Exploration of Differential Privacy Methods and their Applications

Prepared by:

Julian Templeton (Statistics Canada, julian.templeton@statcan.gc.ca)
Loic Muhirwa (Statistics Canada, loic.muhirwa@statcan.gc.ca)
Saptarshi Dutta Gupta (Statistics Canada, Saptarshi.DuttaGupta@statcan.gc.ca)
Benjamin Santos (Statistics Canada, benjamin.santos@statcan.gc.ca)
Rafik Chemli (Statistics Canada, rafik.chemli@statcan.gc.ca)

Prepared for:

Michael Williamson (Public Health Agency of Canada, Michael.Williamson@phac-aspc.gc.ca)

GitHub Repository: https://github.com/PHACDataHub/statscan-phac-diffpriv-collab

Example Webpage: dp-react-app-36sasy4jfa-pd.a.run.app

Executive Summary

Health data has always been at the forefront of discussions regarding the importance of maintaining user privacy. The collection of this data from trusted and reputable parties is imperative such that insights can be derived for diseases and health issues in general, such as linking smoking to lung cancer. Even though collecting this data and disseminating findings from the data benefits society at large, individuals rightfully have concerns as to how their data is used and shared. When this sensitive information is held by or sent to an organization, they must have trust in the measures utilized to keep it secure and keep anything done with the data privatized. This way, even if an individual has cancer from smoking and is in the dataset which exhibits that smoking causes lung cancer, the individual will remain anonymous despite the finding that smoking is linked to this cancer. The research conducted in this project has explored a privacy enhancing technology named Differential Privacy to understand how it can be applied at data ingestion and query dissemination points to enhance the privacy of the outputs and input data.

There are three types of Differential Privacy techniques to test and evaluate. Each of these techniques generate noise, which are just numbers, to then be applied either to query results or to the input data from a respondent. This added noise helps privatize the process by balancing how much a response/query is changed and how much privacy is added. The amount of privacy added is controlled and has strong mathematical guarantees that are more transparent and auditable than traditional techniques. Thus, we explore the impacts of adding more privacy at the cost of worse results and reducing the privacy added for more accurate results.

Global Differential Privacy is the first form of Differential Privacy explored, which adds noise to the outputs of a query made to a database, such as the mean or the sum. This approach can allow analytics to be privatized before being sent to the individual or organization requesting the query. In this setting, the data remains unchanged, where only the results are altered after being calculated from the original dataset. Local Differential Privacy is the second form explored, where noise is added to data when it is collected. This way, a survey response can be altered on the respondent's device before reaching the organization (such as the Public Health Agency of Canada or Statistics Canada). Since the noise is applied on individual datapoints, the resulting dataset after collecting many noisy points will provide noisy query outputs when a query request is made. Therefore, we have found that Local Differential Privacy gives worse detailed results than Global Differential Privacy when queried but can still give summary statistics such as the province which has the most daily smokers. Finally, Shuffle Differential Privacy is tested which is similar to Local Differential Privacy but requires another party to shuffle sets of data before the data reaches its destination.

Each of these approaches can be applied in practice and are actively used by large companies. However, to use on sensitive health data it will require building trust such that a respondent can trust that their responses will remain private even if analytics are shared. An accompanying webpage has been created to help guide a person through how Local Differential Privacy can work in practice (dp-react-app-36sasy4jfa-pd.a.run.app). Overall, both Global and Local Differential Privacy can help the Public Health Agency of Canada in sharing analytics on sensitive dataset with privacy in mind and help privatize the data collection process for sensitive topics in surveys or in sets of collected datasets.

1. Project Scope and Goals

In the era of data-driven decision-making the need to protect individual privacy while extracting meaningful insights from data has become paramount. The proliferation of online surveys, especially in the realm of public health and the sensitive nature of health information, presents a challenging landscape where the privacy of respondents must be preserved without compromising the utility of collected data. Many Privacy Enhancing Technologies (PETs) have been proposed to protect the privacy of individual survey participants, one such technique is Differential Privacy (DP). DP is a rigorous mathematical framework and concept for ensuring the privacy of individuals while allowing useful information to be extracted from datasets. It provides a quantitative measure of how much privacy is preserved when analyzing or sharing sensitive data. The fundamental idea behind this framework is to add carefully calibrated noise to the data or query results in such a way that the statistical properties of the dataset are preserved while protecting individual privacy.

This noise makes it difficult for an adversary to determine whether a particular individual's data is included in the dataset, thereby preventing unauthorized inference of sensitive information. It enables organizations to share and analyze data while minimizing the risk of privacy breaches, promoting responsible data use, and fostering trust among data subjects. The research documented in this report explores the application of three distinct methodologies of DP – Global Differential Privacy (GDP), Local Differential Privacy (LDP), and Shuffle Differential Privacy (SDP) – as safeguards to uphold the confidentiality of respondents participating in online surveys administered by the Public Health Agency of Canada (PHAC). In this research we aim to evaluate the efficacy of DP mechanisms in preserving the confidentiality of respondents in online public health surveys. Through a systematic examination of these methodologies, we seek to delineate their strengths, limitations, and applicability in real-world scenarios, ultimately contributing to the evolving discourse on privacy-preserving data analysis.

2. Background Information

At the heart of the discussion lies the Fundamental Law of Information Recovery, which underscores the delicate balance between data utility and individual privacy. It posits that overly accurate estimates of numerous statistics can lead to a complete erosion of privacy. DP emerges as a pioneering concept to navigate this paradoxical landscape, aiming to glean valuable insights about a population while guaranteeing that no individual's data exerts undue influence on the outcome of analyses.

At its core, DP defines a notion of privacy that guarantees that the outcome of an analysis or query remains nearly unchanged, regardless of whether any single individual's data is included or excluded from the dataset. In other words, it ensures that the presence or absence of any individual's data has a negligible impact on the overall results of data analysis. This process is presented in Figure 1 below.

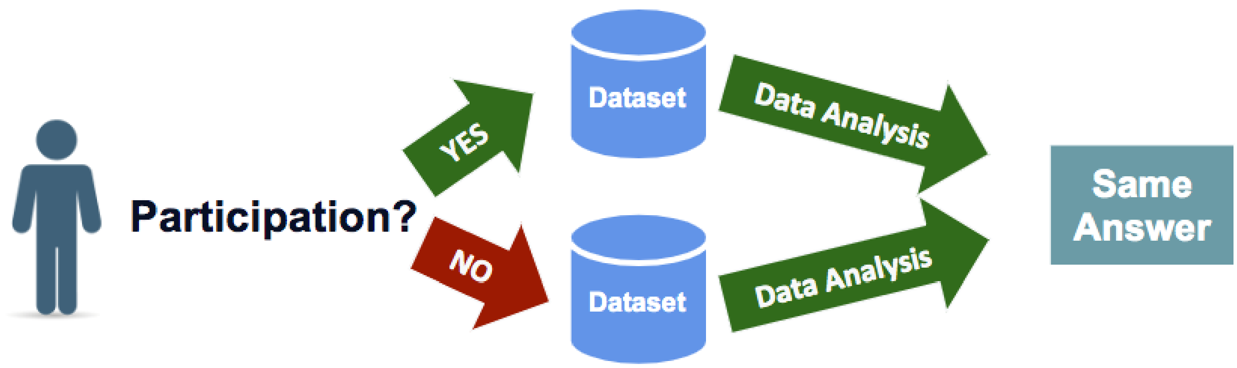


Figure 1 – A high-level overview summarizing the goal of DP.

