Evaluating hypothesis tests on differentially private histogram-based synthetic data

Master of Science Thesis
University of Turku
Department of Computing
Data Analytics
2024

Jan Böhmeke

UNIVERSITY OF TURKU
Department of Computing

Jan Böhmeke: Evaluating hypothesis tests on differentially private histogram-based synthetic data

Master of Science Thesis, 51 p.
Data Analytics
June 2024

Sharing synthetic data that preserves privacy has been suggested as an option for releasing sensitive data without compromising individuals' privacy. The synthetic data should maintain the structure and statistical characteristics of the original data, while ensuring individuals privacy. Differential privacy (DP) effectively assures privacy concerns, while preserving structure and characteristics of the original data. Objectives of this research is to evaluate Students T-test and Mann-Whitney U test empirically to verify if those tests are prone to result in loss of tests validity or decreased power. Empirically demonstrating this is done in terms of Type I and Type II errors. I evaluate the statistical hypothesis tests on sets of additively smoothed DP synthetic data generated from sets of original data. The original data sets are simulated questionnaire data (n=20 000) following 5-point Likert Scale and 10-point Likert Scale and Kaggle Cardiovascular Dataset (n=70 000). The validity of tests was preserved for all privacy budget values ($0.001 \leq \epsilon \leq 100$) and sampled dataset sizes (50,100,500,1000) for all data. The power of the tests was considerably reduced in all cases.

Keywords: Differential Privacy, Synthetic Data, Exponential Mechanism

Contents

List of Figures

List of acronyms

δ	relaxation parameter for (ϵ, δ)-differential privacy
Δu	sensitivity of the scoring function u
ϵ	privacy budget
$\mathbb{N}_0$	Natural numbers starting from 0
μ	mean
σ^2	standard deviation
M	differentially private mechanism
$P(\cdot)$	probability of a random variable $(\cdot)$
X	universe of databases
$\mathbf{g} \circ M$	function composition = g(M(X))
AI	Artificial Intelligence
DP	Differential Privacy
GAN	Generative Adversarial Networks
KL	Kullback-Leibner divergence

1 Introduction

1.1 Context of the Research

As digital information continues to grow at an unprecedented rate, with a projected doubling every three years, the interest for using and sharing the data for research and innovation is also rising [1]. Data has become an essential part of daily life. For example Netflix uses data to recommend users similar movies, based on their watching history. Google uses targeted advertising based on users browsing history, to recommend users similar products what they have browsed before. Pharmacies that share de-identified prescription records with private research firms emphasizes the importance for stronger healthcare data privacy to prevent the re-identification of patients [2].

The rise of Artificial Intelligence (AI) has increased the demand of data to train and generate models for forecasting. Deep learning based generative models such as Generative Adversarial Networks (GANs), can be used to generate synthetic samples. However, due to the high model complexity of deep networks, GANs can easily memorize the training data which can lead to potential privacy disclosure if trained on sensitive data such as patient medical records [3]. Privacy is not too often discussed when using data, which leads to increase risk of disclosing individuals confidential information.

The potential risks and vulnerabilities associated with the sharing or public exposure of individuals data, such as the existence of various techniques to extract private information from anonymized datasets highlights the importance for robust privacy measures in order to protect individuals privacy. Techniques which are used to extract private information from anonymized datasets are for example re-identification attacks [4][5], reconstruction attacks [6] and differencing attacks [6]. It is possible to utilize the existing data within a dataset to deduce missing information or data that has been partially anonymized, in order to reconstruct the dataset. Another approach for extracting information involves linking one dataset with another that contains data about the exact same individuals. Extraction of personal data can also be achieved through the use of auxiliary information. If the auxiliary information matches the data in the dataset, it is possible to single out an individual. [6]

There exists different privacy-preserving techniques to protect personal data. Often used technique is for instance anonymization which means that the identifying information is removed from the dataset. Examples of such techniques include n-confusion, l-diversity and t-closeness [7]. However the primary drawback of these techniques aimed at safeguarding privacy, is that they frequently compromise the accuracy of data analysis. The aim of privacy research is to discover algorithms that offer robust privacy protection for individuals while minimizing the loss of accuracy as much as possible. In certain situations, privacy is assured but not delivered, and instead only basic anonymization is employed which fails to protect against the extraction of personal data. To address this issue, more advanced anonymization mechanisms have been developed and put into practice. [6][7]

The concept of differential privacy (DP) was introduced by Cynthia Dwork in 2006 [8] as a means of anonymizing data and safeguarding the privacy of individuals by injecting random noise into the data. DP provides a solution to the dilemma of extracting useful information about a group while still preserving the confidentiality of an individual's data [8]. There has been proposed several DP methods for deep learning applications. Example of these applications are GANs [3], Convolutional Neural Networks (CNNs) [9] and Reinforcement Learning (RL) [10]. These are used to protect against for example model inversion attacks [4].

Synthetic data, proposed by Rubin in 1993 involves generating data using computer algorithms instead of collecting it from real-world sources [11]. Synthetic data is often discussed as a method for privacy preservation, however this is not the case since it has been demonstrated that synthetic data can leak original information. For example in the worst case scenario, generative models such as GANs [3] can generate copies of the original data it was trained on, which leads to leaking individuals privacy. The utilization of synthetic data offers numerous potential advantages, such as enabling the exploration of scenarios that may be impractical with real-world data, mitigating privacy concerns, and supporting the creation of more diverse and inclusive datasets [12]. Synthetic data has been used in various fields, such as healthcare, finance, and cybersecurity. It is generated artificially and tries to accurately reflect real-world data collected from actual sources. Synthetic data presents both advantages and drawbacks. By synthesizing real world dataset, data could become too noisy and does not accurately depict actual data. One advantage of utilizing synthetic data is that it can be generated in large quantities and be customized for different scenarios. Creating synthetic data once allows it to be used indefinitely [6]. Synthetic data can also be combined with DP to create more private data, called differentially private synthetic data (DP synthetic data) [13][12].

DP synthetic data refers to synthetic data that has been generated in a way that protects the privacy of the individuals represented in the data. The distinction between DP synthetic data and synthetic data without DP can be written in terms of utility, privacy and amount of data-trade off.

The amount of utility, meaning how accurate synthetic data is to have the same statistical properties of real data, differs when adding DP. For DP synthetic data, utility usually suffers compared to synthetic data without DP. [14][13][15]

Since synthetic data offers little to no privacy, adding DP to synthetic data is a solid option to guarantee privacy. However when improving privacy of DP synthetic data, utility of the DP synthetic data usually suffers. This is because adding too much noise reduces utility of synthetic data and in a worse case scenario would render the synthetic data useless, and synthetic data would not give any information about real data [15].

DP synthetic data research focuses on providing algorithms that emphasize utility. When conducting research that is sensitive, it is recommended to use DP synthetic data instead of regular synthetic data to preserve privacy. Upside of DP synthetic data is that it can be used indefinitely and also combined with another dataset and still preserve its privacy guarantee [6]. Downside about DP synthetic data is that either utility or the privacy of the data suffers. Too much privacy could decrease the utility of DP synthetic data, making the DP synthetic data unusable.

It is important to keep the validity of a statistical test such as Student's t-test when it is applied to DP synthetic data [16]. Statistical hypothesis testing is a way to learn about a population using a sample. Appropriate tests for different cases are also needed. For instance for normally distributed data, Student's t-test is a solid option for statistical hypothesis testing [17]. One common use of the Student's t-test is to compare the means of two groups to determine whether they are significantly different. For example a healthcare researcher compares the average pain scores of patients who received a new pain management treatment to those who received standard care. Using a Student's t-test, the researcher analyzes the data from the treatment and control groups to determine if the difference in mean pain scores is statistically significant, indicating the effectiveness of the new treatment.

For DP synthetic data it could be problematic to do hypothesis testing. This is because by adjusting the parameters of DP synthetic data algorithm, the distribution might change entirely. For example normally distributed dataset might turn into uniform distribution, and it is then impossible to do accurate hypothesis testing with Student's t-tests. It is important to carefully design the Student's t-test in order to ensure that it is properly evaluating the relevant properties of the synthetic data. This may involve specifying the appropriate null and alternative hypotheses and choosing the appropriate level of significance for the test.

Agree-Disagree rating scales are popularly used method in social science research questionnaires [18][19]. However such questionnaires need to be adjusted correctly so that they yield good quality data. Agree-Disagree rating scales are susceptible for biases. Some respondents may agree with the statement regardless of its content, if the statement is generally depicted positive. For example "Immigration is good for the economy" may have more Agree votes than "Immigration is bad for the economy" since peoples tendency to be polite [19]. In this thesis main focus relies on 5-point Likert scale data, however 10-point Likert scale data is also briefly described. Since Likert Scale is ordinal data, usually statistical hypothesis testing is done with nonparametric tests such as Mann Whitney U-test or Wilcoxon matched paired tests. However several studies show that parametric tests such as Student's t-test is a valid option to evaluate the validity and utility of DP synthetic data [20][21]. Even though Student's t-test is not usually done for Likert Scale, study [17] concluded that Student's t-test and Mann-Whitney-Wilcoxon (MWW) have similar power for 5-point Likert scale. However, there are significant power differences between the two tests when the distributions are skewed, peaked, or multimodal.

1.2 Contributions of This Research

The research question of this thesis is to test the validity and utility of the Student's t-test on DP synthetic data as well as Mann-Whitney U test, generated using DP Smoothed Histogram algorithm provided by Wasserman and Zhou [22] that has been modified by (Pahikkala et al) [16].

This thesis follows the contributions of this paper [16] where there has been done empirical study of Mann-Whitney U-test on DP synthetic data with several variations of DP synthetic data. In that paper same method was used to empirically evaluate Mann-Whitney U test. That paper also used several other methods focusing solely on simulated parametric data and real-world parametric data. However in this thesis I focus solely on DP synthetic data that has been generated via Additively Smoothed Differentially Private Synthetic Data (AS-DPSD) to evaluate Student's t-test and Mann-Whitney U test. These contributions will add some insight how these two statistical tests behave on this particularly generated DP synthetic data. This thesis also gives insights into the differences of simulated and real world data on Additively Smoothed DP synthetic data. It also provides perspective how data size and privacy budget dependent the Additively Smoothed DP synthetic data method is for simulated and real world data.

1.3 Structure

In this thesis I focus on evaluating statistical hypothesis tests with Type I and Type II errors on histogram-based DP synthetic data.

This thesis is organized to six main sections:

- Chapter 2 introduces DP and its main concepts such as privacy-accuracy dilemma, post-processing and different DP mechanisms.

- Chapter 3 introduces the concept of synthetic data and gives an introduction how to create synthetic data from original data. This chapter also describes different methods to smooth histograms and sample from the smoothed histograms.

- Chapter 4 introduces the concept of DP synthetic data and gives an implementation how to create DP synthetic data with an algorithm that is specifically designed for nonnumeric data.

- Chapter 5, discusses using statistical hypothesis testing for differentially private algorithms and gives a summary for the algorithm designed in chapter 5
- Chapter 6, discusses the results obtained when testing the validity of Student's t-test and Mann-Whitney U test on DP synthetic data.
- Finally, the conclusion of this work, presented in chapter 7, outlines future work directions.

2 Differential Privacy

2.1 Introduction to Differential Privacy

The goal of differential privacy (DP) is to keep each person's data private while still being able to study the whole dataset. DP is a standard way to extract information from a dataset while simultaniously maintaining information [8]. The idea behind DP is that the distribution of the output of a privacy-preserving algorithm should not change significantly whether an individual's data is included in the dataset or not. In DP individuals data cannot be concluded from the output of the algorithm. DP protects data of an individual against several types of attacks, for example re-identification attacks, reconstruction attacks and differencing attacks. In order to achieve DP, carefully designed noise addition to algorithms that produce differentially private dataset is needed [23].

DP aims creating a transformed dataset Z from an input dataset X, by protecting individuals privacy while preserving information. In particular, if changing one entry in the dataset X cannot change the probability distribution drastically, then we can claim that a single individual cannot guess whether he is in the original dataset or not [22]. For numeric datasets you can add noise to perturb the result. However, nonnumerical queries require mechanism that can effectively introduce noise to the dataset while accurately maintaining the discrete set of information. This distinction between numerical and nonnumerical data is crucial in the context of privacy-preserving techniques such as DP, as it points out the need for specialized approaches to protect sensitive data in nonnumerical domains [7][6].

2.2 Definitions and Theorems

Definition 1. *Differential Privacy. A randomized algorithm M satisfies -DP, if for any two datasets x and y satisfying $d(x, y) \leq 1$ and for any possible output O of M, we have*

$$Pr[M(x) = O] \leq exp(\epsilon)Pr[M(y) = O], \tag{2.1}$$

where $Pr[\cdot]$ denotes the probability of an event, ϵ denotes the privacy budget of DP algorithm and d denotes the Hamming distance between the two datasets [23]. Privacy budget ϵ controls the level of privacy. Smaller ϵ value guarantees more privacy, however reduces the accuracy of the dataset and vice versa [24][25].

Definition 2. *Hamming distance. Given two datasets $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$, let $\delta(X, Y)$ denote the Hamming distance between X and Y [22].*

$$\delta(X, Y) = i : X_i \neq Y_i. \tag{2.2}$$

In other words Hamming distance in DP could be defined as the number of pairs of individuals for which the total value changes by less or equal to one.

Definition 3. *Approximate Differential Privacy, also called (ϵ, δ)-DP, is a relaxation of ϵ-DP.*

$$\mathsf{Pr}[F(x) = S] \leq e^{\epsilon}\mathsf{Pr}[F(x') = s] + \delta \tag{2.3}$$

Intuitively, this definition means that with probability $1-\delta$ the same guarantee as pure DP is provided. With probability δ, guarantee does not exist. Hyperparameter δ needs to be adjusted so it does not happen. For example adjusting δ such as $\delta \leq \frac{1}{n^2}$, where n denotes the dataset size [6].

Definition 4. *Suppose a mechanism M provides (ϵ_j, δ_j)-DP for $j = 1, ..., k$*

a) Sequential composition: The sequence of $M_j(X)$ applied on the same X provides $\sum_j \epsilon_j, \sum_j \delta_j$

b) Parallel composition: let D_j be disjoint subsets of the input domain D the sequence of $M_j(X \cap D_j)$ provides $(max(\epsilon_j), (max(\delta_j))$-DP [13]

Composition theorems explore whether privacy guarantee remains intact when the DP-algorithms are used multiple times [8][6]. One of the composition theorem used in DP is Sequential composition. Sequential composition has total privacy cost of $k\epsilon$ since the mechanism M is run k amount of times. Parallel composition is another way to calculate the total privacy cost of multiple data releases. It involves splitting the dataset into separate chunks and applying a differentially private method to each chunk independently. The total privacy cost of parallel composition is only ϵ, since parallel composition is only ran M amount of times [6]. In our case we automatically use parallel composition, since histograms are partitioned into disjointed cells [8]. Given that the chunks are disjoint, each individual's data is confined to a single chunk. Therefore, when the mechanism, denoted as M, is executed k amount of times with a total of k chunks, it ensures that every individual's data is processed only once by the mechanism M [6].

