfellow et al., 2014).

These methods are complex and can be difficult to implement; for example, there is currently no software for synthesis using GANs in R. Methods need to be relatively easy to understand and implement to encourage organisations such as government departments to generate and release synthetic versions of the administrative databases that they hold.

2.5 Metrics for assessing risk in synthetic data

In most data sets, the key variables - those that can re-identify an individual - tend to be categorical. Thus re-identification, in particular, revolves around sample uniqueness (see, for example, Skinner and Elliot, 2002). It is widely accepted that re-identification is meaningless in fully synthetic data sets because individuals in the synthetic data do not necessarily pertain to individuals in the original data. However, re-identification needs to be considered in partially synthetic data. Besides, fully synthetic data sets expressed as contingency tables can be considered to be at risk of a version of re-identification if a high proportion of uniques (cell counts of one) are also unique in the original data.

2.5.1 Re-identification in partially synthetic data

Reiter and Mitra (2009) developed a technique for estimating the risk of re-identification for individuals included in partially synthetic data. They built on the work of Duncan and Lambert (1989) and Fienberg et al. (1997) in taking a probabilistic approach to re-identification. The approach relies on assumptions about intruder knowledge. For example, it may be assumed the intruder knows the synthesis model used but not its parameters; or it may be assumed the intruder knows both the synthesis models and its parameters. It is also assumed that an intruder possesses individuals' key variable values. The idea is to estimate, given this information, the probabilities with which an intruder can link an individual in the synthetic data to an individual in the original data.

This gives a risk of re-identification for each individual and it is this which is the overriding benefit of this risk metric. The ability to estimate a unique risk

probability for each individual is what Willenborg and De Waal (1996) refer to as the "gold-standard" with respect to disclosure risk metrics. Moreover, the individual risk probabilities can be aggregated to obtain an overall measure of risk, such as the expected match risk (Reiter, 2005a)).

This method's usefulness for synthetic administrative databases, however, depends on the way in which the synthesis is carried out. When the data are synthesised at the aggregated level, for example, through a Poisson log-linear model, the direct links between individuals in the original and synthetic data are lost as the data are aggregated, synthesized and disaggregated back to microdata.

2.5.2 Correct Attribution Probability (CAP)

As re-identification is neither meaningful for fully synthetic data nor for any kind of aggregated synthetic data, disclosure risk arguably revolves around attribute disclosure: the amount of information that can be gleaned about individuals' sensitive values.

Taub et al. (2018) developed the *Correct Attribution Probability* (CAP), for measuring attribute disclosure in categorical synthetic data. It compares the empirical distributions of a sensitive variable, conditional on the key variables, obtained from the original and synthetic data.

The $\text{CAP}_{\text{obs},j}$ can be defined as: among the individuals in the original data who share the same key variable values as j (belong to the same cell if the data are tabulated according to its key variables), the proportion who have the same attribute as j for the sensitive variable. Similarly, $\text{CAP}_{\text{syn},j}$ is the same proportion but calculated from the synthetic data.

Let $K_{\text{obs},j}$ denote the key variable values and $T_{\text{obs},j}$ the value of a sensitive variable, for an individual j in the original data $(j = 1, \ldots, n)$; and similarly, $K_{\text{syn},j}$ denote the same $T_{\text{syn},j}$ for an individual j in the synthetic data. Then the CAP scores are given by:

$$\text{CAP}_{\text{obs},j} = \hat{p}(T_{\text{obs},j} \mid K_{\text{obs},j}) = \frac{\sum_{r=1}^{n}[T_{\text{obs},r} = T_{\text{obs},j}, K_{\text{obs},r} = K_{\text{obs},j}]}{\sum_{r=1}^{n}[K_{\text{obs},r} = K_{\text{obs},j}]} \tag{2.6}$$

$$\text{CAP}_{\text{syn},j} = \hat{p}(T_{\text{obs},j} \mid K_{\text{obs},j}) = \frac{\sum_{r=1}^{k}[T_{\text{syn},r} = T_{\text{obs},j}, K_{\text{syn},r} = K_{\text{obs},j}]}{\sum_{r=1}^{k}[K_{\text{syn},r} = K_{\text{obs},j}]}, \tag{2.7}$$

where $[.]$ are Iverson brackets, that is, the value 1 is returned if the proposition inside the bracket is satisfied and 0 otherwise. Taub et al. (2018) recommend, in addition, calculating the baseline CAP score, which is the marginal distribution of the sensitive variable estimated from the original data, that is,

$$\text{CAP}_{\text{b},j} = p(T_{\text{obs},j}) = \frac{1}{n}\sum_{r=1}^{n}[T_{\text{obs},r} = T_{\text{obs},j}]. \tag{2.8}$$

This provides a comparison for when the key variables are entirely independent of the sensitive variable. If the CAP scores from the synthetic data are close to the baseline score, then it suggests the synthetic data are low risk; whereas, if they are close to the original data scores, then it suggests the synthetic data are high risk. As with the Reiter and Mitra (2009) method, the CAP scores can be averaged to give a single risk value for the entire synthetic data.

Hittmeir et al. (2020) highlight the link between CAP scores and the concepts of k-anonymity and L-diversity. For example, if a data set attains 2-anonymity, that is, if at least two individuals observe any observed set of key variable values, then it follows there are at least two values for the sensitive variable, which should yield

lower CAP scores. Similarly, if a data set attains 2-diversity, that is, if there are at least two distinct values for the sensitive variable among individuals with a given key variable pattern, then the maximum CAP score for any individual would be a 1/2.

Owing to synthesis, a set of key variable values observed in the original data may be unobserved in the synthetic data (or *vice versa*), leading to a CAP_{syn} score that is mathematically undefined (a denominator of zero). Taub et al. (2018) describe two approaches for dealing with this: the first is to assign a CAP of zero; the second is to exclude it. Arguably, the latter is to be preferred as it would yield a larger average CAP and hence overestimate - as opposed to underestimate - the disclosure risk of the data set. Hittmeir et al. (2020) propose a third technique, the GCAP, for dealing with this issue, which uses those individuals in the synthetic data for which all but one of the key variable values match (or all but two match and so on). The $\text{GCAP}_{\text{syn},j}$ uses Hamming distance Δ (Hamming, 1950) to consider close matches:

$$\text{GCAP}_{\text{syn},j} = \hat{p}(T_{\text{obs},j} \mid K_{\text{obs},j}) = \frac{\sum_{r=1}^{k}[T_{\text{syn},r} = T_{\text{obs},j},\ \Delta(K_{\text{syn},r}, K_{\text{obs},j}) = \rho_j]}{\sum_{r=1}^{k}[\Delta(K_{\text{syn},r}, K_{\text{obs},j})]}. \quad (2.9)$$

That is, when no exact matches exist, close matches are considered instead.

2.5.3 Bayesian estimation of disclosure risk

Reiter et al. (2014) developed a risk metric that mirrors that of Reiter and Mitra (2009) in the sense that assumptions are made about intruder knowledge. Assumptions are made as to, firstly, the portion of the original data that is known to the intruder before the synthetic data are released (denoted by A); for example, it can be assumed the intruder knows all but one of the individuals' records (an assumption inspired by the differential privacy literature). Secondly, an assumption is made about

the knowledge (denoted by S) that the intruder possesses about the synthesis models.

Let X_i denote the ith row of the original data (individual i's values) that is unknown to the intruder - hence is stochastic as far as the intruder is concerned. The method estimates the posterior distribution of X_i given the synthetic data D_{syn} and the additional information A and S, to determine how accurately the true value of X_i can estimated. That is:

$$p(X_i \mid D_{\text{syn}}, A, S) = \frac{p(D_{\text{syn}} \mid X_i, A, S) \;\; p(X_i \mid A, S)}{\sum_x p(D_{\text{syn}} \mid X_i = x, A, S) \;\; p(X_i \mid A, S)}$$
$$\propto p(D_{\text{syn}} \mid X_i, A, S) \;\; p(X_i \mid A, S). \tag{2.10}$$

More assumptions can be made if required. For example, when calculating the probabilities given in 2.10, the support of X_i can be restricted to, say, records which only differ from the true records by one element.

However, if there are either many individuals in the data set (large n) or many possible values that X_i could take - as would be the case in a large data set - then it becomes computationally challenging to compute 2.10 - particularly the likelihood component $p(D_{\text{syn}} \mid X_i, A, S)$ - even if importance sampling is used, as suggested by Reiter et al. (2014).

2.5.4 Differential Privacy

Differential privacy (DP) (Dwork et al., 2006) is a property of a perturbation (synthesis) mechanism that formally establishes how accurately a given individual's true values can be estimated, given all other $n \quad 1$ individuals' true values are known. DP is a popular area of research at the moment, partly owing to its implementation

by the US Census Bureau to protect 2020 census records.

Rinott et al. (2018) set out how di□erential privacy can be satisfied for contingency tables. In line with Rinott et al.'s notation, let $\mathbf{a} = (a_1, \ldots, a_K)$ *2 A* and $\mathbf{b} = (b_1, \ldots, b_K)$ *2 B* denote a vector of original and synthetic counts (the counts in the original and synthetic data's contingency table), respectively, where *A* and *B* denotes the range of original and synthetic counts. Moreover, $\mathbf{a}$ and $\mathbf{a}^{\prime}$ are *neighbours*, denoted by $\mathbf{a}$ $\mathbf{a}^{\prime}$, if all counts in $\mathbf{a}$ and $\mathbf{a}^{\prime}$ are identical bar one which di□ers by one; this is equivalent to saying the rows of $\mathbf{a}$ and $\mathbf{a}^{\prime}$ di□er by one if expressed in microdata format.

The ϵ-DP definition revolves around the likelihood ratio(s). A perturbation/synthesis mechanism *M* is ϵ-di□erentially private ($\epsilon > 0$) if:

$$\exp(\ \epsilon) \quad \frac{p(M(\mathbf{a}) = \mathbf{b})}{p(M(\mathbf{a}^{\prime}) = \mathbf{b})} \quad \exp(\epsilon), \quad \text{or equivalently,} \quad \left|\log \frac{p(M(\mathbf{a}) = \mathbf{b})}{p(M(\mathbf{a}^{\prime}) = \mathbf{b})}\right| \quad \epsilon, \tag{2.11}$$

8 $\mathbf{a}$ $\mathbf{a}^{\prime}$ *2 A* and *8* $\mathbf{b}$ *2 B*.

This is the definition as given in Rinott et al. (2018), which is a simplification of that given in Dwork et al. (2006) in the case that the range of *M* is discrete. For any $\mathbf{a}$ and $\mathbf{a}^{\prime}$, if this ratio is either too small or too large, then too much information can be gleaned as to which cell the target belongs, that is, the target's true characteristics can be too easily established. The definition considers all potential synthetic data sets in *B*. Thus, DP is not a risk metric for a particular synthetic data set, but rather a property of a synthesis mechanism that guarantees a certain level of privacy.

In general, then, DP guarantees that the generation of synthetic data is largely

unaffected by the presence of a particular individual, and therefore no sensitive information is revealed about this individual.

As mentioned in Section 2.5.4, differential privacy (DP) is a property of a synthesis mechanism. It is a mathematically rigorous approach to risk, that establishes the probability with which an individual's values can be estimated, given all other $n - 1$ individuals' true values are known.

Recall that $\mathbf{a} = (a_1, \ldots, a_K)$ *2 A* and $\mathbf{b} = (b_1, \ldots, b_K)$ *2 B* are arbitrary vectors of original and synthetic counts (the counts in the original and synthetic data's contingency table), respectively; and $\mathbf{a} \quad \mathbf{a}'$ means that $\mathbf{a}$ and $\mathbf{a}'$ are identical except for one count, which differs by one.

As with Quick (2021), suppose $\mathbf{a}$ and $\mathbf{a}'$ differ in their first element only and that $a_1 = a_1' + 1$, such that $||\mathbf{a} \quad \mathbf{a}'|| = 1$ ($a_i = a_i'$ for all $i \quad 2$). Then $\mathbf{a}$ represents the data held by the intruder (all individuals except the unknown target individual) and $\mathbf{a}'$ represents the "completed data", where the target has been added to one of the cells.

2.5.4.1 The multinomial-Dirichlet and DP

As mentioned in section 2.4.3, the multinomial-Dirichlet can be used to generate synthetic data that satisfies DP. Following on from the expression for $p(\mathbf{b} \mid \mathbf{a}, \boldsymbol{\alpha})$

given in (2.3):

$$\begin{aligned}\frac{p(\mathbf{b}\ j\ \mathbf{a},\boldsymbol{\alpha})}{p(\mathbf{b}\ j\ \mathbf{a}^{\emptyset},\boldsymbol{\alpha})} &= \frac{\Gamma(b_1+a_1+\alpha_1)}{\Gamma(a_1+\alpha_1)} \quad \frac{\Gamma(a_1^{\emptyset}+\alpha_1)}{\Gamma(b_1+a_1^{\emptyset}+\alpha_1)}. \\ &= \frac{\Gamma(b_1+a_1+\alpha_1)}{\Gamma(a_1+\alpha_1)} \quad \frac{\Gamma(a_1 \quad 1+\alpha_1)}{\Gamma(b_1+a_1 \quad 1+\alpha_1)} \\ &= \frac{b_1+a_1 \quad 1+\alpha_1}{a_1 \quad 1+\alpha_1} \\ &> 1.\end{aligned}$$

Now, as DP is satisfied if

$$\frac{1}{\exp(\epsilon)} \quad \frac{p(\mathbf{b}\ j\ \mathbf{a},\boldsymbol{\alpha})}{p(\mathbf{b}\ j\ \mathbf{a}^{\emptyset},\boldsymbol{\alpha})} \quad \exp(\epsilon),$$

and as $a_1 \quad 1$ and $b_1 \quad n$, this implies:

$$\frac{b_1+a_1 \quad 1+\alpha_1}{a_1 \quad 1+\alpha_1} \quad \frac{n+\alpha_1}{\alpha_1} \quad \exp(\epsilon),$$

and rearranging gives:

$$\min_i \alpha_i \quad \frac{n}{\exp(\epsilon) \quad 1}.$$

Therefore, if the smallest value in the vector of concentration parameters $\boldsymbol{\alpha}$ (see Section 2.4.3), exceeds $n/(\exp(\epsilon) \quad 1)$, ϵ-DP is satisfied.

2.6 Metrics for assessing utility in synthetic data

Purdam and Elliot (2007) describe two ways in which SDC methods, and therefore also synthetic data sets, affect data utility for an analyst. Firstly, they consider *analytical completeness*, which is the ability to undertake an analysis on the synthetic (perturbed) data; for example, analytical completeness would be lost if a particular analysis cannot even be performed (owing to the absence of certain variables, say). This is related to the concept of uncongeniality (Meng, 1994) in multiple imputation. Secondly, Purdam and Elliot (2007) consider *analytical validity* - a notion introduced by Winkler (2005) - which is the ability to draw conclusions from the synthetic (perturbed) data that are similar to those from the original data.

Snoke et al. (2018) give a detailed description of metrics for assessing analytical validity of synthetic data, specifically. They distinguish between general utility measures, which compare the joint distributions of the observed and synthetic data, such as propensity score matching, and specific utility measures, which compare results pertaining to a certain analysis, such as coverage percentages, obtained from both the observed and synthetic data. These are "specific" to the analyses undertaken; synthetic data may yield valid inferences for some analyses, but not for others. These same two types of data utility measures are referred to by Drechsler and Reiter (2009) as broad and narrow measures, respectively.

2.6.1 General utility measures

General utility metrics aim to measure the similarity between the observed and synthetic data's joint probability distributions. This makes relevant techniques outside of synthetic data that quantify divergences between probability distributions,

such as Hellinger distance and Kullback-Leibler divergence.

2.6.1.1 Distance measures

The Kullback-Leibler divergence measures how a distribution from a fitted model differs from its empirical distribution. Let $\mathbf{p} = (p_1, \ldots, p_K)$ denote the empirical distribution and $\mathbf{q} = (q_1, \ldots, q_K)$ denote the distribution derived from the fitted model. Then the Kullback-Leibler divergence, also known as the relative entropy from $\mathbf{q}$ to $\mathbf{p}$, is defined as (Burnham and Anderson, 2002, Section 2.1, p. 51):

$$I(\mathbf{p}, \mathbf{q}) := \sum_{k=1}^{K} p_k \log\left(\frac{p_k}{q_k}\right). \tag{2.12}$$

The distributions $\mathbf{p}$ and $\mathbf{q}$ can be used to denote the cell probabilities in original and synthetic categorical data, respectively. There are a few issues, however. Firstly, as the Kullback-Leibler divergence is not symmetric, that is, $I(\mathbf{p}, \mathbf{q}) \neq I(\mathbf{q}, \mathbf{p})$, a different Kullback-Leibler divergence is obtained when $\mathbf{p}$ and $\mathbf{q}$ are switched round. Secondly, the Kullback-Leibler divergence is undefined whenever any q_k equals zero, which has complications for zero cell counts.

Similarly, the Hellinger distance between $\mathbf{p}$ and $\mathbf{q}$ is defined as:

$$h(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \| \sqrt{\mathbf{p}} - \sqrt{\mathbf{q}} \| := \frac{1}{2} \sqrt{\sum_{k=1}^{K} (\sqrt{p_k} - \sqrt{q_k})^2}, \tag{2.13}$$

where $\| \cdot \|$ denotes the Euclidean norm (L^2 norm) (Jayram, 2009). Rinott et al. (2018) describe an attractive feature of Hellinger distance in relation to privacy in frequency tables. Unlike distance metrics such as the mean squared error, the loss varies with the size of the cell. For example, with the Hellinger distance, the contribution of a cell

synthesized from, say, 101 to 100 is much less than the contribution from one being synthesized from 1 to 0. These seems appropriate because the latter is likely to be more damaging in actual e□ect.

While these two distance metrics provide a general insight into the similarity between the observed and synthetic data, a low Hellinger distance, for example, does not always imply high utility. Moreover, these metrics are most e□ective when used comparatively, say, to show the rate at which utility changes as risk changes.

2.6.1.2 Propensity Score Matching

Propensity score matching (Rosenbaum and Rubin, 1983), which can assess distinguishability between di□erent kinds of observations, was first used by Woo et al. (2009) to measure utility in a privacy environment, and later formalised by Snoke et al. (2018).

To undertake propensity score matching in a synthetic data setting, an indicator variable is introduced: synthetic records (say) are assigned a value of "1" and the original records a value of "0". A model is then fitted with this newly-created variable as the response and variables in the data as covariates; interactions between variables may also be useful as covariates. Given that the response is binary, a logistic regression model is suitable here; though Snoke et al. (2018) also propose CART. The closer the fitted values from the model (the propensity scores) are to 1/2 (assuming there are an equal number of synthetic and real records), the closer the synthetic records resemble the original records.

Snoke et al. (2018) introduced the metric pMSE, the propensity score mean squared error, which can formally test whether the synthetic and original data have

the same underlying distribution. Jörg Drechsler, speaking at a synthetic data session at the 2020 Royal Statistical Society Conference, noted two downsides with propensity score matching. Firstly, although regarded as a general utility metric, it still depends on the specific model used to compute the propensity scores, hence shares, to a certain extent, the same flaws as specific utility measures; secondly, parametric propensity score models tend to favour parametric synthesis models, and likewise non-parametric models tend to favour non-parametric synthesis models. This can be seen in the second and third rows of Table 10 in Snoke et al. (2018).

Mitra and Reiter (2016) also looked at propensity scores in relation to multiple imputation. As with synthesis, multiple imputation yields $m > 1$ data sets - and hence $m > 1$ individual propensity scores. The paper looked at the best way to combine these individual propensity scores in order to create an overall propensity score. This is analogous to a situation considered in Chapter 4 of the best way to analyse $m > 1$ synthetic categorical data sets given as contingency tables: is it better to average counts first or analyse each contingency table separately and average the results?

The concept of propensity score matching is more suited to data sets with continuous variables, where there is more heterogeneity among observations. Di□erences in categorical data sets correspond to di□erences in cell counts rather than di□erences in individual observations. Unreported investigations have shown that propensity score models struggle to distinguishing between original and synthetic categorical data.

2.6.2 Specific utility measures

The analytical validity of synthetic data can be assessed for specific analyses, for example, by comparing regression coe□cient estimates obtained from models fit to both the observed and synthetic data. The general problem with specific utility measures is that the utility of synthetic data is only established with respect to a particular analysis; and is is typically unknown what analyses are to be performed on the synthetic data. To lessen this issue to a certain extent, Purdam and Elliot (2007) evaluated impact on utility over a variety of analyses.

Simulation studies can be useful to establish the repeated sampling properties of inferences obtained from synthetic data. Performance measures such as bias, mean squared error and confidence interval coverage can be calculated. Morris et al. (2019) provide an excellent overview of how a simulation study, which is essentially a scientific experiment, should be designed, coded and analysed.

2.6.2.1 Confidence interval overlap

One performance measure particularly suited to synthetic data is confidence interval overlap (Karr et al., 2006), which compares the accuracy and length of confidence intervals obtained from the original and synthetic data.

To define confidence interval overlap, let (l_r, u_r) and (l_s, u_s) denote the confidence intervals for a univariate population parameter Q, derived from the real and synthetic data, respectively. Also, let (l_i, u_i) denote the intersection of the two intervals, that is, $l_i = \max(l_r, l_s)$ and $u_i = \min(u_r, u_s)$. Then the confidence interval overlap I_Q for

Q is given as:

$$I_Q = \frac{1}{2}\left(\frac{u_i \quad l_i}{u_r \quad l_r} + \frac{u_i \quad l_i}{u_s \quad l_s}\right). \tag{2.14}$$

Therefore, I_Q is the average of two ratios: the first ratio is the length of the intersection relative to the length of the observed data confidence interval; and the second ratio is the length of the intersection relative to the length of the synthetic data interval. The main advantage of confidence interval overlap is that it is a composite measure that takes into account the length and the accuracy of a confidence interval, whereas, for example, bias only considers the accuracy of a point estimate and coverage does not account for the length of a confidence interval. On the other hand, whether these factors - accuracy and length - are weighted appropriately is open to debate. Valid confidence intervals estimated from synthetic data, that is, confidence intervals that achieve the nominal coverage, are longer than the corresponding confidence intervals estimated from the original data because synthetic data estimates are subject to the uncertainty present in the original data estimates, plus have additional uncertainty from synthesis. However, a synthetic data confidence interval, say, one that is $x\%$ narrower than the original data confidence interval - hence clearly invalid - would yield roughly the same overlap as a confidence interval that is $x\%$ wider. Moreover, either an infinitely wide or infinitely small synthetic data confidence interval would achieve an overlap of 0.5.

2.7 Generating synthetic data in R

2.7.1 The synthpop package

The synthpop R package (Nowok et al., 2016) is a popular way to generate synthetic data via either the traditional route of simulating new values from the posterior predictive distribution or the "plug-in" approach (the default). It was designed as a tool by which synthetic versions of the UK Longitudinal Studies can be made available to analysts as test data.

An attractive feature of synthpop is its ability to generate "bespoke" synthetic data. That is, the synthesizer has control over which variables are synthesized, the order in which variables are synthesized and the synthesis models used for each variable. For synthesis models, the synthesizer has a choice between at least one parametric method and a non-parametric alternative in the form of CART, which is also the default for all variable types.

The package is user-friendly in the sense that it applies combining rules on behalf of the analyst, and can implement general- and specific-utility measures to evaluate the analytical validity of the synthetic data it produces.

Focusing on categorical data, in the first version of synthpop (version 1.1-1) the parametric synthesis model option was multinomial logistic regression, fit via `multinom` function from the nnet package. A newer version of the package (version 1.5-0), however, allows log-linear models to be fitted via IPF, which is undertaken by the `Ipfp` routine from the mipfp package. While this new addition makes synthpop more suited to large categorical data sets such as administrative databases, the post-synthesis evaluations are limited, with pMSE unsuited to categorical data.

2.7.2 The simPop package

The simPop R package can be used to construct synthetic populations. Three approaches are considered; one of these is the fully synthetic data approach described by Raghunathan et al. (2003); the other two approaches - synthetic reconstruction (Beckman et al., 1996) and combinatorial optimisation (CO) - involve *constructing* synthetic populations and are beyond the scope of this thesis.

2.8 The challenges of synthesizing large administrative databases

There are, of course, many challenges that can arise when synthesizing large categorical data sets such as administrative databases. In short, there are too many challenges to cover comprehensively in this thesis. The following are those that have been addressed, through the development of a synthesis approach that uses saturated models.

2.8.1 Large, sparse contingency tables

When presented in microdata format, large categorical data sets have many rows (large n) and often many columns (large p). For categorical data sets expressed in tabular format, there is a subtle difference as to what constitutes a "large" data set. From a practical perspective, the size of the data are no longer governed by rows and columns - but by cells. Neither the number of individuals n nor - to a certain extent - the number of variables p, affects the table's dimensions. Instead, the dimensions are determined by the number of categories across variables. The number of categories

(cells) is likely to be closely related to p, though; typically a larger p would suggest a greater number of cells.

Many existing synthetic data methods - for example, CART and the latent class methods - have been tailored to survey data sets and are applied at the microdata level. When n is large, the central processing unit (CPU) time for these methods can increase greatly, rendering them infeasible. It is more e□cient to synthesize large data sets (large n) at the tabular level. Having a large n in this instance is, in fact, beneficial as it likely reduces the sparsity of the table, that is, it reduces the number of zero cells (see the next subsection), which, in turn, simplifies the model-fitting procedure. (Log-linear models can be hampered by non-existence and non-identifiability issues when certain patterns of zero cells are present; see Fienberg and Rinaldo (2012)).

Yet, when contingency tables are large and sparse, even fitting models at the tabular level can be challenging. The e□ectiveness of algorithms at fitting models in large data sets can be considered in a relative manner: for example, the IPF algorithm is more e□ective than the BFGS algorithm, but there comes a point in turn at which IPF struggles. This is one of the motivating factors behind the use of saturated models proposed in this thesis, which, of course, do not require the use of model-fitting algorithms.

The CPU time not only applies to the synthesis itself, but also to subsequent post-synthesis evaluations. Utility measures such as propensity score matching and risk measures such as those given by Reiter et al. (2014) can be as time-consuming as the synthesis itself.

2.8.2 Random and structural zeros

Categorical data expressed as a contingency table leads to the problem of distinguishing between *random* zeros (also known as *sampling* zeros) and *structural* zeros in the original data. Random zeros are cell counts of zero (zero cell) that arise through chance; that is, an individual with a given set of characteristics could have been observed but was not in the data. Structural zeros arise because a given set of characteristics is inherently not possible, for example, a child aged five attending a secondary school (which are typically for children aged eleven upwards).

Administrative databases and their population-like qualities warrant special consideration with respect to random and structural zeros. It can be argued that any zero count in a finite population must be a structural zero. This issue can be overcome by considering the data as a sample from a super- or future-population.