Embedded within this framework is the understanding that if individuals can trust that their participation in data analysis remains inconsequential to the results, they are more likely to share their information willingly. This notion, encapsulated in Cynthia Dwork and her collaborators' seminal work, awarded them the prestigious 2017 Gödel Prize, underscoring the transformative impact of their contributions. The technical explanations surrounding the DP methods utilized in this work are within the Appendix, where the following discussions remain at a higher-level.

2.1 The Promise of Differential Privacy

DP as articulated by Dwork (2014), encapsulates a profound pledge from data holders to data subjects: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources are available." It embodies a commitment to safeguarding individual privacy amidst the pursuit of data-driven insights, ensuring that participation in surveys or analyses carries reduced risk of harm or exposure.

2.2 Global Differential Privacy (GDP)

In GDP, the privacy guarantee is provided for the dataset as a whole, rather than for individual records within the dataset. This means that any analysis or query performed on the dataset should not reveal sensitive information about any individual participant, even when combined with additional knowledge or external datasets.

The mechanism of GDP typically involves adding noise to a query result that is computed over an entire dataset. This noise ensures that the results of the analysis are not overly influenced by any individual's data and the resulting noisy query cannot be reverse engineered to expose any individual data point, therefore safeguarding individual privacy while still allowing for useful insights to be derived from the data as a whole.

Note that since only the query is differentially private and not the individual data points, this implies that the data holder needs to be trusted by the survey participants.

2.3 Local Differential Privacy (LDP)

In LDP, noise is added to individual data points before they are shared or analyzed. Unlike GDP, which adds noise to a global query, LDP injects noise at the source of the data, i.e., on the survey participant's device or at the data collection point, before any data is transmitted or aggregated. This perturbation process typically involves adding noise sampled from a known distribution to each data point, making it statistically indistinguishable from similar data points but preserving the overall statistical properties of the dataset.

Since the noise is injected before the data reaches the organization that is collecting it, the survey participants do not have to trust the organization.

2.4 Shuffle Differential Privacy (SDP)

In Shuffle DP, privacy guarantees are achieved by shuffling the data before analysis thereby breaking any direct link between an individual's data and their contribution to the dataset. In this model, users generate messages using a local randomizer on their data, similar to the local model. However, in the shuffle model, users trust a central entity to apply a uniformly random permutation on all the messages generated by users.

The process typically involves shuffling the order of data points or perturbing the data in a way that masks the identity of individuals while still allowing for meaningful analysis at the aggregate level. This ensures that any analysis performed on the shuffled dataset does not reveal sensitive information about any specific individual, even when combined with external knowledge or additional datasets.

3. Privacy and Trust

Privacy and trust are intimately related in DP and crucial for protecting sensitive information while still allowing valuable data analysis. Before delving into how trust and privacy intersect in this context, it's essential to understand that privacy itself is a multifaceted concept, with various definitions depending on the context and stakeholders involved. When we talk about defining privacy for a specific application, the aim is to capture the core essence of what individuals consider private and what they perceive as breaches of their privacy. This definition often hinges on the concerns and preferences of data providers, reflecting their priorities regarding privacy breaches.

DP offers a unique and specific perspective on privacy, departing from deterministic definitions commonly associated with cryptographic methods, in which privacy is breached by an adversary solving a computationally intractable deterministic problem. Crucially, the definition of privacy in DP is probabilistic rather than absolute. Instead of guaranteeing that no information about an individual is leaked, it provides a level of uncertainty or "plausible deniability". This probabilistic approach acknowledges that complete privacy may be unattainable but aims to limit the risk of privacy breaches to an acceptable level.

Now, when it comes to trust in DP, the distinction between global and local DP becomes significant. In the global approach, data is shared with a data curator or holder who then applies privacy-preserving mechanisms to protect the dataset before releasing aggregated or analyzed information. In this scenario, survey participants must trust the organization collecting the data since it has access to the true answers provided by individuals. On the other hand, LDP offers a more decentralized approach. Here, noise is added to individual responses before they even reach the organization collecting the data. This means that survey participants do not need to trust the data curator or holder since their responses are already protected before being aggregated.

However, while one can prove that LDP is being applied and this information can be verified by any user, the average technical knowledge of a respondent will likely be insufficient for them to do so. Thus, although the user can be shown how their values have been adjusted, they still need to trust that what is being seen will actually be sent. To do so, trust must be built between the user and data holder. Part of building this trust is to provide a transparent view into how this is implemented and ensure that anyone can audit the implementation. While the average user may not be able to audit this properly, the fact that it is auditable can help build confidence that the solution is correctly implemented, and that the data being sent is the augmented data. This is one potential solution to an open problem. Furthermore, the user needs to trust that when the data is sent, it will be used appropriately for the intended purpose and not will not be subject to attacks aimed at reverse engineering the original outputs. Thus, not only is having security mechanisms appropriately set important, so is protecting the privacy of the data if opportunities arise that may help violate a user's privacy.

The relationship between privacy and trust in DP underscores the importance of balancing data utility with individual privacy concerns. By adopting a probabilistic definition of privacy and implementing PETs like DP, organizations can foster trust while still deriving valuable insights from sensitive data.

4. Simulated Experiments

To understand the effects of applying the different forms of DP, we created a simulation environment which, given a set of input parameters, will output a variety of results to analyze how well each DP method performed. Prior to discussing the simulation environment, the data used within the experiments will be outlined. Next, the system design for the experiments will be discussed at a high-level to understand how each test works and how comparisons are made. This will also detail the outputs from the system which are used to directly compare each DP method. A set of experiments will be presented and compared. This will lead into a discussion of the overall trends found when applying each technique, any issues faced, and how these approaches can be used in practice.

4.1 Data Used

For this work we selected a dataset that is relevant to all involved by ensuring that it is both health-related and topical for a National Statistical Office. To this end, we utilize Statistics Canada's Canadian Community Health Survey (CCHS) Public Use Microdata File (PUMF) dataset. This is an open dataset with a variety of features pertaining to the health of Canadians. Any results within this work are only done to test the DP approaches and are not meant to derive any additional insights from the data itself. It consists of 113,290 responses with 16 columns following a cleaning process. The cleaning process conducted has remapped custom codes used within the columns to be a sequential set of categorical values for discrete features and has maintained continuous values as provided. For example, each response comes from a Canadian province where the GEO_PRV column will contain the value 24 for Quebec and 35 for Ontario. These get remapped to a number between 1 and 13 based on the alphabetical ordering of the 13 provinces. This is conducted for all discrete values, where the remapped columns are outlined within the data preprocessing notebook called data.ipynb. Furthermore, certain responses depend on previous responses. For example, the SMK_005 feature asks whether a person smokes "daily", "occasionally", "not at all", "don't know", or "refusal". Based on the answer, a respondent may not need to answer the following smoking questions, such as SMK_015, which asks whether the person has smoked every day within the last 30 days. An overview of how the preprocessed data appears is presented in Figure 2 below.