Theorem 1. *Post-Processing Theorem: Let M be an ϵ-differentially private mechanism and g be an arbitrary mapping from the set of possible outputs to an arbitrary set. Then, $g \circ M$ is ϵ-differentially private.[26]*

The post-processing property ensures the safety of conducting any arbitrary computations on the output of a differentially private mechanism. Additionally, it guarantees that the privacy protection offered by the mechanism remains intact, eliminating any possibility of compromising the privacy protection of the mechanism. Post-processing is used in differentially private algorithms to reduce noise and improve the accuracy of their results [6].

In DP, noise needed to calibrate the privacy budget ϵ depends on the querys sensitivity. Sensitivity of a function measures the influence one individual can have on the output of the query [27]. There exists different variants of sensitivity. For example smooth sensitivity, global sensitivity and local sensitivity. The most widely used variant of sensitivity is global sensitivity [6][27].

Definition 5. *For a query $f : D \rightarrow \mathbb{R}$, the global sensitivity of f is defined as*

$$GS(f) = \max_{x,x':d(x,x')\leq 1} |f(x) - f(x')| \tag{2.4}$$

[6]

Definition 6. *Local sensitivity of a function $f : \mathcal{D} \rightarrow \mathbb{R}$ at $x : \mathcal{D}$ is defined as:*

$$LS(f,x) = \max_{x':d(x,x')\leq 1} |f(x) - f(x')| \tag{2.5}$$

[6]

The biggest distinction between global and local sensitivity is that the query f and the actual dataset x affects local sensitivity, where in global sensitivity only query f affects to global sensitivity [6].

Histograms have sensitivity of 1 by default, since they are partitioned into disjointed cells. Since addition or removal of single element has sensitivity of 1 [8], global sensitivity is automatically defined. Global sensitivity works well when queries have relative lower sensitivity values, such as counting or sum queries. For queries such as median, average, the global sensitivity yields high values comparing with true answers. For these type of problems local sensitivity is needed [28].

2.3 Differentially Private mechanisms

Various mechanisms exist that satisfy DP. However, what sets these mechanisms apart is their effectiveness on different types of data. For instance, data that is mostly numerical is for Laplace or Gaussian Mechanism suitable. Nonnumerical data on the other hand may rely on Exponential Mechanism to make the dataset differentially private [6].

Definition 7. *Laplace mechanism: For a function $f(x)$ which returns a number, the following definition of satisfies $(\epsilon, 0)$-DP:*

$$F(x) = f(x) + \mathsf{Lap}(\frac{s}{\epsilon}) \tag{2.6}$$

[6]

where s is the sensitivity of f, and $\mathsf{Lap}(S)$ denotes sampling from the Laplace distribution with center 0 and scale S. [6]

Definition 8. *The Gaussian Mechanism satisfies (ϵ, δ)-DP by adding Gaussian noise with zero mean and variance, σ^2, such that*

$$F(x) = f(x) + \mathcal{N}(\sigma^2) \tag{2.7}$$

$$\textit{where } \sigma^2 = \frac{2s^2 \log(1.25/\delta)}{\epsilon^2} \tag{2.8}$$

where s is the sensitivity of f, and $\mathcal{N}(\sigma^2)$ denotes sampling from the Gaussian distribution with center 0 and variance σ^2.

The Gaussian mechanism can be utilized in the same manner as the Laplace mechanism for real-valued functions $f : D \to \mathbb{R}$, and it is simple to compare the outcomes of both mechanisms for a given value of ϵ. It differs from the Laplace mechanism by introducing Gaussian noise instead of Laplacian noise. It serves as an alternative approach to the Laplace mechanism. The Gaussian mechanism does not satisfy pure ϵ-DP, but does satisfy (ϵ, δ)-DP [6].

Definition 9. *Exponential Mechanism; 1. The analyst selects a set $\mathcal{R}$ of possible outputs 2. The analyst specifies a scoring function $u : \mathcal{D} \times \mathcal{R} \rightarrow \mathbb{R}$ with global sensitivity Δu 3. The exponential mechanism outputs $r \in \mathcal{R}$ with probability proportional to:*

$$exp\left(\frac{\epsilon u(x,r)}{2\Delta u}\right) \tag{2.9}$$

[6]

Exponential Mechanism was invented for situations where the "best" response should be chosen but adding noise directly could destroy the value [8]. The analyst defines which element is the “best” by specifying a scoring function that outputs a score for each element in the set, and also defines the set of things to pick from. The mechanism provides DP by approximately maximizing the score of the element it returns. However to satisfy DP, the Exponential Mechanism sometimes returns an element from the set which does not have the highest score [6].

The biggest practical difference between the Exponential Mechanism and other mechanisms such as Laplace mechanism and Gaussian Mechanism is that the output of the Exponential Mechanism is always a member of the set $\mathcal{R}$. Exponential Mechanism is used in scenarios where for example Laplace Mechanism would destroy the outcome. For example in Likert 5-scale data Laplace Mechanism should not be used, since it could turn answer: "Neither agree or disagree" to "Disagree" [6].

Usually the problems in Exponential Mechanism arise when talking about local DP and time complexity. Local DP as already told has higher time complexity compared to global DP. This is why there has been developed many variants of Exponential Mechanism such as Base 2 - Exponential Mechanism [29], Concentrated DP with Exponential Mechanism [30], Joint Exponential Mechanism combined with Top-k sampling [31] and Multiplicative Weights Exponential Mechanism [32].

Report noisy-max is an alternative version of Exponential Mechanism. Instead of applying score function s of value $exp\left(\frac{\epsilon u(x,r)}{2\Delta u}\right)$, Report noisy-max algorithm produces score function s with $u(x, r) + \mathsf{Lap}(\frac{s}{\epsilon})$, however it is $n\epsilon$-differentially private, thus releases more information than Exponential Mechanism [6]. We can conclude that for nonnumeric queries the best way to maintain accuracy and same time apply DP is Exponential Mechanism.

2.4 Different Types of Attacks

There exists different types of attacks to get de-identified information from a dataset. Example of these attacks are re-identification attacks, reconstruction attacks and differencing attacks.

Re-identification attacks are type of attacks that re-identify individuals from "anonymized" dataset. One example of re-identification attacks are linkage attacks. Linkage attacks are attacks, that match anonymized records from a different dataset with non-anonymized records from another dataset. An example of this is the Netflix Prize competition. In this competition, Netflix made available a training dataset consisting of movie records that had been anonymized. These records were linked to similar users in the Internet Movie Database, which allowed for the identification of the IMDB users even though their records had been at least partially anonymized [8]. DP fixes this issue because if dataset is differentially private, it doesn´t have any effect on auxiliary information. For instance if Netflix's training data is differentially private, it is close to impossible to create a linkage with it with another dataset because other part of the linkage is differentially private, say in other words not the same data, even if the data the user inputted to both parts of the linkage is the same [8][6]. Another example of re-identification attack is for example the following: A person gives an discrete answer to some non-differentially-private, yet anonymized questionnaire, which includes age. Now if adversary has auxiliary information about persons age, it can search by age all of the persons. And vice versa, if age is anonymized, and adversary knows precisely the answer, it can conclude the age.

Another type of attack is a differencing attack. It takes into account multiple statistics that include the target's data to get sensitive information about that person. Differencing attack means that if an adversary already knows auxiliary information about individual from the dataset, it can use that information to direct the attack to the individual person. For example if an adversary knows that some person is 30 years old and has blonde hair, then he searches those attributes from the dataset [6].

A reconstruction attack is a type of privacy attack on aggregate data that reconstructs a significant portion of a raw dataset. If we know some information about the dataset, we can reconstruct it like a sudoku puzzle. For example: given I know x-y things about the dataset x, I can conclude y things because I know some x already. A type of reconstruction attack is for example model-inversion attack. Model-Inversion attack is an attack which purpose is to recover images from a facial recognition system [4].

Sometimes DP is not enough however to protect against re-identification attacks. Paper [33] proposed a new way type of re-identification attack which uses noisy-sample mean to breach in to the dataset. However there exists continously newer methods to protect against re-identification attacks. For example paper [5] proposed an algorithm which combines k-anonymity and pure DP to emphasize the security which DP provides.

2.5 Privacy-Accuracy dilemma

In DP there is a dilemma between privacy and accuracy. The more private the data is, the less accurate it becomes and vice versa. The ϵ parameter in the definition of DP is called the privacy budget. Small ϵ values have more noise in the outputs, which leads to stronger privacy protection. In contrast to small ϵ values, larger values of ϵ introduce less noise in the outputs, resulting in less privacy [6].

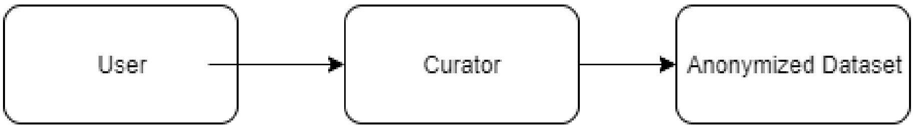


Figure 2.1: Central model of DP.

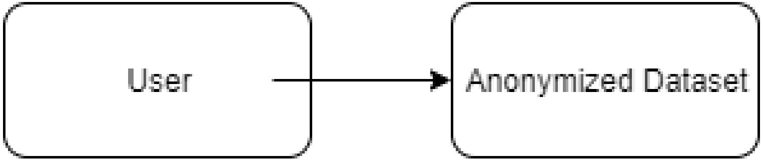


Figure 2.2: Local model of DP.

Adjusting privacy budget usually depends on the granularity of the data. If very sensitive data, that can harm individuals privacy is used, it is better to adjust privacy budget such that it maintains higher privacy over higher accuracy. Privacy budget also depends on the variant of DP. For instance Concentrated DP, (ϵ, δ)-DP and $(\epsilon, 0)$-DP do not have the same way to adjust privacy budget. [34]

In DP the challenge is to add enough noise to satisfy the definition of DP, but not so much that the answer becomes too noisy to be useful. For this process, basic mechanisms have been developed. Mechanisms try to answer the questions: what kind of noise and how much noise to use. For example Laplace mechanism tends to work better in pure DP and Gaussian mechanism better in approximate DP [6].

2.6 Central version and Local version of DP

Central model of DP is a model where there is a curator who applies DP to a single dataset [5]. An alternative to the central model of DP is the local model of DP. In local model of DP, there is no need for trusted data curator, since the data is already made differentially private before it arrives to the data curator. One example of local model of DP is Randomized Response Mechanism [6].

Local model has also significant disadvantages over a central model. The accuracy of the differentially private data is usually much lower in local model than in central model [6]. Illustration of central model is shown in Figure 2.1. Illustration of local model is seen in figure 2.2. Example of local model versus central model: Central model: Question data is gathered from 1000 individuals, each one of the individuals answer 10 questions. Now to implement central DP to such data, we need to use any mechanism M to the whole dataset $M(1000i \times 10q)$. Local model: Question data is gathered from 1000 individuals, each one of the individuals answer 10 questions. Now to implement local DP to such data, we need to use a mechanism M such that for single individuals $1i$ answer $1q - 10q$, the answer is already differentially private before it gets combined with other individuals data [6].

2.7 Summary

In previous chapters I introduced the concept of DP. I explained why is it needed for privacy preservation of data. How it protects against certain types of attacks such as reconstruction attacks, re-identification attacks and differencing attacks. I also introduced several variants such as the (ϵ, δ)-DP variant and concluded that $(\epsilon, 0)$-DP is the most suitable approach for the experiments that will be conducted in this thesis. I also introduced two models of DP, central and local model. Explanation regarding the hyperparameters which control the privacy-accuracy dilemma was also briefly described. Mechanisms were also introduced, emphasizing Exponential mechanism which will be the main mechanism in this thesis.

3 Synthetic Data

Synthetic data was proposed by Rubin in 1993 [11]. In its simplest terms synthetic data produces arbitrary data which elements are not original, since it is generated using statistical models from original data. The main idea of synthetic data is to produce data similar to original data, so that original data is protected and synthetic data could be safely used for example, for studies. However it is shown that for example GANs [3] using original data as a training set, can generate copies of the data it was trained on. This has a potential to compromise personal privacy. Synthetic data is a new form of data that is generated from existing data. Synthetic data can be used to fill in gaps in the current dataset or to create a dataset that never existed before. Synthetic data can be generated either from scratch or by sampling from existing dataset. In this thesis I focus primarily on generating synthetic data by sampling from existing dataset. Synthetic data is pivotal in the modern age since as the data continues to grow in exponential rate, and companies use data to sell more products, privacy is often neglected. Synthetic data gives an arbitrary version of original data, but still preserves the same properties as the original data if done correctly.

Synthetic data has several valuable applications in computer vision training. Usually computer vision training needs large amount of accurately labeled data, which could cost a lot. With the help of synthetic data, it is possibly to generate labeled data as a solution. In these situations privacy is not in the main concern because synthetic data is intended to complement real world data [12]. Synthetic data also allows to investigate data with same causal structure as original data but with modified distributions, which enables research under different circumstances [12].

Given a dataset X, a synthetic dataset Y is a new dataset that has the same structure as X , but whose elements are not original. Now we could argue that by randomly sampling $x \in X$ from the original dataset would produce synthetic data Y where information of the people in the original dataset is secured. However this is not the case since there still are patterns from the original distribution.

3.1 Histograms

One of the most common synthetic representation is a histogram that is used in this thesis. Histogram has many good qualities for DP, for example it automatically satisfies parallel composition [6]. Every "bin" in a histogram is determined by a potential value associated with a data attribute. Since it is not possible for a single row to possess multiple values for an attribute simultaneously, the definition of these bins ensures their distinctness and disjointness [6]. There exists different ways to create differentially private data from histograms. One way to release differentially private numerical data is to generate differentially private histograms and then synthesize numerical data [15].

The difference between synthetic representation and synthetic data is that synthetic representation is not in the same shape as original data [6]. Histogram is an example of synthetic representation. To make the synthetic representation in the same shape as original data, synthetic data needs to be made from synthetic representation. The biggest advantage of synthetic representation is that, we can answer infinitely many queries without additional privacy budget [6].