There may also be ambiguity as to whether a zero cell is a random or structural zero. A certain combination of values may be improbable but not impossible. For example, it could be argued that a child aged five attending a secondary school *is* possible in special circumstances. It is, therefore, not always straightforward to identify a data set's structural zeros, especially when there are many cells.

2.8.3 No defined sampling frame

Unlike survey and census data, the population from which administrative data originate is not always well-defined. This has complications for generating - and subsequently obtaining inferences from - synthetic administrative data.

The NHS Patient Register, for example, which holds records for every person ever registered with a General Practitioner (GP) in England and Wales since the NHS was

founded in July 1948, has several issues when used for statistical analysis (O□ce for National Statistics, 2016). These issues relate to the target population (the people of England and Wales). Firstly, there is under-coverage because some people have never registered with a GP (the hidden portion of the population); and over-coverage because, though e□ort is taken to clean the data, inevitably there will be people included who have either died or migrated. Secondly, the database has been built up over many years; but the population of England and Wales is not closed, and changes within the population since 1948 are not captured. Similar issues regarding populations exist in other administrative databases; for example, the English School Census held by the Department for Education, includes school pupils who attend state schools, but excludes those who attend privately-funded schools.

A related issue is the assumption that observations (individuals, units) are independent - an assumption intrinsic to the majority of statistical analyses - may be less realistic in administrative databases. As an example, in the English School Census, the presence of one siblings is likely to suggest the presence of another sibling, as parents commonly send their children to the same school. Similarly, administrative databases are more likely than traditional data sources to include multiple records for the same individual; this may occur, for example, if a child switches school during the school year.

When the population in question cannot be clearly defined, it restricts the type of synthetic data that can be produced. It is not possible to generate fully synthetic data in the sense of Raghunathan et al. (2003), which requires the generation of a synthetic population. A partially synthetic data approach is still valid, though, as this involves replacing the existing values and does not require assumptions about

the underlying population. While questions would remain about how to accurately obtain inferences from synthetic administrative data - how to obtain estimates for uncertainty, for example - this is analogous to the problem of obtaining inferences from administrative data more generally, for which the challenges are well-documented (see Hand, 2018).

Finally, risk evaluations often make implicit assumptions about the underlying population. For example, these often revolve around estimating the probability that an individual who is unique in the sample is also unique in the population (Skinner et al., 1994). These notions of sample uniqueness and population uniqueness become less clear when dealing with administrative data.

2.9 The developments of this thesis

The term "administrative database" encompass a wide variety of data sets. Each database is unique in the challenge it poses from a synthesis perspective. The methodological developments to generating synthetic data proposed in this thesis have addressed the following important challenges:

1. The synthesis of "big" data sets. The most obvious difference between administrative databases and survey data sets is their size. The difficulty in fitting models - and also in simulating from such models - calls for an alternative solution. It is worth noting that the ability of model-fitting algorithms lies on a scale. Although some algorithms are more capable than others with dealing with large data sets, there usually comes a point when any model breaks down. An exception, however, are saturated models. These do not require the use of model-fitting algorithms to obtain fitted values, and so can be "fit" to any

size of data set. Chapter 3 sets out the use of saturated models for synthesizing categorical data; and Chapter 4 considers the case of generating $m > 1$ synthetic data sets, as is often required.

2. Evaluating risk and utility *a priori*. A secondary consideration in the synthesis of large data sets is the ability to carry out post-synthesis checks to evaluate risk and utility. This thesis has explored the possibility of evaluating risk and utility prior to synthesis and tailoring the synthesis towards a particular risk or utility metric. While this does not entirely remove the need for post-synthesis evaluations, it makes the process more e□cient and user-friendly for the synthesizer. They do need to be continually re-running the synthesis to obtain an insight into risk and utility. Chapter 3 sets out the notion of this *a priori* approach to synthesis, which is further explored in Chapters 4 and 5.

3. Simultaneous use of under- and over-dispersion when synthesizing categorical data. The thesis has focused on ideal distributional properties of synthesis models for tabular data. Existing approaches for tabular data tend to use the Poisson distribution. However, the Poisson distribution's variance increases with the mean, which causes more noise to be applied to larger counts than smaller counts; yet larger counts are lower risk, thus require *less* noise than smaller counts. Chapter 5 explores the use of the discretized gamma family distribution for synthesis; this is a distribution which is over-dispersed for large counts and under-dispersed for small counts, and thus particularly suited to synthesis.

Chapter 3

Using saturated count models to synthesize categorical data

As a means of protecting privacy, the philosophy of synthetic data methods typically differ from that of other SDC methods. Synthetic data methods create new data; SDC methods perturb or suppress the original data. The synthesis mechanism introduced here relies on the notion that, more philosophically in line with other SDC methods, synthetic data can be generated by adding noise to the original data.

3.1 The motivation for using saturated models

Usually, the purpose of statistical modelling is either inferential, for instance, to estimate a population parameter such as a regression coefficient, or predictive. In these instances, too many model parameters yield uninformative estimates from an inferential perspective, and a model that is too rigid from a prediction perspective. Hence there is a need to strive for parsimony. But modelling with the intention of

generating synthetic data is neither inferential nor predictive. Instead, the objective is to capture the underlying distribution governing the original data so that the properties of the data are preserved in the synthetic data - the model itself is not of interest.

As with imputation models, when a synthesis model is less complex than the analyst's model subsequently fitted to the synthetic data, then the analyst's model and the synthesis model are said to be *uncongenial* (Meng, 1994). For example, suppose an all two-way interaction log-linear model is used for synthesis, and that an analyst later fits a model to the synthetic data that includes a three-way interaction. Then the synthetic data would not support this analysis as the three-way associations were irrevocably lost at the modelling step. There is a loss of analytical completeness. Therefore, to preserve relationships, over- is preferable to under-fitting in synthesis models, to capture as many relationships as possible. Using saturated models preserves all relationships in the data, thus avoiding unnecessarily distorting joint distributions when modelling.

Secondly, the time taken to undertake the synthesis computationally is substantially reduced. The computational time comprises the time taken to fit the model plus the time taken to draw the synthetic values - and using saturated models eliminates the former.

Thirdly, the synthetic counts' expected values are equal to the original counts, due to the absence of smoothing. This allows expected values of risk and utility metrics to be derived analytically, as demonstrated later in the thesis, which means that the synthesizer knows *a priori*, that is, prior to generating the synthetic data, what the risk and utility of the resulting synthetic data are likely to be.

3.2 The synthesis mechanism

This mechanism fits a saturated count model to the original data's contingency table. Let $f_1, f_2, \ldots, f_K$ denote the observed counts; then the corresponding synthetic counts $f_1^{\text{syn}}, f_2^{\text{syn}}, \ldots, f_K^{\text{syn}}$ are generated by simulating from:

$$f_i^{\text{syn}} \quad X_i \quad i = 1, 2, \ldots, K \tag{3.1}$$

where X_i is a count distribution with mean f_i. The method, then, e□ectively assumes the synthetic counts are stochastic with expectations equal to the original counts.

3.2.1 The Poisson model

The Poisson distribution is the most natural choice for modelling the counts (the most natural choice for X in 3.1). Besides, statistical models often assume that individuals' observations are independent. In microdata format, this translates into assuming the rows of the microdata set are independent; as a contingency table, it translates into assuming cell counts are Poisson distributed, that is, it assumes each individual belongs to one - and only one - cell, independently.

Suppose then that a saturated Poisson model is fit; then each synthetic count f_i^{syn} $(i = 1, \ldots, K)$ is obtained by simulating from:

$$f_i^{\text{syn}} \mid \mu_i \quad \text{Poisson}(\mu_i)$$
$$\text{with} \quad \mu_i = f_i, \tag{3.2}$$

where f_i is the corresponding original count.

Properties of the synthesis mechanism then relate directly to properties of the Poisson distribution. For example, the Poisson distribution's probability mass function (PMF) gives the probability that an arbitrary synthetic count f^{syn} equals N_2, given that the original count f equals N_1:

$$p(f^{\text{syn}} = N_2 \mid f = N_1) = \frac{\exp(-N_1) N_1^{N_2}}{N_2!},$$

where N_1 and N_2 are non-negative integers.

Before moving on to two-parameter count distributions, there are alternative one-parameter count distributions that can be used rather than the Poisson. Another well-known one-parameter distribution is the geometric, which is not usually associated with modelling, as it tends to be parameterized probabilistically: the number of successes until the first failure (or *vice versa*). Yet, as with the Poisson, it can be parameterized in terms of the mean; see Rigby et al. (2019). When parameterized in this way, a geometric random variable with mean μ has variance $1 + \mu$; hence the variance is always greater than the Poisson, and which can be useful if the Poisson uncertainty is insufficient to mask the true original counts.

Incidentally, see Rigby et al. (2019) for more about all the distributions mentioned henceforth in the thesis, including their parameterizations. This book is written by the creators of the GAMLSS (Generalized Additive Models for Location, Scale and Shape) approach. The distributions used in the GAMLSS framework are particularly useful here, as explored later in the thesis.

3.2.2 Two-parameter count distribution models

The Poisson variability (variance equal to the mean), may not provide sufficient

protection to at-risk original counts. The variability can be increased - without introducing bias - by using an overdispersed count distribution in place of the Poisson. The two distributions considered here are the negative binomial (NBI) and Poisson-inverse Gaussian (PIG) distributions. Both are continuously mixed Poisson distributions, where the gamma and inverse-Gaussian are the respective mixing distributions.

As well as the mean parameter (denoted by μ), these two-parameter distributions have a shape parameter (when using Rigby et al.'s parameterizations), denoted by σ, that controls the variance. The notion is that σ is used as a tuning parameter: the synthesizer sets σ to control the noise applied to the original counts. As demonstrated later on, σ can be tuned according to a specific risk or utility metric.

The ith synthetic count f_i^{syn} $(i = 1, \ldots, K)$ is then generated by simulating from either:

$$f_i^{\text{syn}} \mid \mu_i, \sigma \quad \text{NBI}(\mu_i, \sigma) \quad \text{or} \quad f_i^{\text{syn}} \mid \mu_i, \sigma \quad \text{PIG}(\mu_i, \sigma)$$
$$\text{with} \quad \mu_i = f_i.$$

As with the Poisson, the NBI and PIG's PMFs gives the probability that an arbitrary synthetic count f^{syn} equals N_2, given the original count f equals N_1. Firstly, for the NBI:

$$p(f^{\text{syn}} = N_2 \mid f = N_1, \sigma) = \frac{\Gamma(N_2 + 1/\sigma)}{\Gamma(N_2 + 1)\ \Gamma(1/\sigma)} \left(\frac{\sigma N_1}{1 + \sigma N_1}\right)^{N_2} \left(\frac{1}{1 + \sigma N_1}\right)^{1/\sigma}, \tag{3.3}$$

and similarly for the PIG:

$$p(f^{\text{syn}} = N_2 \mid f = N_1, \sigma) = \left(\frac{2c}{\pi}\right)^{1/2} \frac{N_1^{N_2} \exp(1/\sigma) K_{N_2 - 1/2}(c)}{(c\sigma)^{N_2} N_2!}, \tag{3.4}$$

$$\text{where} \quad c^2 = \frac{1}{\sigma^2} + \frac{2N_1}{\sigma} \quad \text{and} \quad K_\lambda(t) = \frac{1}{2} \int_0^1 x^{\lambda - 1} \exp\left\{ -\frac{1}{2} t(x + x^{-1}) \right\} dx$$

is the modified Bessel function of the third kind. Although the NBI and PIG distributions have different PMFs (and differing higher moments), the first two moments, the mean and variance, are identical:

$$\mathsf{E}[f^{\text{syn}} \mid f = N_1, \sigma] = N_1, \text{ and } \mathsf{Var}[f^{\text{syn}} \mid f = N_1, \sigma] = N_1 + \sigma N_1^2. \tag{3.5}$$

This clearly shows how the parameter σ controls the variance of the synthetic counts. As with the one-parameter distributions, there is an array of two-parameter count distributions that can be used here. These include the double Poisson, generalized Poisson and Waring distributions; the latter is from the geometric family and can be derived by mixing the geometric with the beta.

3.2.3 Additive smoothing to deal with zero counts (introducing α)

There is a downside with using saturated models that needs addressing: as synthetic counts corresponding to original counts of zero have a mean of zero, original counts of zero cells are always synthesized to zero, resulting in too many zeros in the synthetic data. That is, synthetic counts of zero comprise all original counts of zero, plus non-zero original counts that become zero through simulation. An excess of zeros can

a□ect the risk and utility of the synthetic data.

With regards to risk, the issue is not so much with the zeros themselves, which are relatively low risk, but with what can be deduced from the non-zero cell counts. It follows that any non-zero synthetic count must have originated from a non-zero original count. So, given a non-zero synthetic count, an attacker can ascertain that the original count was *at least one*.

The addition of a pseudocount $\alpha > 0$ (which despite its name is not typically an integer) to all random zeros in the original data (structural zeros should remain zero) opens the possibility at least that zero original counts are synthesized to non-zero. For example, when the Poisson model is used and $\alpha > 0$ is added, the probability that a random zero $f = 0$ is synthesized to $f^{\text{syn}} = N_2$ is:

$$p(f^{\text{syn}} = N_2 \mid f = 0,\ \alpha) = \frac{\exp(-\alpha)\ \alpha^{N_2}}{N_2!}.$$

The `syn.catall` function in synthpop (Nowok et al., 2016), which uses a saturated multinomial to produce synthetic data, allows the synthesizer to specify a Dirichlet prior; this is analogous to the addition of a pseudocount proposed here. As with this method, α can be added to *every* original count instead of just zeros, which may be altogether smoother and result in less distortion between variables' relationships.

Another alternative, as suggested by a reviewer of Jackson et al. (2022, Early view), is the use of the pseudo-Bayes estimator. The addition of the same $\alpha > 0$ to all random zeros essentially assumes all random zeros among the original counts are equally likely to be non-zero. However, it can be argued that some random zeros in the original data are more (or less) likely to be non-zeros than others; for example, some zeros pertain to higher order marginal counts that are also zero. This can be accounted for,

to a certain extent, by smoothing the original counts through fitting, for example, an all two-way interaction log-linear model. But the benefits of using saturated models would be lost. The pseudo-Bayes estimator, as presented in Chapter 12 of Bishop et al. (1975), provides an alternative to adding constant α. A set of prior cell probabilities (denoted by $\boldsymbol{\lambda}$) can be selected using, for example, external information. Based on these prior probabilities, the original counts can be re-weighted to provide a set of adjusted counts. A saturated model can then be applied as before, but using instead these adjusted counts. Hence, while the original counts are smoothed, they are not smoothed through modelling decisions (setting interactions to zero), but through the choice of $\boldsymbol{\lambda}$. This means, however, that the fundamental challenge is just transferred from choosing α to choosing $\boldsymbol{\lambda}$. Neither does this strategy account for structural zeros, so the approach would need to be adapted. In general, this pseudo-Bayes estimator involves further consideration and is a substantial research question in its own right.

3.2.4 Properties of the synthesis mechanism

While most count distributions - including the Poisson, NBI and PIG - are degenerate when the mean is zero, practically this does not a□ect the method: whenever an original count is zero, the synthetic count is also zero. Conveniently, this feature naturally accounts for structural zeros, which rightly remain zero. Random zeros, on the other hand, do need to be accounted for; hence the proposed solution in Section 3.2.3 (the α parameter).

This synthesis mechanism produces completely synthesized data using the terminology of Raab et al. (2016). However, somewhat confusingly, as a synthetic population is not created, the data are partially synthetic in the sense of Reiter (2003),

rather than fully synthetic in the sense of Raghunathan et al. (2003); incidentally, Drechsler (2018) seeks to clear up some of the confusion surrounding the term "fully synthetic" data sets. Finally, the synthesis is via the "plug-in approach" (Reiter and Kinney, 2012), that is, the Bayesian posterior predictive distribution is not used: synthetic counts are simulated directly from the fitted model.

The "sample" size of the synthetic data n_{syn} (the grand total), the sum of the synthetic counts, is stochastic; and, as these counts are independent random variables whose means sum to n, n_{syn} also has mean n (and is Poisson distributed if the Poisson is used). Yet it need not be the case that $E(n_{\text{syn}}) = n$. As Raab et al. (2016) demonstrate, for completely synthesized data, valid inferences can still be obtained when $n_{\text{syn}} \neq n$. As n_{syn} increases, the cell proportions in any generated synthetic data set - obtained by dividing each count by n_{syn} - tends toward the proportions in the original data.

Of course, in addition to n_{syn}, all synthetic counts have an unbiasedness property: their expectations are equal to the original counts. As explained later on, this allows expected properties of the synthetic data to be obtained.

The method, in effect, removes the unwelcome uncertainty that arises through model choice and then injects it accordingly by increasing σ. There is effectively only one source of variability in the method: the uncertainty that arises through simulation. This contrasts with almost every other synthesis method. For example, when an unsaturated log-linear model is used for synthesis, the synthetic counts' means are estimated thus have uncertainty attached. Minimizing the sources of uncertainty allows the synthesizer to have greater control of the variability in the mechanism.

Most synthesis methods require non-trivial modelling decisions, that fall on the

part of the synthesizer. There are advantages to the development of methods that relieve the burden placed on the synthesizer, who may not be a trained modeller. In a similar way, the modelling aspect can be time-consuming, especially in large data sets. This method generates synthetic data quickly, irrespective of the data's size.

3.3 Satisfying risk and utility metrics *a priori*

The upshot of this synthesis mechanism is that there are two tuning parameters, σ and α, which can be tuned according to a certain risk or utility metric. As saturated models are used, the expectations and variances of synthetic counts of size k are all identical. This allows many metrics to be expressed analytically.

3.3.1 The τ metrics

The following τ metrics are an example of a simple set of metrics that can be represented analytically in terms of σ and α:

$\tau_1(k) :=$ The proportion of counts of size k in the synthetic data.

$\tau_2(k) :=$ The proportion of counts of size k in the original data.

$\tau_3(k) :=$ The proportion of original counts of size k that are synthesized to k.

$\tau_4(k) :=$ The proportion of synthetic counts of k that originated from a count of k.

The expected values of these τ metrics can be derived analytically, as demonstrated below for when the Poisson model is used for synthesis. Although first note that, as structural zeros should not be involved in the synthesis, when $k = 0$ the τ metrics

refer to random zeros, only; for example, $\tau_1(0)$ is the proportion of *random* zeros in the synthetic data.

3.3.1.1 The metrics τ_1 and τ_2

The metric $\tau_1(k)$ is the proportion of cells of size k in the synthetic data, that is,

$$\begin{aligned}\tau_1(k) = p(f^{\text{syn}} = k) &= \sum_{j=0}^{1} p(f^{\text{syn}} = k \mid f = j)\ \ p(f = j) \quad k = 0, 1, 2, \ldots \\ &= \frac{\exp(\ \ \alpha)\alpha^k}{k!}\ \ \tau_2(0) + \sum_{j=1}^{1} \frac{\exp(\ \ j)j^k}{k!}\ \ \tau_2(j),\end{aligned} \tag{3.6}$$

where $\tau_2(k)$ is the proportion of original counts of size k, that is,

$$\tau_2(k) = p(f = k) \quad k = 0, 1, 2, \ldots\ . \tag{3.7}$$

3.3.1.2 The metric τ_3

The metric $\tau_3(k)$ is the proportion of original counts of k that are synthesized to k:

$$\begin{aligned}\tau_3(k) = p(f^{\text{syn}} = k \mid f = k) &\quad k = 0, 1, 2, \ldots \\ &= \begin{cases} \exp(\ \ \alpha)\alpha^k/k! & \text{if } k = 0 \\ \exp(\ \ k)k^k/k! & \text{if } k \quad 1 \end{cases}\end{aligned} \tag{3.8}$$

3.3.1.3 The metric τ_4

The metric $\tau_4(k)$ is the proportion of synthetic counts of k that originated from a count of k:

$$\tau_4(k) = p(f = k | f^{\text{syn}} = k) \quad k = 0, 1, 2, \ldots \tag{3.9}$$

The metric $\tau_4(k)$ can be expressed in terms of the other τ metrics:

$$\tau_4(k) = p(f = k | f^{\text{syn}} = k) = \frac{p(f^{\text{syn}} = k | f = k) \quad p(f = k)}{p(f^{\text{syn}} = k)} = \frac{\tau_3(k) \quad \tau_2(k)}{\tau_1(k)}$$

$$= \begin{cases} \exp(\quad \alpha)\alpha^k \quad \tau_2(0) \Big/ \left(\exp(\quad \alpha)\alpha^k \quad \tau_2(0) + \sum_{j=1}^{1} \exp(\quad j)j^k \quad \tau_2(j) \right) & \text{if } k = 0 \\ \exp(\quad k)k^k \quad \tau_2(k) \Big/ \left(\exp(\quad \alpha)\alpha^k \quad \tau_2(0) + \sum_{j=1}^{1} \exp(\quad j)j^k \quad \tau_2(j) \right) & \text{if } k \quad 1. \end{cases}$$

The infinite sum need not be calculated in the above expressions. Instead, it su□ces to take the sum over the set:

$$R = \{j \; | \; \tau_2(j) > 0\}, \tag{3.10}$$

that is, the set of cell sizes that are observed in the original data. Note, the contribution from any integer not in R is zero in $\tau_1(k)$ and $\tau_4(k)$.

While the variance of the τ metrics can be approximated analytically, empirical results have shown that the variance is typically negligible, especially in large data sets with many cells.

The metrics τ_1, τ_2 and τ_4, unlike τ_3, are conditional on the distribution of original counts, that is, the proportion of zeros, ones, twos, *et cetera*. To illustrate, suppose all

Table 3.1: The metrics τ_1, τ_2, τ_3 and τ_4 and whether they depend on the original data, the synthesis model or both.

	The original data	The synthesis model
τ_1	✓	✓
τ_2	✓	
τ_3		✓
τ_4	✓	✓

original counts are equal to one and that the Poisson distribution is used to generate synthetic counts. Then $\tau_3(1) = \exp(\quad 1)$, which is given by the Poisson's probability mass function; and $\tau_4(1) = 1$ because a synthetic count of one must have originated from a count of ones. Now, when the original counts comprise a range of non-zero values, then $\tau_3(1) = \exp(\quad 1)$ remains unchanged, but $\tau_4(1) \quad 1$ because a synthetic count of one could have originated from any non-zero original count. Similarly, τ_1, τ_3 and τ_4 - but not τ_2 - depend on the synthesis model (see Table 3.1).

As uniques - cell counts of one - are often considered to be most at risk of disclosure, an important value with respect to risk is $\tau_4(1)$: the proportion of uniques in the synthetic data which were also unique in the original data. This is arguably more important than $\tau_3(1)$: the proportion of uniques in the original data which are also unique in the synthetic data. This is because the former assumes knowledge of the synthetic data, which an attacker has access to; whereas the latter assumes knowledge of the original data, which an attacker cannot access.

3.3.1.4 Using σ and α to tune the τ metrics

As an example, suppose the synthesizer wishes to reduce $\tau_4(1)$, the τ metric most associated with risk. There are two ways in which they can do so. The first way is to increase σ, which increases the variance of the synthetic counts. The second way

is to increase α because original counts of zero with α added are much more likely to be synthesized to one than to any other non-zero value; for example, when $\alpha = 0.1$ and the Poisson model is used, a zero count is exactly twenty times more likely to be synthesized to one than two, which increases the number of uniques in the synthetic data and thereby decreases $\tau_4(1)$. In most instances it will be preferable to reduce $\tau_4(1)$ by increasing σ rather than α, as σ adds noise (variance) to synthetic counts whereas α adds bias.

Below, $\tau_4(1)$ is expressed analytically for when the NBI or PIG distributions are used. This shows the relationship between σ and α and $\tau_4(1)$. These expressions can be easily derived from their respective PMFs, as demonstrated earlier for the Poisson. Firstly, when the NBI is used:

$$\tau_4(1) = \frac{\tau_2(1)}{(1+\sigma)^{1+1/\sigma}} \Bigg/ \left(\frac{\alpha \quad \tau_2(0)}{(1+\alpha\sigma)^{1+1/\sigma}} + \sum_{j=1}^{1} \frac{j \quad \tau_2(j)}{(1+j\sigma)^{1+1/\sigma}} \right),$$

and secondly when the PIG is used:

$$\tau_4(1) = \left[\frac{\exp(\quad c_1)}{c_1} \quad \tau_2(1) \right] \Bigg/ \left[\frac{\alpha \quad \exp(\quad c_\alpha)}{c_\alpha} \quad \tau_2(0) + \sum_{j=1}^{1} \frac{j \quad \exp(\quad c_j)}{c_j} \quad \tau_2(j) \right],$$

where $c_j^2 = \sigma^{\ 2} + 2j\sigma \quad 1$; and note that, when $\lambda = 1/2$, the Bessel function of the third kind has a closed formed expression: $K_{1/2}(t) = (\pi/2t)^{1/2}\exp(\quad t)$.

Alternatively, a synthesizer may wish that the synthetic data contain the same proportion of zeros as the original data, that is, $\tau_1(0) = \tau_2(0)$. When $\alpha = 0$, the inequality $\tau_1(0) \quad \tau_2(0)$ holds; but, as α increases, the expected di□erence between $\tau_1(0)$ and $\tau_2(0)$ narrows (until it eventually reverses).

It is possible to derive the α required such that $\tau_1(0) = \tau_2(0)$. That is, when the

NBI is used for synthesis, for a given σ , the α required is:

$$\begin{aligned}
\tau_2(0) &= \tau_1(0) \\
\tau_2(0) &= \frac{1}{(1+\sigma\ \alpha\ )^{1/\sigma}}\ \ \tau_2(0) + \sum_{j=1}^{1} \frac{1}{(1+\sigma\ j)^{1/\sigma}}\ \ \tau_2(j) \\
(\ )\ (1+\sigma\ \alpha\ )^{1/\sigma} &= \left(1 \quad \frac{1}{\tau_2(0)} \sum_{j=1}^{1} \frac{1}{(1+\sigma\ j)^{1/\sigma}}\ \ \tau_2(j)\right)^{1} \\
(\ )\ \alpha &= \frac{1}{\sigma}\left[\left(1 \quad \frac{1}{\tau_2(0)} \sum_{j=1}^{1} \frac{1}{(1+\sigma\ j)^{1/\sigma}}\ \ \tau_2(j)\right)^{\sigma} \quad 1\right];
\end{aligned}$$

and similarly for the PIG:

$$\begin{aligned}
\tau_2(0) &= \tau_1(0) \\
\tau_2(0) &= \exp(1/\sigma \quad c_\alpha)\ \ \tau_2(0) + \sum_{j=1}^{1} \exp(1/\sigma \quad c_j)\ \ \tau_2(j) \\
(\ )\ c_\alpha &= \frac{1}{\sigma} \quad \log\left(1 \quad \frac{1}{\tau_1(0)} \sum_{j=1}^{1} \exp(1/\sigma \quad c_j)\ \ \tau_2(j)\right) \\
(\ )\ \alpha &= \frac{1}{2}\left\{\sigma\ \left[\frac{1}{\sigma} \quad \log\left(1 \quad \frac{1}{\tau_1(0)} \sum_{j=1}^{1} \exp(1/\sigma \quad c_j)\ \ \tau_2(j)\right)\right]^2 \quad \frac{1}{\sigma}\right\}.
\end{aligned}$$

3.3.2 Loss functions

As mentioned in Section 2.6, Snoke et al. (2018) distinguish between general utility measures, which compare the joint distributions of the observed and synthetic data, and specific utility measures, which compare results pertaining to a certain analysis.