	ID	GEO_PRV	GEODGHR4	DHH_SEX	DHHGMS	DHHGAGE	GEN_005	GEN_015	GEN_020	GEN_025	SMK_005	SMK_015	SMK_020	SMK_030	HWTDGHTM	HWTDGWTK
0	0	2	37	0	4	0	6	7	0	8	2	3	0	5	1.651	74.25
1	1	0	77	1	4	14	4	3	3	8	2	3	0	5	1.727	108.00
2	2	3	25	0	1	9	6	3	5	2	2	3	2	4	1.600	60.75
3	3	1	68	0	4	10	6	7	3	5	2	3	2	4	1.676	81.00
4	4	8	46	0	1	9	1	1	0	0	2	3	2	4	1.753	63.00
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
113285	113285	8	62	0	3	2	2	5	2	8	0	4	0	4	1.600	44.10
113286	113286	8	46	1	1	8	3	3	5	6	2	3	2	4	1.753	123.75
113287	113287	6	96	1	1	11	6	7	0	3	2	3	2	4	1.727	101.25
113288	113288	0	2	1	1	9	6	3	0	6	2	3	2	4	1.829	73.35
113289	113289	8	85	0	4	7	3	7	4	0	2	1	0	5	1.753	69.75

Figure 2 – An overview of the preprocessed PUMF data.

There are a few other important notes regarding the data which have an impact on how the DP methods are applied. The first is that not all columns will be modified following the application of a DP method. For instance, a respondent may have their responses regarding whether or not they smoke changed, but the demographic information surrounding the response should remain static. This allows the data to still be properly stratified to properly derive statistics based on that information. The stratification refers to performing selected queries on groupings of the data. For example, we can analyze the counts of daily smokers, or any other variable, for each province or by each province and each age group. Thus, any column which may be used for stratification should remain unaltered. By having the chance of the response values being changed, the user privacy can be protected without interfering with the groupings which an analyst aims to explore. Within the simulation environment we allow a user to control which features remain unchanged throughout the tests, but the list we utilize is presented in Table 1 below.

Column English Name	Column Key
Province	GEO_PRV
Sex	DHH_SEX
Geographical zone (ex: Eastern Regional)	GEODGHR4
Marital status	DHHGMS
Age	DHHGAGE
Response identifier	ID

Table 1 – Static features utilized within the simulation program.

The second important note is that each response has a corresponding weight associated to it. That is, each response's value is multiplied by an accompanying weight value derived by the statisticians responsible for the survey. For example, if Ontario is expected to provide the most responses, it may be weighed lower than provinces expected to count for few responses. This means that a single response may value several responses. This is important since the DP process itself must utilize these weights when considering the amount of noise to be added to a single response. If one response for a given province is weighted at 5.0, this means that adjusting this response with noise is equivalent to adding noise to 5 responses. Thus, the noise being added must be scaled based on the corresponding weight values provided for the data. The act of selecting weights is a massive topic that will not be discussed but note that this PUMF dataset contains weight values for the responses which we utilize within all the experiments. This includes any querying being performed such that a response with a weight of 5.0 is considered as 5 times the appropriate amount within the calculation being performed (such as the mean or the sum). Figure 3 illustrates an example for select provinces of how the actual amount of responses is translated following the weighing of each response within the province. For example, Quebec only received 24,125 responses, but is weighed as if though it received 7,173,161 responses.

Answer Categories	Code	Frequency	Weighted Frequency	%
NEWFOUNDLAND AND LABRADOR	10	3,291	459,931	1.5
PRINCE EDWARD ISLAND	11	1,928	129,507	0.4
NOVA SCOTIA	12	4,811	821,776	2.6
NEW BRUNSWICK	13	3,706	648,259	2.1
QUEBEC	24	24,125	7,173,161	22.9
ONTARIO	35	33,511	12,235,212	39.1
MANITOBA	46	5,481	1,060,834	3.4
SASKATCHEWAN	47	4,835	917,474	2.9

Figure 3– Example of weighing responses for select provinces.

4.2 Simulation Environment

To provide a robust testing environment, we designed a simulation program which can take a variety of inputs and test them on LDP, SDP, and GDP. The program accepts a configuration YAML file as input which then specifies the run parameters. For instance, the weights for the dataset can be multiplied by a scaler and the system can test multiple epsilon values within a single run. Thus, a single run can output a substantial amount of analytics to test how well each DP method works. The system does utilize a user-specified random seed at initialization, but unfortunately the Python library used for the DP noise generation, OpenDP, cannot be made reproducible. Even within the library's test cases, only the expected output range from the noise functions is asserted, rather than a concrete output. To circumvent this, the system needs to be run k times with the outputs from the system averaged from these k runs. This helps provide higher confidence in the results and reduce the variance observed between runs. Figure 4 overviews how the tests can be conducted, where the data can be stratified for certain columns and tests within the pipeline can apply DP to each of the groupings.

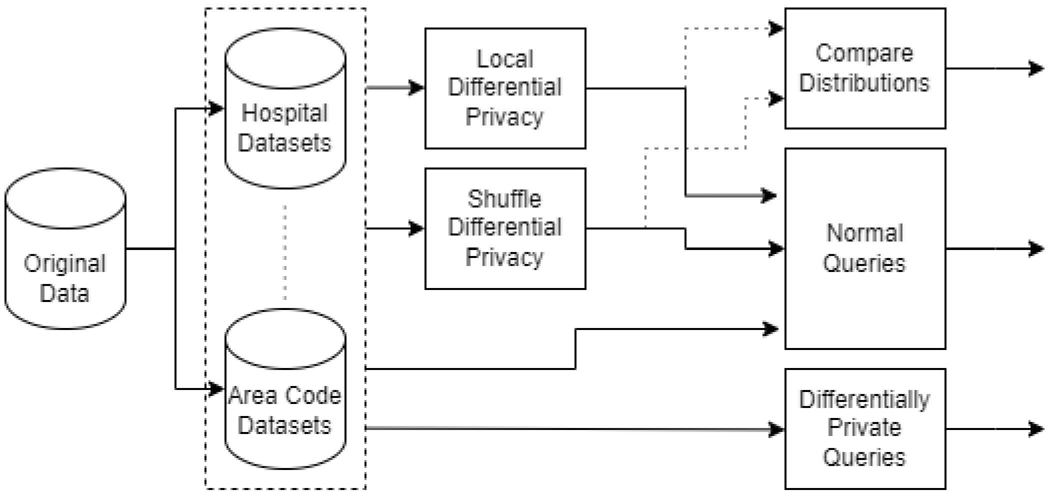


Figure 4 – Example of the general simulation pipeline with fake groupings.