3.2 Sampling from a histogram

Synthetic representation is created as follows. First we want to ensure that all numerical values are non-negative. Then we need to sum each count in histogram bins so they sum to 1 and treat them as probabilities. The last phase is to produce new samples based on these probabilities. New samples are obtained from the distribution by randomly selecting a bin of the histogram, with the choice being influenced by he probabilities assigned to each option [6]. This last phase can be done for example with NumPy-librarys random.choice method, which will be the main method in this thesis to create new samples from a histogram.

4 DP synthetic data

DP synthetic data is a combination of DP and synthetic data which enables disclosing data that is analytically useful while preserving the privacy of individuals in the data. DP synthetic data is used in variety of different applications such as binary data, categorical data, continous data and network data [13]. There exists various ways how DP synthetic data can be generated, one example is via supervised learning model like in the paper [25]. Other data generation methods are also possible for DP synthetic data like GANs [3] and various histogram methods such as smoothed histogram method and perturbed histogram methods like in the paper [22].

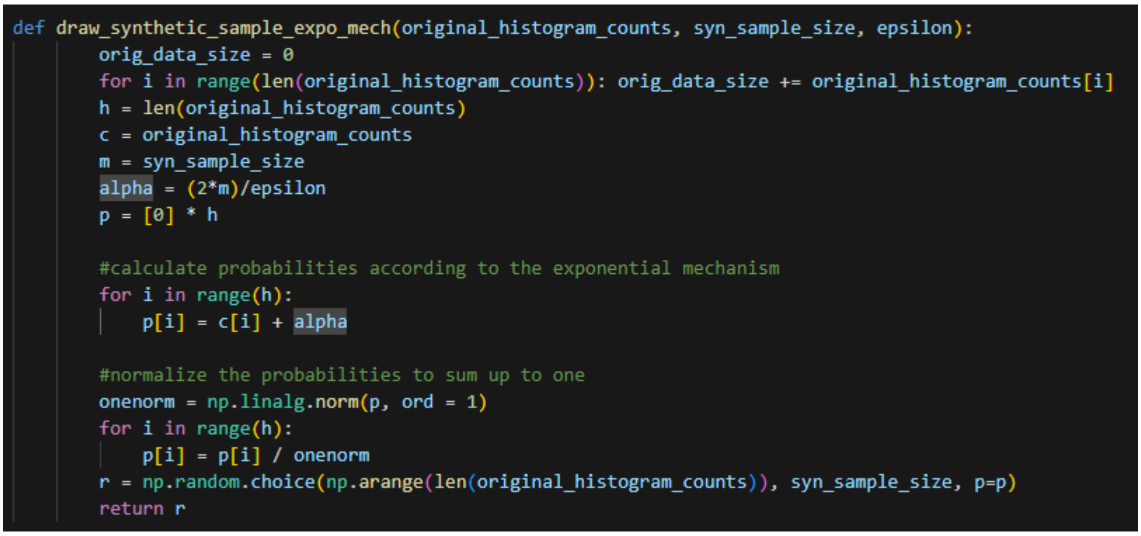
Synthetic data usually almost always loses utility with the addition of DP. The goal is to reduce as little utility as possible. In other words, the addition of DP should not reduce the usefulness of synthetic data. Adding too much noise could potentially make the synthetic data unusable for certain situations that need to retain as much utility as possible. One of these situations could be for example cancer diagnostic of patients. Some datasets however would rather have strong privacy guarantee on the cost of utility. DP synthetic data is useful for such applications when people need to only play around the dataset and not having fear of revealing sensitive information [13].

4.1 Creating DP synthetic data

There exists different types of differentially private made histograms, from which differentially private synthetic data can be sampled. The paper [22] introduces two different DP histograms. The first method draws observations from a smoothed histogram and the second method proposed in the paper draws observations from a randomly perturbed histogram. For histogram based DP synthetic data methods the histogram can be made differentially private by adding for instance Laplace noise to each count in the histogram. This satisfies DP definition because of parallel composition [6]. The emphasis of this thesis is on the algorithm of Pahikkala et al., where the main focus is smoothed histogram method. Smoothing from a perturbed histogram is not the main goal in this thesis [16].

4.2 Additively Smoothed Differentially Private Synthetic Data

Additively Smoothed Differentially Private Synthetic Data (AS-DPSD) method generates synthetic data by drawing data from the probability distribution determined by a histogram in which the probabilities of the bins are proportional to $c_i + \frac{2m}{\epsilon}$. The number of original data in every i histogram bin is denoted as c_i and the amount of synthetic data drawn is denoted as m. The approach is similar to the one in the paper [22]. Unlike the other considered DP methods, the utility of this method is inversely proportional also to the amount synthetic data drawn. Therefore, in our experiments the amount of synthetic data drawn generated is considerably smaller than the original data [16]. The AS-DPSD-algorithm is illustrated in Figure 4.1.



```
def draw_synthetic_sample_expo_mech(original_histogram_counts, syn_sample_size, epsilon):
        orig_data_size = 0
        for i in range(len(original_histogram_counts)): orig_data_size += original_histogram_counts[i]
        h = len(original_histogram_counts)
        c = original_histogram_counts
        m = syn_sample_size
        alpha = (2*m)/epsilon
        p = [0] * h

        #calculate probabilities according to the exponential mechanism
        for i in range(h):
            p[i] = c[i] + alpha

        #normalize the probabilities to sum up to one
        onenorm = np.linalg.norm(p, ord = 1)
        for i in range(h):
            p[i] = p[i] / onenorm
        r = np.random.choice(np.arange(len(original_histogram_counts)), syn_sample_size, p=p)
        return r
```

Figure 4.1: AS-DPSD algorithm

4.2.1 Hyperparameters

The AS-DPSD algorithm is heavily dependent on the hyperparameters e.g privacy budget and the amount of synthetic data drawn from the histogram. The amount m of synthetic data sampled from the histogram usually should be at most $m = \frac{n}{10}$, where n denotes the number of original data. It also depends on the type and shape of the data, e.g if the data is continuous or discrete.

In the intermediate phase (Figure 4.3), it can be seen how different hyperparameters influence the outcome of the probabilities. Based on these probabilities the synthetic sample is drawn. This is shown in Figure 4.4.

In figure 4.3 it is seen that the lower the privacy budget ϵ is, the more the normal distribution tends to look like uniform distribution. This leads to too much noise and as a conclusion the synthetic dataset can not give any insight about the original dataset, thus becomes useless. In the figure it can also be seen that when the synthetic sample size is 100, it seems that ϵ value does not seem to have difference and both tend to follow normal distribution. However due to the sample size being

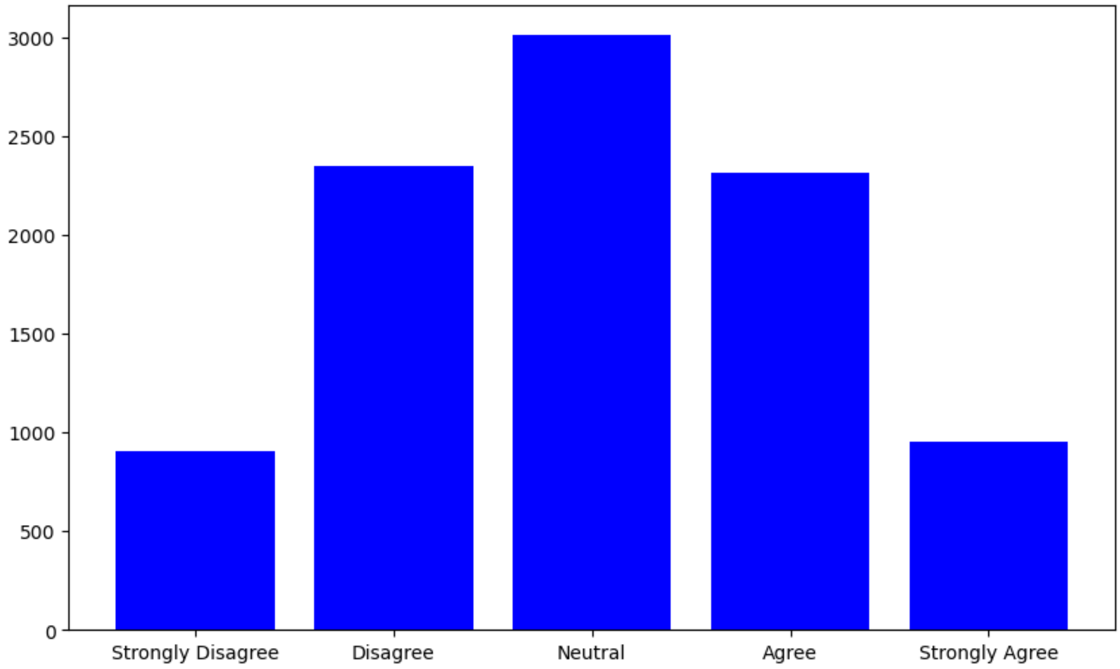


Figure 4.2: Original data distribution with size of 10000

only 100, it does not produce valid results.

Utility of the AS-DPSD algorithm is dependent on epsilon, number of synthetic data drawn, number of original data and number of bins. Illustration of how every hyperparameter and bin effects the utility is shown in Chapter 4.2.2. Figures 4.5-4.6 illustrate the simulated Likert data which is later used in the thesis.

4.2.2 Effect of bins

In this subsection I demonstrate what kind of effect do bins have on the data. Figures 4.12-4.16 were produced by AS-DPSD from Cardiovascular Dataset with different hyperparameters and bin sizes and Figure 4.11 shows the original dataset. Hyperparameters ϵ was set to (0.001, 0.1, 1) and synthetic sample size was set to 1000.

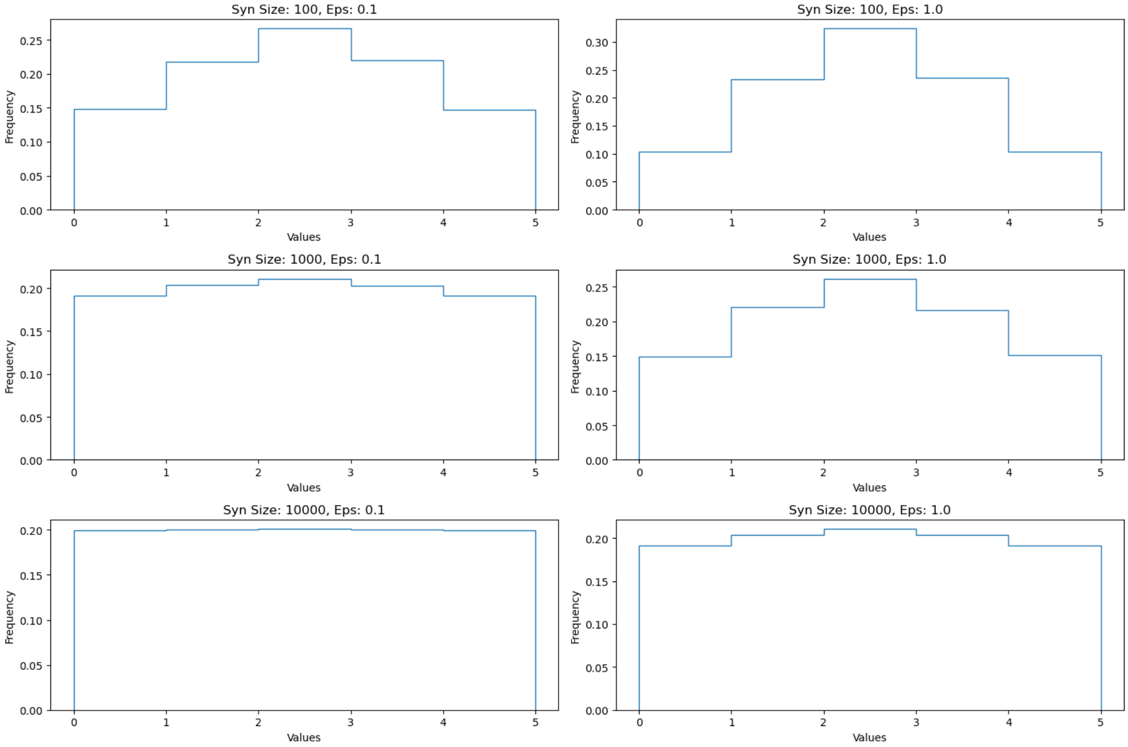


Figure 4.3: Probabilities of the AS-DPSD histogram. Original data is shown in figure 4.2.