As with the τ metrics, utility metrics, particularly general utility metrics, can be obtained - or at least approximated - analytically *a priori*.

Loss functions are widely used throughout statistics, for example, model param-

eters are typically estimated through minimising a loss function. A similar notion holds when estimating utility: utility can be maximized through minimizing some loss function.

3.3.2.1 The mean squared error loss function

The *quadratic loss function* (*squared error*) is arguably the most well-known; its expectation (known as the *mean squared error*), is given as:

$$\begin{aligned} L_1 &= E\left[\sum_{k=1}^{K}(f_k \quad f_k^{\text{syn}})^2\right] \\ &= \sum_{k=1}^{K}\left\{f_k^2 \quad 2f_k E\left(f_k^{\text{syn}}\right) + E\left[\left(f_k^{\text{syn}}\right)^2\right]\right\} \qquad (3.11) \\ &= \sum_{k=1}^{K}\left[f_k^2 \quad 2f_k E\left(f_k^{\text{syn}}\right) + \left[E\left(f_k^{\text{syn}}\right)\right]^2 + \text{Var}\left(f_k^{\text{syn}}\right)\right] \qquad (3.12) \\ &= \sum_{k=1}^{K}\left[\text{Var}\left(f_k^{\text{syn}}\right)\right]. \qquad (3.13) \end{aligned}$$

Note that equation (3.11) follows by the linearity of expectation, (3.12) follows by the definition of variance and (3.13) from the fact that, when $\alpha = 0$, $E(f_k^{\text{syn}}) = f_k$. Substituting in the NBI and PIG's variance function gives,

$$L_1(\sigma, \nu, m) \quad \sum_{k=1}^{K} f_k + \sigma f_k^2. \qquad (3.14)$$

Now, L_1 can also be expressed in terms of the proportion of zeros, ones, twos, etc. - $\tau_2(0), \tau_2(1), \tau_2(2), \ldots$ - and the number of cells K,

$$L_1(\sigma, \nu, m) \quad \sum_{j2R} \left[K \quad \tau_2(j) \quad (j + \sigma j^2)\right]. \tag{3.15}$$

where R is as defined in (3.10).

3.3.2.2 The 0-1 loss function

For an arbitrary original count f and its corresponding synthetic count f^{syn}, the expected value of the *0-1 loss function*, denoted by L_2, is given as

$$L_2 = E[I(f \neq f^{syn})], \tag{3.16}$$

where I is an indicator function. Now, L_2 is closely related to the metric $\tau_3(k)$ because for an original count of k (that is, $f = k$), $L_2 = p(f \neq f^{syn}) = 1 \quad p(f = f^{syn}) = 1 \quad \tau_3(k)$. This reinforces the close links - and the inherent trade-o - between risk and utility, because while maximum utility is attained when a loss function is minimized, maximum privacy (lowest risk) is attained when a loss function is maximized.

3.3.3 Marginal synthetic counts

As original counts are synthesized at the lowest level of aggregation, when saturated count models are used for synthesis, all marginal counts in the synthetic data's contingency table - that relate to sum of synthetic counts - are stochastic. Their means are equal to the corresponding marginal counts in the original data (they are unbiased); their variances depend on the count distribution used for synthesis.

Applying more variance - more noise - when synthesizing the individual cells leads, unsurprisingly, to greater variance in the marginal counts. Take, for instance, the synthetic data sample size n_{syn}, which is the sum of all K cells in the table (the grand total), that is, $n_{\text{syn}} = \sum_{k=1}^{K} f_k^{\text{syn}}$. The mean of n_{syn} is equal to n and, as counts are synthesized independently, its variance is equal to:

$$\text{Var}(n_{\text{syn}}) = \text{Var}\left(\sum_{k=1}^{K} f_k^{\text{syn}}\right) = \sum_{k=1}^{K} \text{Var}\left(f_k^{\text{syn}}\right) \quad \sum_{k=1}^{K} f_k + \sigma f_k^2,$$

which is exactly equal to the expression for L_1 above. As above, this can alternatively be expressed in terms of the $\tau_2(k)$ values,

$$\text{Var}(n_{\text{syn}}) \quad \sum_{j2R} K \quad \tau_2(j) \quad (j + \sigma j^2),$$

and where, again, it su□ces to take the sum over R.

In large tables (large K), the central limit theorem establishes that n_{syn} is normally distributed with mean and variance given above. This can be used, say, to calculate the probability that n_{syn} will be within d of n:

$$\begin{aligned} p(jn_{\text{syn}} \quad nj < d) &= p(n_{\text{syn}} < n + d) \quad p(n_{\text{syn}} < n \quad d) \\ &\Phi\left[\frac{(n+d) \quad n}{\left(\sum_{k=1}^{K} f_k + \sigma f_k^2\right)^{1/2}}\right] \quad \Phi\left[\frac{(n \quad d) \quad n}{\left(\sum_{k=1}^{K} f_k + \sigma f_k^2\right)^{1/2}}\right] \\ &= 2\Phi\left[\frac{d}{\left(\sum_{k=1}^{K} f_k + \sigma f_k^2\right)^{1/2}}\right] \end{aligned}$$

where Φ is the cumulative distribution function (CDF) of the standard normal

distribution. This probability may be useful, for example, in population data where true marginal counts are known and analysts may reject large deviations between n and n_{syn}.

The calculations carried out above can be applied to any marginal count - not just n_{syn} - by summing over the relevant subset of cells. For those count distributions such as the Poisson, for which there are identities for sums of random variables, it is possible to calculate exact distributions of marginal counts. Nevertheless, when K is large, such as in administrative databases the normal approximate would su□ce.

3.4 Linking the mechanism to differential privacy

Assume for simplicity that the pseudocount $\alpha > 0$ is added to *all* original counts (not just to zeros). When the Poisson is used the set of synthetic counts $\mathbf{b}$ is obtained by simulating from:

$$b_i \mid a_i, \alpha \quad \text{Poisson}(a_i + \alpha) \quad \text{for} \quad i = 1, \ldots, K. \tag{3.17}$$

As with the multinomial in Section 2.5.4.1, suppose $\mathbf{a}$ and $\mathbf{a}^{\emptyset}$ di□er in their first element only, that is, $a_1 = a_1^{\emptyset} + 1$, such that $||\mathbf{a} \quad \mathbf{a}^{\emptyset}|| = 1$; then

$$\frac{p(\mathbf{b} \mid \mathbf{a}, \alpha)}{p(\mathbf{b} \mid \mathbf{a}^{\emptyset}, \alpha)} = \exp(\quad 1)\left(\frac{a_1 + \alpha}{a_1 \quad 1 + \alpha}\right)^{b_1}. \tag{3.18}$$

This quantity is bounded below, with the minimum occurring when $b_1 = 0$; but is unbounded above as b_1 can take any integer up to infinity. Therefore, for all $\mathbf{a}$ $\mathbf{a}^{\emptyset}$ and $\mathbf{b}$:

$$\exp(\quad 1) \quad \frac{p(\mathbf{b} \, j \, \mathbf{a}, \boldsymbol{\alpha})}{p(\mathbf{b} \, j \, \mathbf{a}^{\emptyset}, \boldsymbol{\alpha})} < 1 \, . \tag{3.19}$$

Therefore, ϵ-DP clearly cannot be satisfied. However, (ϵ, δ)-probabilistic DP can be satisfied; that is, ϵ and δ can be found such that:

$$\frac{1}{\exp(\epsilon)} \quad \frac{p(\mathbf{b} \, j \, \mathbf{a}, \boldsymbol{\alpha})}{p(\mathbf{b} \, j \, \mathbf{a}^{\emptyset}, \boldsymbol{\alpha})} \quad \exp(\epsilon) \quad \text{with probability } 1 \quad \delta. \tag{3.20}$$

Firstly, considering the left hand side of the inequality in (3.20):

$$p\left(\frac{1}{\exp(\epsilon)} \quad \frac{p(\mathbf{b} \, j \, \mathbf{a}, \boldsymbol{\alpha})}{p(\mathbf{b} \, j \, \mathbf{a}^{\emptyset}, \boldsymbol{\alpha})}\right) = p\left(b_1 \quad \frac{1 \quad \epsilon}{\log\left(\frac{a_i + \alpha}{a_i \quad 1 + \alpha}\right)}\right). \tag{3.21}$$

When $\epsilon > 1$, this probability is equal to 1. When $\epsilon \quad 1$, this probability can be obtained from the Poisson's CDF. The use of the CDF to satisfy relaxations of DP was considered by Balle and Wang (2018) (though within a Gaussian framework). The probability in (3.21) is at its smallest when $a_1 = \max_i a_i$. Similarly, considering the right hand side of the inequality in (3.20):

$$p\left(\frac{p(\mathbf{b} \, j \, \mathbf{a}, \boldsymbol{\alpha})}{p(\mathbf{b} \, j \, \mathbf{a}^{\emptyset}, \boldsymbol{\alpha})} \quad \exp(\epsilon)\right) = p\left(b_1 \quad \frac{\epsilon + 1}{\log\left(\frac{a_i + \alpha}{a_i \quad 1 + \alpha}\right)}\right), \tag{3.22}$$

which is at its smallest when $a_1 = \min_i a_i$. Therefore, for a given ϵ and for all $\mathbf{a}$ $\mathbf{a}^{\emptyset}$,

$$\delta = \max\left\{ F\left(\frac{1 \quad \epsilon}{\log\left(\frac{\max a_i + \alpha}{\max a_i \quad 1+\alpha}\right)} \right), 1 \quad F\left(\frac{1+\epsilon}{\log\left(\frac{\min a_i + \alpha}{\min a_i \quad 1+\alpha}\right)} \right) \right\}$$

is the probability that DP fails (where F is the Poisson's CDF). This is closely related to the concept of (ϵ, δ)-probabilistic DP (Dwork and Roth, 2014).

Similarly, using the NBI rather the Poisson (the PIG could of course also be used) - again with the slight di□erence that α is applied to all original counts - each synthetic count b_i $(i, \ldots, K)$ is obtained by simulating from:

$$b_i \, j \, a_i, \alpha, \sigma \quad \text{NBI}(a_i + \alpha, \sigma).$$

Once again supposing $\mathbf{a}$ and $\mathbf{a}^{\emptyset}$ di□er in their first element only and that $a_1 = a_1^{\emptyset} + 1$; then

$$\frac{p(\mathbf{b} \, j \, \mathbf{a}, \sigma, \alpha)}{p(\mathbf{b} \, j \, \mathbf{a}^{\emptyset}, \sigma\alpha)} = \left\{ \frac{(a_1 + \alpha)[1 + \sigma(a_1 \quad 1 + \alpha)]}{(a_1 \quad 1 + \alpha)[1 + \sigma(a_1 + \alpha)]} \right\}^{b_1} \quad \left[\frac{1 + \sigma(a_1 \quad 1 + \alpha)}{1 + \sigma(a_1 + \alpha)} \right]^{1/\sigma}.$$

This expression is again unbounded (b_1 can be any non-negative integer). Following the same procedure as used for the Poisson, for a given ϵ, (ϵ, δ)-probabilistic DP can be satisfied by setting:

$$\delta = \max\left\{ F\left(\frac{\epsilon \quad \frac{1}{\sigma}\log\left[\frac{1+\sigma(a_1 \quad 1+\alpha)}{1+\sigma(a_1+\alpha)}\right]}{\log\left\{\frac{(a_1+\alpha)[1+\sigma(a_1 \quad 1+\alpha)]}{(a_1 \quad 1+\alpha)[1+\sigma(a_1+\alpha)]}\right\}} \right), 1 \quad F\left(\frac{\epsilon \quad \frac{1}{\sigma}\log\left[\frac{1+\sigma(a_1 \quad 1+\alpha)}{1+\sigma(a_1+\alpha)}\right]}{\log\left\{\frac{(a_1+\alpha)[1+\sigma(a_1 \quad 1+\alpha)]}{(a_1 \quad 1+\alpha)[1+\sigma(a_1+\alpha)]}\right\}} \right) \right\}$$

where F is now the NBI's CDF.

For when the Poisson and NBI are used, δ was found empirically, for various

$\min_i a_i$, $\max_i a_i$, α, ϵ and σ (for the NBI). The results are given in Tables 3.2 and 3.3.

In general, there is an inverse relationship between ϵ and δ: a smaller ϵ necessitates a larger δ to achieve (ϵ, δ)-probabilistic DP. Similarly, decreasing α - or decreasing σ in the NBI - necessitates either a smaller ϵ or larger δ.

For the values of $\min_i a_i$ and $\max_i a_i$ considered, while larger $\min_i a_i$ imply smaller ϵ and/or smaller δ, ϵ and δ are unaffected by $\max_i a_i$. The latter needs further investigation. It may just be due to the values considered. If, for example, much greater minimum and maximum values are considered - as may be seen in census tables - then $\max_i a_i$ may play a part.

The link to differential privacy deserves further consideration. Essentially, DP assumes an intruder is trying to locate the cell to which just one individual belongs (knowing the locations of all the other individuals in the data). The reason why the multinomial family of distributions can satisfy ϵ-DP - but the count distributions cannot - is because every synthetic count is bounded, that is, there is a maximum value that any synthetic count can take, namely n. With count distributions there is a chance - however small - that a very large count can be obtained. To see how large counts cause the DP definition to fail, suppose in an intruder's data set - which, of course, is the actual data set minus the target individual - a certain cell has a count of 1. Then suppose in the synthetic data - generated using the Poisson - this cell has a count of 5. As it is relatively much more likely that this count originated from a cell with a count of 2 than from a count of 1 (11.7 times more likely), the intruder can infer that this cell is a likely origin of the target.

It is interesting how with DP and its strong intruder knowledge assumption, disclosure risk is deemed to be at its greatest when the scope for potential movement

Table 3.2: The δ such that (ϵ, δ)-probabilistic DP is achieved when the Poisson is used, for various $\min_i a_i$, $\max_i a_i$, α and ϵ.

		$\alpha = 1$			$\alpha = 2$			$\alpha = 3$		
		$\max_i a_i$								
	$\min_i a_i$	5	7	10	5	7	10	5	7	10
$\epsilon = 0.25$	1	0.59399	0.59399	0.59399	0.35277	0.35277	0.35277	0.37116	0.37116	0.37116
	3	0.37116	0.37116	0.37116	0.38404	0.38404	0.38404	0.39370	0.39370	0.39370
	5	0.39370	0.39370	0.39370	0.27091	0.27091	0.27091	0.28338	0.28338	0.28338
$\epsilon = 0.5$	1	0.32332	0.32332	0.32332	0.35277	0.35277	0.35277	0.21487	0.21487	0.21487
	3	0.21487	0.21487	0.21487	0.23782	0.23782	0.23782	0.15276	0.15276	0.15276
	5	0.15276	0.15276	0.15276	0.16950	0.16950	0.16950	0.11192	0.11192	0.11192
$\epsilon = 1$	1	0.32332	0.32332	0.32332	0.18474	0.18474	0.18474	0.11067	0.11067	0.11067
	3	0.11067	0.11067	0.11067	0.06809	0.06809	0.06809	0.04262	0.04262	0.04262
	5	0.04262	0.04262	0.04262	0.02700	0.02700	0.02700	0.01726	0.01726	0.01726
$\epsilon = 1.5$	1	0.14288	0.14288	0.14288	0.03351	0.03351	0.03351	0.02136	0.02136	0.02136
	3	0.02136	0.02136	0.02136	0.00545	0.00545	0.00545	0.00363	0.00363	0.00363
	5	0.00363	0.00363	0.00363	0.00096	0.00096	0.00096	0.00065	0.00065	0.00065

Table 3.3: The δ such that (ϵ, δ)-probabilistic DP is achieved when the NBI is used, for various $\min_i a_i$, $\max_i a_i$, α and ϵ.

			$\alpha = 1$			$\alpha = 2$			$\alpha = 3$		
			$\max_i a_i$								
		$\min_i a_i$	5	7	10	5	7	10	5	7	10
	$\epsilon = 0.5$	1	0.55279	0.55279	0.55279	0.55279	0.55279	0.55279	0.14493	0.14493	0.14493
		3	0.12953	0.12953	0.12953	0.11145	0.11145	0.11145	0.04085	0.04085	0.04085
		5	0.03850	0.03850	0.03850	0.03505	0.03505	0.03505	0.01275	0.01275	0.01275
	$\epsilon = 1$	1	0.55279	0.55279	0.55279	0.37390	0.37390	0.37390	0.08242	0.08242	0.08242
$\sigma = 2$		3	0.04085	0.04085	0.04085	0.03557	0.03557	0.03557	0.00711	0.00711	0.00711
		5	0.00477	0.00477	0.00477	0.00437	0.00437	0.00437	0.00076	0.00076	0.00076
	$\epsilon = 1.5$	1	0.55279	0.55279	0.55279	0.37390	0.37390	0.37390	0.03690	0.03690	0.03690
		3	0.01574	0.01574	0.01574	0.01378	0.01378	0.01378	0.00117	0.00117	0.00117
		5	0.00363	0.00363	0.00363	0.00096	0.00096	0.00096	0.00065	0.00065	0.00065
	$\epsilon = 0.5$	1	0.38096	0.38096	0.38096	0.38096	0.38096	0.38096	0.08126	0.08126	0.08126
		3	0.06044	0.06044	0.06044	0.05196	0.05196	0.05196	0.01838	0.01838	0.01838
		5	0.01569	0.01569	0.01569	0.01379	0.01379	0.01379	0.00508	0.00508	0.00508
	$\epsilon = 1$	1	0.38096	0.38096	0.38096	0.26840	0.26840	0.26840	0.03878	0.03878	0.03878
$\sigma = 5$		3	0.01838	0.01838	0.01838	0.01614	0.01614	0.01614	0.00274	0.00274	0.00274
		5	0.00182	0.00182	0.00182	0.00156	0.00156	0.00156	0.00026	0.00026	0.00026
	$\epsilon = 1.5$	1	0.38096	0.38096	0.38096	0.26840	0.26840	0.26840	0.01751	0.01751	0.01751
		3	0.00671	0.00671	0.00671	0.00526	0.00526	0.00526	0.00044	0.00044	0.00044
		5	0.00025	0.00025	0.00025	0.00020	0.00020	0.00020	0.00002	0.00002	0.00002
	$\epsilon = 0.5$	1	0.26247	0.26247	0.26247	0.19223	0.19223	0.19223	0.04704	0.04704	0.04704
		3	0.03201	0.03201	0.03201	0.02827	0.02827	0.02827	0.00958	0.00958	0.00958
		5	0.00768	0.00768	0.00768	0.00701	0.00701	0.00701	0.00252	0.00252	0.00252
	$\epsilon = 1$	1	0.26247	0.26247	0.26247	0.19223	0.19223	0.19223	0.02059	0.02059	0.02059
$\sigma = 10$		3	0.00958	0.00958	0.00958	0.00806	0.00806	0.00806	0.00134	0.00134	0.00134
		5	0.00089	0.00089	0.00089	0.00074	0.00074	0.00074	0.00012	0.00012	0.00012
	$\epsilon = 1.5$	1	0.26247	0.26247	0.26247	0.15544	0.15544	0.15544	0.00934	0.00934	0.00934
		3	0.00342	0.00342	0.00342	0.00265	0.00265	0.00265	0.00021	0.00021	0.00021
		5	0.00012	0.00012	0.00012	0.00009	0.00009	0.00009	0.00001	0.00001	0.00001

between original and synthetic counts is great. This appears counterintuitive as statistical disclosure control methods nearly always seek to reduce risk by increasing the divergence from the original counts. It emphasizes the importance of considering intruder knowledge when quantifying disclosure risk and highlights the advantages of metrics such as that of Reiter and Mitra (2009) which are conditional on intruder knowledge.

3.5 Empirical synthesis when $m = 1$

3.5.1 The ESCsub data

The English School Census (ESC) is an administrative database that holds information about pupils in state-funded schools. Every school term the Department for Education (DfE) requests that all nursery, primary and secondary schools, which are fully or partly funded by the state, submit details about the school and its pupils. This is just one example of an administrative database held by a government department; other examples include, but are not limited to, the Patient Register (held by the Department of Health) and the Customer Information System (held by the Department for Work and Pensions).

For obvious reasons, access to the ESC data, as well as to other administrative databases, is highly restricted. However, in previous work conducted by the ONS, a carefully constructed data set using publicly available sources was created to be used as a substitute to the ESC, in order to develop synthesis methods for administrative data. These data were used here as the basis for generating a synthetic database.

The data were constructed by the ONS using public 2011 census output tables

involving various combinations of local authority, sex, age and ethnicity[1]. Language attributes from the census were also included and artificially expanded to match with categories in the ESC. In addition, school phase attributes were incorporated, some adjustments for migration were applied, and non-response and invalid categories were added to various variables, again taking publicly available information from the census.

The two variables measured at the school level were ignored for this illustration, which focused instead on the remaining five variables measured at the pupil level. Henceforth, this data set is referred to as the ESCsub where "sub" denotes substitute. Table 3.4 summarises the variables present in the ESCsub illustration. The data comprise $n = 8,190,870$ pupils over $p = 5$ categorical variables, giving rise to a multi-way contingency table with $K = 326 \quad 20 \quad 4 \quad 19 \quad 7 = 3.5 \quad 10^6$ cells. A more detailed description of the data's origin is available at Blanchard et al. (2022). The breakdown of the cell counts are given in Table 3.5; only 333,660 (9.6$\%$) are non-zero - so the data are sparse. There are no structural zeros.

So while the data are in a sense simulated, this was done using real data sources and care was taken to ensure that the resulting data reflect, at the very least, the typical structure present in the ESC. As such, this was a good example to use to demonstrate our synthesis method and a similar performance is expected when the method is applied to the actual ESC, as well as other similar large categorical administrative databases. Importantly, the data were not generated from a statistical model and thus do not favour a particular synthesis method.

[1]Specifically, information from the following public sources were used to create the data: `http://www.nomisweb.co.uk/census/2011; http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-user-guide/quality-and-methods/quality/quality-measures/response-and-imputation-rates/index.html; https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2014`

Table 3.4: The ESCsub's variables and their numbers of categories.

Variable	Type	# Categories
Area Code/ Geography (V)	Categorical	326
Ethnicity (W)	Categorical	20
Sex (X)	Categorical	4
Age (Y)	Categorical	19
Language (Z)	Categorical	7

Table 3.5: Distribution of cell sizes in the ESCsub data.

Cell count	Frequency	% of cells
0	3,134,980	90.38
1	119,917	3.46
2	51,412	1.48
3	25,952	0.75
4	19,450	0.56
5	13,076	0.38
6	10,345	0.30
7	7,947	0.23
8	7,077	0.20
9	5,809	0.17
10	5,163	0.15
11	67,512	1.95
Total	3,468,640	100

3.5.2 The synthesis

The synthesis was carried out in R (version 3.6.3) using the methods described earlier in this chapter. The Poisson, NBI and PIG models were compared by examining how the synthetic counts deviate from the original counts, by computing summaries of risk and utility. This evaluation also includes a comparison of parameter estimates obtained from a log-linear analysis, which was performed on both the original and the synthetic data.

Just $m = 1$ data set was generated for each synthesis model. The CPU times to carry this out were 0.2, 0.3 and 162 seconds for the Poisson, NBI and PIG models, respectively. The PIG model took notably longer, although is still fast compared to other methods, such as the conditional approaches (Section 2.3.1) that synthesize the data at the microdata level.

Let V, W, X, Y and Z denote the five variables in the data, and let f_{vwxyx} denote the cell count of a particular cell in the cross-classified table corresponding to category v *2* V, w *2* W , x *2* X, y *2* Y and z *2* Z. A synthetic count was then drawn for this cell by,

$$f^{\text{syn}}_{vwxyz} \quad \text{Poisson}(f_{vwxyz}),$$

when the Poisson synthesis model was used; or, when either of the two-parameter distributions were used,

$$f^{\text{syn}}_{vwxyz} \quad \text{NBI}(f_{vwxyz}, \sigma) \quad \text{or} \quad f^{\text{syn}}_{vwxyz} \quad \text{PIG}(f_{vwxyz}, \sigma).$$

For the Poisson, the only parameter to be set was α, the pseudocount added to

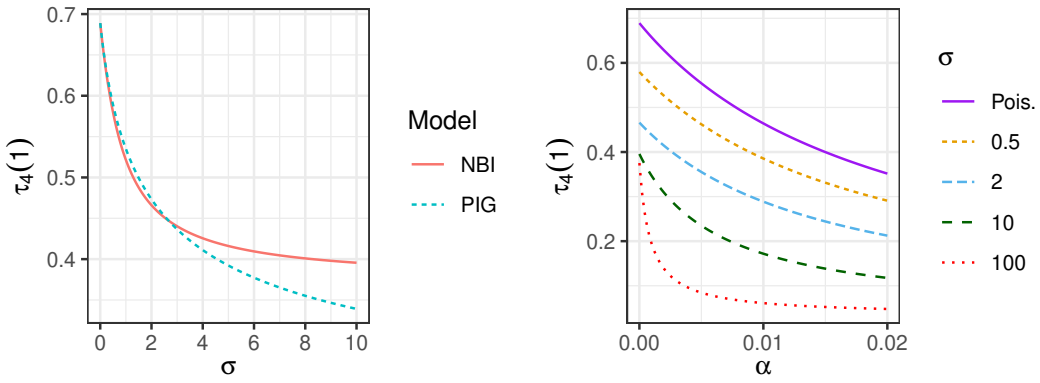

Figure 3.1: Left plot: $\tau_4(1)$ as a function of σ for when the NBI (solid line) and PIG (dashed line) models were used ($\alpha = 0$). Right plot: $\tau_4(1)$ as a function of α for different σ values when the NBI was used.

random zeros in the original data. For the two-parameter count models, there was the additional parameter σ to consider.