Once the input data is loaded and the weights are adjusted for each sample based on the user input, the sensitivity values needed for the queries being performed are computed. Since each query with a specific epsilon value will have a corresponding sensitivity value, these are calculated based on the provided epsilon. We support both the sum and mean queries within the system, and thus only compute the sensitivity for the sum query since it will utilize more noise than the mean query. This ensures that only the most expensive tested query option is considered, presenting only the results from when more privacy budget is used. First, LDP is applied to the columns within the dataset. The mechanism used for discrete values is randomized response while the mechanism for continuous values is a Laplace distribution. Gaussian distributions are another option which can be considered, but the system only supports Laplace noise as of writing.

Once LDP has been applied, the SDP process begins. Here, the implementation is custom-built based on the approach detailed within Scott (2021). The goal of this approach is to use the shuffling algorithm to mimic the shuffler's role within the SDP process. Within the applied approach, noise is picked a binomial distribution and applied to each column from a received response. The responses are then shuffled between each other, mimicking the described shuffling step. While the expected behaviour of SDP is to be better than LDP, we later observe that the implementation within the simulation pipeline is performing poorly with behaviours not actively reflecting what is expected of varying epsilon values. Thus, all results observed in future sections for this SDP implementation should consider this and expect better results with a different implementation. However, from discussions with the subject matter experts, the shuffler required for the SDP process may be a dealbreaker in terms of its practicality. Having another party receive such sensitive data may be infeasible within the health domain while maintaining user trust.

Following the outputs of the LDP and SDP processes, three different datasets will be sent through the query process. The original data, the LDP dataset, and the SDP dataset will all be grouped by the specified stratification variables. These groups are passed to the query functions to calculate both the sums and means for every column which DP can affect (such as the daily smoking column SMK_005.0). Note that we can consider the LDP and SDP datasets as synthetic versions of the original data since the data recipient only knows of the privacy preserving versions of that data rather than the actual responses.

With the sums and means computed for the datasets, the original data is also passed into the query functions with GDP being applied. GDP applies Laplace noise to the outputs of the query results where categorical columns are rounded to their nearest integer value in range. Not all columns benefit from both the mean and sum being computed, thus we consider the sum useful for discrete variables, such as the total number of daily smokers per grouping, and the mean useful for continuous values, such as the height and weight of a respondent per grouping.

Once all queries are computed, they are saved in an appropriate output format per DP noise type with the absolute errors recorded for how each DP method's query results compare to the query results from the original data without applying DP. Similarly, since LDP and SDP can be considered synthetic representations of the original data, we compare and save the distributions of these datasets with the original data through correlation plots. This allows the simulation program to highlight how well the approaches maintain the original distributions despite altering the data itself. Ideally, the same trends observed within the original data can still be found within the altered data for sufficiently high epsilon values. The KSComplements are also computed and output to evaluate how well each column's shape is maintained by the synthetic data.

All outputs from the system are compiled within unique output folders, where an additional aggregation script is available to average the results of k different tests with the same input values. Therefore, the results can be more accurate and reduce the variability observed from the lack of reproducibility within OpenDP.

4.3 Experiment Results

To understand the impacts of using LDP, SDP, and GDP, we have run simulations with a variety of input parameters. However, there are many outputs for each run, making it infeasible to report each individually. Therefore, this section will highlight one set of runs utilizing five different epsilon values. Of the five values tested, only the outputs from four will be presented in terms of query performance and three will be highlighted regarding the LDP and SDP correlations to the original data. While all query results are available for the tests, only six will be presented within this section. All runs will occur five times, with the averages over the five runs being presented for the query results and only one run's output being analyzed for the distribution comparisons between LDP, SDP, and the original data. The run parameters utilized are outlined in Table 2, where all tests in this section are stratified based on the response's province.

Parameter Name	**Value(s) Utilized**
Stratification Variable(s)	GEO_PRV ("Province")
Response Weight Scaler	1.0
Number of Runs	5
Epsilon Values	5.99 3.0 1.0 0.8 0.6
Static Columns (see Table 1)	GEO_PRV DHH_SEX GEODGHR4 DHHGMS DHHGAGE ID
Query Results Tracked in this Section	Mean of HWTDGHTM ("Height") Mean of HWTDGWTK ("Weight") Sum of SMK_005.0 ("Daily smoker", code 1 in dataset) Sum of GEN_005.2 ("Good perceived health", code 3 in dataset) Sum of GEN_025.0 ("Work is not at all stressful", code 1 in dataset) Sum of GEN_025.3 ("Work is quite a bit stressful", code 4 in dataset)

Table 2 – Simulation key parameters and outputs.

4.3.1 Evaluating Query Results

First, we will compare the performance of each DP method based on their outputs from the sums and means for epsilon values of 5.99, 1.0, 0.8, and 0.6. These values are selected to highlight the impact of larger and smaller epsilon values. Since SDP only supports epsilon values under 6.0 within the simulation program, 5.99 will highlight the impacts of little privacy being added. The remaining epsilon values start at 1.0 and decrease by 0.2 to view the iterative balancing between the privacy added and the utility lost. Each query is applied to each feature for every province and uses the original survey weights within the PUMF file. The means presented are for the weights and heights since the sums of these values are irrelevant to compute. Each discrete value will only present the sums since the means will not matter (i.e., the counts of each will be analyzed).

Each test will provide the query results themselves within bar charts with varying y-axis values due to the drastic increases observed when significant noise is added. The outputs from the query results are the average of five separate runs with the same input parameters. Following the outputs of a given query, a brief discussion point will summarize the results, where the discussion section will outline the general findings from all tests performed. The discussions will reference tables within the appendix which contain the absoluter error values for each DP method when compared to the original query results, for each epsilon tested.