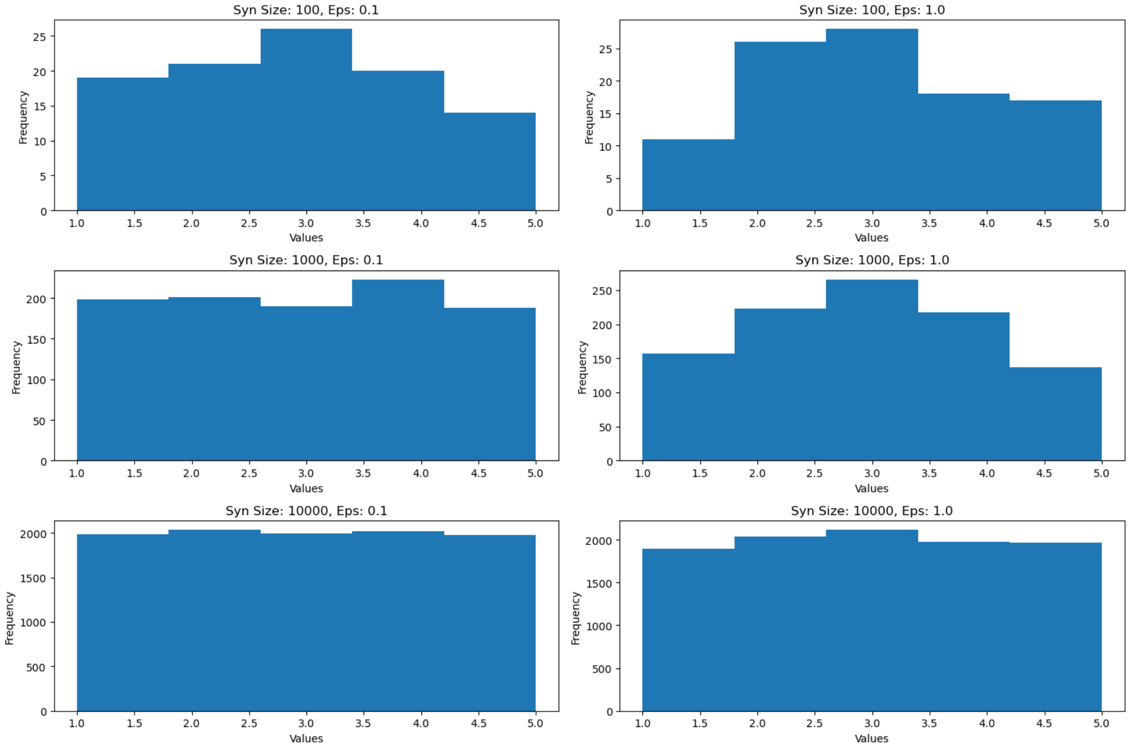


Figure 4.4: Illustration of final output of the AS-DPSD algorithm. Original data is shown in figure 4.2, and the probabilities from which the synthetic data is drawn is shown in figure 4.3

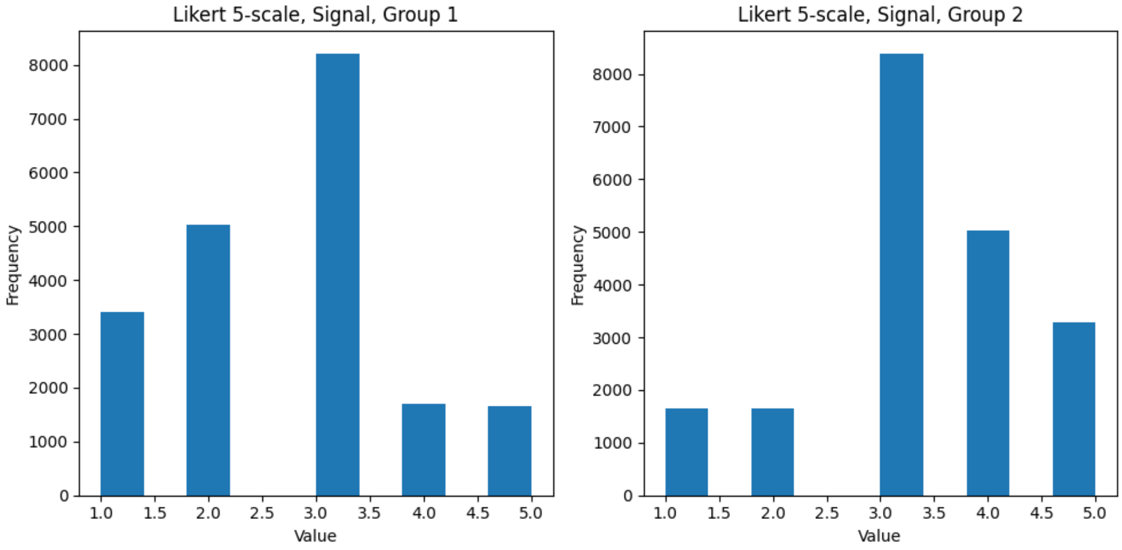


Figure 4.5: Simulated Likert 5-Scale data

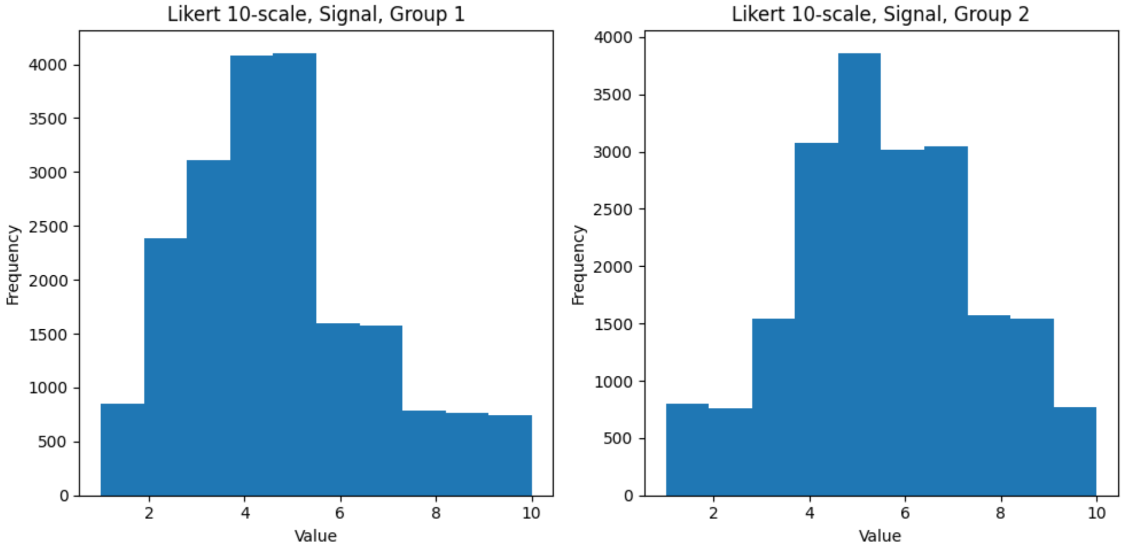


Figure 4.6: Simulated Likert 10-Scale data

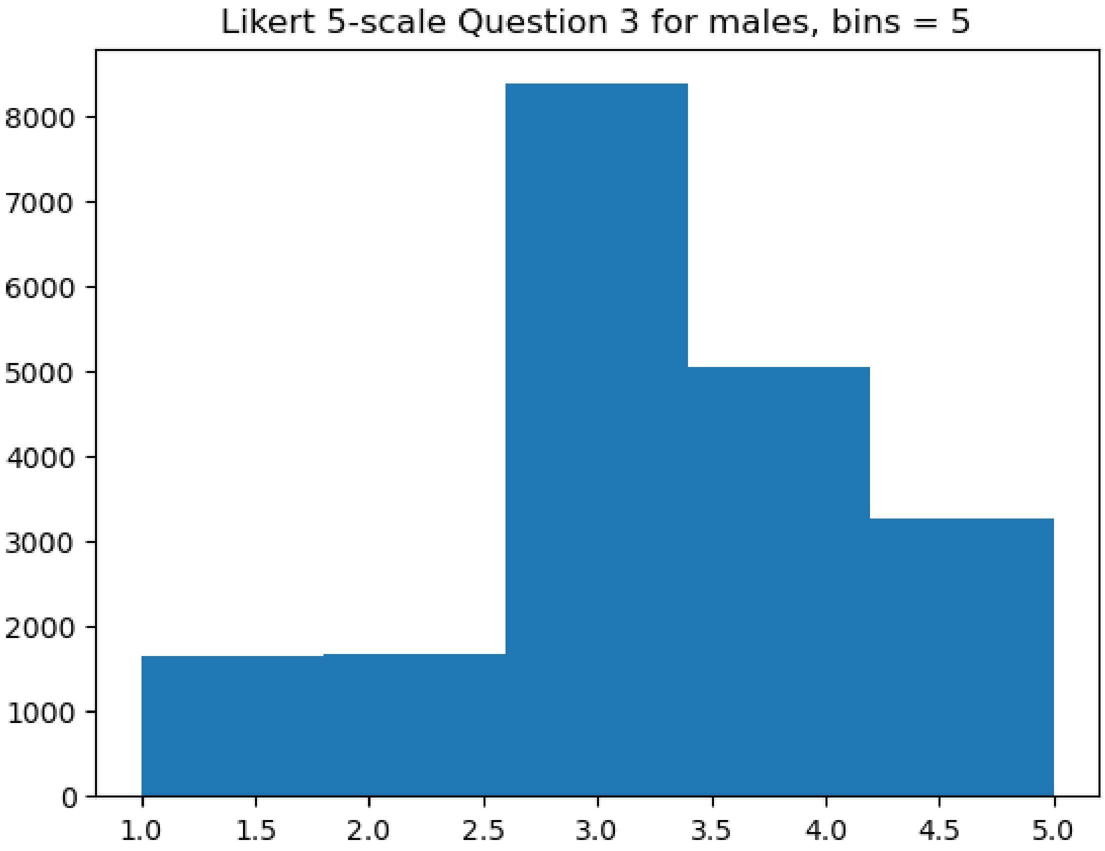


Figure 4.7: Likert 5-scale without AS-DPSD algorithm, bins set to 5

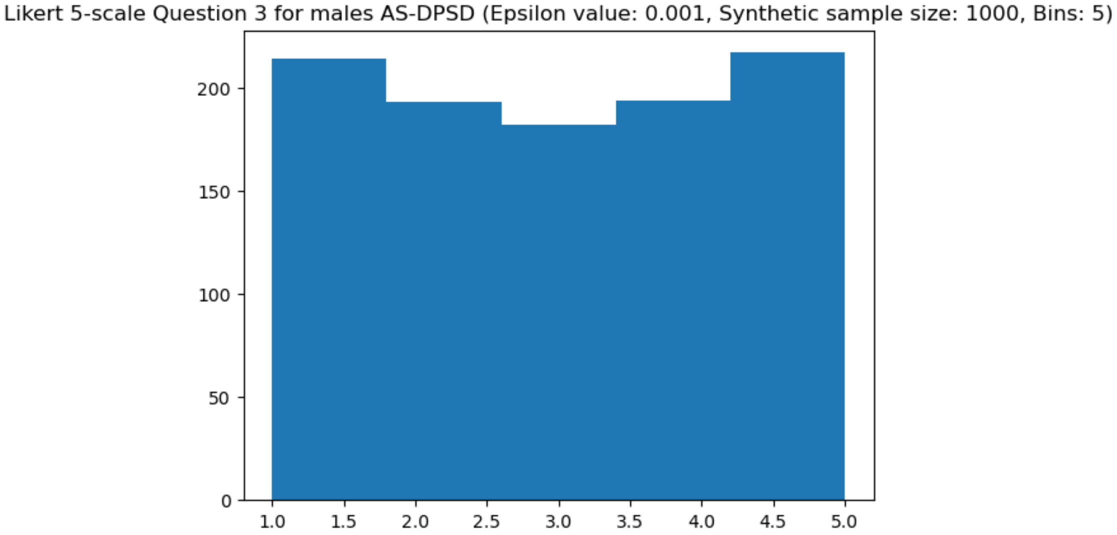


Figure 4.8: Likert 5-scale question 3 for males, synthetic sample size 1000, epsilon 0.001

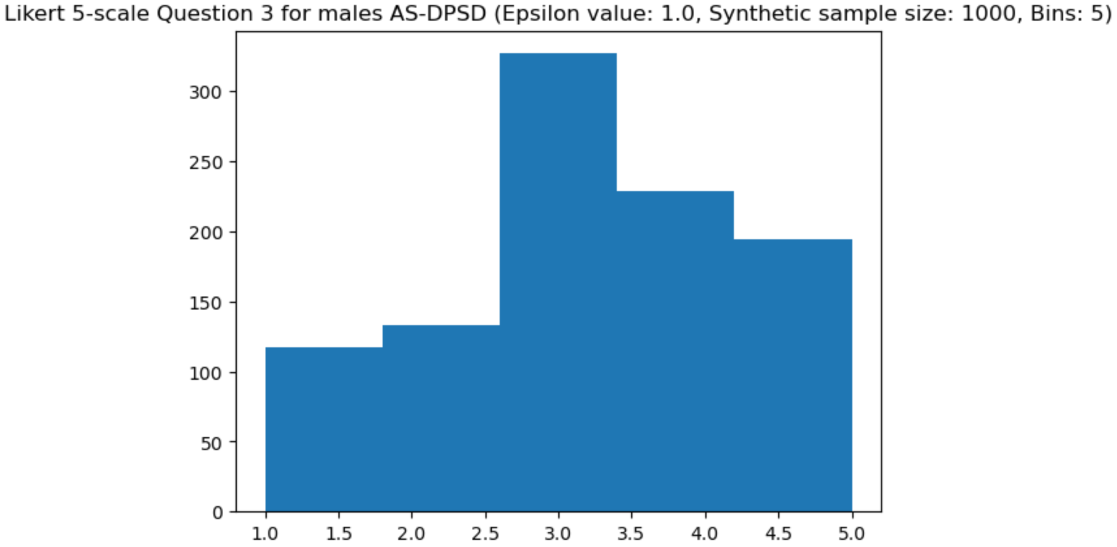


Figure 4.9: Likert 5-scale question 3 for males, synthetic sample size 1000, epsilon 1

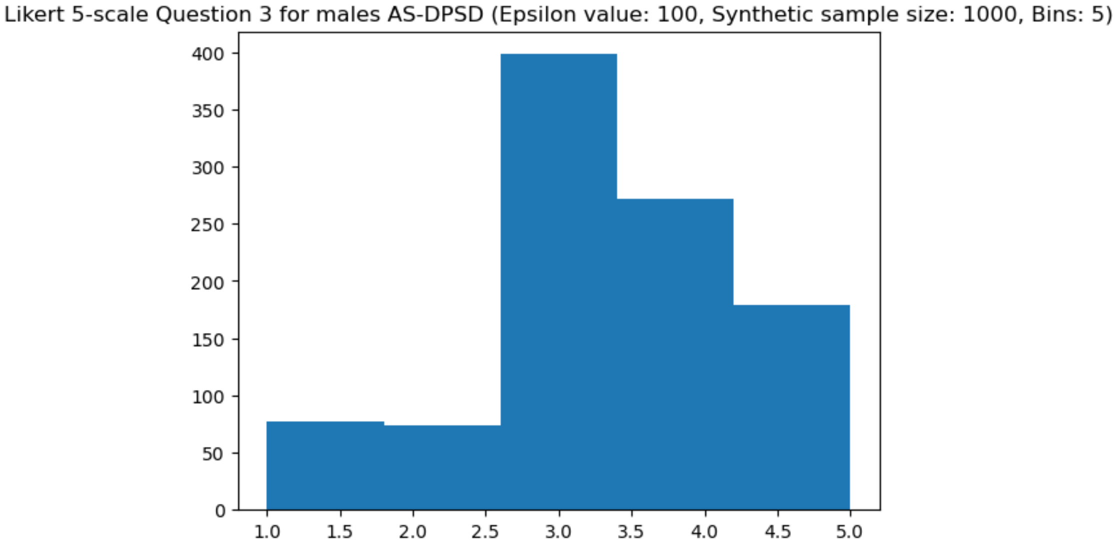


Figure 4.10: Likert 5-scale question 3 for males, synthetic sample size 1000, epsilon 100

Cardiovascular dataset

In the following graphs its easier to see what kind of effect different bin sizes have on the AS-DPSD algortihm. Looking at Figures 4.15, 4.16 and it can be seen that having more bins makes it less accurate even though other hyperparameters are set same, thus can be concluded that by decreasing the number of bins in AS-DPSD makes the algorithm more accurate, which makes it easier to generate private synthetic data.