As mentioned previously, one of the appealing features of using these saturated synthesis models is that it allows the synthesizer to determine properties of the synthesis model *a priori*, thus reducing the amount of empirical evaluation necessary during the synthesis. For illustration, Figure 3.1 (left) compares the effect of σ on the risk metric $\tau_4(1)$ for the NBI and PIG models. For large σ, the risk levels off for the NBI but continues to fall away for the PIG. Figure 3.1 (right) also looks at $\tau_4(1)$, but at the combined effect of σ and α when the NBI is used. For all σ, $\tau_4(1)$ falls as α increases and $\tau_4(1)$ is always lower for the NBI than for the Poisson.

3.5.3 Descriptive summaries of risk and utility

Table 3.6 gives the proportion of cell counts in the synthesized tables that are within $p\%$ of their original size, for different σ and α. The first block of results considers all cells, and the second block only considers non-zero original cells. The reason for

splitting the results into these two blocks is two-fold. Firstly, as the data have a high proportion of zero counts, including the zeros clouds the results somewhat, making it less clear to see the effect of increasing σ. Secondly, as the zero counts do not pertain to an actual individual, they are, in a sense, less risky than the non-zero counts and therefore arguably less important.

Smaller values of p can be viewed as summaries of risk while larger values of p measures of utility. To elaborate, if a large proportion of original and synthetic cell counts are very close, say within 0.5% $(p = 0.5)$ of each other, then the synthetic data could be considered to be high risk. If, on the other hand, few original and synthetic cell counts are within, for example, 50% $(p = 50)$ of each other, then this is likely to indicate low utility. As an example, the 0.927 value in the top-left corner of the table means that when $\alpha = 0$ and $\sigma = 0$, 92.7$\%$ of all cell counts in the synthetic data were within 0.5$\%$ of the corresponding count in the original data. The Poisson model had the greatest utility but the greatest risk. There was little to choose between the NBI and PIG models based on these summaries. As expected, greater σ or α lead to greater divergences in original and synthetic cell sizes.

The similarities between the NBI and PIG models are also highlighted in Figures 3.2 and 3.3, which plot the synthetic versus original counts and also percentage differences (between synthetic and original counts) versus original counts. While the Poisson model's points ($\sigma = 0$ cases) were close to the 45 line - which indicates strong correlation between synthetic and original counts - this correlation reduces as σ increases in both the NBI and PIG models. Even a relatively small value of $\sigma = 0.01$ introduced noticeable dispersion around the 45 line. The right panels display a funnel shape, that is, percentage differences were greater for smaller counts than for larger

counts. This is an ideal profile for balancing risk and utility, as the riskiest individuals are the ones corresponding to small cell counts, and these cell counts require the most movement during synthesis. On the other hand, large counts are relatively low risk, and proportional changes to large counts will have a more significant impact on utility, thus relatively less perturbation is desired.

Table 3.7 presents empirical values for the τ metrics, again for varying σ and α. The expected values are known prior to synthesis, though a small difference occurs, owing to simulation noise. But, for cell sizes that are prevalent in the original data - such as zeros and ones - this error is negligible. For example, the empirical value obtained for $\tau_3(1)$ when the Poisson model was used ($\sigma = 0, \alpha = 0$) is 0.3674, which is almost identical to the expected value, exp(-1)=0.3679.

Table 3.7 also illustrates the suitability of α in reducing risk. The values for $\tau_4(1)$ are substantially lower when $\alpha = 0.02$ than when $\alpha = 0$; for example, when the Poisson model is used, $\tau_4(1)$ is 0.352 compared to 0.689. For a given α, the NBI and PIG models almost always have a lower risk than the Poisson model when considering the $\tau_3(1)$ and $\tau_4(1)$ metrics. It is particularly interesting to note varying profiles between synthesis models and these metrics. For example, if one model has a lower $\tau_3(1)$ value than another, then this is not necessarily the case when comparing the corresponding $\tau_4(1)$ value. To illustrate, consider the case when $\alpha = 0$ and $\sigma = 10$. For the NBI, the value of $\tau_3(1)$ is $0.0711 < 0.1532$ the value for the PIG. But for $\tau_4(1)$, with these same parameter values, the value under the NBI is $0.3910 > 0.3387$ the value under the PIG. The specific choice of synthesis model to use would depend on the synthesizer's range of permitted values for τ_3, and τ_4, and choosing the model that best satisfies these requirements.

Table 3.6: The proportion of synthetic counts within $p\%$ of the corresponding original counts. The table includes results for both the NBI and the PIG, for different σ and α. The upper block of results considers all original cell counts, while the lower block considers only non-zero original cells. For $\alpha = 0.02$, whenever a zero count was synthesized to a non-zero count, although the percentage difference was not estimable (zero denominator), it was deemed to be greater than 50% for the purpose of this table.

		Proportion of synthetic cell counts within $p\%$ of the original									
		NBI					PIG				
	p	0.5	1	5	10	50	0.5	1	5	10	50
		All original cell counts									
	σ										
$\alpha = 0$	0 (Pois.)	0.927	0.927	0.931	0.935	0.967	0.927	0.927	0.931	0.935	0.967
	0.1	0.924	0.924	0.926	0.928	0.961	0.925	0.925	0.926	0.928	0.961
	0.5	0.920	0.920	0.920	0.922	0.946	0.921	0.921	0.921	0.923	0.949
	1	0.917	0.917	0.917	0.918	0.937	0.918	0.918	0.919	0.920	0.942
	2	0.914	0.914	0.914	0.914	0.928	0.916	0.916	0.916	0.917	0.935
	5	0.910	0.910	0.910	0.910	0.918	0.913	0.913	0.913	0.914	0.927
	10	0.907	0.907	0.907	0.908	0.912	0.911	0.911	0.911	0.912	0.921
$\alpha = 0.02$	0 (Pois.)	0.909	0.910	0.913	0.917	0.949	0.909	0.910	0.913	0.917	0.949
	0.1	0.907	0.907	0.908	0.910	0.943	0.907	0.907	0.908	0.910	0.943
	0.5	0.902	0.902	0.903	0.904	0.928	0.903	0.903	0.903	0.905	0.931
	1	0.899	0.899	0.900	0.901	0.920	0.901	0.901	0.901	0.902	0.924
	2	0.896	0.896	0.896	0.897	0.911	0.899	0.899	0.899	0.900	0.918
	5	0.893	0.893	0.893	0.893	0.901	0.896	0.896	0.896	0.897	0.909
	10	0.891	0.891	0.891	0.891	0.896	0.895	0.895	0.895	0.895	0.905
		Non-zero original cell counts									
	σ										
$\alpha = 0$	0 (Pois.)	0.242	0.245	0.280	0.327	0.658	0.242	0.245	0.280	0.327	0.658
	0.1	0.214	0.215	0.226	0.252	0.592	0.217	0.218	0.229	0.256	0.598
	0.5	0.167	0.167	0.173	0.187	0.437	0.177	0.177	0.182	0.197	0.468
	1	0.136	0.136	0.140	0.150	0.347	0.153	0.153	0.157	0.167	0.395
	2	0.102	0.102	0.105	0.111	0.253	0.128	0.128	0.131	0.138	0.324
	5	0.059	0.059	0.061	0.064	0.145	0.097	0.097	0.099	0.104	0.238
	10	0.037	0.037	0.038	0.040	0.089	0.076	0.076	0.077	0.081	0.183
$\alpha = 0.02$	0 (Pois.)	0.242	0.245	0.279	0.326	0.657	0.242	0.245	0.279	0.326	0.657
	0.1	0.215	0.215	0.226	0.253	0.593	0.215	0.216	0.227	0.254	0.598
	0.5	0.167	0.167	0.172	0.186	0.437	0.175	0.176	0.181	0.196	0.468
	1	0.137	0.137	0.141	0.151	0.348	0.153	0.153	0.157	0.167	0.396
	2	0.102	0.102	0.105	0.111	0.253	0.128	0.128	0.130	0.138	0.324
	5	0.061	0.061	0.062	0.065	0.147	0.096	0.096	0.098	0.103	0.237
	10	0.037	0.037	0.037	0.039	0.088	0.076	0.076	0.077	0.081	0.182

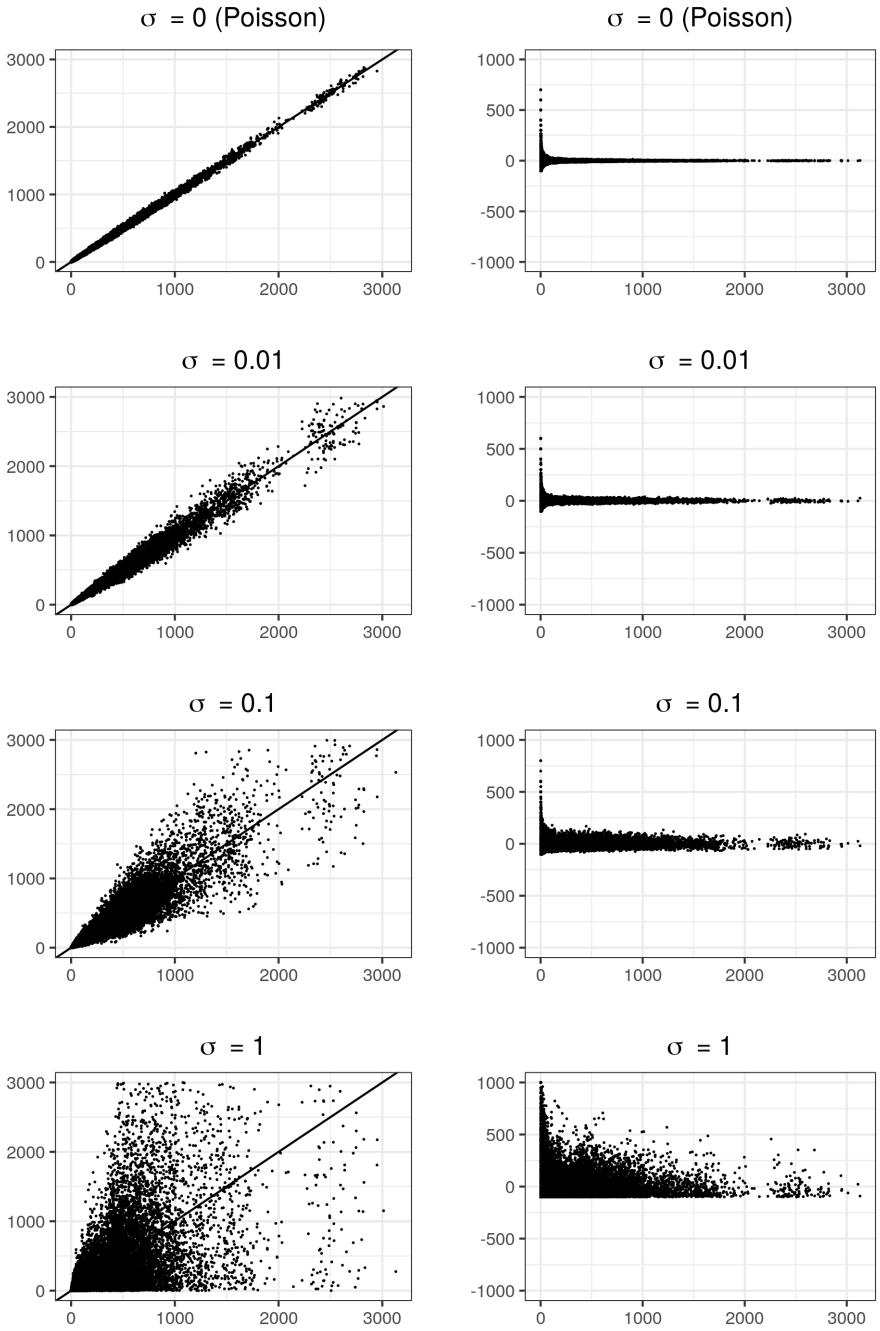

Figure 3.2: Left hand plots: the synthetic counts versus the original counts for different σ (for the NBI with $\alpha = 0$). Right hand plots: original and synthetic counts' percentage differences versus the original counts. Original counts of zero are omitted.

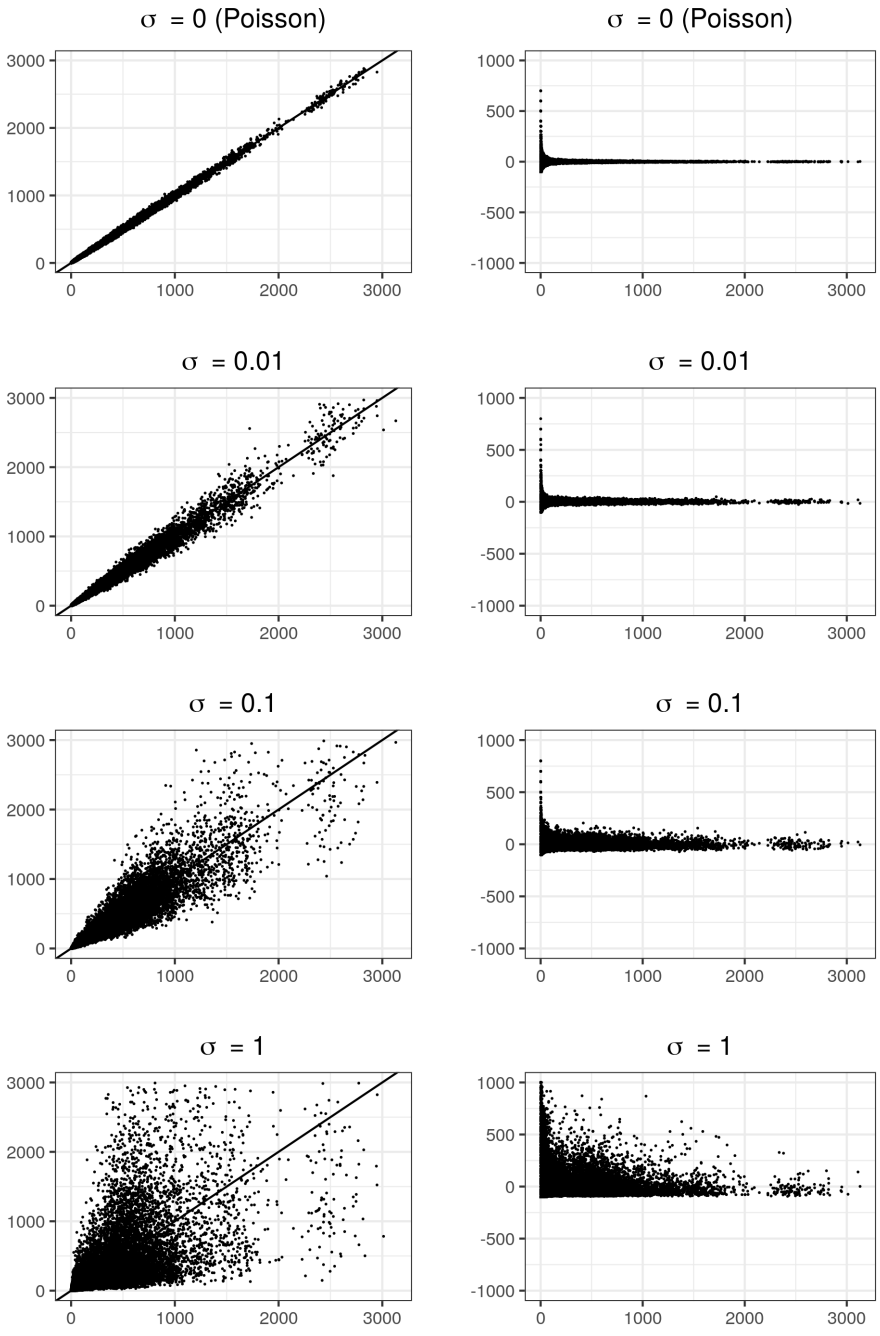

Figure 3.3: As Figure 3.2 but for when the PIG was used rather than the NBI.

Table 3.7: Empirical values obtained for the τ metrics.

		NBI				PIG			
	k	0	1	2	3	0	1	2	3
	σ				$\tau_1(k)$				
$\alpha = 0$	0 (Pois.)	0.9190	0.0184	0.0135	0.0086	0.9190	0.0184	0.0135	0.0086
	0.1	0.9204	0.0183	0.0130	0.0085	0.9203	0.0184	0.0130	0.0084
	0.5	0.9256	0.0177	0.0117	0.0077	0.9243	0.0181	0.0121	0.0078
	1	0.9317	0.0166	0.0105	0.0068	0.9280	0.0179	0.0111	0.0072
	5	0.9587	0.0098	0.0054	0.0035	0.9422	0.0167	0.0086	0.0053
	10	0.9713	0.0064	0.0033	0.0022	0.9500	0.0156	0.0072	0.0042
$\alpha = 0.02$	0 (Pois.)	0.9013	0.0359	0.0136	0.0086	0.9013	0.0359	0.0136	0.0086
	0.1	0.9024	0.0360	0.0133	0.0084	0.9024	0.0361	0.0133	0.0085
	0.5	0.9078	0.0352	0.0120	0.0077	0.9065	0.0357	0.0122	0.0078
	1	0.9139	0.0339	0.0107	0.0069	0.9101	0.0355	0.0115	0.0072
	5	0.9415	0.0259	0.0063	0.0036	0.9251	0.0330	0.0093	0.0053
	10	0.9550	0.0212	0.0047	0.0024	0.9337	0.0307	0.0084	0.0045
					$\tau_3(k)$				
$\alpha = 0$	0 (Pois.)	1	0.3674	0.2701	0.2231	1	0.3674	0.2701	0.2231
	0.1	1	0.3489	0.2457	0.1976	1	0.3538	0.2484	0.1974
	0.5	1	0.2964	0.1874	0.1340	1	0.3090	0.2022	0.1468
	1	1	0.2499	0.1489	0.1024	1	0.2779	0.1677	0.1197
	5	1	0.1144	0.0618	0.0403	1	0.1895	0.0981	0.0654
	10	1	0.0724	0.0378	0.0248	1	0.1532	0.0740	0.0466
$\alpha = 0.02$	0 (Pois.)	0.9804	0.3648	0.2695	0.2247	0.9804	0.3648	0.2695	0.2247
	0.1	0.9802	0.3499	0.2450	0.1950	0.9801	0.3494	0.2466	0.1984
	0.5	0.9803	0.2950	0.1876	0.1391	0.9803	0.3090	0.1981	0.1436
	1	0.9804	0.2515	0.1498	0.1052	0.9803	0.2782	0.1689	0.1218
	5	0.9812	0.1172	0.0620	0.0429	0.9810	0.1888	0.0959	0.0629
	10	0.9819	0.0711	0.0374	0.0255	0.9817	0.1524	0.0750	0.0486
					$\tau_4(k)$				
$\alpha = 0$	0 (Pois.)	0.9835	0.6893	0.2974	0.1943	0.9835	0.6893	0.2974	0.1943
	0.1	0.9820	0.6603	0.2811	0.1742	0.9821	0.6648	0.2827	0.1760
	0.5	0.9764	0.5788	0.2372	0.1304	0.9778	0.5890	0.2484	0.1416
	1	0.9701	0.5203	0.2108	0.1125	0.9739	0.5369	0.2232	0.1243
	5	0.9427	0.4043	0.1710	0.0858	0.9593	0.3919	0.1694	0.0929
	10	0.9305	0.3910	0.1677	0.0851	0.9513	0.3387	0.1521	0.0822
$\alpha = 0.02$	0 (Pois.)	0.9831	0.3516	0.2935	0.1957	0.9831	0.3516	0.2935	0.1957
	0.1	0.9817	0.3357	0.2735	0.1733	0.9817	0.3350	0.2745	0.1751
	0.5	0.9759	0.2898	0.2316	0.1348	0.9774	0.2991	0.2403	0.1379
	1	0.9696	0.2567	0.2066	0.1141	0.9735	0.2712	0.2168	0.1259
	5	0.9419	0.1562	0.1469	0.0890	0.9584	0.1979	0.1520	0.0887
	10	0.9293	0.1162	0.1172	0.0799	0.9503	0.1715	0.1320	0.0808

3.5.4 Testing specific utility through log-linear model analysis

The synthesizer does not know, of course, what analyses users of the synthetic data would perform. Among the variables included in the data, which are best described as demographic, there is no obvious response variable, so analysts may be interested in associations between variables. Therefore, a log-linear analysis was chosen as a suitable way to test the specific utility of synthetic data generated with the synthesis method described in Section 3. It is difficult to obtain parameter estimates and parameters' standard error estimates for the full five-variable data, since large amounts of memory and storage are required - the same problem faced when fitting unsaturated synthesis models. To relieve some of this pressure, the all two-way interaction model was fitted to three of the data's five variables, ethnicity, age and language, resulting in 608 parameters.

The confidence interval overlap metric (defined in 2.14) was used to measure similarities between estimates. Combining rules are required to obtain valid parameter estimates and standard errors from synthetic data, even when just $m = 1$ synthetic data set is generated. This is because there are always two sources of uncertainty in synthetic data that need to be accounted for: the sampling uncertainty inherent in the original data, and the uncertainty owing to synthesis. To simplify the analysis - after all, the purpose here is just to evaluate the utility of the synthetic data - the original data were assumed to constitute a simple random sample drawn from a super-population, and sufficiently large for a large sample approximation to hold. This allowed the estimator T_s^{plug} to be used (see Table 2.1), which provides valid variance estimates for large samples when analysing just $m = 1$ completely synthesized data set. When $m = 1$ ($\bar{v}_m = v$) and $n = n_{\text{syn}}$, the estimator simplifies to $2v$, that is, the

Table 3.8: How σ and α a□ect the trimmed mean (top and bottom 10% excluded) percentage di□erence between log-linear parameter estimates obtained from the observed and synthetic data. The trimmed mean was used to subdue the e□ect of huge percentage di□erences arising through the presence of zero counts. For clarity, for an arbitrary original log-linear parameter estimate q and its corresponding synthetic estimate q^{syn}, the percentage di□erence was calculated by $100 \quad (q^{\text{syn}} \quad q)/q$.

	$\sigma = 0$ (Pois.)	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 5$	$\sigma = 10$
The NBI model							
$\alpha = 0$	-1.7	3.9	-12.0	-0.1	-18.7	10.6	-108.4
$\alpha = 0.005$	-23.5	-30.9	-31.7	-34.8	-34.0	14.8	-32.9
$\alpha = 0.01$	-32.5	-33.2	-33.7	-38.9	-47.8	41.0	-64.7
$\alpha = 0.015$	-38.8	-39.8	-47.7	-52.3	-47.6	-20.3	-3.5
$\alpha = 0.02$	-37.1	-34.0	-44.4	-40.9	-27.7	-42.0	-33.5
The PIG model							
$\alpha = 0$	-1.7	-2.2	16.2	11.2	-33.0	-6.6	187.4
$\alpha = 0.005$	-23.5	-21.7	-28.4	-21.7	-33.0	-73.6	-990.4
$\alpha = 0.01$	-32.5	-29.6	-31.7	-48.6	-42.3	25.9	-399.3
$\alpha = 0.015$	-38.8	-26.4	-36.6	-37.6	-20.6	-64.4	-504.0
$\alpha = 0.02$	-37.1	-47.8	-40.2	-20.6	-40.5	-76.2	425.2

variance estimate from the synthetic data, doubled.

Finally, in log-linear models, estimability issues can arise through the presence of zero counts in the data. This can lead to issues surrounding non-existence and non-identifiability of estimates (Fienberg and Rinaldo, 2012). But no serious model fitting issues arose in this particular example. There were some parameters included in the model with a true value of *1*. For such parameters, R returned a large negative value, typically in the vicinity of -20.

Figures 3.4 and 3.5 present boxplots of confidence interval overlap values for the log-linear model parameters across the di□erent synthesis models. They demonstrate

how increasing σ and α causes utility to fall away. For example, irrespective of α, whenever $\sigma = 10$, the median confidence interval overlap is zero. In general, a high proportion of the overlap values are equal to 1/2. This can be seen, for example, in the centre and right plots of Figure 3.5, where several of the upper quartiles are equal to 1/2.

Figure 3.6 and Table 3.8 presents (trimmed) mean percentage differences between synthetic and observed parameter estimates for various σ and α. Even setting α small can have an adverse effect on utility. For example, when $\sigma = 0$ (the Poisson model), increasing α from 0 to 0.005 causes the (trimmed) mean percentage difference in estimates to fall from -1.7% to 23.5%, thus demonstrating the bias caused by $\alpha > 0$. The general trend is that increasing α and σ results in larger percentage differences.

3.5.5 Balancing risk and utility

A key question a synthesizer would have is: which synthesis method offers the best balance between utility and risk? To address this, the risk-utility trade-off from each generated synthetic data set can be plotted. An example is displayed in Figure 3.7. Privacy has been measured on the y-axis via $1 \quad \tau_4(1)$, that is, (1 - risk), and utility on the x-axis by mean confidence interval overlap. The original data sit at the point (1,0), that is, maximum utility and minimum privacy. All points must lie within the unit square $[0, 1] \quad [0, 1]$ and, in general, the further from the origin, the better the synthetic data. For instance, when two points lie on the same horizontal line, it suggests that for the same level of privacy, one achieves a greater level of utility than the other.

This visualisation offers a convenient way to compare the performance of different

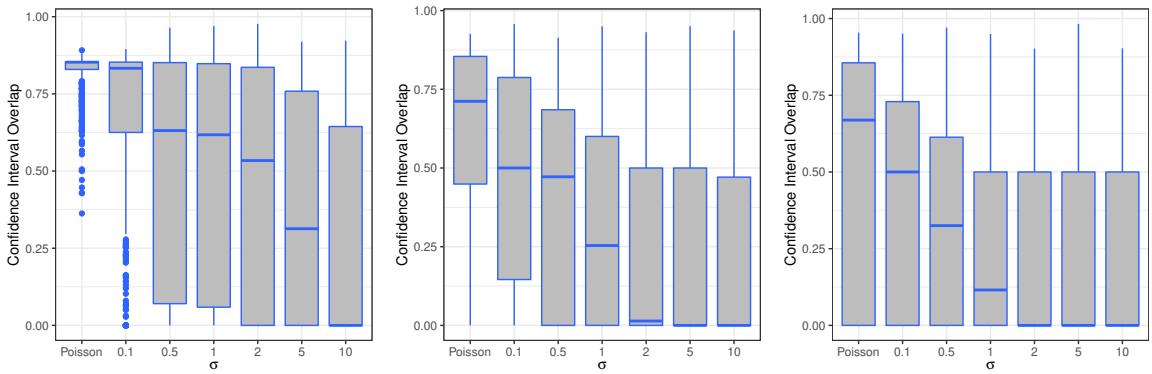

Figure 3.4: Boxplots showing how σ and α a□ect log-linear parameters' confidence interval overlap when the NBI is used. The left frame is the case where $\alpha = 0$; the middle frame where $\alpha = 0.01$; and the right frame where $\alpha = 0.02$

synthesis models. For example, it may be possible for one synthetic data set to strictly dominate another: the PIG model with $\sigma = 10$ provides greater utility and lower risk than the NBI model with $\sigma = 100$. The choice depends on the priorities of the data holder and users. For example, it may be that synthetic data can only be released if $\tau_4(1)$ is less than 0.5, in which case the synthetic data with the highest utility that satisfies this requirement could be released, here this would be the PIG model with $\sigma = 10$. Alternatively, it may be that only data with a utility value of at least 0.5 would be deemed useful enough for release, in which case the synthetic data generated under a NBI model with $\sigma = 0.1$ would be chosen.