Comparing Means for the Heights in Provinces

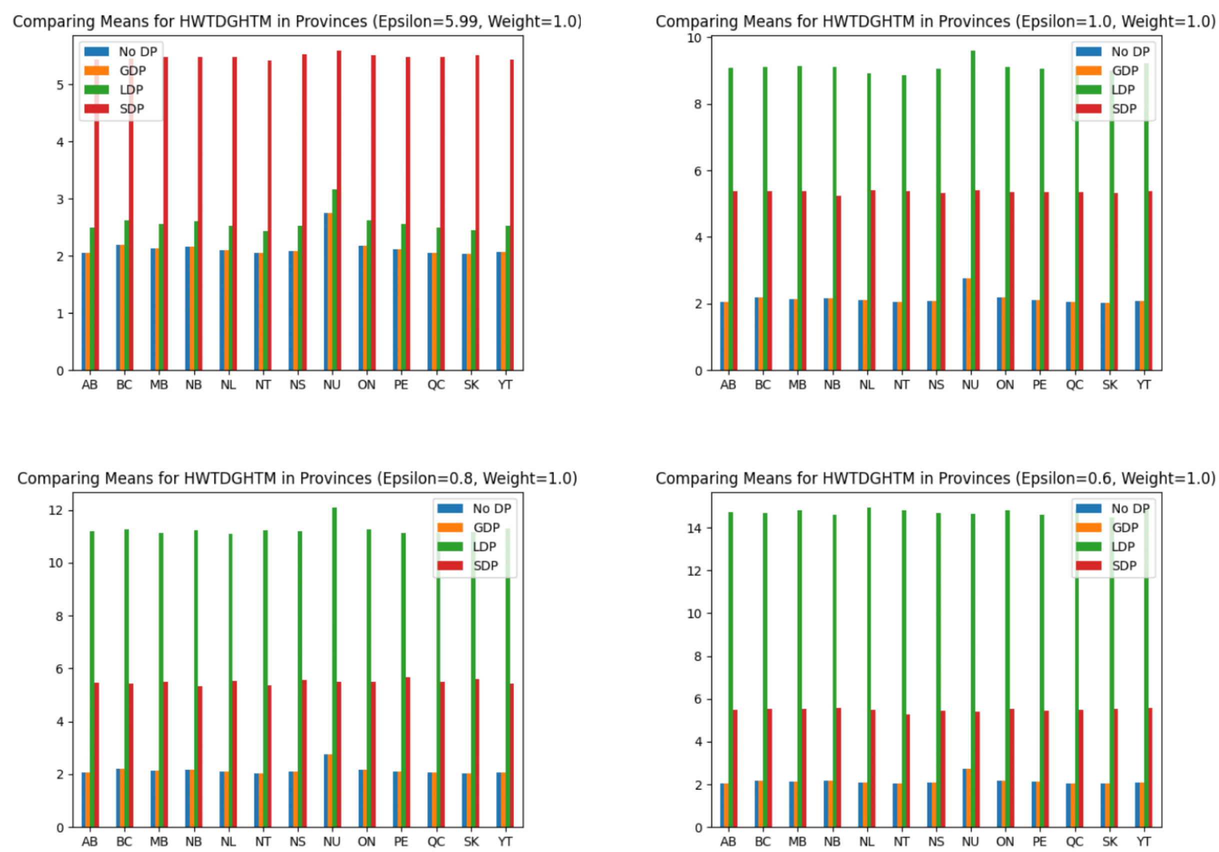


Figure 5 – Query results for the means for heights in provinces.

From Figure 5, we can make the following observations. As the epsilon values decrease for the average heights in each province, we observe a key difference between LDP and GDP. When LDP is applied with an epsilon of 5.99, the results are close to when no DP is used. However, the values drastically change as the epsilon values decrease, which is expected. GDP maintains very close results through each epsilon value, where the amount of error introduced between the average heights slightly increases as epsilon decreases (see [Appendix 7.1] for all absolute error values). In practice we may want to adjust the GDP solution for the mean to introduce more noise. However, LDP can introduce substantial noise to the individual responses which results in absurd statistics. Although we cannot say that the average height in Ontario is several times higher than the actual average, the trends within the data do highlight that outside of epsilon 0.6, Nunavut has the highest average height. Therefore, certain statistics can still be derived when significant noise is added since LDP can maintain the general trends within the data. SDP maintains a similar distribution of error regardless of the epsilon values being provided which indicates potential issues with the implemented solution in the pipeline. We have expected this to generally perform better than LDP and this may be true but is not captured properly in the implementation.

Comparing Means for the Weights in Provinces

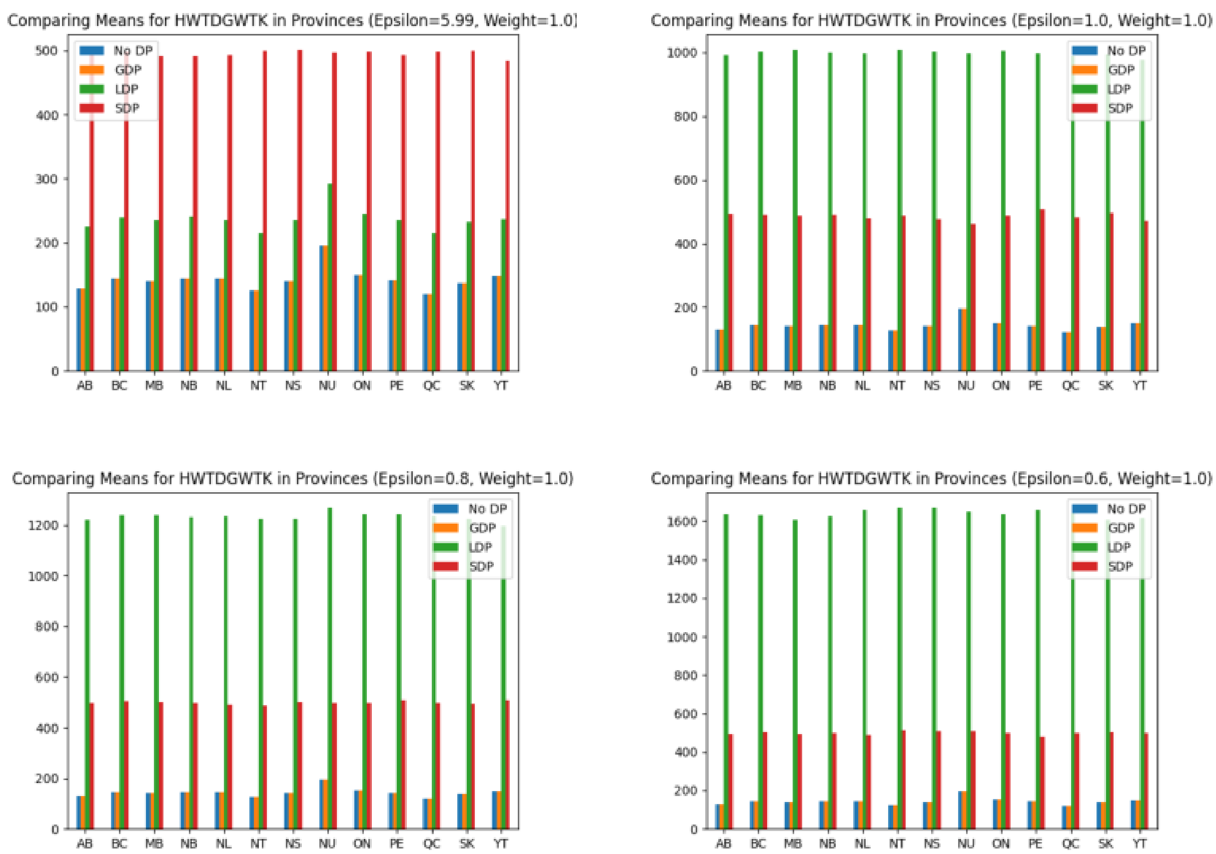


Figure 6 – Query results for the means for weights in provinces.

The trends found within the means of the weight values within Figure 6 are similar to those found with the heights (Figure 5). The GDP errors within [Appendix 7.2] remain very low but are higher than what has been observed with the heights. Similarly, SDP does not demonstrate the expected variance based on the selected epsilon, likely due to its implementation. LDP increases the average weight substantially more than the average height but observes the same findings found when analyzing the LDP height results. The general trends tend to remain similar, but not identical, for the epsilon values. The larger epsilon values retain these well, whereas lower epsilon values can result in issues such as Nunavut not weighing the most on average.