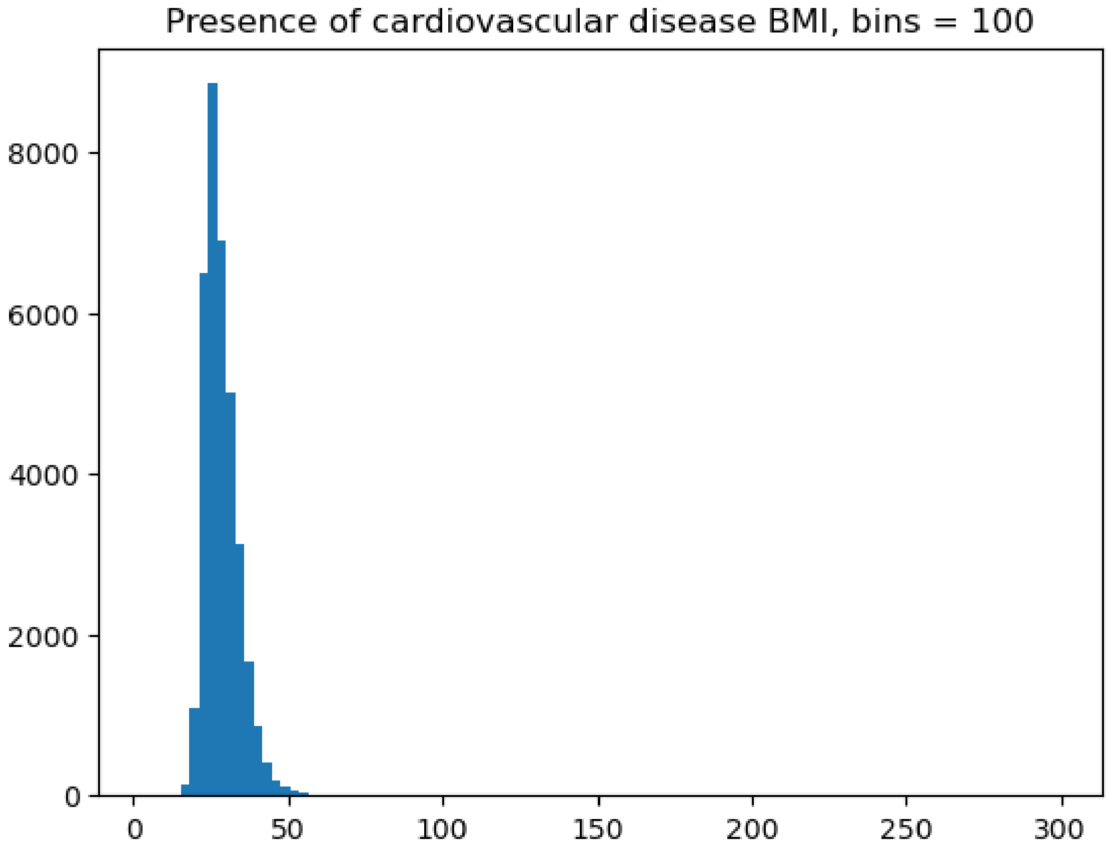


Figure 4.11: Cardiovascular Dataset Presence of Cardiovascular disease without AS-DPSD algorithm, bins set to 100

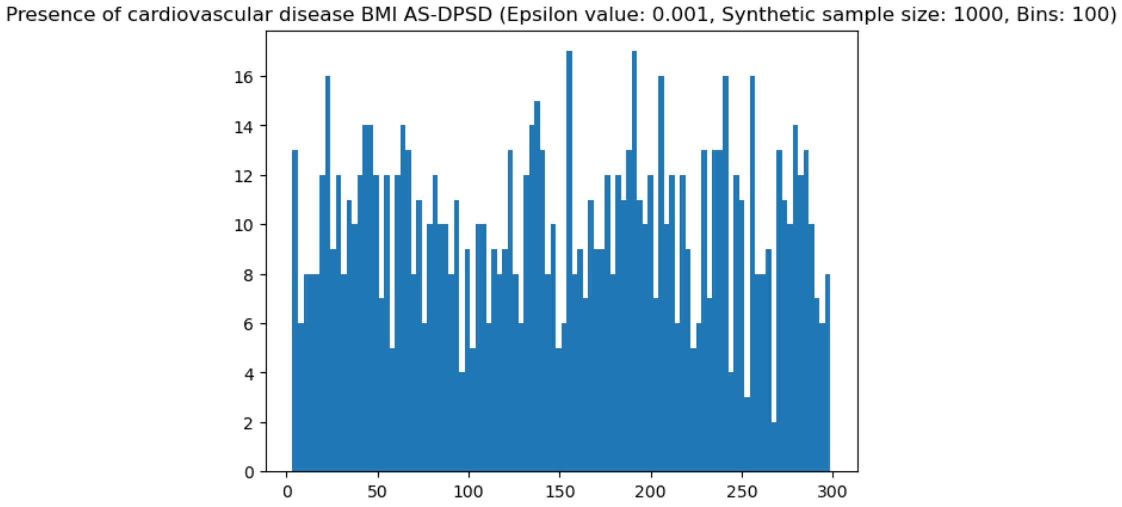


Figure 4.12: Cardiovascular Dataset AS-DPSD Presence of Cardiovascular disease Synthetic sample size 1000, bins 100, epsilon 0.001

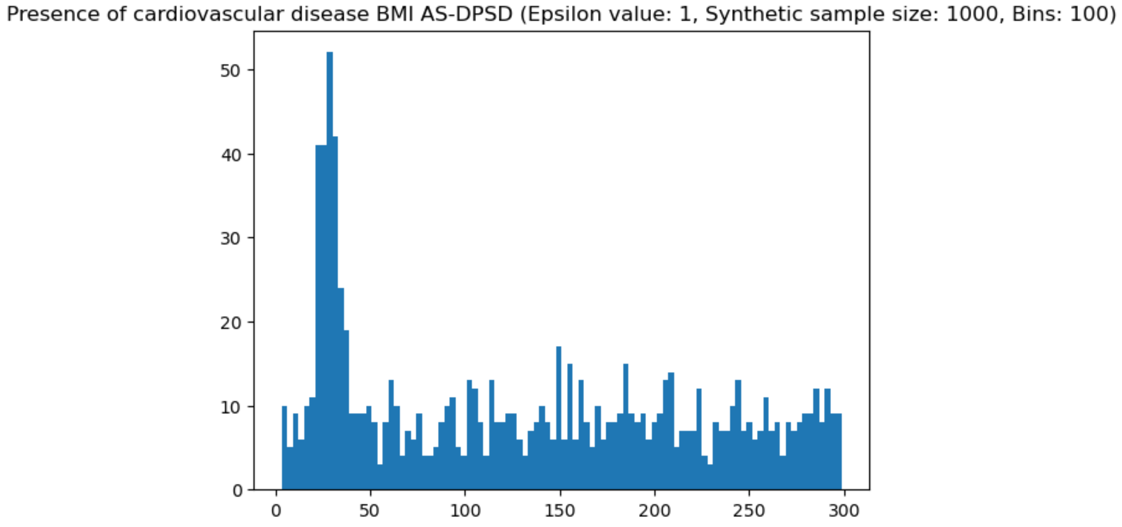


Figure 4.13: Cardiovascular Dataset AS-DPSD Presence of Cardiovascular disease Synthetic sample size 1000, bins 100, epsilon 1

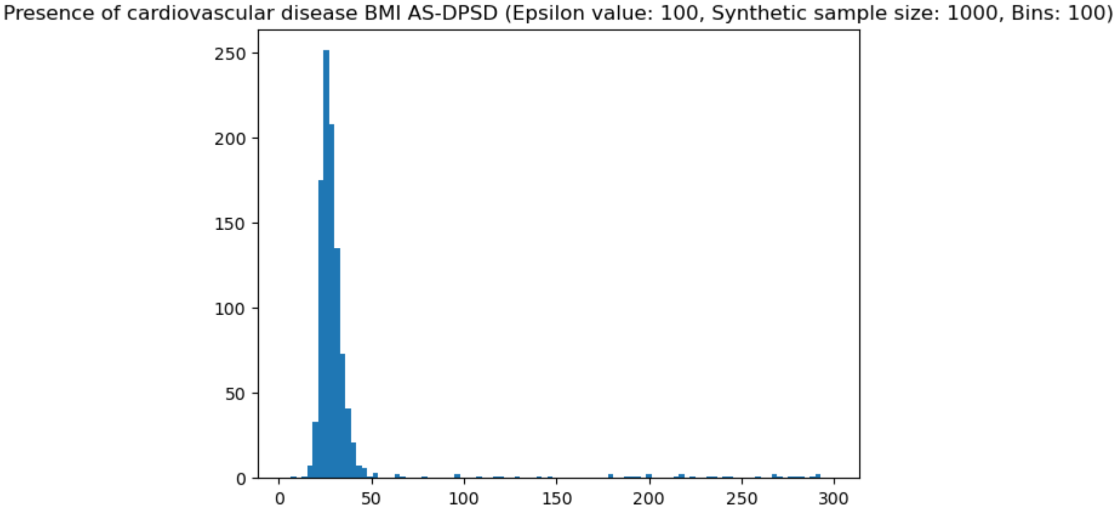


Figure 4.14: Cardiovascular Dataset AS-DPSD Presence of Cardiovascular disease Synthetic sample size 1000, bins 100, epsilon 100

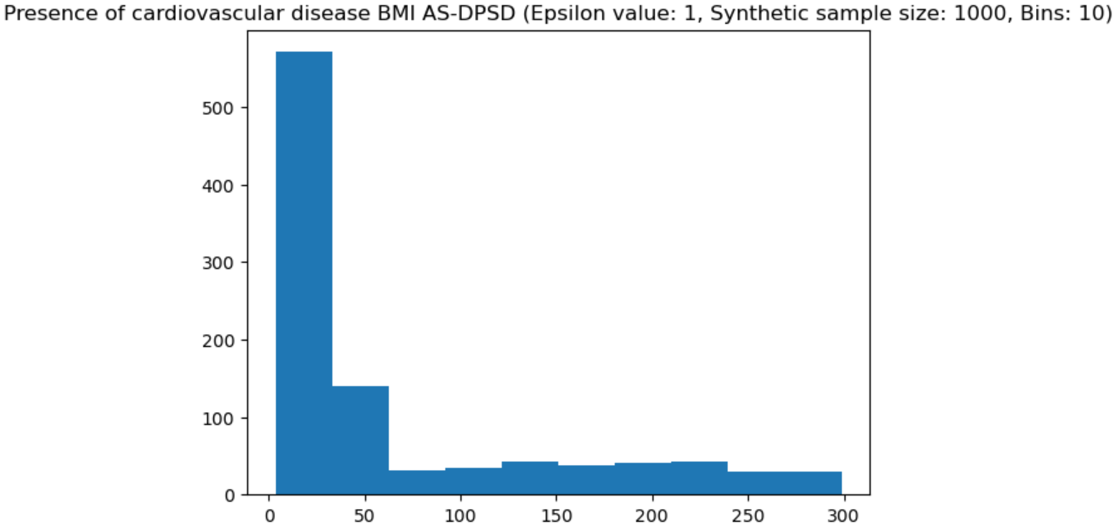


Figure 4.15: Cardiovascular Dataset AS-DPSD Presence of Cardiovascular disease Synthetic sample size 1000, bins 10, epsilon 1

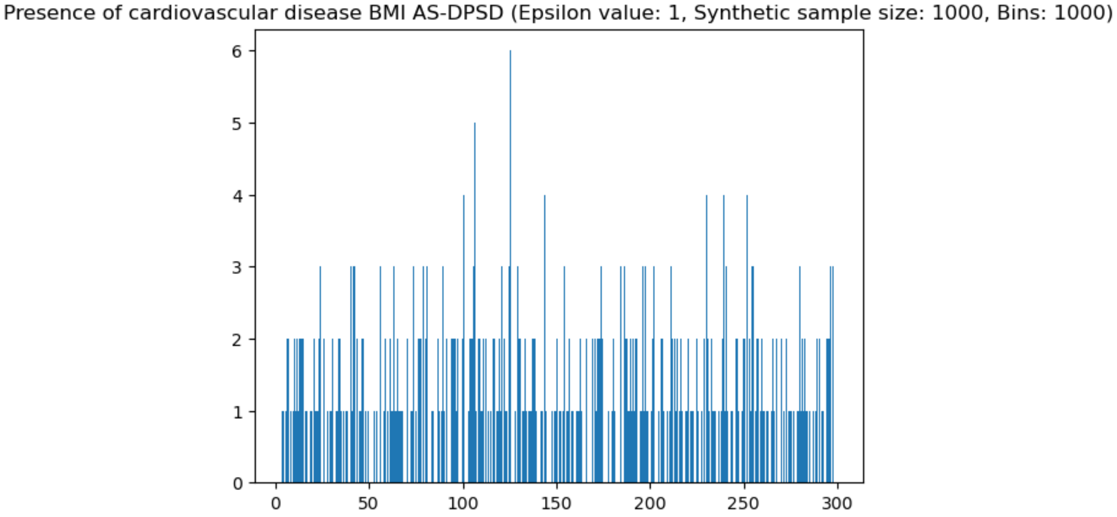


Figure 4.16: Cardiovascular Dataset AS-DPSD Presence of Cardiovascular disease Synthetic sample size 1000, bins 1000, epsilon 1

5 Statistical Hypothesis Testing

5.1 Parametric Hypothesis Testing

Parametric hypothesis testing relies on specific probability distributions and assumptions population characteristics using data collected from samples. One of the parametric test is Student´s T-test. The fundamental concept in parametric hypothesis testing is to select appropriate test statistics, such as p-value or t-statistic [17][20].

Another concept is the p-value, which describes how much results deviate from null hypothesis. Smaller p-value indicates stronger evidence and larger p-value suggests there is less evidence for the conclusion of null hypothesis. Practical example would be to have 2 groups with different test scores (people who study and people who do not study). Conducting a Student's t-test with p-value of 0.05 we have a stronger evidence that studying helps (null hypothesis in this scenario is "there is no significant difference in test scores between the study group and the non-study group). Alternative hypothesis in this case is "Students who study perform better on the test". So when the p-value is less than α, null hypothesis is rejected in favor of alternative hypothesis. In this thesis $\alpha = 0.05$ is used as a threshold value for the significance level in parametric hypothesis tests.