In practice, a range of di□erent metrics for utility and privacy can be created and feed into determining which synthesis method is chosen. This decision is also likely to be application specific.

3.5.6 The Poisson versus overdispersed count distributions

The intention is that the synthesizer would usually use the NBI or PIG distributions,

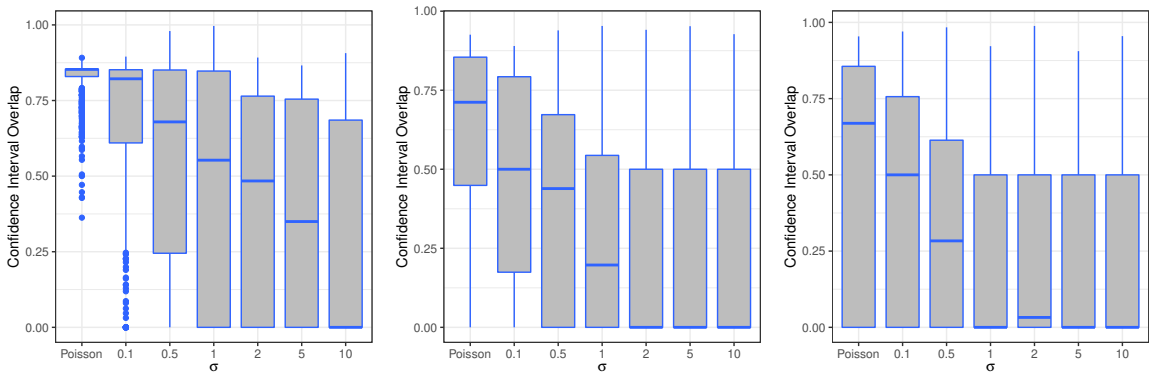

Figure 3.5: As Figure 3.4 but for when the PIG is used.

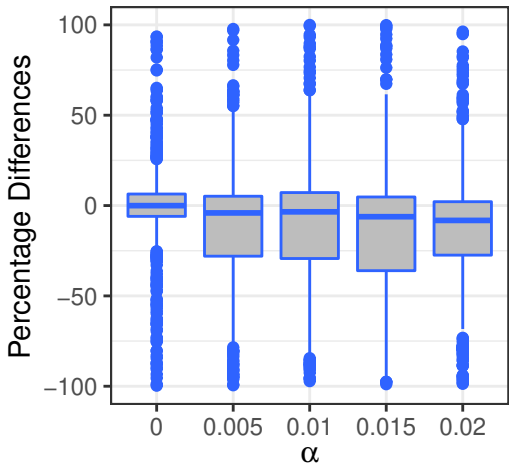

Figure 3.6: Boxplots showing how α a□ects percentage di□erences in parameter estimates obtained from the observed and synthetic data.

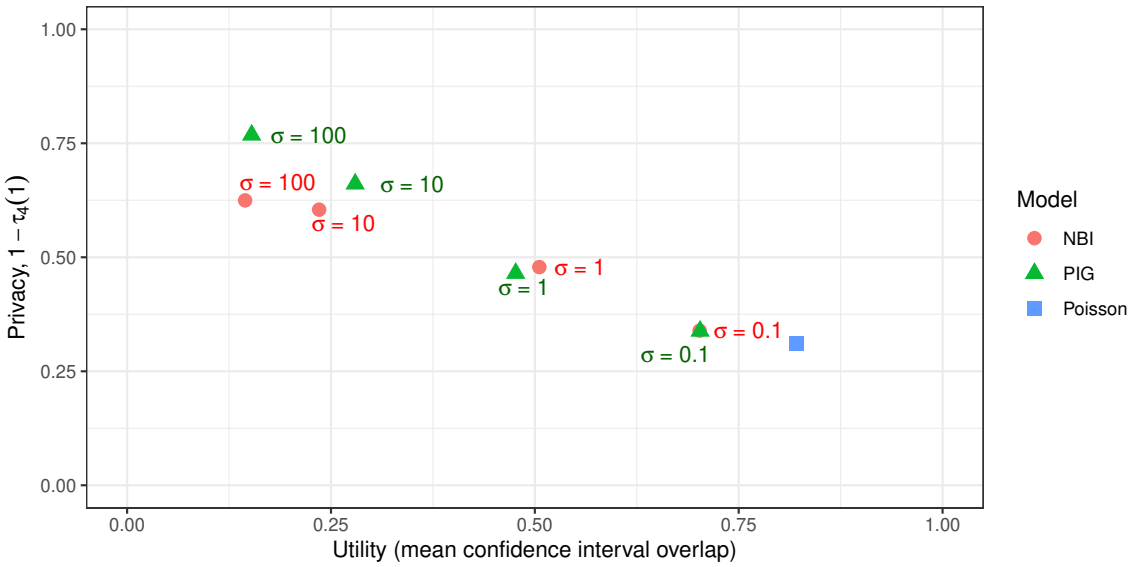

Figure 3.7: This plot, which is resemblant of a product possibility frontier in economics, provides a visual representation of the risk-utility trade-o for di erent σ ($\alpha = 0$).

rather than the Poisson, which is far too limited to be used in practice. In general, the NBI and PIG models give similar results, yet this is to be expected as both share the same variance function. Nevertheless, there are some marked di erences between the two, especially when σ is large; for example, when $\sigma = 10$ and $\alpha = 0$, there are substantially fewer zeros in the synthetic data when the PIG is used than when the NBI is used ($\tau_1(0)$ values of 0.950 and 0.971, respectively). There is, of course, scope to use other count distributions here; those with a di erent variance function would have an entirely di erent profile altogether. Moreover, both the NBI and PIG are also both limited in that they can only model overdispersion and not underdispersion, hence the variance is always greater than the Poisson - and this may be unnecessary for certain parts of the data deemed to be at lower risk of disclosure. The double Poisson distribution, for example, which can be used to model underdispersed count data,

allows the variance to be set lower than in the Poisson.

3.5.7 Summary

As data sets becomes bigger, the ability - from a computational perspective - to fit models requiring model-fitting algorithms diminishes. The use of saturated models set out in this chapter allows data sets to be synthesized e□ciently irrespective of their size, thus is suited to large administrative databases. Also, further e□ciency is achieved through the use of the scale parameter of overdispersed count distributions, such as the negative binomial and Poisson-inverse Gaussian, to apply noise to original counts. In the mechanism described, this scale parameter σ becomes a tuning parameter, and which, along with α (the psuedocount added to random zeros) can be tuned to set the value of a risk or utility metric *a priori*. The empirical example illustrates how σ and α can be adjusted to produce synthetic data with di□erent levels of risk and utility.

However, this chapter only considered the case of generating $m = 1$ synthetic data set. But $m > 1$ data sets can - and, sometimes, *must* - be generated. In the same way, using this framework means that certain properties can also be found analytically. It e□ectively introduces another parameter for the synthesizer to set, thus providing further flexibility.

Chapter 4

Extending the mechanism to generate $m > 1$ synthetic data sets

The original inferential frameworks for fully and partially synthetic data sets (Raghunathan et al., 2003; Reiter, 2003) relied on the generation of $m > 1$ synthetic data sets, because they required the computation of the between-synthesis variance b_m (see below). When the original data is not a simple random sample, $m > 1$ synthetic data sets must be generated, highlighting the importance of considering the $m > 1$ case.

When the original data constitute a simple random sample, and the data are completely synthesized, valid inferences can be obtained from $m = 1$ synthetic data set (Raab et al., 2016). In the previous section, this assumption was taken, though mainly out of convenience. In this instance, while $m > 1$ data sets are not intrinsic to obtaining *valid* inferences, the *quality* of inferences - for example, the width of confidence intervals - can, nevertheless, be improved upon by increasing m - but at the expense of higher risk. It is less a question, therefore, of which m allows valid

inferences to be obtained, but rather a question of which value of m is optimal with respect to the risk-utility trade-o□?

In either case, m can be viewed as a tuning parameter, and, as with the other tuning parameters σ and α, expected risk and utility profiles can be derived analytically, *a priori*. When saturated models are used for synthesis, ignoring the small bias arising from $\alpha > 0$, simulation error is the only source of uncertainty - and increasing m reduces simulation error. The notion is that $m > 1$ may allow a more favourable position in relation to the risk-utility trade-o□ than when $m = 1$; in short, it increases the number of options available to the synthesizer.

The use of parallel processing can substantially reduce the central processing unit (CPU) time when generating multiple data sets. Besides, the CPU time taken is typically negligible anyway; the synthesis presented in Section 3.5 took 0.3 seconds for the NBI with $m = 1$ on a typical laptop running R.

4.1 Obtaining inferences from $m > 1$ data sets

4.1.1 Analysing the $m > 1$ data sets before averaging the results

As mentioned in Section 2.3.2, when analysing multiple synthetic data sets the analyst considers each data set separately before later combining inferences. Most combining rules are based on the following three quantities (Drechsler, 2011):

$$\bar{q}_m = \frac{1}{m}\sum_{l=1}^{m} q^{(l)}, \qquad b_m = \frac{1}{(m \quad 1)}\sum_{l=1}^{m}(q^{(l)} \quad \bar{q}_m)^2, \qquad \bar{v}_m = \frac{1}{m}\sum_{l=1}^{m} v^{(l)}.$$

When using the synthesis method described in the previous chapter, partially - rather than fully - synthetic data sets are generated, because a synthetic population is not constructed and sampled from, as stipulated in Raghunathan et al. (2003). Hence, the following estimator T_p (Reiter, 2003), is valid when estimating Var$(\hat{Q})$,

$$T_p = \frac{b_m}{m} + \bar{v}_m.$$

The sampling distribution (if frequentist) or posterior distribution (if Bayesian) of $\hat{Q}$ is a t-distribution with $\nu_p = (m \quad 1)\,(1 + m\bar{v}_m/b_m)^2$ degrees of freedom. Often, ν_p is large enough for the t-distribution to be approximated by a normal distribution. However, when the between-synthesis variability is much larger than the within-synthesis variability, that is, when b_m is much larger than $\bar{v}_m$ - as may happen when large amounts of noise are applied to protect sensitive records - then ν_p is crucial to obtaining valid inferences.

If the original data constitute a simple random sample and are large enough to support a large sample assumption, T_s is also valid:

$$T_s = \bar{v}_m\Big(\frac{n_{\text{syn}}}{n} + \frac{1}{m}\Big) \quad \bar{v}_m\Big(1 + \frac{1}{m}\Big).$$

The large sample assumption facilitates the use of a normal distribution for the sampling distribution (or the posterior distribution) of $\hat{Q}$ when T_s is used to estimate the variance. The notion is that, in large samples, b_m can be replaced with $\bar{v}_m$. It is di□cult to assess, however, when a large sample assumption is reasonable, because it also depends on the specific analysis being undertaken on the synthetic data, that is, it depends on the analysis's su□cient statistic(s).

The estimators T_p and T_s assume that $n_{\text{syn}} = n$ (or that n_{syn} is constant across the m synthetic data sets in the case of T_s). When using count models as opposed to multinomial models, n_{syn} is stochastic and this assumption is violated. However, in a simulation study unreported here, the effect of varying n_{syn} was found to have a negligible effect on the validity of inferences, for example, confidence intervals still achieved the nominal coverage. Nevertheless, in some cases, new estimators may be required; such estimators may introduce weights $w_1 \ldots, w_m$ that relate to $n_{\text{syn}}^{(1)}, \ldots, n_{\text{syn}}^{(m)}$, the sample sizes of the m synthetic data sets.

4.1.2 Averaging the $m > 1$ data sets before analysing them

When faced with multiple categorical data sets, analysts (and attackers) may either pool or average the data sets *before* analysing them. This is feasible only with contingency tables, as they have the same structure across the $m > 1$ data sets. There are several advantages to doing so. Firstly, it means that analysts only have to undertake their analyses once rather than multiple times, thus leading to reduced computational time. Note, although averaging leads to non-integer "counts", standard software such as the `glm` function in R can typically cope with this and still allow models to be fit. Secondly, model-fitting in aggregated data is often hampered by the presence of zero counts, but either averaging or pooling reduces the proportion of zero counts, since it only takes one non-zero across the $m > 1$ data sets to produce a non-zero when averaged or pooled.

If the data sets are averaged (pooled), then the Raab et al. (2016) estimator T_s would have to be used to estimate variances, so this approach would not be suitable if the conditions required for this estimator do not hold.

When the data sets are averaged and the NBI or PIG is used for synthesis, for a given original count $f_i = N$ ($i = 1, \ldots, K$), the corresponding mean synthetic cell count $\bar{f}_i^{\text{syn}}$ has mean and variance,

$$E(\bar{f}_i^{\text{syn}}) = N \quad \text{and} \quad \text{Var}(\bar{f}_i^{\text{syn}}) = \frac{1}{m}\big(N + \sigma N^2\big), \tag{4.1}$$

as the synthetic data sets are independent.

Thus, for a given original count, the variance of the corresponding mean synthetic count is inversely proportional to m, and linearly related to σ. This means that the minimum obtainable variance when σ alone is tuned - which is achieved as $\sigma \to 0$ and the NBI tends towards its limiting distribution, the Poisson - is N/m. On the other hand, increasing m can essentially take the variance to zero. If m is too large, though, the original counts are simply returned when averaged, which, of course, renders the synthesis worthless. This, perhaps, suggests the suitability of m as a tuning parameter in cases where the original counts are large and there is relatively low risk, such that a relatively small variance suffices.

4.2 Introducing the $\tau_3(k, d)$ and $\tau_4(k, d)$ metrics

When multiple synthetic data sets are generated and the mean synthetic count calculated - which is no longer always an integer - it becomes more suitable to consider the proportion of synthetic counts *within a certain distance of* original counts of k. To allow this, the metrics $\tau_3(k)$ and $\tau_4(k)$ can be extended to $\tau_3(k, d)$ and $\tau_4(k, d)$,

respectively:

$$\tau_3(k,d) \coloneqq p(\mathbf{j} f^{\text{syn}} \quad k\mathbf{j} \quad d\,\mathbf{j}\, f = k), \quad \tau_4(k,d) \coloneqq p(f = k\,\mathbf{j}\, \mathbf{j} f^{\text{syn}} \quad k\mathbf{j} \quad d).$$

The metric $\tau_3(k,d)$ is the probability that a cell count of size k in the original data is synthesized to within d of k; and $\tau_4(k,d)$ is the probability that a cell count within d of k in the synthetic data originated from a cell of k. Unlike k, $d > 0$ does not need to be an integer. By extending the $\tau_1(k)$ metric, such that $\tau_1(k,d)$ is the proportion of synthetic counts within d of k, it follows that $\tau_3(k,d)\tau_2(k) = \tau_4(k,d)\tau_1(k,d)$.

The $\tau_3(k)$ and $\tau_4(k)$ metrics are then special cases of $\tau_3(k,d)$ and $\tau_4(k,d)$, respectively (the case where $d = 0$). For small k, these $\tau(k,d)$ metrics are intended primarily as risk metrics, because they are dealing with uniques or near uniques. However, when d is reasonably large, $\tau_3(k,d)$ and $\tau_4(k,d)$ are, perhaps, better viewed as utility metrics, because they are dealing with the proportion of small counts that are synthesized to much larger counts (which impacts utility).

When $m > 1$ is su□ciently large, tractable expressions for the $\tau_3(k,d)$ and $\tau_4(k,d)$ metrics can be obtained via the Central Limit Theorem (CLT), as the distribution of each mean synthetic count can be approximated by a normal distribution, with mean and variance as given in (4.1). That is, given an original count $f_i = N$ ($i = 1, \ldots, K$), when m is large, the distribution of the corresponding mean synthetic cell count $\bar{f}_i^{\text{syn}}$ is given as:

$$\bar{f}_i^{\text{syn}}\ \mathbf{j}\ f_i = N, \sigma, m \quad \text{Normal}(N, (N + \sigma N^2)/m).$$

This can be used to approximate $\tau_3(k, d)$ and $\tau_4(k, d)$:

$$\begin{aligned}
\tau_3(k,d) &= p(j\bar{f}^{\text{syn}} \quad kj \quad d\,j\, f = k), \\
&= p(\bar{f}^{\text{syn}} < k + d\,j\, f = k) \quad p(\bar{f}^{\text{syn}} < k \quad d\,j\, f = k), \\
&= \Phi\left(\frac{(k+d) \quad k}{\sqrt{(k+\sigma k^2)/m}}\right) \quad \Phi\left(\frac{(k \quad d) \quad k}{\sqrt{(k+\sigma k^2)/m}}\right) \\
&= 2\Phi\left(\frac{d}{\sqrt{(k+\sigma k^2)/m}}\right) \quad 1, \qquad (4.2)
\end{aligned}$$

$$\begin{aligned}
\tau_4(k,d) &= p(f = k\,j\,j\bar{f}^{\text{syn}} \quad kj \quad d) \\
&= \frac{\tau_3(k,d) \quad \tau_2(k)}{\sum_{i=0}^{1} p(jf^{\text{syn}} \quad kj \quad d\,j\, f = i) \quad p(f = i)} \\
&= \frac{\left[2\Phi\left(d\Big/\sqrt{(k+\sigma k^2)/m}\right) \quad 1\right] \quad \tau_2(k)}{\sum_{i=1}^{\infty}\left[\Phi\left((k+d \quad i)\Big/\sqrt{(i+\sigma i^2)/m}\right) \quad \Phi\left((k \quad d \quad i)\Big/\sqrt{(i+\sigma i^2)/m}\right)\right] \quad \tau_2(i)} \qquad (4.3)
\end{aligned}$$

where Φ is which is used to denote the cumulative distribution function (CDF) of the standard normal distribution.

4.3 Empirical synthesis when $m > 1$

The ESCsub data was again synthesized; this time just for the NBI (with $\alpha = 0.01$) but using a range of σ (0, 0.1, 0.5, 2 and 10) and m values (from 1 to 50). The function `rNBI` from the R package gamlss.dist Stasinopoulos et al. (2007) was again used to generate the synthetic counts.

4.3.1 Measuring risk

The $\tau_3(1, d)$ and $\tau_4(1, d)$ metrics (that is, setting $k = 1$), introduced in Section 4.2, were used as risk metrics. Figure 4.1 shows that either increasing m or decreasing σ increases $\tau_3(1, d)$ and $\tau_4(1, d)$ and hence risk. There is an initial fall in the $\tau_3(1, 0.1)$ curves as m increases initially, suggesting lower not higher risk. However, this is just owing to the small d: for example, when $d = 0.1$, the only way to obtain a mean synthetic count within 0.1 of k when, say $m = 5$, is by obtaining a one in each of the five synthetic data sets, compared to just once when $m = 1$.

When m is large, the $\tau_3(k, d)$ and $\tau_4(k, d)$ metrics can be approximated analytically through (4.2), which relies on the CLT. There is uncertainty in both the empirical values (owing to simulation error) and the analytical values (owing to the normal approximation), though the divergences between the empirical and analytical values are small.

In general, then, increasing m or decreasing σ increases risk. This is also shown visually in Figure 4.2, which demonstrates how m and σ can be used in tandem to adjust risk. Here, $\tau_3(1, 0.1)$ is used as the z-axis (risk) but any $\tau_3(k, d)$ or $\tau_4(k, d)$ would give similar results.

4.3.2 Measuring utility

As saturated models are used, increasing m (for a given σ) causes the mean synthetic counts to tend towards the original counts. This can be seen in the Hellinger and Euclidean distances given in Figure 4.3, which show an improvement in general utility when either increasing m or reducing σ.

These measures are equally relevant to risk, too, hence Figure 4.3 reiterates that

risk increases with m. It is fairly trivial, however, that reducing simulation error increases risk and utility. It is more useful to gain an insight into the *rate* at which risk and utility increase with m, that is, the shape of the curves. For example, Figure 4.3, shows that increasing m has greater effect when $\sigma = 1$ than when $\sigma = 0.1$.

The utility of synthetic data can also be assessed for specific analyses by, for example, comparing regression coefficient estimates obtained from a model fit to both the observed and synthetic data. While such measures only assess the synthetic data's ability to support a particular analysis, they nevertheless can be a useful indicator to, for example, the required m needed to attain a satisfactory level of utility.

Here, the estimand of interest is the slope parameter from the logistic regression of age Y (aged 9 = 0, 10 = 1) on language X. A subset of the data were used, as just two of the language variable's seven categories were considered, while the age variable was dichotomised. When estimated from the original data, β_1 - which is a log marginal odds ratio - was equal to -0.0075 with a 95% confidence interval of $(-0.0151, -0.0001)$. Note that, in order to estimate this, it was assumed that the original data constituted a simple random sample drawn from a much larger population. It is hugely doubtful whether such an assumption would be reasonable in practice, but the purpose here was just to evaluate the ability of the synthetic data to produce similar conclusions to the original data.

The analysis was undertaken in the two ways described in Section 4.1. Firstly, the $m > 1$ synthetic data sets were analysed separately and variance estimates were obtained through the estimator T_p. Secondly, the $m > 1$ synthetic data sets were pooled into one data set prior to the analysis and variance estimates were obtained through the estimator T_s.

As can be seen in Figure 4.4, the estimates from T_p were noticeably larger than those from T_s, for small m. This was worrying for the validity of T_s - and the confidence intervals subsequently computed using T_s - especially since the sampling distributions of the estimates were not approximated by a normal distribution, but by a t-distribution with ν_p degrees of freedom, thus widening confidence intervals further. This suggests that the large sample approximation that T_s relies on was not reasonable in this case.

The confidence interval overlap results are presented in Table 4.1. The top frame gives the overlap values from when the data sets are analysed separately, and the bottom frame gives the results from when the data sets are pooled. It can be seen that increasing m broadly results in an increase in the overlap; and that the overlap tends towards 1 as the original and synthetic data confidence intervals converge. The confidence intervals computed using T_s are less robust as those using T_p, which is evident in the zero overlap when $m = 20$ and $\sigma = 10$. This is because, unlike the variance estimator T_p, T_s only considers the within-synthesis variability $\bar{v}_m$, not the between-synthesis variability b_m.

4.3.3 Tuning m and σ in relation to the risk-utility trade-o

The plots in Figure 4.5 show how m and σ can be tuned in tandem to produce synthetic data sets that sit favourably within the risk-utility trade-o. These trade-o plots, though, depend on the metrics used to measure risk and utility. Here, risk was measured by either $\tau_4(1, 0.5)$ or $\tau_4(1, 0.75)$, and utility by either confidence interval overlap (using T_p) or Hellinger distance. The Hellinger distances were standardised onto the interval of [0,1] (by dividing by the largest Hellinger distance observed and

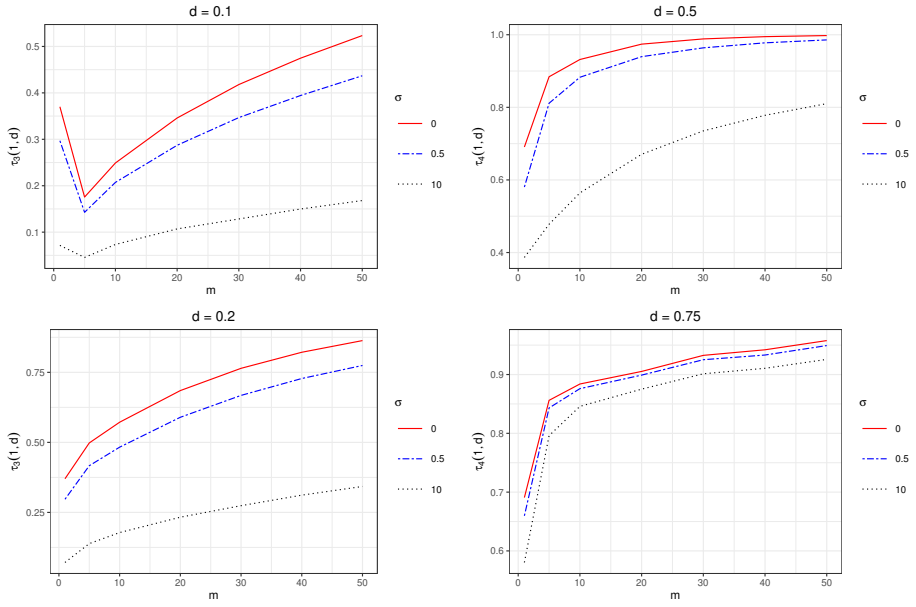

Figure 4.1: The left hand plots give the empirical values of $\tau_3(1, d)$ for $d = 0.1$ and 0.2; the right hand plots give the empirical values of $\tau_4(1, d)$ for $d = 0.5$ and 0.75.

then subtracting from 1, so that 1 and 0 represent maximum and minimum utility, respectively).

It is possible to strictly dominate synthetic data sets over others, that is, obtain lower risk *and* greater utility values. For example, looking at the top-left plot, synthetic data sets generated with $m = 50,\ \sigma = 2$ have higher risk but lower utility than when $m = 20,\ \sigma = 0.5$. These visual trade-o□s are plotted using the empirical results, so are subject to variation from simulation; the confidence interval overlap values, in particular, can be volatile, especially when σ is large.

The intention is that the synthesizer produces such plots before releasing the data. Furthermore, as many metrics can be expressed analytically when using saturated models, they can be produced before the synthetic data is even generated.

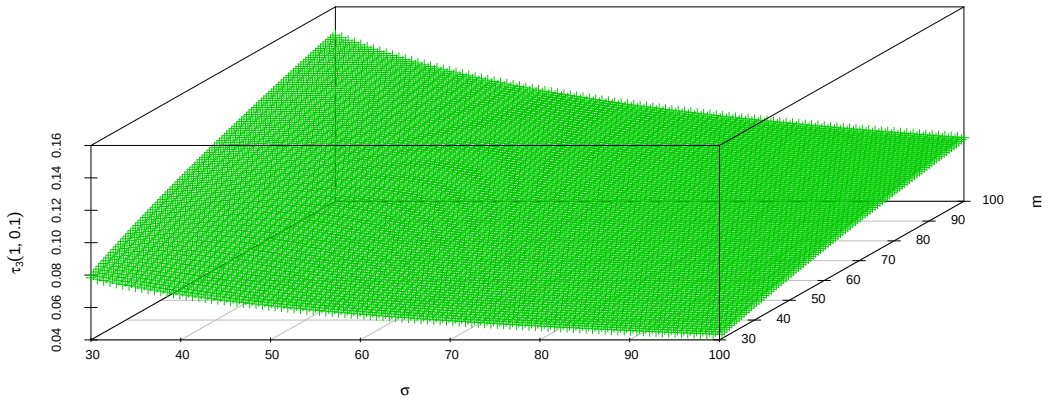

Figure 4.2: The expected $\tau_3(1, 0.1)$ values for m and σ greater than 30.