Comparing Counts of Daily Smokers in Provinces

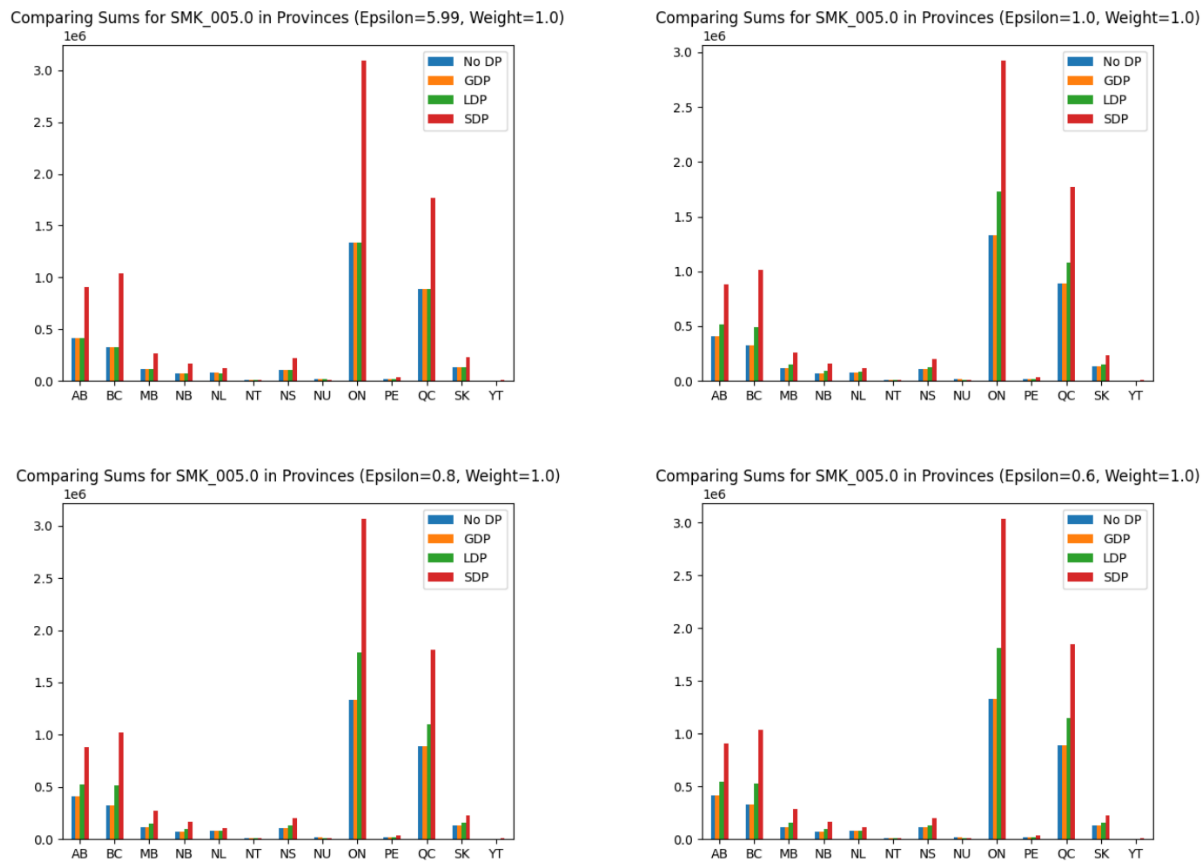


Figure 7 – Query results for the counts of daily smokers in provinces.

When comparing the counts results to the results from the means, new insights can be found on the behaviors of LDP and GDP. First, the general trends in the data are represented much better for all provinces within Figure 7. This is due to the counts of each province for daily smokers varying drastically based on the populations of the provinces and how they are weighed. Unlike the continuous valued height, the counts of these values will not alter the meaning of the value being analyzed, instead the discrete selection may change for the target column. This implies that so long as a sufficient number of responses are provided for the categories, the LDP will not have a substantial enough impact to change the facts derived from the analysis. For example, in all cases, Ontario has the most daily smokers. This occurs despite the ~477,312 responses being added to this group (from [Appendix 7.3]). In fact, observing the absolute errors for each epsilon value and the original counts within the bar plots, these values tend to scale based on the input size. This ensure that the Northwest Territories and Yukon, which have a low total amount of daily smokers, has noise added proportional to the original amount. By doing so, we better maintain the trends of the results and can more precisely understand information such as the top five provinces with daily smokers. This can then be compared to the top five average weights to see if there is a correlation between the two.

GDP performs similarly well, where the counts are off by a small margin, but a larger amount than the means. In fact, these generally indicate that any daily smoker being in or out of the dataset will still result in the same outcomes being determined while adding more privacy to the results.