Normally parametric hypothesis testing is not used in ordinal data, however, based on study [17] normally distributed Likert scale survey has only 5 percent difference compared to nonparametric Mann Whitney-Wilcoxon and another paper [20] says assumption of normality does not need to hold when conducting parametric hypothesis testing with Student's t-test. In this thesis I use Student's t-test to conduct the experiment, where X is normally distributed and I want to show that $X \sim X'$.

Definition 10. *Independent two-sample Student's t-test:*

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{2}{n}}}$$

where

$$s_p = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}.$$

A hypothesis test can have two types of errors: type I and type II. A type I error occurs if the test incorrectly rejects H_0 when it is in fact true. A type II error occurs if the test fails to reject H_0 when the alternative hypothesis is true [24].

In this thesis all Student's t-test experiments were done using SciPy (v1.10.1).

5.1.1 Students T-test

Student's t-test also known as t-test, is widely used statistical test to compare groups' means for a particular variable. It was proposed by Mr. William Gosset, who published his work under the pseudonym "Student", thus the name Student's t-test is commonly used in literature. Student's t-test is similar to the z-test in the way that a Student's t-test may apply to a single sample or two-sample situations [35]. Usually Student's t-test is used for normally distributed continuous data, however [20] and [17] conducted otherwise. The paper [20] demonstrated the validity of Student's t-test by simulating extremely non-Normal data. For example [17] showed that for multimodal distributions Student's t-test was more powerful when the strong multimodal distribution was compared with a skewed or peak distribution, with power advantages up to 26 percent. Paper [17] showed that for Likert scale Type I error rate was close to the nominal value of 5 percent for all sample sizes and for all combinations of distributions. When conducting Student's t-test we want also know which variation of Student's t-test should be the most suitable for the testing. There exists 3 variations of Student's t-test and they all excel in specific scenarios [35]

Usually conducting tests with Student's t-test we want to first know if the distribution we evaluate is normally distributed or not. So we want to know if the assumption of normality holds, this is conducted by using Shapiro-Wilk test to show the normality of it first. Student's t-test is said to be parametric because it is assumed it follows normal distribution for the data.

5.2 Nonparametric hypothesis testing

Statistical hypothesis tests such as Mann-Whitney U test, Wilcoxon two-sample test and Kruskal-Wallis tests are said to be nonparametric since is assumed it does not follow a no specific distribution where in the case Student's t-test does [20]. DPSD Likert scale examples in figure 5.3 are ordinal data and could work better for nonparametric tests such as Wilcoxon, Mann-Whitney U test and Kruskal-Wallis.

5.2.1 Mann-Whitney U test

The purpose of Mann-Whitney U test is to specify if two groups from the same distribution. Similarly like Student's T-test. However Student's T-test assumes data is normally distributed where Mann-Whitney U test does not.

Mann-Whitney U test calculates the U statistic for each group U_x, U_y. Mann-Whitney U test is defined mathematically:

Definition 11.

$$U_x = n_x n_y + \frac{n_x(n_x+1)}{2} - R_x \tag{5.1}$$

$$U_y = n_x n_y + \frac{n_y(n_y+1)}{2} - R_y \tag{5.2}$$

[36]

where n_x, n_y represents the data and R_x, R_y represents the sum of ranks assigned on groups U_x, U_y.

5.3 Type I and Type II errors

A Type I error is an incorrect rejection of a true null hypothesis. This type of error is also called a false positive. The probability of a Type I error can be reduced by decreasing the significance level (e.g., from 0.05 to 0.01). Type I error comes from if statistical hypothesis test such as Mann Whitney U-test says that there exists difference between two samples when there actually is no difference [24]. Unlike Type I error, Type II error occurs if a statistical hypothesis test fails to reject false hypothesis. In this thesis we focus on the null hypothesis H_0, which is that two samples or groups are drawn from the same distribution. An illustration of the possible outcomes of statistical hypothesis tests are shown in Figure 5.1.

	H_0 is true	H_0 is false
Reject H_0	Type I error	TP
Fail to reject H_0	TN	Type II error

Figure 5.1: Illustration of all of the possible outcomes of statistical tests.

6 Results

This chapter consists of evaluating parametric tests such as the Student's t-test on DP synthetic data generated from Likert 5-scale data, Likert 10-scale data and Cardiovascular dataset BMI data. Type I and Type II errors are used to test whether statistical hypothesis testing performed on differentially private synthetic data is likely lead to loss of tests validity or decreased power. I will also conduct the evaluation on nonparametric tests such as Mann-Whitney U test on Likert 5-scale data, Likert 10-scale data and Cardiovascular dataset BMI data.

6.1 Experimental Evaluation

Set of experiments was performed when using DP synthetic datasets generated from both simulated and real-world datasets. This is to evaluate the utility of the Student's t-test on DP synthetic data empirically. In the following subsections, we present the datasets, the implementation details of the DP synthetic data generation method used (Additively Smoothed Differentially Private Synthetic Data) and the experiments conducted. The original dataset in this thesis is Kaggle Cardiovascular dataset [37]. This dataset consists of two variables, one is binary and the other one is a continuous variable. The binary variable represents the label of two groups (healthy and non-healthy). The continuous variable on the other hand is used to compare the groups (healthy and nonhealthy) with Student's t-test and Mann-Whitney U test. Simulated data in this thesis is 5-point Likert scale, 10-

point Likert scale. Similarly like the original dataset implementation, this dataset was split into binary variable and ordinal variable when evaluating the power (Type II error) of Student's t-test and Mann-Whitney U test. We also evaluate the utility (Type I error) of Student's t-test and Mann-Whitney U test with 5-point Likert scale and 10-point Likert scale AS-DPSD data. More accurate description of how the simulated data is generated is shown in the Chapter 6.2.

6.2 Implementations

In this thesis the data is generated using Pythons built-in Random library. The type of data which was generated follows mainly 5-point Likert scale, where the choices are Neither agree or disagree, disagree, agree, strongly agree and strongly disagree. I used these answers to be generated since they are commonly used in question data -related queries [18]. I also take a brief look into 10-point Likert scale. I also choose one parametric test and one non-parametric tests for evaluation of the DP synthetic data. The main focus relies on evaluating the utility of Student's t-test, however MW u-test is also used for evaluation of the same data. In this thesis the data was simulated very careful way in which it still resembles normal distribution even though the data itself is ordinal.

6.2.1 Simulated Data

This procedure goes as follows: First we initialize c which has length of histogram count vector h. Then we initialize m which is the number of synthetic data to be drawn. The privacy budget is ϵ which has the values of $[0.05, 1]$. Original data has the size of 10000, from which AS-DPSD algorithm takes m amount of data which is 50, 100, 500, 1000. Sample sizes greater than 5, do not require the assumption of normality and will yield nearly correct answers even for manifestly nonnormal and asymetric distributions like exponentials [21].

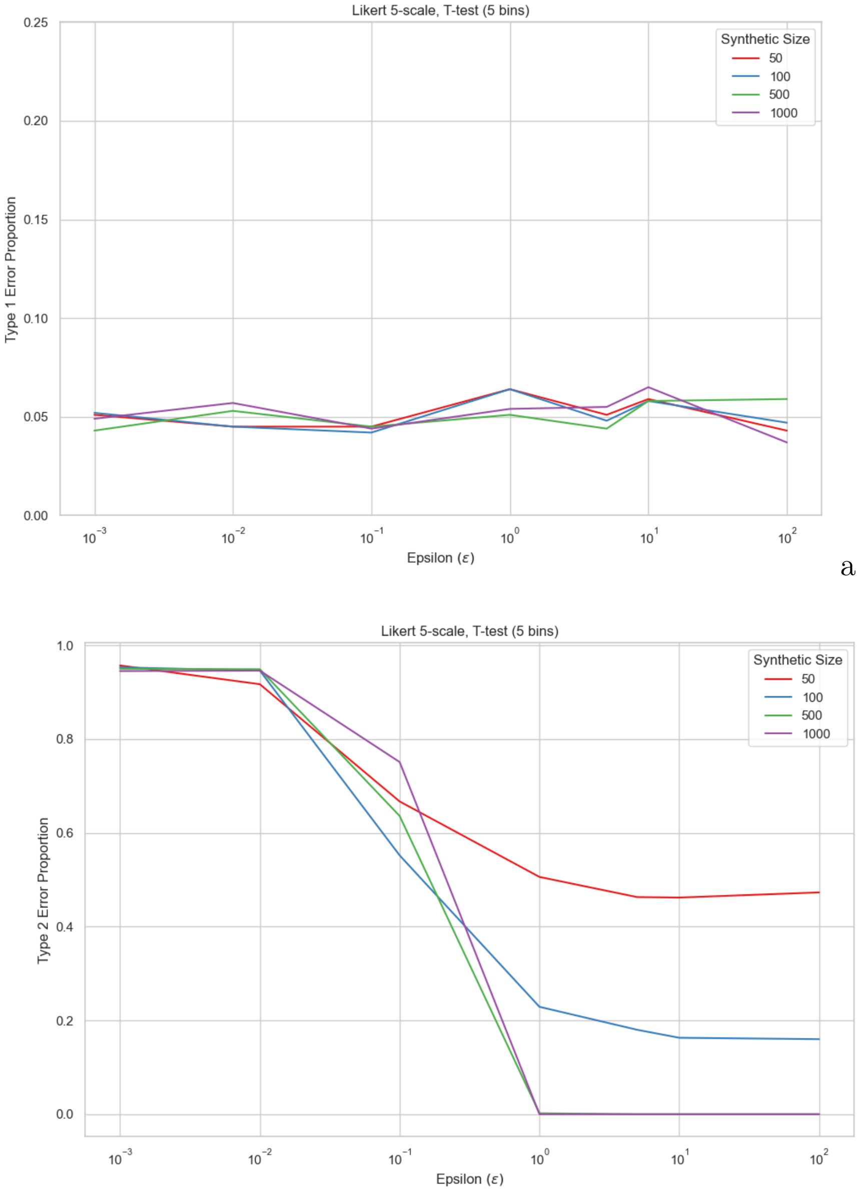
We can see from this simulated Likert Data that all of these behave similarly. In this particular scenario Type II error starts to fall of when the dataset size is 100 and the epsilon is 1 or above. This is also predictable because mostly commonly in the DP-literature the privacy budget is set to be between 0.01 and 1. This can be seen for example in the paper [15], where different groups competing in Differential Privacy Synthetic Data Challenge used ϵ values ranging from 0.01 to 1.

It is also important to categorize the Student's t-test and Mann-Whitney U test evaluation between power and utility. This is done by generating simulated signal data and nonsignal data. In signal data the data is sampled from two distinct distributions while in nonsignal data the data is sampled from the same distribution. Experiment is done in this way because we want to show Type I for nonsignal data and Type II errors for signal data. Illustration of different types of errors is shown in Figure 5.1 and in Figure 5.2.

6.2.2 Results with original data before AS-DPSD

The results of Student's t-test and Mann-Whitney U test with the original data before AS-DPSD algorithm were the following. The p-value of the Student's t-test for the Nonsignal 5-point Likert scale before AS-DPSD algorithm was ≈ 0.46 and for Nonsignal 10-point Likert Scale before AS-DPSD algorithm was ≈ 0.06. The p-value of the The Mann-Whitney U test for the Nonsignal 5-point Likert Scale was ≈ 0.35 and for Nonsignal 10-point Likert Scale before AS-DPSD algorithm was ≈ 0.08.

Student's t-test for the Signal 5-point Likert scale before AS-DPSD algorithm was 0 and for Signal 10-point Likert Scale before AS-DPSD algorithm was $\approx 6.21 \times 10^{-262}$. The Mann-Whitney U test for the Signal 5-point was 0 and for Signal 10-point Likert Scale before AS-DPSD algorithm was $\approx 7.21 \times 10^{-296}$.



a) Non-Signal Data

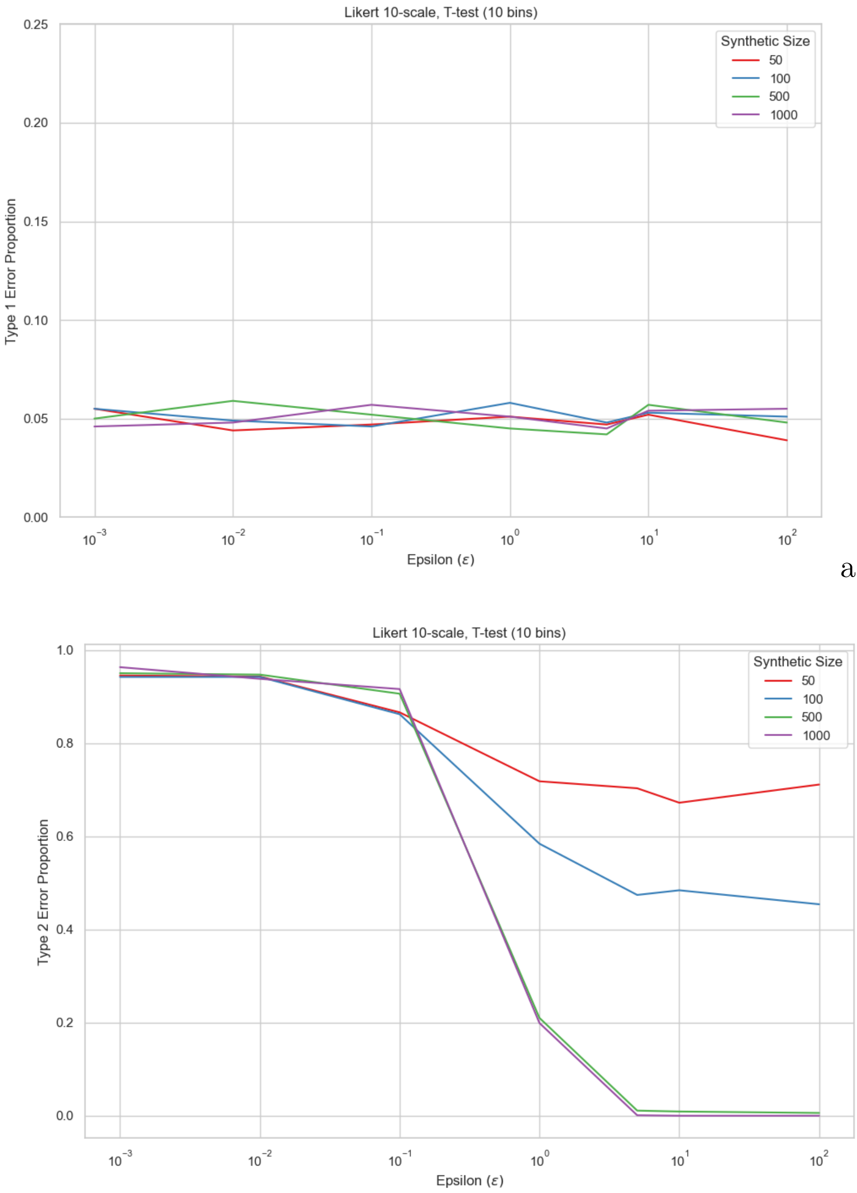
b) Signal Data

Figure 6.1: Student's T-test for 5-point Likert Scale data

For both signal data, all of the p-values before AS-DPSD algorithm are below 0.05 and for non-signal data the values are above 0.05.