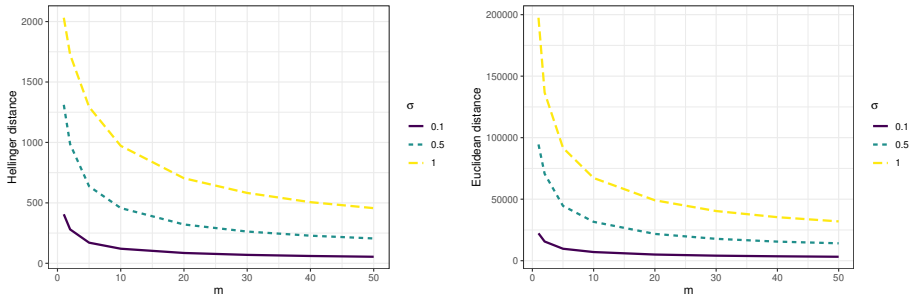

Figure 4.3: The Hellinger and Euclidean distances as m increases, for various values of σ. These plots have been created using the cell counts rather than the cell probabilities, though the two are proportional.

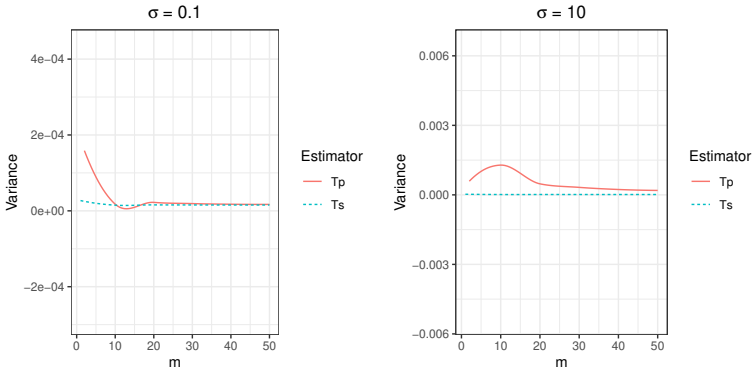

Figure 4.4: The values of the estimators T_p and T_s. For small m, T_p is larger than T_s, before converging for larger m. The estimator T_s remains fairly constant across m.

Table 4.1: The confidence interval overlap results from when: (i) the data sets were analysed separately and T_p was used to estimate confidence intervals; and (ii) the data sets were pooled and T_s was used to estimate confidence intervals.

	$m = 2$	$m = 5$	$m = 10$	$m = 20$	$m = 30$	$m = 40$	$m = 50$
The overlap when the data sets were analysed separately and T_p used							
$\sigma = 0$	0.883	0.901	0.950	0.992	0.990	0.994	0.983
$\sigma = 0.1$	0.533	0.692	0.822	0.898	0.913	0.925	0.917
$\sigma = 0.5$	0.536	0.635	0.778	0.843	0.878	0.909	0.923
$\sigma = 2$	0.000	0.587	0.667	0.726	0.716	0.742	0.780
$\sigma = 10$	0.522	0.535	0.554	0.583	0.604	0.623	0.638
The overlap when the data sets were pooled and T_s used							
$\sigma = 0$	0.881	0.905	0.951	0.988	0.990	0.994	0.983
$\sigma = 0.1$	0.700	0.317	0.802	0.942	0.904	0.920	0.915
$\sigma = 0.5$	0.221	0.344	0.653	0.789	0.864	0.915	0.967
$\sigma = 2$	0.020	0.436	0.856	0.775	0.825	0.809	0.906
$\sigma = 10$	0.000	0.664	0.454	0.000	0.078	0.258	0.465

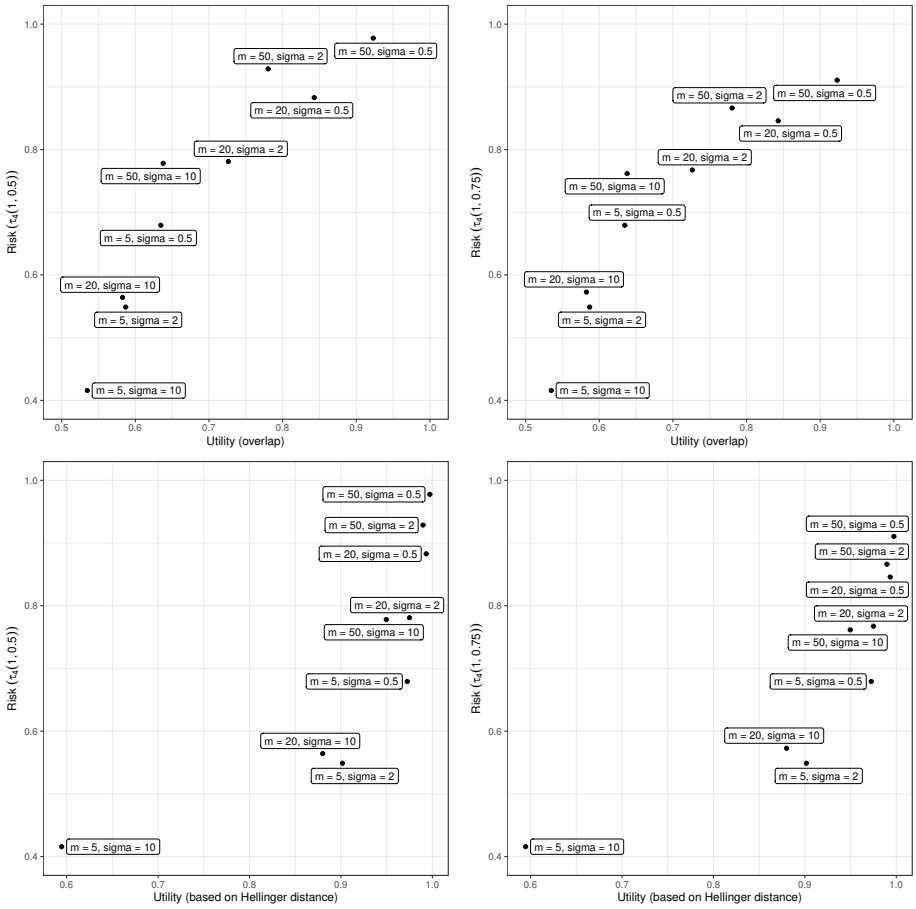

Figure 4.5: Risk-utility trade-o plots to show where various synthetic data sets are located with respect to the risk-utility trade-o. The optimal position in each plot - that is, the lowest risk and the highest utility - is the bottom right corner. To measure risk, the metrics $\tau_4(1, 0.5)$ and $\tau_4(1, 0.75)$ were used. To measure utility, the confidence interval overlap and Hellinger distance were used.

4.3.4 Summary

The synthesis mechanism introduced in Chapter 3 only considered the case of generating $m = 1$ data set. Unless the original data are a simple random sample, however, $m > 1$ data sets must be generated to properly quantify the uncertainty from synthesis; and even if a simple random sample assumption is reasonable, generating $m > 1$ synthetic data sets may allow a more favourable risk and utility profile than $m = 1$. Either way, m can be viewed as a tuning parameter. This chapter has looked at the considerations for $m > 1$, such as extending the $\tau_3(k)$ and $\tau_4(k)$ metrics to $\tau_3(k, d)$ and $\tau_4(k, d)$, and also considering how best to analyse $m > 1$ categorical data sets.

While the use of σ and m allow the synthesizer to set the synthetic counts' variance, they cannot control where the variability falls. The synthesizer does not necessarily want the variability to manifest itself in, say, a heavy right tail in the synthesis distribution's probability mass function. Some movement is required in synthetic counts to reduce risk, but large movements are unnecessary and may have an adverse effect on the data's utility. The use of three-parameter count distributions would provide the synthesizer with control over the skewness in addition to the variance.

Chapter 5

Further approaches to synthesis: the discretized gamma family distribution

5.1 The benefits of using three-parameter count distributions

The use of two-parameter count distributions allow the synthesizer to set the variance, but not, for instance, the skewness or kurtosis, which are also play an important role when producing synthetic data. When the variance is increased to lower risk, it is not desirable for the additional variability to manifest itself in heavier tails: for synthesis, it is preferable to use a light-tailed distribution. Firstly, this is because when a count distribution has a heavy left tail - bearing in mind that count distributions are truncated at zero - it increases the probability of obtaining synthetic counts of zero. Despite providing a natural form of protection - when considering the data in microdata format, the effect of a non-zero being synthesized to a zero is that the

observation is removed from the data - too many zeros can impact negatively on utility: for example, analysts' models may struggle in the presence of an excess of zero counts. To correct the imbalance, it may be necessary to set α (the pseudocount mentioned in Section 3) greater than zero, but this introduces bias. Rather than relying too much on α, it is more e□cient to reduce the number of zeros through using a distribution with a light left tail. Secondly, in a similar way it is preferable, too, to have a light right tail, to reduce the probability of returning excessively large synthetic counts.

Before moving on to the discretized gamma family distribution, which is explored in the remainder of this chapter, the shape of the distribution can be improved upon by using a three-parameter count distribution such as the Delaporte, instead of the NBI and PIG.

5.1.1 The Delaporte distribution

The Delaporte distribution is an explicit count distribution named after Pierre Delaporte, who used the distribution to analyse insurance claims relating to road accidents (Delaporte, 1960). Rather than the *gamma*-Poisson mixture that results in the negative binomial, it is a mixture between the *shifted gamma* and Poisson distributions.

Rigby et al. (2019) introduced a parameterisation whereby, in line with the Poisson, NBI and PIG, the mean of a Delaporte random variable Y is equal to $\mu > 0$. As with the NBI, there is a parameter $\sigma > 0$, but additionally there is a third parameter ν *2* $(0, 1)$.

Its probability mass function is given as:

$$f(y|\mu,\sigma,\nu) = \frac{\exp(\quad \mu\nu)}{\Gamma(1/\sigma)}[1+\mu\sigma(1\quad \nu)]^{\ 1/\sigma} S \tag{5.1}$$

$$\text{where} \quad S = \sum_{j=0}^{y} \binom{y}{j} \frac{\mu^y \nu^{y\ \ j}}{y!} \left[\mu + \frac{1}{\sigma(1\quad \nu)}\right]^{\ j} \Gamma\left(\frac{1}{\sigma}+j\right).$$

The mean and variance are given as:

$$\mathsf{E}[Y \mid \mu,\sigma,\nu] = \mu \quad \text{and} \quad \mathsf{Var}[Y \mid \mu,\sigma,\nu] = \mu + \sigma(1\quad \nu)^2\mu^2.$$

The Poisson(μ) is the limiting distribution as $\nu \ ! \ 1$, for fixed σ; and the limiting distribution as $\nu \ ! \ 0$ is NBI(μ, σ). The parameters can be conditioned upon so that μ controls the mean (first moment), σ controls the variance (second moment) and ν controls the skewness (third moment).

For a given mean and variance (location and scale), the shape of the Delaporte can be adjusted to reduce the heaviness of the tails. This can be seen in Figure 5.1, which gives three Delaporte distributions with the same means and variances but di□erent shapes; the blue dashed line, for example, has a smaller probability at zero as well as a lighter right tail, which would result in fewer zero synthetic counts and fewer unnecessarily large synthetic counts.

Moreover, Figure 5.2 shows how this additional parameter ν can be used to adjust risk. It shows how ν a□ects the $\tau_4(1)$ curve - the proportion of uniques in the synthetic data that were also unique in the original data - when the Delaporte distribution is used to synthesize the ESC_{sub} data.

Other three-parameter distributions give similar flexibility. One example is the Sichel distribution, which is a mixture between the generalised inverse-Gaussian

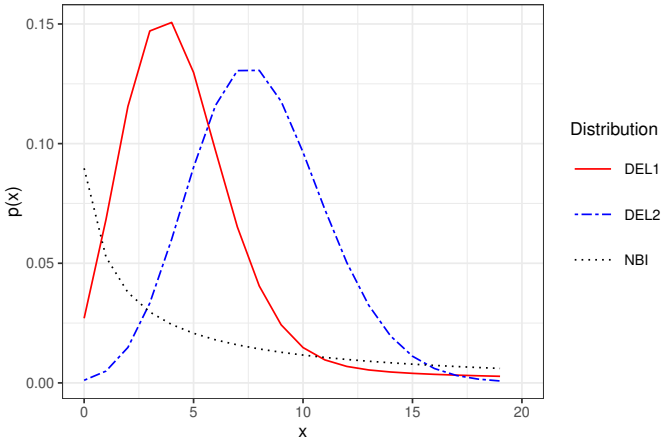

Figure 5.1: The probability mass functions of three Delaporte distributions with the same mean and variance (10 and 510, respectively). The dotted black line is also essentially an NBI distribution, a limiting case of the Delaporte, as $\nu \to 1$.

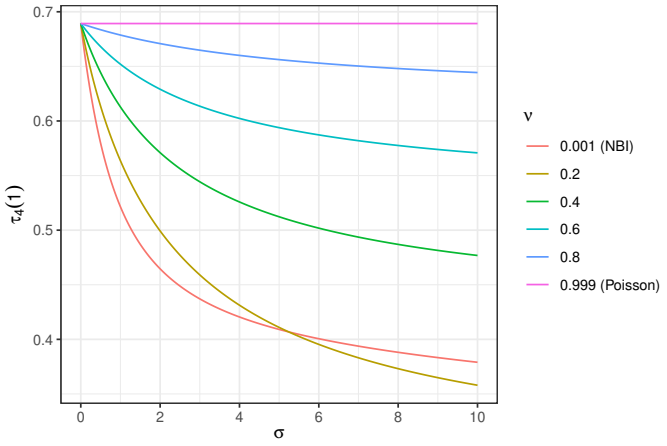

Figure 5.2: How σ and ν affect $\tau_4(1)$ when the Delaporte is used for synthesis.

distribution and the Poisson, and can be viewed as an extension to the PIG.

5.2 The limitations of Poisson-based distributions for synthesis

In a synthesis context, the main issue with many explicit count distributions is that they are overdispersed. Distributions such as the NBI, PIG, Sichel, Delaporte, *et cetera*, are continuously mixed Poisson distributions (as previously mentioned, the mixing distributions are the gamma, inverse-Gaussian, generalised inverse Gaussian, shifted gamma distributions, respectively). Mixing Poisson distributions naturally adds variance because, e□ectively, the Poisson mean is assumed to be stochastic, not fixed. To illustrate, suppose X Poisson(λ) which has mean and variance $\lambda > 0$. Now suppose that $X \mid \gamma$ Poisson($\lambda\gamma$), where γ is a continuous mixing distribution with mean equal to 1 and variance equal to σ^2. Then it follows that:

$$E(X) = E[E(X \mid \gamma)] = \lambda$$

$$\text{Var}(X) = E[\text{Var}(X \mid \gamma)] + \text{Var}[E(X \mid \gamma)] = E(\lambda\gamma) + \text{Var}(\lambda\gamma) = \lambda + \lambda^2\sigma^2 > \lambda,$$

hence mixing Poisson distributions adds variance.

For small original counts, these distributions are fit for purpose: the overdispersion can be utilised to mask the original counts' true values. For larger counts, however, which tend to be lower risk, the variance is often too large, resulting in too much noise being applied. The ideal relationship is one where smaller counts are overdispersed and larger counts are underdispersed. One practical - albeit inelegant - solution is to

specify a unique σ_i for each count ($i = 1, \ldots, K$), or at least specify a unique σ for "large counts" and "small counts". However, this would increase the complexity of the mechanism.

5.3 The use of the discretized gamma family distribution

5.3.1 Creating count distributions through discretization

Flexible count distributions can be produced by discretizing a continuous distribution defined on the interval $(0, \infty)$, an "underused" method according to Rigby et al. (2019). Let W denote a continuous random variable defined on $(0, \infty)$, with probability density function (PDF) $f_W(w)$ and cumulative distribution function (CDF) $F_W(w)$. Then the corresponding discretised version of W, Y, has probability mass function (PMF):

$$f_Y(y) = p(Y = y) = p(y < W < y + 1) = F_W(y + 1) \quad F_W(y) \quad \text{for} \quad y = 0, 1, 2, \ldots. \tag{5.2}$$

If W has an explicit CDF then Y has both an explicit PMF and CDF (Rigby et al., 2019). The properties of a discretized continuous distribution, such as its mean and variance, are not exactly the same as those of the continuous version, owing to the discretization process, but similar enough to allow a rough comparison to be drawn.

Discretization can produce count distributions that are underdispersed. As an example, a well-known two-parameter continuous distribution on $(0, \infty)$ is the gamma

distribution. One such parameterization gives the mean as μ and the variance as $\mu^2\sigma^2$ (hence $\sigma > 0$ here is the coe□cient of variation). The variance-mean ratio is given as $\mu\sigma^2$, thus setting $\sigma < \mu^{-1/2}$ returns a ratio that is less than one, something which cannot be achieved through the majority of standard count distributions.

However, the gamma also demonstrates the need for further flexibility - additional parameters - because the variance-mean ratio is still an increasing function of μ, thus larger counts always have greater variance than smaller counts.

5.3.2 The gamma family (GAF) distribution

The gamma family (GAF) distribution is an extension to the gamma distribution, defined on the interval $(0, \infty)$. It has three parameters ($\mu > 0$, $\sigma > 0$ and $-\infty < \nu < \infty$) rather than two, and, importantly, this third parameter ν can be used to model the variance-mean relationship. The GAF distribution's PDF and CDF are given as:

$$f_W(w \mid \mu, \sigma, \nu) = \frac{w^{\sigma_1^{-2}-1}\exp(-w\sigma_1^{-2}\mu^{-1})}{(\sigma_1^2\mu)^{\sigma_1^{-2}}\Gamma(\sigma_1^{-2})} \qquad \text{for} \quad y > 0$$

$$F_W(w \mid \mu, \sigma, \nu) = \frac{\gamma(\sigma_1^{-2}, w\mu^{-1}\sigma_1^{-2})}{\Gamma(\sigma_1^{-2})} \qquad \text{for} \quad y > 0$$

where $\sigma_1 = \sigma\mu^{\nu/2-1}$ and $\gamma(a,x) = \int_0^x t^{a-1}e^{-t}dt$ is the lower incomplete gamma function. Its mean and variance are:

$$E[Y \mid \mu, \sigma, \gamma] = \mu \quad \text{and} \quad \text{Var}[Y \mid \mu, \sigma, \gamma] = \sigma^2\mu^{\nu}. \tag{5.3}$$

When $\nu < 0$, the variance is a decreasing function of the mean. Figure 5.3 displays the variance-mean relationship for three GAF distributions with di□erent ν values. For a

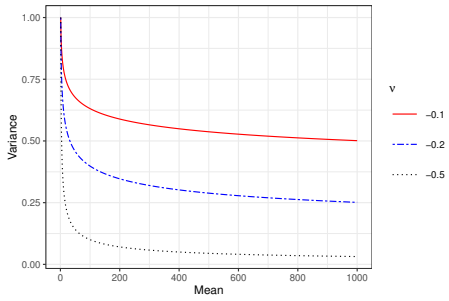

Figure 5.3: The variance-mean relationship for three GAF distributions with different ν values (with $\sigma = 1$).

given μ, the relationship between the variance and (negative) ν is one of exponential decay (or exponential growth if $\nu > 0$). That is, the parameter ν controls the rate at which the variance falls away, that is, the rate of decay.

The quantity σ^2 can be thought of as the variance when $\mu = 1$, which has an intuitive representation in a synthesis context: it is the variability applied to uniques (original counts of one). Therefore, both parameters, σ and ν, are easily interpretable in a synthesis context, which is not the case for many count distributions, and is a powerful argument in favour of the GAF.

5.3.3 The discretized gamma family distribution (DGAF)

The discretized gamma family distribution (DGAF) is the discretized version of the GAF distribution. When using (5.2) to discretize, it will return a distribution that is right skewed relative to the continuous version (see the lower left plot in Figure 5.4). This is most noticeable when the mean is small: for example, when $\mu = 1$, the entire density below the mean in the continuous version is transformed to zero in the discretized version. Consequently, the mean and variance of the DGAF are substantially different to those of the GAF (given in 5.3), which in turn diminishes

the synthesizer's ability to determine expected properties of the synthetic data, *a priori*.

The problem with the discretization mechanism given in (5.2) is that the midpoint of each interval is not equal to the subsequent discretised value. Therefore, a slight change to the discretizion mechanism is proposed:

$$f_Y(y) = \begin{cases} F_W(1/2) & \text{if } y = 0 \\ F_W(y+1/2) \quad F_W(y \quad 1/2) & \text{if } y = 1, 2, \ldots \end{cases} \tag{5.4}$$

which results in the DGAF having the following PMF:

$$f_Y(y) = \begin{cases} \frac{\gamma(\sigma_1^2,(y+1/2)\mu \ \ {}^1\sigma_1{}^2)}{(\sigma_1{}^2)} & \text{if } y = 0 \\ \frac{\gamma(\sigma_1^2,(y+1/2)\mu \ \ {}^1\sigma_1{}^2)}{(\sigma_1{}^2)} \quad \frac{\gamma(\sigma_1^2,(y \ \ 1/2)\mu \ \ {}^1\sigma_1{}^2)}{(\sigma_1{}^2)} & \text{if } y = 1, 2, \ldots ; \end{cases}$$

and which can be viewed in the bottom panel of Figure 5.4. To examine whether the GAF's mean and variance properties given in (5.3) also hold for the DGAF, the means and variances of DGAF distributions with various μ, σ and ν were found through Monte Carlo simulation. These values are presented in Table 5.1: the largest relative discrepancies between the means of the GAF(μ,σ,ν) and DGAF(μ,σ,ν) distributions occur when μ is small, while the largest di□erences between the variances occur when σ is small. The means and variances are su□ciently similar to allow the mean and variance of the DGAF to be approximated by those of the corresponding GAF, thereby facilitating an *a priori* approach to synthesis; that is, if Y DGAF(μ,σ,ν) then

$$E[Y \mid \mu, \sigma, \nu] \quad \mu \quad \text{and} \quad \text{Var}[Y \mid \mu, \sigma, \nu] \quad \sigma^2\mu \ {}^{\nu}. \tag{5.5}$$

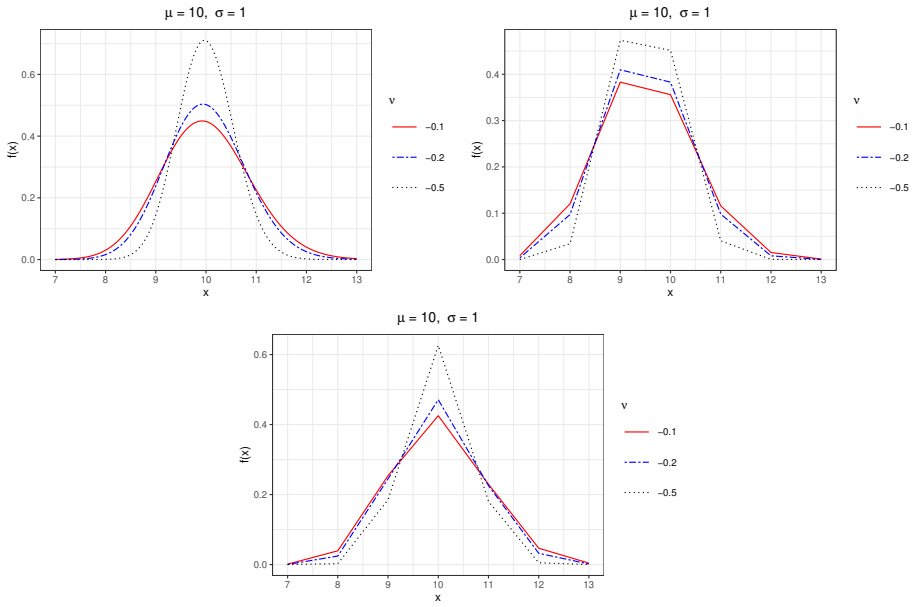

Figure 5.4: The PDFs of GAF and DGAF distributions with $\mu = 10$. The top-left plot is the GAF; the top-right and bottom plots are the DGAF discretized using (5.2) and (5.4), respectively.

E□ectively, when drawing a DGAF random variate, the discretization mechanism given in (5.2) takes a GAF random variate and always rounds *down* to the nearest integer, whereas that given in (5.4) rounds to the nearest integer.

5.3.4 Additive smoothing with the DGAF

A disadvantage, however, when using the DGAF is that additive smoothing - adding $\alpha > 0$ to non-structural zero original counts - is less e□ective. For a given $\nu < 0$ the variance-mean relationship of the DGAF, as with the GAF, decays exponentially. Its variance is equal to σ^2 when $\mu = 1$ and tends to infinity as μ tends to zero. Interpolating between 0 and 1, it follows that when $\mu = 0.01$, say - as is the case when $\alpha = 0.01$ - the variance is huge, that is, the distribution becomes highly overdispersed

Table 5.1: The mean and variance of various DGAF(μ,σ,ν) distributions when discretized using the mechanism given in (5.4) alongside the mean and variance for the corresponding GAF(μ,σ,ν) distributions in parentheses.

		Mean of DGAF(μ,σ,ν) dist.		Var. of DGAF(μ,σ,ν) dist.	
		$\sigma = 2$	$\sigma = 10$	$\sigma = 2$	$\sigma = 10$
$\nu = 0$	$\mu = 1$	0.96 (1)	1.00 (1)	4.11 (4)	100.31 (100)
	$\mu = 10$	10.00 (10)	9.99 (10)	4.08 (4)	100.13 (100)
	$\mu = 20$	20.00 (20)	20.00 (20)	4.09 (4)	100.20 (100)
$\nu = -1$	$\mu = 1$	0.96 (1)	0.99 (1)	4.12 (4)	99.08 (100)
	$\mu = 10$	10.00 (10)	10.00 (10)	0.48 (0.4)	10.09 (10)
	$\mu = 20$	20.00 (20)	20.00 (20)	0.27 (0.2)	5.09 (5)

and has very heavy tails.

Now, for the NBI, setting $\alpha = 0.01$ means, in practice, that roughly one in a hundred zero counts are synthesized to one. However, for the DGAF, owing to the larger variance, far fewer zeros - typially, less than one in a thousand - are synthesized to non-zeros; and when they are, unlike with the NBI, they are synthesized to a range of non-zero counts, not just mainly ones. Thus, as the role of α is to reduce the proportion of synthetic counts of zero, it is less effective when using the DGAF. This can be seen in Figure 5.5, which compares the PMFs of the DGAF and NBI distributions when $\mu = 0.01$. For the NBI (dotted line), the probability of obtaining a one is slightly less than 0.01 and the probabilities for all counts greater than one are effectively zero. For the DGAF (solid line) all probabilities are effectively zero, though the right hand plot shows that larger counts are relatively more likely from the DGAF. Using the DGAF, then, may necessitate a larger α, which would induce greater bias into the synthesis.