<u>Comparing Counts of People with Good Perceived Health in Provinces</u>

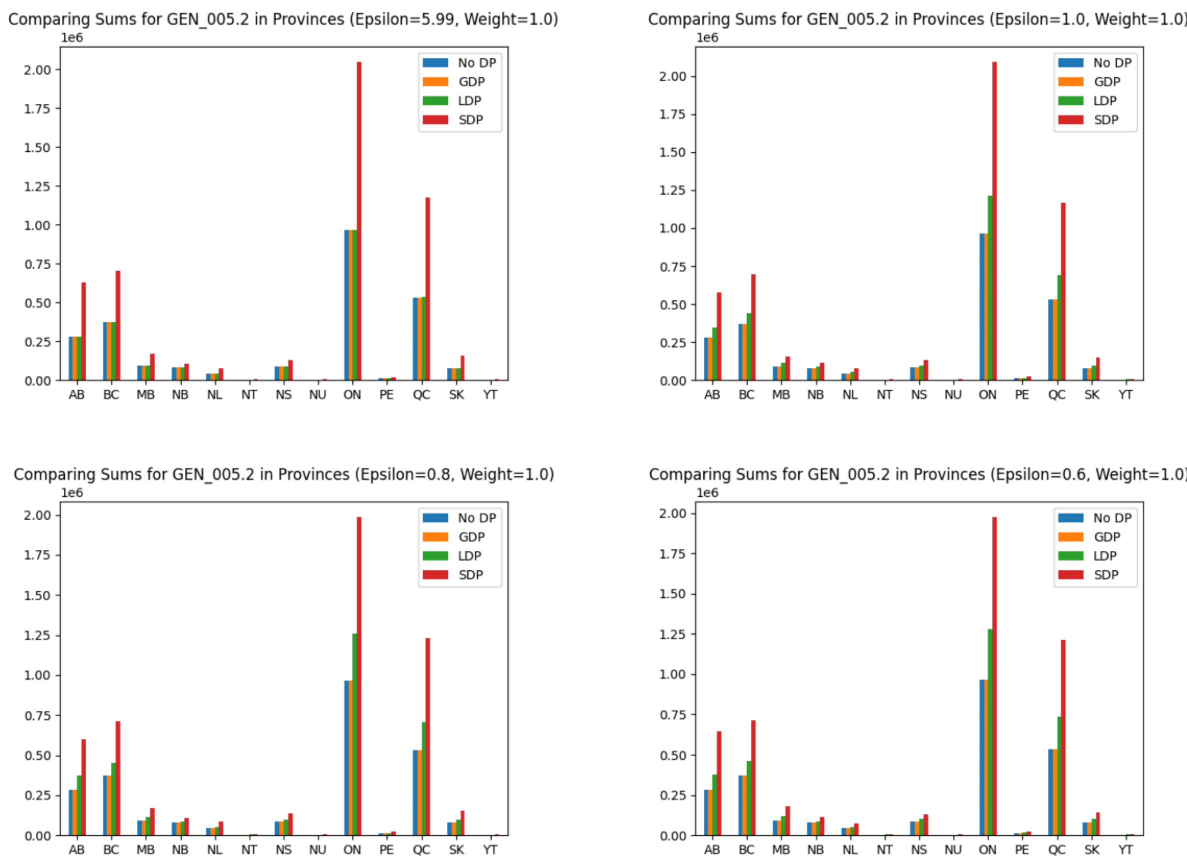


Figure 8 – Query results for the counts of people with good perceived health in provinces.

The results from Figure 8 follow the same patterns observed when analyzing daily smoker counts. While the general number of people who perceive their health as good is smaller than daily smokers, the DP mechanisms behave in the same way. This indicates consistency in how LDP and GDP will behave with varied epsilon values and further demonstrate that adding more privacy will result in a corresponding loss in utility.

Comparing Counts of People who Believe that Work is not Stressful at all in Provinces

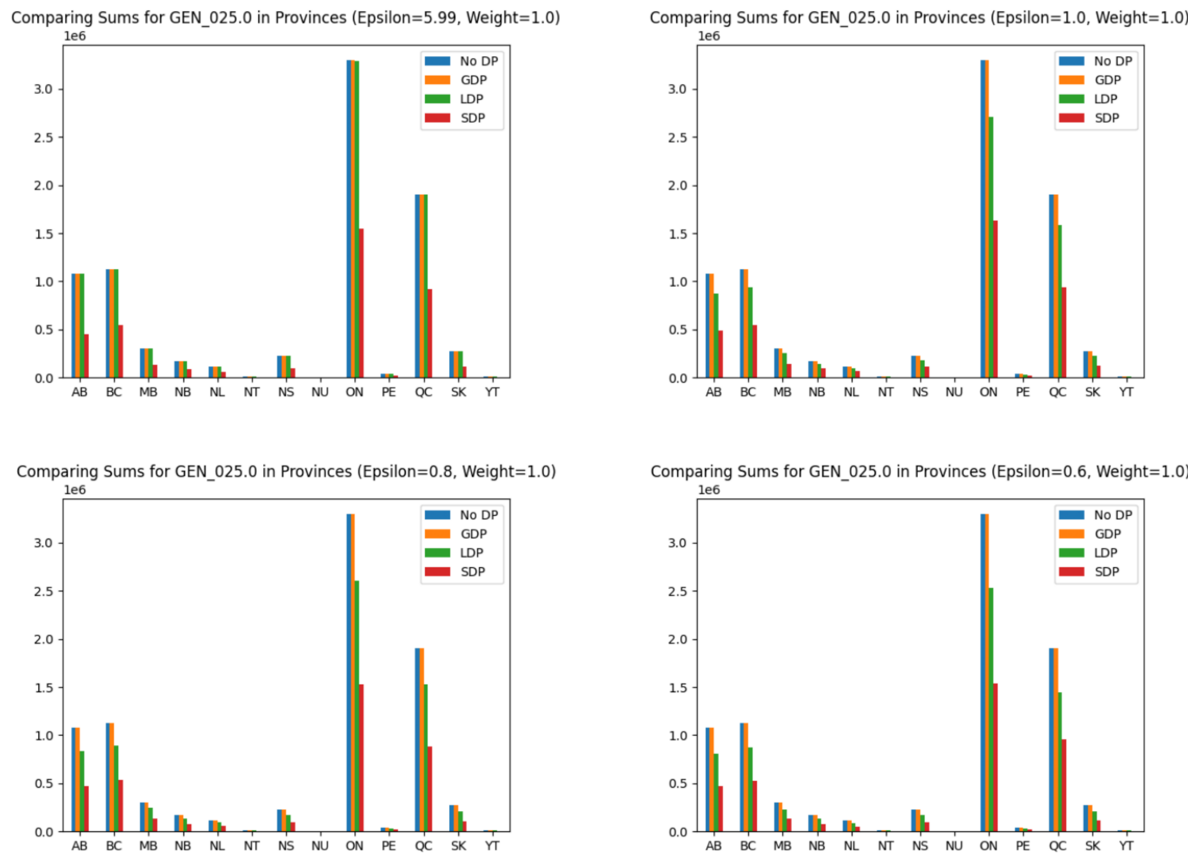


Figure 9 – Query results for the counts of people who do not find work stressful at all in provinces.

Like the other count plots, the results in Figure 9 highlight the same observable patterns. However, this time the counts are being reduced in the LDP process. Despite affecting the totals of those who find work not at all stressful, this still maintains the same trends found with the original data and the same observations as epsilon changes. The same scaling of noise values based on the input size is also observed.

Comparing Counts of People who Believe that Work Quite a bit Stressful in Provinces