6.2.3 Data generation

Nonsignal data was generated to evaluate the validity of Student´s T-test and Mann-Whitney U test with Type I errors. Note that test is valid, for example Student´s



a) Non-Signal Data

b) Signal Data

Figure 6.2: Student's T-test for 10-point Likert Scale data

T-test is valid, if the proportion of Type I error is below significance level (0.05). Signal data on the other hand was generated to evaluate the power of statistical hypothesis tests in this thesis. The power of a statistical hypothesis test is the probability of correctly rejecting a false null hypothesis that is below significance level [16]. Likert 5-scale nonsignal data was generated by using Numpys Random-library. Generation consists of creating two groups of data "male" and "female" and creating a scale with options "Strongly Disagree", "Disagree", "Neutral", "Agree", "Strongly Agree". The weights were put [0.1, 0.2, 0.4, 0.2, 0.1] for the 5-scale data, from which 10000 males and 10000 females were sampled evenly. Likert 5-scale signal data was generated similar way as the nonsignal data, but male and females having different weights. For instance females having weights [0.2, 0.3, 0.5, 0.1, 0.1] and males [0.1, 0.1, 0.5, 0.3, 0.2]. Nonsignal and signal Likert 10-scale data follows the same protocol.

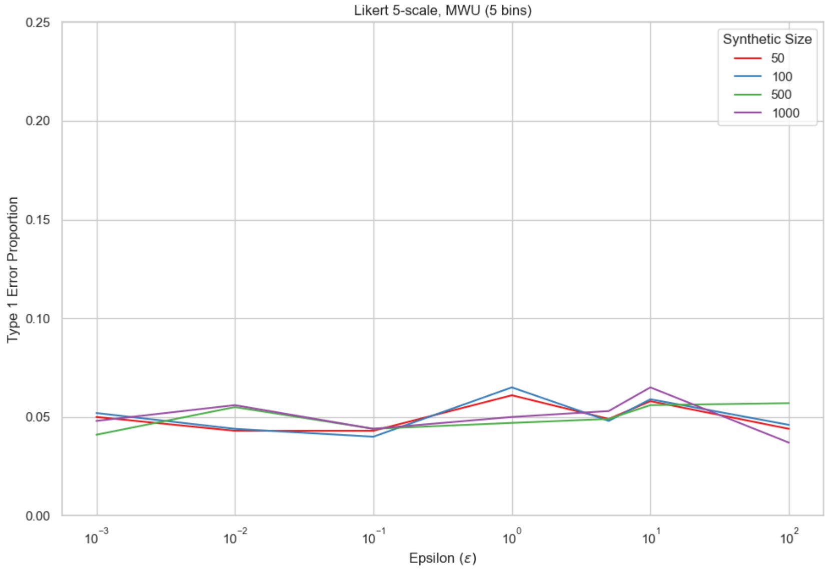
6.2.4 Real world data

For the real world data, the experimentation was used with Kaggle Cardiovascular dataset. Comparison was done by calculating body-mass index (BMI) from weight and height columns. Then the experimentation was done by splitting the data into two. One group that have cardiovascular disease and another group who do not have cardiovascular disease (yes-disease and non-disease). Null hypothesis H_0 in this case is "The BMI level for individuals with the presence of cardiovascular disease and the ones with absence cardio-vascular disease originate from the same distribution" [16]. Experiments were repeated 1000 times to compute the proportion of Type II error.

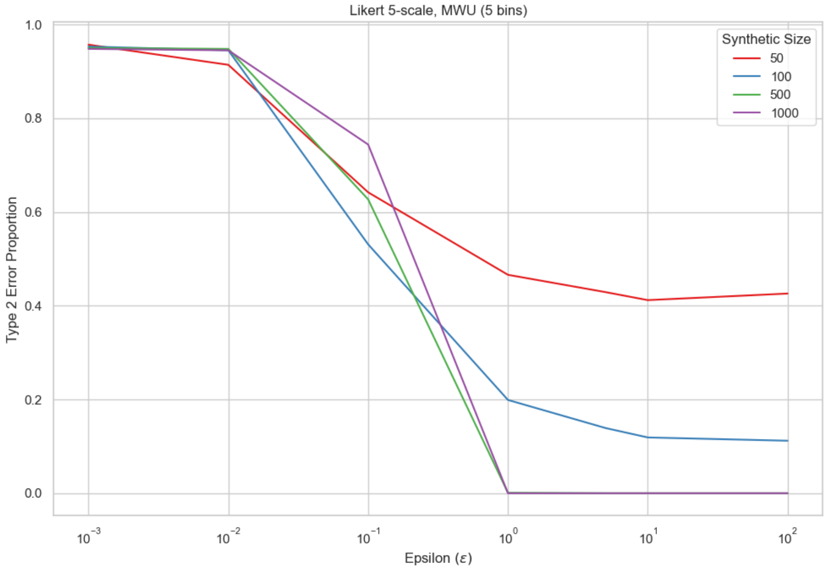
6.2.5 Results with data after AS-DPSD algorithm

The Figures 6.1-6.4 have very similar results on both Mann-Whitney U test and Student's T-test. There is almost no difference between Likert 5-scale and Likert 10-scale on Type I error neither for Mann-Whitney U test nor Student's T-test. However for Type II error, Likert 5-scale reaches Type II error at 10^0 for synthetic size 1000 for both Mann-Whitney U test and Student's T-test. Likert 10-scale reaches Type II error at around 7 for both Mann-Whitney U test and Student's t-test.

We can see from the Figure 6.5 that the difference between Mann-Whitney U



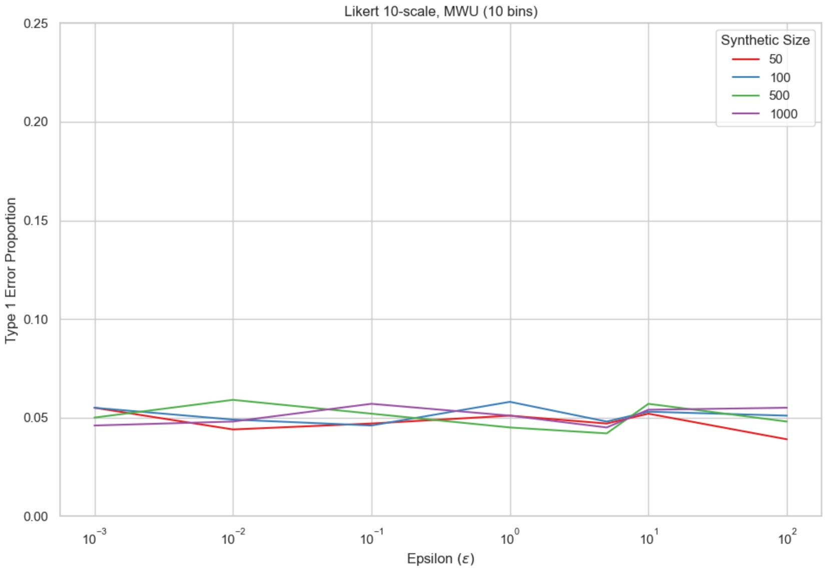
a) Mann-Whitney U test Type I error for 5-point Likert Scale data



b) Mann-Whitney U test Type II error for 5-point Likert Scale data

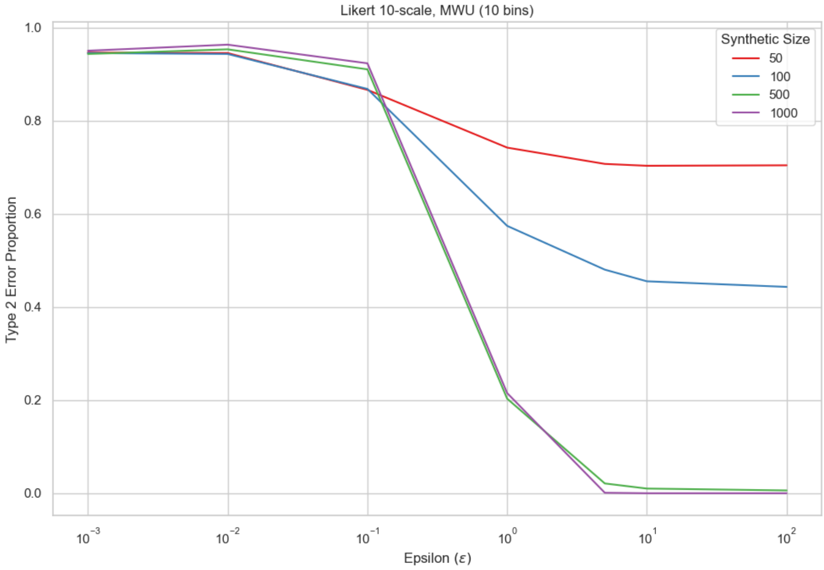
Figure 6.3: Mann-Whitney U test evaluation with Type I and Type II errors on AS-DPSD Exponential Mechanism with 5-point Likert scale scale data.

test and Student's T-test is that the Type II error starts to fall sooner towards 0.05 and for Student's T-test much later. However both tests reach 0.05 mark on 10^2. From Figure 6.6 the impact of bin size can be seen. The decrease in power is larger



a) Non-Signal Data

10-point Likert Scale Mann-Whitney U test

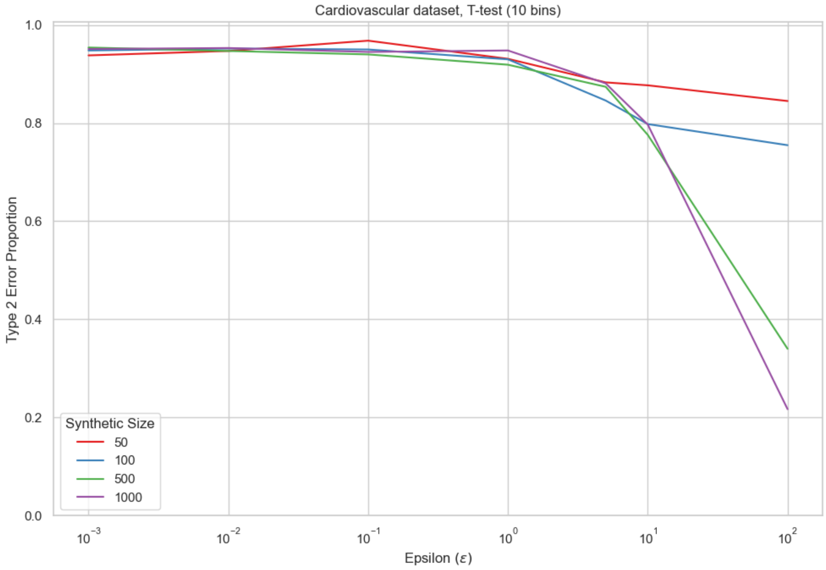


b) Signal Data

10-point Likert Scale Mann-Whitney U test

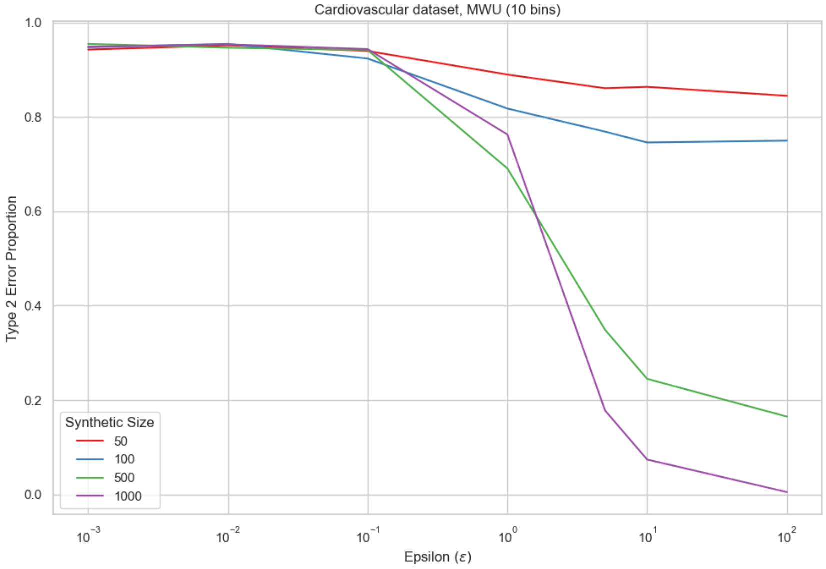
Figure 6.4: Mann-Whitney U test evaluation with Type I and Type II errors on AS-DPSD Exponential Mechanism with 10-point Likert Scale

on Figure 6.6(a) than on Figure 6.6(b). Similar results as in Figure 6.6 can be seen in Figure 6.7.