Although α is less effective, there would not necessarily be an inflated proportion

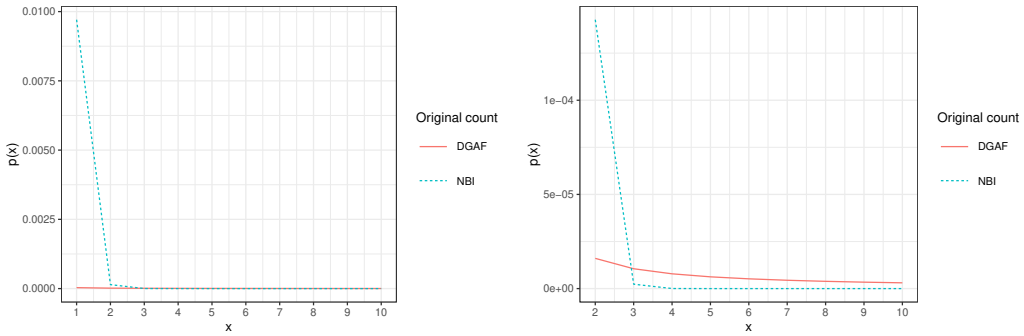

Figure 5.5: The PMFs, when the mean is 0.01 ($\alpha = 0.01$), of the DGAF ($\sigma = 1$ and $\nu = \ \ 0.25$) and NBI ($\sigma = 2$) distributions. The left and right hand plot gives the probabilities for counts greater than zero and one, respectively. The probabilities of obtaining a zero, excluded here for visual purposes, are 0.9998 and 0.9901 for the DGAF and NBI, respectively.

of zero synthetic counts when using the DGAF compared with the NBI. The number of zeros is o□set by the fact that fewer non-zero original counts are synthesized to zero.

5.3.5 Setting the tuning parameters

Each data set is unique in its confidentiality and structure. In categorical data this uniqueness is generally captured by the vector of original counts. While the original counts are not interchangeable - they conform to a structure that is governed by the contingency table - when using saturated models, in particular, they can be decomposed - and characterised by - the proportion of zeros, ones, twos, *et cetera*, in addition to the number of counts (cells) K. After omitting structural zeros, administrative databases, for example, due to their size and sparsity, typically have large K and a substantial proportion of both zero counts and large counts. In contrast, survey data sets usually have relatively small K, fewer zeros and fewer very large

counts. In general, there is more heterogeneity in the range of counts observed in administrative data than in survey data.

In practice, the setting of the synthesis mechanism's tuning parameters is a policy decision made by the synthesizer. Essentially, the decision hinges on the confidentiality and structure. The synthesizer needs to establish the level of risk deemed to be acceptable and be familiar with the structure of the data - the proportion of zeros, ones, twos, *et cetera*.

A rough guide to the setting of the tuning parameters is as follows. A good starting point is the setting of m, the number of synthetic data sets. If the original data are not a simple random sample (SRS), $m > 1$ would be required to obtain valid inferences (see, for example, Drechsler, 2011); though $m = 1$ may suffice if a SRS assumption is reasonable (see Raab et al., 2016). The synthesizer can next focus on small counts, which tend to be those most at-risk, and establish an acceptable level of protection. The parameter σ is effective here: increasing σ adds variance - and hence protection - to all synthetic counts, though small counts in particular would be largely unaffected by later changes made to the parameter ν. A relatively large m would typically necessitate a relatively large σ to offset the additional risk. The parameter ν is intended to be used for fine-tuning: it can be used to reduce the noise applied to larger counts without largely affecting the smaller counts. Finally, the parameter α is primarily intended as a way to allow zero counts to be synthesized to non-zeros; in some instances $\alpha = 0$ would suffice.

The synthesizer can use the *a priori* nature of the synthesis - that is, expected values of risk and utility metrics - to gain an insight into what tuning parameter values to use. The objective is to find a combination of values that yields synthetic

data with high utility but low risk. This can be achieved by assessing risk and utility over a range of possible values and selecting the best option.

5.3.6 Combining risk and utility

In the synthetic data literature, risk and utility are often treated as separate entities; for instance, post-synthesis the synthesizer evaluates first one then the other. As risk and utility are intertwined - maximizing one typically minimizes the other - they can be considered simultaneously through the utility-risk ratio (UR ratio), where a utility function is divided by a risk function (or *vice versa*). As with the risk and utility metrics themselves, the UR ratio can be expressed analytically as functions of the synthesis mechanism's tuning parameters.

A key consideration is the rate at which the UR ratio changes with respect to the tuning parameters. If the UR ratio is increasing, for example, it suggests the gains in utility outweigh the additional risk. The synthesizer can tune a parameter, for example, such that the UR ratio is maximized, which would occur when its gradient is zero. Hence the calculation of partial derivatives could be useful here.

To illustrate the notion, if risk is measured by $1 \quad \tau_3(1)$ and utility by the inverse of the mean squared error loss $1/L_1$, then the UR ratio is:

$$\mathsf{UR}(\sigma,\nu) = \frac{\mathsf{Utility}(\sigma,\nu)}{\mathsf{Risk}(\sigma,\nu)} = \frac{1/L_1(\sigma,\nu)}{1 \quad \tau_3(1\,j\,\sigma,\nu)}$$

when $m = 1$ and $\alpha = 0$. This UR ratio is shown graphically in the left-hand plot of Figure 5.6 (using the byssinosis data used in the next section). The UR ratio in this instance has a negative gradient, suggesting increasing σ leads to a less favourable position in relation to the utility-risk trade-o□. The UR ratio, however, depends on

how risk and utility are measured; the use of other risk and utility functions would yield di□erent results.

Another way to present this notion of combining risk and utility is through a utility-privacy plot, where utility is plotted against privacy. An example is given in the right-hand plot of Figure 5.6, where utility is measured by $1/L_1$, standardized such that 0 and 1 represent minimum and maximum utility, respectively; and privacy is measured by $1 \quad \tau_3(1)$, such that all possible values lie within the unit square $[0,1] \quad [0,1]$. The optimal position is at the point furthest from the origin, that is, the point (1,1). Figure 5.6 shows that decreasing ν shifts the curve to the right, allowing both higher utility and higher risk, suggesting that smaller ν is better in this instance. Yet, care needs to be taken because other metrics would give di□erent results. For comparison, Figure 5.6 also includes the curve for when the Poisson-inverse Gaussian (PIG) is used for synthesis. The PIG is "dominated" by the DGAF distribution in the sense that the DGAF has both higher utility and privacy.

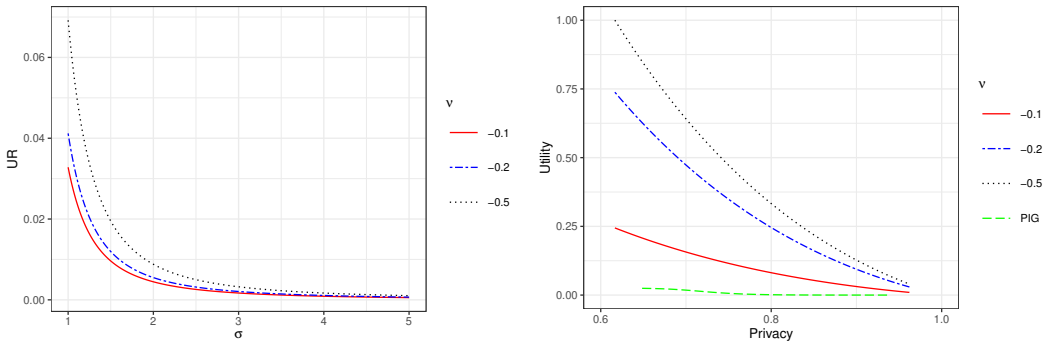

Figure 5.6: Left: the UR ratio against σ for various ν. Right: a utility-privacy plot where utility is measured by $1/L_1$ and privacy by 1-$\tau_3(1)$. The optimal position is at the point (1,1), which represents maximum utility and privacy.

5.4 Application: syntheses of two contrasting categorical data sets

The effectiveness of synthesis methods depend on the nature of the data being synthesized. Categorical data sets exist in a variety of forms: those originating from surveys tend to have relatively few cells. Administrative categorical data sets, on the other hand, tend to be much larger and sparser, with a higher proportion of zero counts; they are more likely to contain structural zeros than survey data sets as they are not prepared with statistical analysis in mind; and they are also more likely to contain very large counts (in excess of 1000, say), as they can be more akin to census data in terms of n, the number of individuals included. The effectiveness of the DGAF, therefore, is demonstrated over two contrasting data sets. In addition to the ESC_{sub} data set, a large administrative database described in Section (3.5), a data set derived from a survey of US cotton-industry workers relating to the respiratory disorder, byssinosis, was also synthesized.

5.4.1 The byssinosis data set

Byssinosis, also known as Brown Lung, is a respiratory disorder caused by dust from fibres such as cotton. Higgins and Koch (1977) introduced a categorical data set arising from a survey of US cotton-industry workers. The survey had $n = 5419$ respondents and $p = 6$ variables. Among these variables, there was a clear response - whether workers suffered from byssinosis - in addition to five categorical explanatory variables relating to workers' race, sex, smoking status, length of employment and exposure to dust in the workplace. Table 5.2 provides a summary of these variables.

Table 5.2: A summary of the variables in the byssinosis data set (Higgins and Koch, 1977).

Name	Description	Categories
BYSSINOSIS (Y)	Worker a□ected by byssinosis?	yes=1, no=0
RACE (X_1)	Worker's race	non-white=1, white=0
SEX (X_2)	Worker's sex	female=1, male=0
SMOKING (X_3)	Worker's smoking status	smoker=1 non-smoker=0
EMPLOYMENT (X_4)	Worker's employment in years	< 10=2, 10-20=1, > 20=0
DUST (X_5)	Worker's workplace's dustiness	most=2, less=1, least=0

The data can be expressed as a contingency table with 144 cells, of which 44 have a count of zero, and none of which are structural zeros.

By contrast, the ESC_{sub} data set, of course, has records for approximately $n = 8.2 \quad 10^6$ schoolchildren, and, as a contingency table, has approximately $3.5 \quad 10^6$ cells, around 90% of which are zeros (again, none of which are identified as structural zeros).

Figure 5.7 compares the cell sizes in the byssinosis and ESC_{sub} data sets. The ESC_{sub} data set has a substantial proportion of zero counts, along with some very large counts (note the faint mark for the > 100 bar). The cell sizes in the byssinosis data, on the other hand, are more homogeneous; for example, there is roughly the same number of counts in the 2-5 range as there are in the 21-50 range.

5.4.2 Generating the synthetic data sets

Both data sets were synthesized using the DGAF distribution and - for comparison - the NBI distribution, using a range of tuning parameter values (for the DGAF, the various combinations of $\sigma = 0.5$, 1 and 2 and $\nu = 0$, -0.25 and -0.5; and for the NBI, $\sigma_{\text{NBI}} = 0.5$, 1 and 2). For both distributions, three values of m were considered (3, 5 and 10) and $\alpha = 0.01$ was added to all zero original counts (as no structural zeros

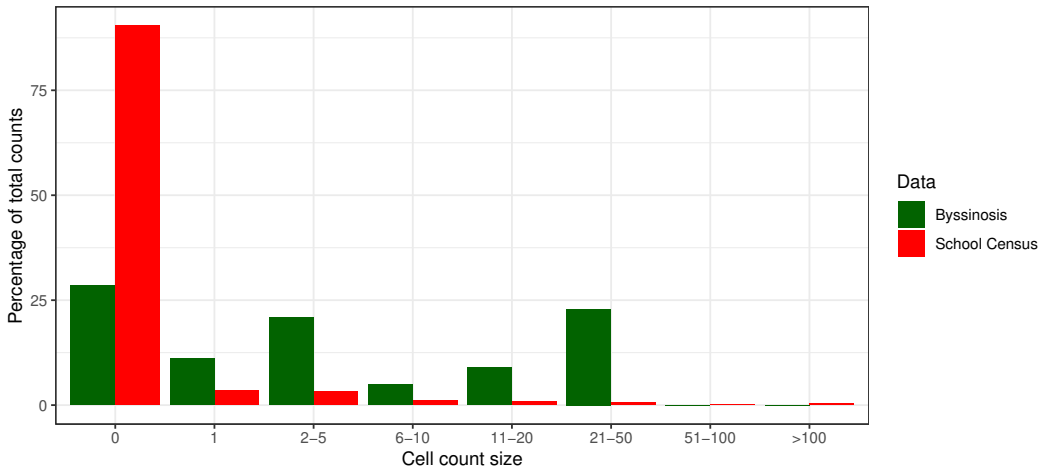

Figure 5.7: A comparison of the cell sizes in the byssinosis and ESC_{sub} data sets.

were identified).

The function `rGAF` from the `R` package gamlss.dist (Stasinopoulos et al., 2007) was used to generate gamma family random variates, which in turn were discretized using the mechanism in (5.4). Similarly, the function `rNBI` from gamlss.dist was used for the NBI.

The byssinosis synthetic data sets were generated instantly, that is, the CPU time was negligible. The synthesis was reasonably quick for the ESC_{sub}, too, with a CPU time of 66 seconds for the DGAF and 41 seconds for the NBI.

5.4.3 The results

The wild fluctuations in Figure 5.8 clearly demonstrate how the NBI applies too much noise when synthesizing large counts. The DGAF, by contrast, applies roughly the same noise as the NBI to small counts, but far less variability to large counts. This immediately suggests the suitability of the DGAF for synthesis, where the general objective is to perturb small counts without over-perturbing large counts. The top

two plots of Figure 5.8 also show the role of the parameter ν when using the DGAF, which is to reduce the noise applied to larger counts (the variability when $\nu = \ 0.1$ is greater than the variability when $\nu = \ 0.5$).

Before evaluating utility it is arguably more important to consider risk. When evaluating risk in categorical synthetic data it is useful to focus on original counts of one (uniques). The bar chart in the top-right of Figure 5.9 shows the range of synthetic counts obtained from original counts of one for the ESC_{sub} data set; the bar height at one gives the empirical value of the metric $\tau_3(1)$, the proportion of ones that were synthesized to one, which was lower for the DGAF than for the NBI (0.16 vs. 0.25). The bars in Figure 5.9 would always be roughly the same regardless of the data being synthesized: as the number of cells increases, these proportions would tend towards their true values - which can be derived analytically, for example, for the DGAF with a given σ and ν, the true value of $\tau_3(1)$ is the probability of obtaining a one from a DGAF(1, σ, ν) distribution - as simulation error is reduced.

Arguably more important from a risk perspective is $\tau_4(1)$, the proportion of synthetic counts of one that originated from a one. This proportion *does* change - and quite considerably - with the data being synthesized, as demonstrated in the bar charts in the top-right of Figures 5.10 and 5.11. For the byssinosis data set, the empirical $\tau_4(1)$ value was lower for the DGAF than for the NBI (0.30 vs. 0.67); while the reverse occurred for the ESC_{sub} data set (0.47 vs. 0.34). These differences stem from the nature of the original data, particularly the number of zeros, in addition to the differing distributional properties. Firstly, the DGAF has a much lower $\tau_4(1)$ value than the NBI when synthesizing the byssinosis data set, mainly because of its lower $\tau_3(1)$ value: fewer ones were synthesized to one, so it roughly follows that

fewer synthetic ones originated from ones. For the ESC_{sub} data, $\tau_4(1)$ was higher for the DGAF than the NBI, which was because, as discussed in Section 5.3.4, α is less effective at converting zero original counts into non-zero synthetic counts - which can be seen in the top-left bar chart of Figure 5.9 - resulting in fewer zeros being synthesized to ones, thus fewer ones overall, thus increasing the probability that a synthetic count of one originated from a one. The difference between the $\tau_4(1)$ values is smaller for the ESC_{sub} data set than it is for the byssinosis data set suggesting that the DGAF outperforms the NBI in relation to this metric.

The lower bar charts in Figures 5.9, 5.10 and 5.11 further demonstrate the superiority of the DGAF with dealing with larger counts. For example, Figure 5.9 shows that, for original counts in the range 100 to 200, nearly all corresponding synthetic counts were also in the range 100 to 200. Similarly, Figures 5.10 and 5.11 show that a substantial proportion of synthetic counts in the range 100 to 200 originated from counts in the range 100 to 200.

5.4.4 Evaluating specific utility

Aitkin et al. (2009) fitted a logistic regression (logit) model to the byssinosis data set to assess the factors associated with developing byssinosis. After omitting insignificant covariates through a process of backward elimination, and dichotomizing the EMPLOYMENT and DUST variables by collapsing the "10-20" and "$>$20" and "less" and "least" categories, respectively, the reduced model had six parameters: an intercept, four main effects relating to each explanatory variable except RACE, and an interaction between DUST and SEX.

This analysis was repeated on the synthetic data sets for various tuning parameter

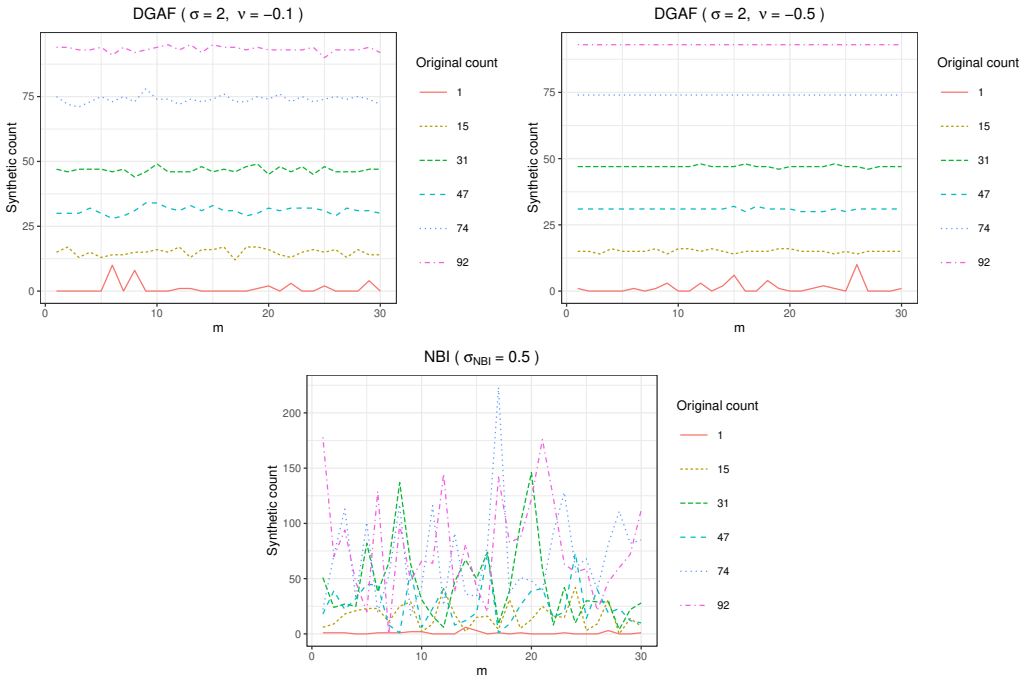

Figure 5.8: For a selection of original counts in the byssinosis data, $m = 30$ corresponding synthetic counts. The top plots relate to the DGAF and the lower plot relates to the NBI.

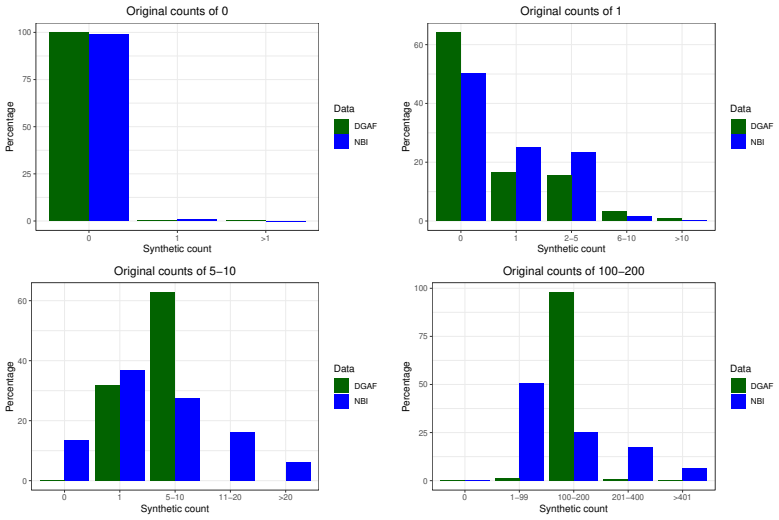

Figure 5.9: The range of synthetic counts obtained for a given size of original count when synthesizing the ESC_{sub} data set (the corresponding plots for the byssinosis data set are similar). The top left plot relates to original counts of 0; the top right to original counts of 1; the bottom plots to original counts of 5-10 and 100-200.

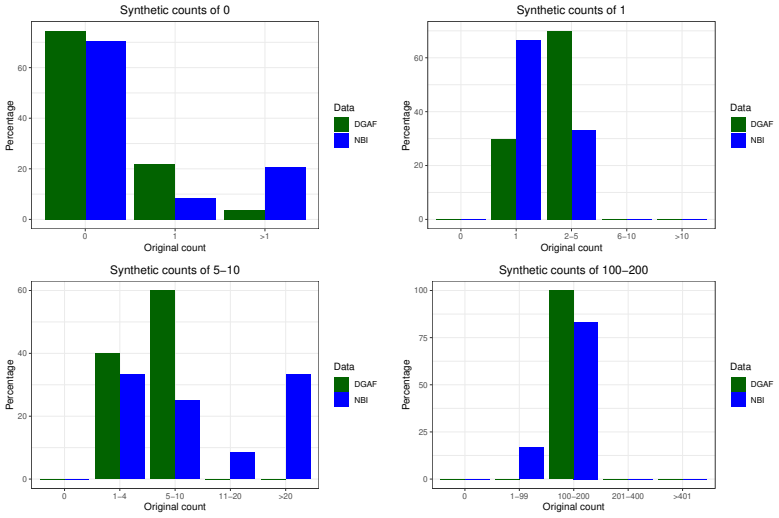

Figure 5.10: The range of original counts obtained for a given size of synthetic count when synthesizing the byssinosis data set. The top left plot relates to synthetic counts of 0; the top right to original counts of 1; the bottom plots to synthetic counts of 5-10 and 100-200.

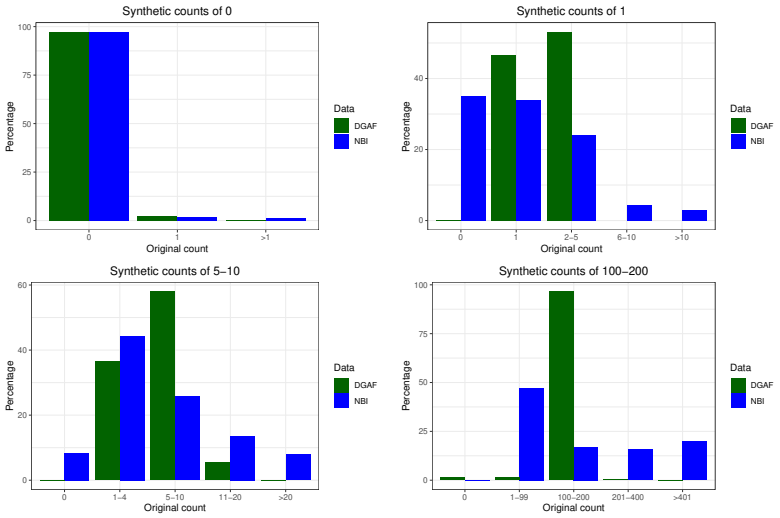

Figure 5.11: As per Figure 5.10 but for the ESC_{sub} data set.

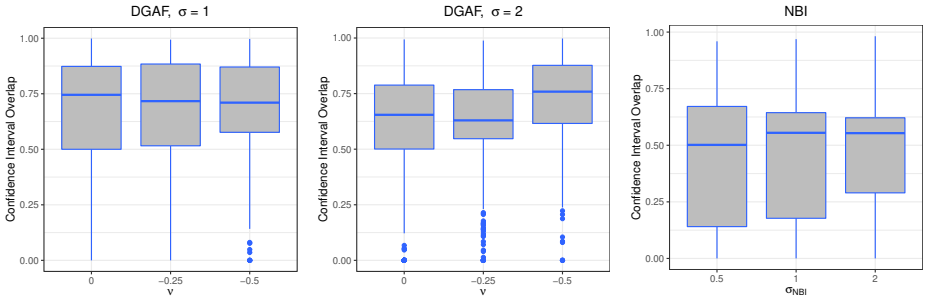

Figure 5.12: The parameter estimates' overlap for a model fitted to the original and synthetic ESC_{sub} data sets. The left and middle plots are for the DGAF and the right plot is for the NBI.

values. For each parameter, the $m > 1$ point estimates $q_1, \ldots, q_m$ were averaged to give an overall estimate $\bar{q}_m$, with some estimates that were essentially infinite discarded to prevent biased averages. The variance of $\bar{q}_m$ was estimated through $T_p = b_m/m + \bar{v}_m$ (Reiter, 2003), where b_m is the sample variance of the $m > 1$ point estimates, and $\bar{v}_m$ is the mean of the point estimates' variances. As this data set was reasonably large, the estimates' sampling distributions were approximated by normal distributions.

The confidence interval overlap (see Karr et al., 2006; Snoke et al., 2018) for the parameter estimates obtained from the original and synthetic data using the DGAF are given in Table 5.3. In general, for the DGAF reducing the variance by either increasing m, decreasing σ or making ν more negative, yields estimates that are closer to those obtained from the original data; for example, looking at β_2, when $\sigma = 0.5$, $\nu = \ \ 0.5$ and $m = 10$, the overlap was 0.990, very close to 1. The overlap values for the NBI are notably lower than those of the DGAF, with values ranging from 0.03 to 0.89 compared with 0.73 to 1.00.

In a similar way, a model was fitted to the original and synthetic ESC_{sub} data sets. Unlike the actual ESC data, which contains, for example, variables relating to pupils' exam results, the ESC_{sub} data set lacks a realistic response variable, so the analysis used in Chapter 3 was repeated.

The confidence interval overlap for the log-linear model parameter estimates are given as boxplots in Figure 5.12. The median overlap, for example, when using the DGAF was always higher than that of the NBI. Interestingly, altering the DGAF or NBI's tuning parameter appears to have little e□ect on the overlap; for example, for the DGAF with $\sigma = 1$, contrary to what one would expect, the median overlap decreases as ν decreases. This can be attributed to simulation uncertainty.