a) Signal Data

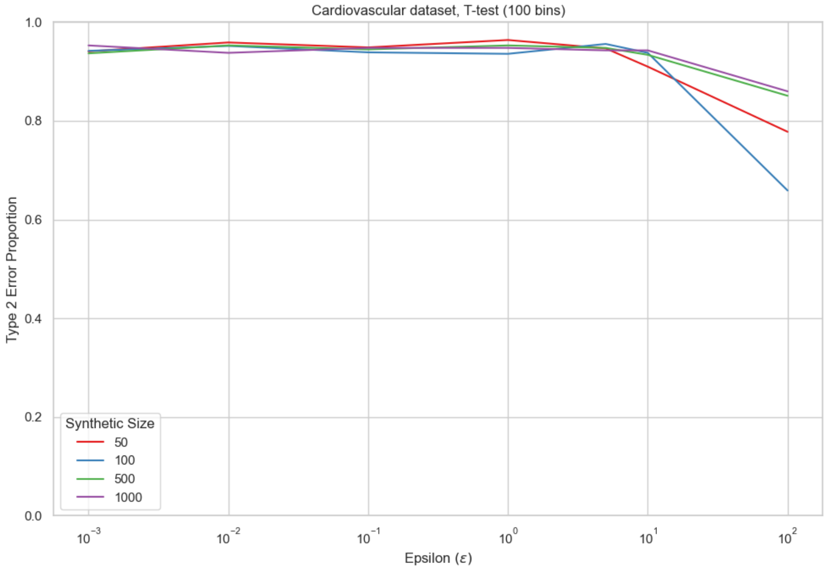
cardiovascular dataset Student's t-test



b) Signal Data

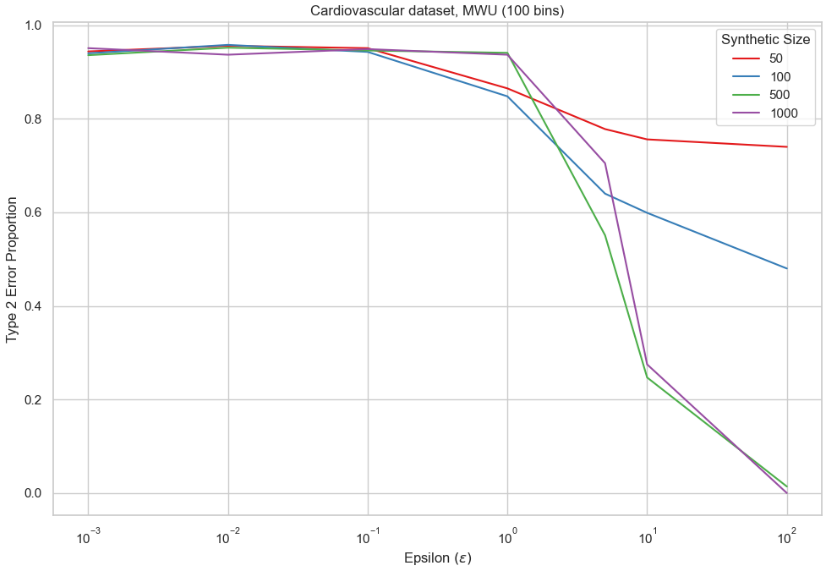
cardiovascular dataset Mann-Whitney U test

Figure 6.5: Cardiovascular dataset 10 bins



a) Signal Data

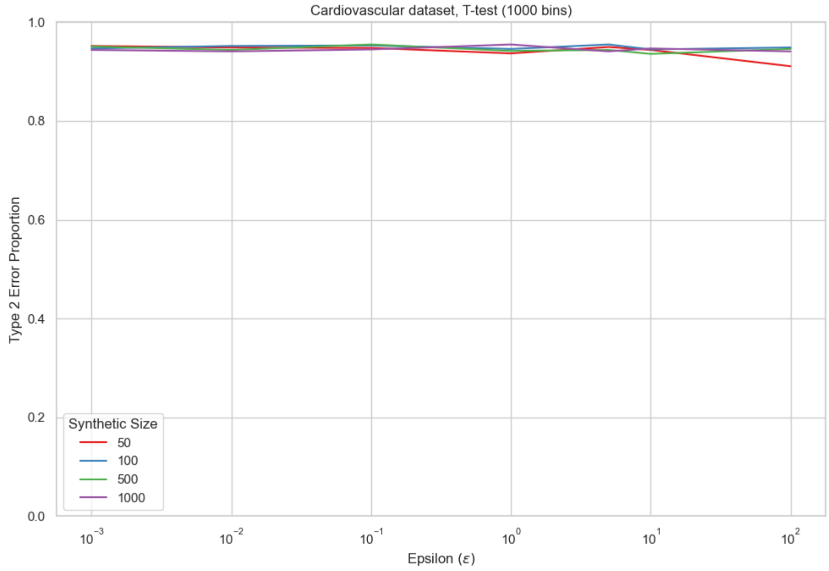
cardiovascular dataset Student's t-test



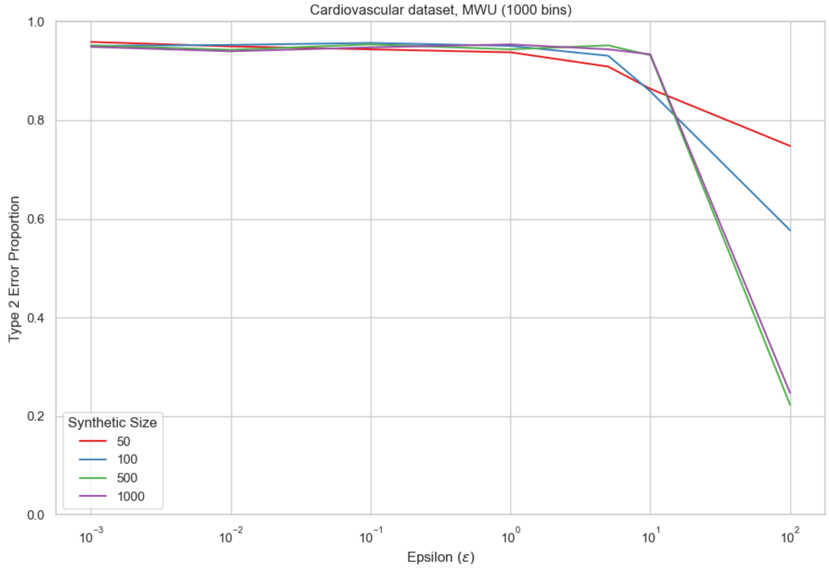
b) Signal Data

cardiovascular dataset Mann-Whitney U test

Figure 6.6: Cardiovascular dataset 100 bins



a) Signal Data cardiovascular dataset Student's t-test



b) Signal Data cardiovascular dataset Mann-Whitney U test

Figure 6.7: Cardiovascular dataset 1000 bins

7 Conclusion

The aim of this thesis was to study the validity and utility of the Student's t-test and Mann Whitney U test on DPSD that is sampled from AS-DPSD-algorithm. In this study it was shown that Mann-Whitney U test and Student´s T-test performed on Likert 5-scale data led to decreased power for ϵ values being $\leq 10^0$ for both tests. Mann-Whitney U test and Student's T-test performed on Likert 10-scale led to decreased power for ϵ values being $\leq 10^1$ for both tests. Mann-Whitney U test and Student's T-test got similar Type I errors on both Likert 5-scale data and Likert 10-scale data, where the p-value being around 0.05. Note that the test is valid if the Type I error is less or equal than the significance level. Looking at Figure 6.2(a) for instance it can be seen that the Type I error is not always less than 0.05 for all ϵ values. By using AS-DPSD one should carefully consider the hyperparameters of the algorithm or when simulating data. This is more crucial when the simulated data is ordinal scale. However according the results of this thesis and paper [18], it should not have too much effect when using statistical hypothesis test to ordinal scale data.

7.1 Future works

In future works it would be beneficial to include the evaluation of utilities of other parametric and non-parametric tests used with the same AS-DPSD (Additively Smoothed Differentially Private Synthetic Data) - algorithm. Also more real datasets are needed, since in this thesis I only delved into Kaggles Cardiovascular dataset. The variety of hyperparameters is needed with the combination of dataset size and the type of simulated data. Given only few Likert type of datas with really specific distributions were used, it would be beneficial to see different distributions, other than normal distributions. Smaller and larger Likert scales could be used in future experiments.

References

[1] K. N. Cukier, V. Mayer-Schönberger, and M. Pitici, "The rise of big data: How it's changing the way we think about the world", 2014. [Online]. Available: `https://api.semanticscholar.org/CorpusID:156813417`.

[2] K. Emam, F. Dankar, R. Vaillancourt, T. Roffey, and M. Lysyk, "Evaluating the risk of re-identification of patients from hospital prescription records", *The Canadian journal of hospital pharmacy*, vol. 62, pp. 307–19, Jul. 2009. DOI: `10.4212/cjhp.v62i4.812`.

[3] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, *Differentially private generative adversarial network*, 2018. arXiv: `1802.06739 [cs.LG]`.

[4] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015.

[5] R. Ratra, P. Gulia, and N. Gill, "Evaluation of re-identification risk using anonymization and differential privacy in healthcare", *International Journal of Advanced Computer Science and Applications*, vol. 13, Jan. 2022. DOI: `10.14569/IJACSA.2022.0130266`.

[6] J. P. Near and C. Abuah, *Programming Differential Privacy*. 2021, vol. 1. [Online]. Available: `https://uvm-plaid.github.io/programming-dp/`.

[7] D. Desfontaines and B. Pejó, *Sok: Differential privacies*, 2022. arXiv: 1906.01337 [cs.CR].

[8] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy", *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, Aug. 2014, ISSN: 1551-305X. DOI: 10.1561/0400000042. [Online]. Available: https://doi.org/10.1561/0400000042.

[9] N. W. Remerscheid, A. Ziller, D. Rueckert, and G. Kaissis, *Smoothnets: Optimizing cnn architecture design for differentially private deep learning*, 2022. arXiv: 2205.04095 [cs.CV].

[10] B. Wang and N. Hegde, "Privacy-preserving q-learning with functional noise in continuous spaces", in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[11] D. Rubin, "Discussion: Statistical disclosure limitation. journal of official statistics", vol. 9, pp. 461–468, 1993.

[12] J. Jordon, L. Szpruch, F. Houssiau, *et al.*, *Synthetic data – what, why and how?*, 2022. arXiv: 2205.03257 [cs.LG].

[13] C. M. Bowen and J. Snoke, "Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge", *Journal of Privacy and Confidentiality*, vol. 11, no. 1, Feb. 2021. DOI: 10.29012/jpc.748. [Online]. Available: https://journalprivacyconfidentiality.org/index.php/jpc/article/view/748.

[14] P. Nanayakkara, J. Bater, X. He, J. Hullman, and J. Rogers, *Visualizing privacy-utility trade-offs in differentially private data releases*, 2022. arXiv: 2201.05964 [cs.CR].

[15] C. Bowen and F. Liu, “Comparative study of differentially private data synthesis methods”, *Statistical Science*, vol. 35, Feb. 2016. DOI: 10.1214/19-STS742.

[16] I. Montoya Perez, P. Movahedi, V. Nieminen, A. Airola, and T. Pahikkala, “Does differentially private synthetic data lead to synthetic discoveries?”, *arXiv e-prints*, arXiv–2403, 2024.

[17] J. de Winter and D. Dodou, “Five-point likert items: T test versus mann-whitney-wilcoxon”, *Practical Assessment, Research and Evaluation*, vol. 15, pp. 1–16, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:260517918.

[18] W. Saris, M. Revilla, J. Krosnick, and E. Shaeffer, “Comparing questions with agree/disagree response options to questions with item-specific response options”, *Survey Research Methods*, vol. 4, pp. 61–79, May 2010. DOI: 10.18148/srm/2010.v4i1.2682.

[19] M. Revilla, W. Saris, and J. Krosnick, “Choosing the number of categories in agree-disagree scales”, *Sociological Methods amp Research*, vol. 43, pp. 73–97, Feb. 2014. DOI: 10.1177/0049124113509605.

[20] T. Lumley, P. Diehr, S. Emerson, and L. Chen, “The importance of the normality assumption in large public health data sets”, *Annual review of public health*, vol. 23, pp. 151–69, Feb. 2002. DOI: 10.1146/annurev.publhealth.23.100901.140546.

[21] G. Norman, “Likert scales, levels of measurement adn the “laws” of statistics”, *Advances in health sciences education : theory and practice*, vol. 15, pp. 625–32, Feb. 2010. DOI: 10.1007/s10459-010-9222-y.

[22] L. Wasserman and S. Zhou, “A statistical framework for differential privacy”, *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 375–389,

2010, ISSN: 01621459. [Online]. Available: http://www.jstor.org/stable/29747034 (visited on 05/20/2023).

[23] V. A. E. Farias, F. T. Brito, C. Flynn, J. C. Machado, S. Majumdar, and D. Srivastava, *Local dampening: Differential privacy for non-numeric queries via local sensitivity*, 2022. arXiv: 2012.04117 [cs.CR].

[24] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy", in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '18, Toronto, Canada: Association for Computing Machinery, 2018, pp. 475–489, ISBN: 9781450356930. DOI: 10.1145/3243734.3243818. [Online]. Available: https://doi.org/10.1145/3243734.3243818.

[25] B. A. Kopp, J. Allen, A. Becker, *et al.*, "Microsoft smartnoise differential privacy machine learning case studies", 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232375099.

[26] K. Zhu, P. V. Hentenryck, and F. Fioretto, *Bias and variance of post-processing in differential privacy*, 2020. arXiv: 2010.04327 [cs.LG].

[27] M. Arapinis, D. Figueira, and M. Gaboardi, "Sensitivity of Counting Queries", in *International Colloquium on Automata, Languages, and Programming (ICALP)*, Rome, Italy, Jul. 2016. [Online]. Available: https://hal.science/hal-01713317.

[28] T. Zhu, G. Li, W. Zhou, and P. S. Yu, "Differential privacy and applications", in *Advances in Information Security*, 2017.

[29] C. Ilvento, *Implementing the exponential mechanism with base-2 differential privacy*, 2020. arXiv: 1912.04222 [cs.CR].

[30] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds", vol. 9985, Nov. 2016, pp. 635–658, ISBN: 978-3-662-53640-7. DOI: `10.1007/978-3-662-53641-4_24`.

[31] A. M. medina, M. Joseph, J. Gillenwater, and M. Ribero, "A joint exponential mechanism for differentially private top-k set", in *NeurIPS 2021 Workshop Privacy in Machine Learning*, 2021. [Online]. Available: `https://openreview.net/forum?id=BjBeRB3NqG`.

[32] M. Hardt, K. Ligett, and F. Mcsherry, "A simple and practical algorithm for differentially private data release", in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25, Curran Associates, Inc., 2012. [Online]. Available: `https://proceedings.neurips.cc/paper_files/paper/2012/file/208e43f0e45c4c78cafadb83d2888cb6-Paper.pdf`.

[33] D. Su, H. Huynh, Z. Chen, and W. Lu, "Re-identification attack to privacy-preserving data analysis with noisy sample-mean", Aug. 2020, pp. 1045–1053. DOI: `10.1145/3394486.3403148`.

[34] C. Dwork, N. Kohli, and D. Mulligan, "Differential privacy in practice: Expose your epsilons!", *Journal of Privacy and Confidentiality*, vol. 9, Oct. 2019. DOI: `10.29012/jpc.689`.

[35] A. Al-Achi, "The student's t-test: A brief description", 2019. [Online]. Available: `https://api.semanticscholar.org/CorpusID:155548180`.

[36] N. Nachar, "The mann-whitney u: A test for assessing whether two independent samples come from the same distribution", *Tutorials in Quantitative Methods for Psychology*, vol. 4, Mar. 2008. DOI: `10.20982/tqmp.04.1.p013`.

[37] S.Ulianova, "Cardiovascular disease dataset — kaggle", 2019.