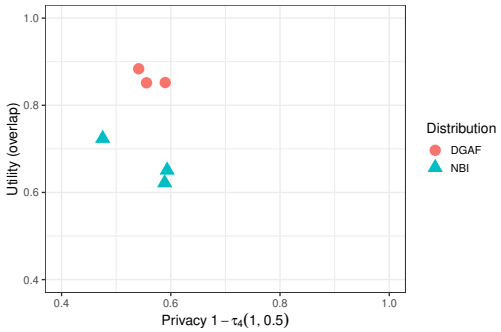

Figure 5.13: A utility-privacy trade-off plot when synthesizing the byssinosis data. The optimal position - that is, the lowest risk and the highest utility - is the top-right corner (the point (1,1)).

The results of these two analyses further emphasize how higher utility can be achieved with the DGAF. Yet higher utility is meaningless without considering the increase in risk. Figure 5.10 shows that $\tau_4(1)$ - a measure of risk - is also lower for the DGAF than the NBI when synthesizing the byssinosis data. Figure 5.13, for example, gives a utility-privacy plot for the byssinosis data, where overlap (utility) is plotted against $1 \quad \tau_4(1, 0.5)$ (risk). Points relating to the DGAF (denoted by circles), are situated further away from the origin than the points relating to the NBI (denoted by triangles), suggesting a better risk-utility trade-off.

5.4.5 Summary

Existing approaches for synthesizing data at the tabular level - including even the approach set out in Chapters 3 and 4 - tend to use the Poisson or Poisson-based distributions. Yet these distributions' variances are always greater than or equal to the the mean, and, as a result, more noise is applied to larger counts than smaller counts. But this is contrary to the objective of data synthesis, where larger counts are typically lower risk than smaller counts, and therefore require less perturbation.

Table 5.3: The overlap for the parameter estimates from the original and synthetic data.

			β_1	β_2	β_3	β_4	β_5	β_6
		$\sigma = 0.5, \nu = 0$	0.970	0.987	0.843	0.984	0.996	0.857
		$\sigma = 0.5, \nu = 0.25$	0.978	0.970	0.955	0.959	0.992	0.945
		$\sigma = 0.5, \nu = 0.5$	0.965	0.992	0.922	0.862	0.979	0.951
		$\sigma = 1, \nu = 0$	0.958	0.984	0.802	0.925	0.920	0.824
	$m = 3$	$\sigma = 1, \nu = 0.25$	0.981	0.782	0.918	0.890	0.992	0.884
		$\sigma = 1, \nu = 0.5$	0.950	0.802	0.913	0.887	0.965	0.983
		$\sigma = 2, \nu = 0$	0.949	0.801	0.807	0.891	0.906	0.787
		$\sigma = 2, \nu = 0.25$	0.976	0.915	0.946	0.732	0.967	0.918
		$\sigma = 2, \nu = 0.5$	0.865	0.926	0.808	0.816	0.966	0.825
		$\sigma = 0.5, \nu = 0$	0.973	0.986	0.907	0.967	0.993	0.909
		$\sigma = 0.5, \nu = 0.25$	0.966	0.990	0.973	0.979	0.991	0.950
		$\sigma = 0.5, \nu = 0.5$	0.956	0.996	0.931	0.907	0.975	0.942
		$\sigma = 1, \nu = 0$	0.968	0.974	0.858	0.984	0.947	0.867
DGAF	$m = 5$	$\sigma = 1, \nu = 0.25$	0.993	0.934	0.891	0.980	0.994	0.888
		$\sigma = 1, \nu = 0.5$	0.964	0.858	0.951	0.882	0.993	0.990
		$\sigma = 2, \nu = 0$	0.920	0.903	0.805	0.938	0.868	0.789
		$\sigma = 2, \nu = 0.25$	0.889	0.914	0.926	0.845	0.894	0.883
		$\sigma = 2, \nu = 0.5$	0.922	0.794	0.836	0.919	0.933	0.835
		$\sigma = 0.5, \nu = 0$	0.980	0.996	0.919	0.960	0.957	0.927
		$\sigma = 0.5, \nu = 0.25$	0.992	0.963	0.966	0.994	0.998	0.967
		$\sigma = 0.5, \nu = 0.5$	0.975	0.990	0.955	0.977	0.981	0.962
		$\sigma = 1, \nu = 0$	0.960	0.934	0.921	0.992	0.955	0.926
	$m = 10$	$\sigma = 1, \nu = 0.25$	0.994	0.945	0.939	0.970	0.987	0.943
		$\sigma = 1, \nu = 0.5$	0.994	0.967	0.935	0.915	0.984	0.933
		$\sigma = 2, \nu = 0$	0.942	0.806	0.806	0.938	0.867	0.801
		$\sigma = 2, \nu = 0.25$	0.947	0.971	0.895	0.846	0.945	0.904
		$\sigma = 2, \nu = 0.5$	0.927	0.861	0.888	0.972	0.954	0.850
		$\sigma = 0.5$	0.810	0.862	0.821	0.719	0.887	0.773
	$m = 3$	$\sigma = 1$	0.251	0.031	0.809	0.619	0.736	0.756
		$\sigma = 2$	0.641	0.186	0.850	0.664	0.661	0.824
		$\sigma = 0.5$	0.746	0.808	0.852	0.781	0.890	0.837
NBI	$m = 5$	$\sigma = 1$	0.679	0.416	0.785	0.701	0.593	0.599
		$\sigma = 2$	0.616	0.642	0.794	0.590	0.701	0.782
		$\sigma = 0.5$	0.801	0.877	0.853	0.720	0.863	0.850
	$m = 10$	$\sigma = 1$	0.747	0.559	0.821	0.817	0.803	0.668
		$\sigma = 2$	0.711	0.695	0.857	0.626	0.763	0.786

This chapter, through focusing on the discretized gamma family distribution, explores the ideal properties of a distribution used for synthesis. The DGAF provides the synthesizer with control of the variance-mean relationship, which allows noise to be applied more finely, thereby producing synthetic data with greater utility for a given level of risk.

Chapter 6

Discussion

While this thesis has focused on the specific case of using a saturated synthesis model, the benefits of this approach carry over to when a saturated model is *not* used. Expected properties of synthetic data can still be derived or estimated, *a priori*, when, for example, original counts are smoothed through a log-linear model. The difference - downside - in this instance is that expected synthetic counts would not be equal to the original counts, so metrics cannot be neatly expressed analytically. This may not be a problem, though, if metrics are built into software. Similarly, the benefits of the DGAF - and its ability to control the mean-variance relationship - would carry over, because the notion of perturbing larger counts less than smaller counts remains.

Chapters 3 and 4, in setting out the use of overdispersed count distributions, used the negative binomial and Poisson-inverse Gaussian distributions for synthesis. Yet such distributions - as well as other Poisson-based distributions and the Poisson itself - are sub-optimal. Simply too much noise is applied to larger cell counts. The DGAF, given its flexibility surrounding the mean-variance relationship, the intuitiveness of

its parameters and not a great deal of extra overhead, is, quite frankly, more e□ective.

On a di□erent note, the estimator T_s (Raab et al., 2016) used throughout this thesis has a rather contentious history and continues to divide opinion in the privacy community. Its advantage is obvious: it eliminates the need for $m > 1$. Its two assumptions - two assumptions that are not required by its alternative T_p (Reiter, 2003) - are that the original data are a simple random sample and are su□ciently large. The former, I think, is not too much of an issue; many data sets requiring synthesis *are* simple random samples. It does need consideration in relation to administrative data, though I believe there are solutions around this: for example, a weighted subsample could be taken from the administrative data. The large sample assumption, however, I have found to be more problematic. It is unclear when such as assumption is reasonable, especially in relation to categorical data and contingency tables. It appears to be linked to a wider issue with tabular data analysis that stretches beyond synthetic data. For example, it is often unclear whether "sample size" in tabular data sets should refer to the number of individuals n or the number of cells K. This has implications when using, for example, the Bayesian Information Criterion (BIC), which depends on n. Thus in turn can a□ect model selection. Incidentally, I encountered this issue when working on a project in the area of Multiple Systems Estimation, where selecting interactions is intrinsic to the estimates obtained for the size of the unknown population. Using K rather than n is, I think, more intuitive in categorical data, given that it is the counts that are modelled.

Throughout the empirical syntheses in this thesis, only variables at the pupil-level were considered. The actual English School Census, though - as with other administrative databases - has a hierarchical structure, with variables at a higher level

of aggregation. Incorporating these higher-level variables into the synthesis presents challenges from a modelling perspective. The use of saturated models within a contingency table structure would naturally preserve relationships to a certain extent, but an alternative would be to use a multi-level model.

This thesis has focused on the large, categorical nature of administrative databases. Yet there are other considerations; some of which a□ect the ability to implement these methods in practice.

For example, administrative databases may contain continuous variables, in addition to categorical ones, and, of course, continuous data cannot be tabulated in the same way as categorical data. Lindsey (1974) o□ers a way forward here, though. Continuous observations can be viewed as observations from a multinomial distribution and essentially treated as categorical. Or observations from a continuous variable can be grouped, to provide a distinct set of categories. Intervals may need to be created to categorize the continuous values. Incidentally, the interval widths would be associated with risk: the smaller the interval, the greater the risk. Moreover, an advantage of categorizing in this way is that the synthesizer can reduce the proportion of zero cells by increasing or decreasing interval widths as appropriate. Note that post-synthesis the data may need to be converted back to continuous, which can easily be achieved by, for example, taking a draw from a uniform distribution where the endpoints are the limits of the interval.

Moreover, skip patterns may present a problem in practice. These are a series of questions that are only answered given responses from previously asked questions; for example, in relation to the English School Census any questions relating to GCSEs are only relevant to pupils in secondary schools. Essentially, skip patterns lead to

structural zeros. A solution to this could be to split the table into parts and synthesize each part separately, including structural zeros in the formulation. This would also allow more variability to be applied to more sensitive parts of the table.

Unlike the multinomial-Dirichlet synthesis method, the Poisson and NBI models cannot satisfy ϵ-DP - at least not without truncating the range of obtainable synthetic counts. Yet, from a practical viewpoint, (ϵ, δ)-DP can provide similar risk guarantees as well as greater flexibility. It needs to be remembered that the choice of ϵ and δ are both policy decisions. For a given ϵ, setting $\delta = 0$ provides, of course, a stricter risk guarantee than $\delta > 0$. Nevertheless, comparing between di□erent ϵ and δ is not straightforward: which is to be preferred, say, $\epsilon = 1$ and $\delta = 0$ or $\epsilon = 0.5$ and $\delta = 0.0001$? Utilising the cumulative distribution functions of the Poisson and NBI allow δ and ϵ to be found - and hence tuned - analytically.

There is no panacea for synthetic data generation. A compromise always needs to be struck between risk, utility and, in the case of large data sets, computational time. Di□erent methods are, of course, suited to di□erent data types and sizes. CART models, for example, are e□ective in synthesizing data sets, comprising a mix of continuous and categorical variables, and can easily be carried out via the excellent R package. Yet when n is large, demands on memory makes it challenging, computationally, to implement many methods. In such instances, for categorical data it is more e□cient to undertake the synthesis at the tabular level. Among these approaches, the advantage of using a saturated model - rather than, for example, an all two-way interaction log-linear model - is three-fold. They are quick, eliminate the need to make modelling decisions and support an *a priori* approach to synthesis.

The demand - and need - for near-real-time data was highlighted during the Covid-

19 pandemic. Synthetic data sets are a way to facilitate this, as mentioned in a recent interview with the UK's Deputy National Statistician (Tarran, 2022). Moreover, there is a simultaneous interest in new data sources, such as administrative data. Thus marrying the two - developing synthetic data methods for administrative databases - is a topical and important area of research. The synthesis approach set out in this thesis provides a way to undertake this, and will hopefully inspire further research in this area.

Bibliography

Abowd, J. M. and Vilhuber, L. (2008) How Protective Are Synthetic Data? In *Privacy in Statistical Databases 2008, Proceedings* (eds. J. Domingo-Ferrer and Y. Saygın), 239–246. Berlin, Heidelberg: Springer Berlin Heidelberg.

Aitkin, M., Francis, B., Hinde, J. and Darnell, R. (2009) *Statistical modelling in R*. Oxford Statistical Science Series. Oxford: Oxford University Press.

Asmussen, S. and Edwards, D. (1983) Collapsibility and response variables in contingency tables. *Biometrika*, 70, 567–578.

Balle, B. and Wang, Y.-X. (2018) Improving the Gaussian mechanism for di□erential privacy: Analytical calibration and optimal denoising. In *Proceedings of the 35th International Conference on Machine Learning* (eds. J. Dy and A. Krause), vol. 80 of *Proceedings of Machine Learning Research*, 394–403. PMLR. URL: https://proceedings.mlr.press/v80/balle18a.html.

Bates, A. G., Spakulová, I., Dove, I. and Mealor, A. (2019) ONS methodology working paper series number 16 - Synthetic data pilot. URL: https://www.ons.gov.uk/methodology/

methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot.

Beckman, R. J., Baggerly, K. A. and McKay, M. D. (1996) Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30, 415–429.

Bethlehem, J. G., Keller, W. J. and Pannekoek, J. (1990) Disclosure Control of Microdata. *Journal of the American Statistical Association*, 85, 38–45.

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. New York, USA: The MIT Press.

Blanchard, S., Jackson, J. E., Mitra, R., Francis, B. J. and Dove, I. (2022) A constructed English School Census substitute. URL: 10.17635/lancaster/researchdata/533.

Boser, B. E., Guyon, I. M. and Vapnik, V. N. (1992) A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (ed. D. Haussler), 144–152.

Breiman, L. (1996) Bagging Predictors. *Machine Learning*, 24, 123–140.

— (2001) Random Forests. *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. A. (1984) *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall / CRC.

Broyden, C. G. (1970) The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6, 76–90.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York, NY: Springer New York, 2nd edn.

Caiola, G. and Reiter, J. P. (2010) Random Forests for Generating Partially Synthetic, Categorical Data. *Transactions on Data Privacy*, 3, 27–42.

Couper, M. P., Singer, E., Conrad, F. G. and Groves, R. M. (2008) Risk of Disclosure, Perceptions of Risk, and Concerns about Privacy and Confidentiality as Factors in Survey Participation. *Journal of O cial Statistics*, 24, 255–275.

Dalenius, T. (1977) Towards a methodology for statistical disclosure control. *Statistisk tidskrift*, 5, 429–444.

Dalenius, T. and Reiss, S. P. (1982) Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference*, 6, 73–85.

Delaporte, P. J. (1960) Quelques problèmes de statistiques mathématiques poses par l'Assurance Automobile et le Bonus pour non sinistre. *Bulletin Trimestriel de l'Institut des Actuaires Français*, 227, 87–102.

Deming, W. E. and Stephan, F. F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11, 427–444.

Drechsler, J. (2010) Using Support Vector Machines for Generating Synthetic Datasets. In *Privacy in Statistical Databases 2010, Proceedings* (eds. J. Domingo-Ferrer and E. Magkos), vol. 6344, 148–161. Springer.

— (2011) *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation*, vol. 201 of *Lecture Notes in Statistics*. New York, NY: Springer Science & Business Media.

— (2018) Some Clarifications Regarding Fully Synthetic Data. In *Privacy in Statistical Databases 2018, Proceedings* (eds. J. Domingo-Ferrer and F. Montes), 109–121. Cham: Springer International Publishing.

Drechsler, J. and Reiter, J. (2009) Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey. *Journal of O cial Statistics*, 25, 589–603.

Drechsler, J. and Reiter, J. P. (2010) Sampling with Synthesis: A New Approach for Releasing Public Use Census Microdata. *Journal of the American Statistical Association*, 105, 1347–1357.

— (2011) An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, 55, 3232–3243.

Duncan, G. and Lambert, D. (1989) The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistics*, 7, 207–217.

Duncan, G. T., Elliot, M. and Juan Jose Salazar, G. (2011) *Statistical Confidentiality Principles and Practice*. Springer Statistics for Social and Behavioral Sciences. New York, NY: Springer Science & Business Media.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., Roehrig, S. F. et al. (2001) Disclosure limitation methods and information loss for tabular data. In *Disclosure*

and Data Access: Theory and Practical Applications for Statistical Agencies (eds. P. Doyle, J. Lane, J. Theeuwes and L. Zayatz), 135–166. Elsevier, Amsterdam.

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006) Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography* (eds. S. Halevi and T. Rabin), 265–284. Berlin, Heidelberg: Springer Berlin Heidelberg.

Dwork, C. and Roth, A. (2014) The Algorithmic Foundations of Di□erential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 211–407. URL: `http://dx.doi.org/10.1561/0400000042`.

Elliot, M., Skinner, C. and Dale, A. (1998) Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive E□ects of Geographical Detail on Disclosure Risk. *Research in O cial Statistics*, 1, 53–67.

Fienberg, S. E. and Makov, U. E. (1998) Confidentiality, Uniqueness, and Disclosure Limitation for Categorical Data. *Journal of O cial Statistics*, 14, 385–397.

Fienberg, S. E., Makov, U. E. and Sanil, A. P. (1997) A Bayesian Approach to Data Disclosure: Optimal Intruder Behavior for Continuous Data. *Journal of O cial Statistics*, 13, 75–89.

Fienberg, S. E. and Rinaldo, A. (2012) Maximum Likelihood Estimation in Log-Linear Models. *The Annals of Statistics*, 40, 996–1023.

Fletcher, R. (1970) A new approach to variable metric algorithms. *The Computer Journal*, 13, 317–322.

Fuller, W. (1987) *Measurement Error Models*. Wiley series in probability and mathematical sciences. John Wiley & Sons, Canada.

Goldfarb, D. (1970) A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24, 23–26.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, 2672–2680.

Gouweleeuw, J. M., Kooiman, P. and De Wolf, P. (1998) Post Randomisation for Statistical Disclosure Control: Theory and Implementation. *Journal of O cial Statistics*, 14, 463–478.

Hamming, R. W. (1950) Error Detecting and Error Correcting Codes. *The Bell System Technical Journal*, 29, 147–160.

Hand, D. J. (2018) Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 555–605.

Higgins, J. E. and Koch, G. G. (1977) Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. *International Statistical Review / Revue Internationale de Statistique*, 45, 51–38.

Hittmeir, M., Mayer, R. and Ekelhart, A. (2020) A Baseline for Attribute Disclosure Risk in Synthetic Data. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, 133–143.

HM Government (2018) Help Shape Our Future: The 2021 Census of Population and Housing in England and Wales. URL: `https://assets.publishing.`

service.gov.uk/government/uploads/system/uploads/attachment_data/file/765089/Census2021WhitePaper.pdf.

Hu, J., Akande, O. and Wang, Q. (2021) Multiple Imputation and Synthetic Data Generation with NPBayesImputeCat. *The R Journal*, 13, 90–110.

Hu, J., Reiter, J. P. and Wang, Q. (2014) Disclosure Risk Evaluation for Fully Synthetic Categorical Data. In *Privacy in Statistical Databases 2014, Proceedings* (ed. J. Domingo-Ferrer), 185–199. Cham: Springer International Publishing.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E. S., Seri, G. and Wolf, P. (2010) Handbook on Statistical Disclosure Control. *ESSnet on Statistical Disclosure Control*.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K. and De Wolf, P.-P. (2012) *Statistical Disclosure Control*. Wiley Series in Survey Methodology. Chichester, UK: John Wiley & Sons.

Information Commissioner's O□ce (2020) Guide to the General Data Protection Regulation (GDPR). URL: https://ico.org.uk/media/for-organisations/guide-to-the-general-data-protection-regulation-gdpr-1-0.pdf.

Jackson, J. E., Mitra, R., Francis, B. J. and Dove, I. (2022, Early view) Using saturated count models for user-friendly synthesis of categorical data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Jayram, T. S. (2009) Hellinger Strikes Back: A Note on the Multi-party Information Complexity of AND. In *Approximation, Randomization, and Combinatorial*

Optimization. Algorithms and Techniques (eds. I. Dinur, K. Jansen, J. Naor and J. Rolim), 562–573. Berlin, Heidelberg: Springer Berlin Heidelberg.

Kaloskampis, I. (2019) Synthetic data for public good. URL: https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil, A. P. (2006) A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *The American Statistician*, 60, 224–232.

Kokosi, T., De Stavola, B., Mitra, R., Frayling, L., Doherty, A., Dove, I., Sonnenberg, P. and Harron, K. (2022) An overview on synthetic administrative data for research. *International Journal of Population Data Science*, 7.

Lang, J. B. (1996) On the Comparison of Multinomial and Poisson Log-Linear Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 253–266.

Lindsey, J. K. (1974) Construction and Comparison of Statistical Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36, 418–425.

Little, R. J. (1993) Statistical Analysis of Masked Data. *Journal of O cial Statistics*, 9, 407–426.

Machanavajjhala, A., Kifer, D., Gehrke, J. and Venkitasubramaniam, M. (2007) L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1, 3–es. URL: https://doi.org/10.1145/1217299.1217302.

Manrique-Vallier, D. and Hu, J. (2018) Bayesian non-parametric generation of fully

synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 635–647.

Mateo Sanz, J. M. and Domingo Ferrer, J. (1998) A Comparative Study of Microaggregation Methods. *Qüestiió*, 22, 511–526.

McLachlan, G. and Peel, D. (2000) *Finite Mixture Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, 1st edn.

Meng, X.-L. (1994) Multiple-Imputation Inferences with Uncongenial Sources of Input. *Statistical Science*, 9, 538–558.

Mitra, R. and Reiter, J. P. (2016) A comparison of two methods of estimating propensity scores after multiple imputation. *Statistical Methods in Medical Research*, 25, 188–204.

Morris, T. P., White, I. R. and Crowther, M. J. (2019) Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38, 2074–2102.

Nowok, B., Raab, G. M., Dibben, C. et al. (2016) synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, 74, 1–26.

O□ce for National Statistics (2016) Patient Register: quality assurance of administrative data used in population statistics, Dec 2016. Online. URL: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/methodologies/patientregisterqualityassuranceofadministrativedatausedinpopulation\statisticsdec2016.

Purdam, K. and Elliot, M. (2007) A Case Study of the Impact of Statistical Disclosure Control on Data Quality in the Individual UK Samples of Anonymised Records. *Environment and Planning A*, 39, 1101–1118.

Quick, H. (2021) Generating Poisson-distributed di erentially private synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184, 1093–1108.

Raab, G. M., Nowok, B. and Dibben, C. (2016) Practical Data Synthesis For Large Samples. *Journal of Privacy and Confidentiality*, 7, 67–97.

Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003) Multiple Imputation for Statistical Disclosure Limitation. *Journal of O cial Statistics*, 19, 1–16.

Raper, S. (2020) Leo Breiman's "two cultures". *Significance*, 17, 34–37.

Reiter, J. P. (2003) Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181–188.

— (2005a) Estimating Risks of Identification Disclosure in Microdata. *Journal of the American Statistical Association*, 100, 1103–1112.

— (2005b) Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168, 185–205.

— (2005c) Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *Journal of Statistical Planning and Inference*, 131, 365–377.

— (2005d) Using CART to Generate Partially Synthetic Public Use Microdata. *Journal of O cial Statistics*, 21, 441–462.

Reiter, J. P. and Drechsler, J. (2010) Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Statistica Sinica*, 405–421.

Reiter, J. P. and Kinney, S. K. (2012) Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of O cial Statistics*, 28, 583.

Reiter, J. P. and Mitra, R. (2009) Estimating Risks of Identification Disclosure in Partially Synthetic Data. *Journal of Privacy and Confidentiality*, 1.

Reiter, J. P. and Raghunathan, T. E. (2007) The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association*, 102, 1462–1471.

Reiter, J. P., Wang, Q. and Zhang, B. (2014) Bayesian Estimation of Disclosure Risks for Multiply Imputed, Synthetic Data. *Journal of Privacy and Confidentiality*, 6. URL: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/635.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z. and De Bastiani, F. (2019) *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R*. Boca Raton, FL: CRC Press.

Rinott, Y., O'Keefe, C. M., Shlomo, N., Skinner, C. et al. (2018) Confidentiality and Di erential Privacy in the Dissemination of Frequency Tables. *Statistical Science*, 33, 358–385.

Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal e□ects. *Biometrika*, 70, 41–55.

Rubin, D. (1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Rubin, D. B. (1993) Statistical Disclosure Limitation. *Journal of O cial Statistics*, 9, 461–468.

Schneier, B. (2007) Why 'Anonymous' Data Sometimes Isn't. Wired. URL: https://www.wired.com/2007/12/why-anonymous-data-sometimes-isnt/.

SDAP Working Group (2019) *Handbook on Statistical Disclosure Control for Outputs*. URL: https://ukdataservice.ac.uk/app/uploads/thf_datareport_aw_web.pdf.

Shanno, D. F. (1970) Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24, 647–656.

Si, Y. and Reiter, J. P. (2013) Nonparametric Bayesian Multiple Imputation for Incomplete Categorical Variables in Large-Scale Assessment Surveys. *Journal of Educational and Behavioral Statistics*, 38, 499–521.

Skinner, C., Marsh, C., Openshaw, S. and Wymer, C. (1994) Disclosure control for census microdata. *Journal of O cial Statistics*, 10, 31–51.

Skinner, C. and Shlomo, N. (2008) Assessing Identification Risk in Survey Microdata Using Log-linear Models. *Journal of the American Statistical Association*, 103, 989–1001.

Skinner, C. J. (1992) On identification disclosure and prediction disclosure for microdata. *Statistica Neerlandica*, 46, 21–32.

Skinner, C. J. and Elliot, M. J. (2002) A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 855–867.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A. (2018) General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181, 663–688.

Stasinopoulos, D. M., Rigby, R. A. et al. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, 23, 1–46.

Sweeney, L. (2002) k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10, 557–570.

Tarran, B. (2022) "i want to be ready to answer questions we don't yet know we need to know": An interview with Alison Pritchard. *Significance*, 19, 42–44.

Taub, J., Elliot, M., Pampaka, M. and Smith, D. (2018) Di□erential Correct Attribution Probability for Synthetic Data: An Exploration. In *Privacy in Statistical Databases 2018, Proceedings* (eds. J. Domingo-Ferrer and F. Montes), vol. 11126, 122–137. Springer (Lecture Notes in Computer Science).

Templ, M. (2017) *Statistical Disclosure Control for Microdata Methods and Applications in R*. Cham, Switzerland: Springer Nature.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A. and Sijtsma, K. (2008) Multiple Imputation of Incomplete Categorical Data Using Latent Class Analysis. *Sociological Methodology*, 38, 369–397.

Willenborg, L. and De Waal, T. (1996) *Statistical Disclosure Control in Practice*, vol. 111. Springer Science & Business Media, 1 edn.

Winkler, W. E. (2005) Re-Identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata. *Research Report Series (Statistics)*, 9, 1–14.

Woo, M.-J., Reiter, J. P., Oganian, A. and Karr, A. F. (2009) Global Measures of Data Utility for Microdata Masked for Disclosure Limitation. *Journal of Privacy and Confidentiality*, 1.