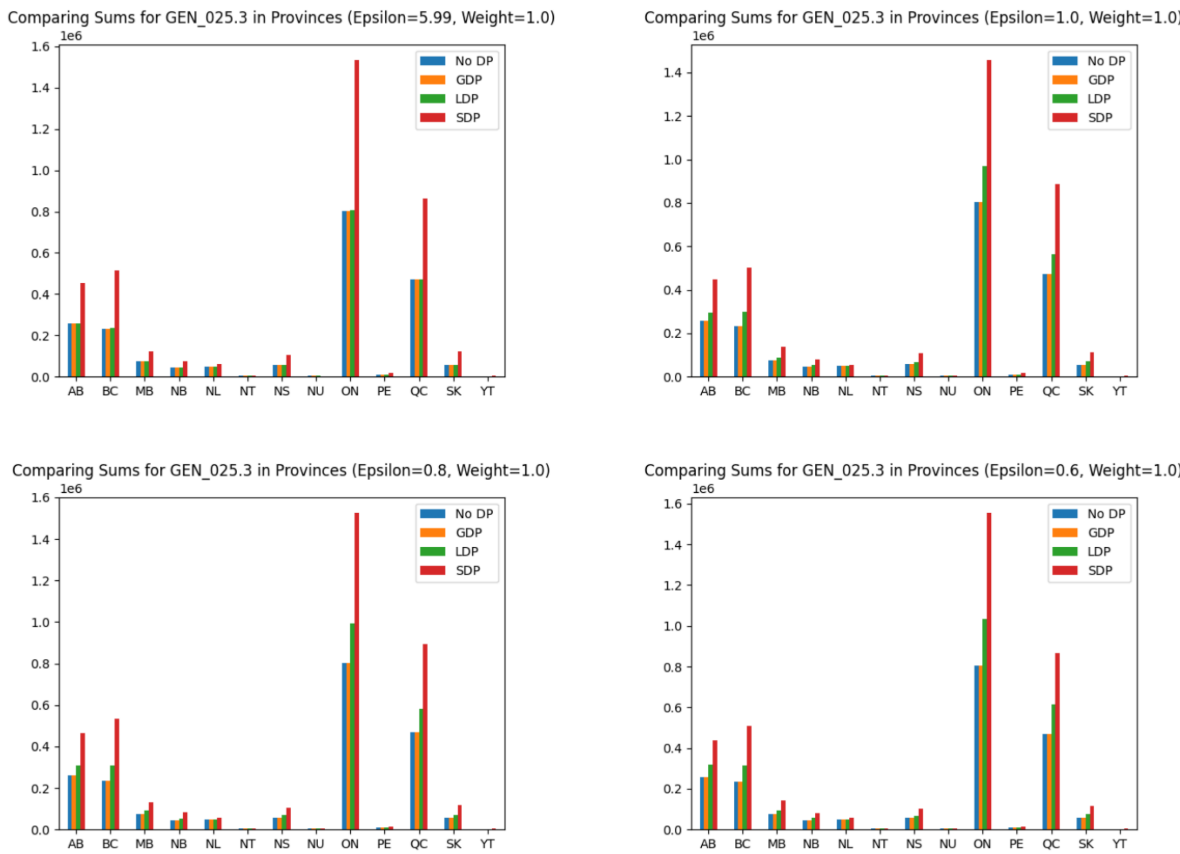


Figure 10 – Query results for the counts of people who find work quite a bit stressful in provinces.

The final queries from Figure 10 further highlight the consistency in GDP and LDP's behavior when applied to groupings with enough responses. Outside of the observation that more people generally find work not at all stressful, rather than quite a bit stressful (from Figure 9), the DP implementations yield the same properties. As with the previous sections, the absolute values of the errors can be found in the appendix ([Appendix 7.6]).

4.3.2 Evaluating LDP and SDP Dataset Quality

As previously described, the outputs from the LDP and SDP process can be considered as synthetic datasets due to the changes being made directly to the responses themselves. Thus, this subsection will directly compare the resulting LDP and SDP datasets to the original dataset through correlation plots and column shape evaluations with the KSComplements. A result is considered good if the columns affected by LDP and SDP follow the same correlations and shape as the original data. For both of these metrics, a perfect score of 1.0 implies perfect correlation and shape capture and a score of 0.0 is the lowest possible score. Since each datapoint is arbitrarily augmented with noise, the ideal scenario is that the overall data maintains a similar distribution as the original, despite the changes to the values themselves. For example, the previous query results highlighted that although the heights of the LDP process can become absurdly large, the general trends can still be accurately analyzed when a sufficient balance between privacy and utility is established (see Figure 5).

As a reference point, the correlations of features for the original data are presented in Figure 11. A set of comparison plots will evaluate how well each column matches between this correlation plot and the synthetic dataset (where the left correlation plots will be for an LDP dataset and the right will be for an SDP dataset).

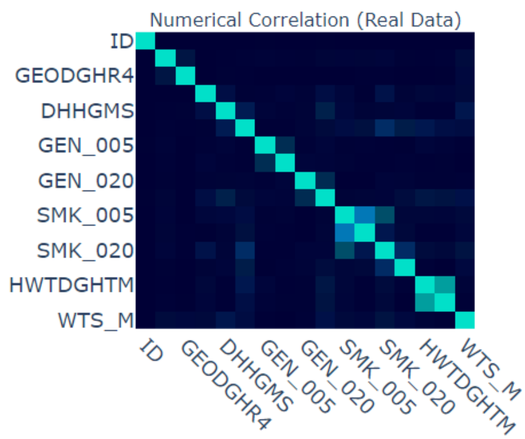


Figure 11 – The correlations of features for the original dataset.

Using the correlations and column shapes of the original data as reference, the remainder of this subsection will directly analyze the correlations and column shapes of the LDP and SDP datasets for epsilon 5.99, 1.0, and 0.6. This will help understand how the two approaches are comparing and the impacts of epsilon on the quality of the augmented data. From the query results it is expected that the SDP results will not be as meaningful due to implementation issues, however these will still be reported. The presented results are each from the first of the five runs performed for the set of experiments conducted within this section.

Note that the average scores which are included are not a good metric for evaluation on their own since they consider static features. The focus will be based on the observable differences, and their intensity, within the plots alongside relevant scores.

Comparing LDP and SDP Datasets to the Original Dataset for Epsilon 5.99

This first test directly compares the LDP and SDP datasets for an epsilon value of 5.99. Figure 12 displays the correlation comparisons to the original data and Figures 13 and 14 present the column shape plots.

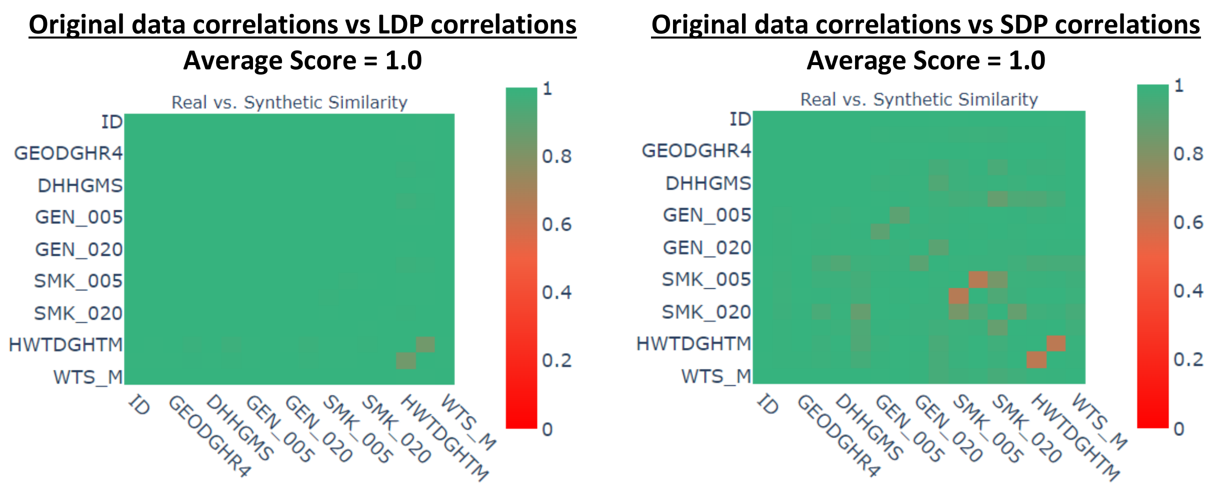


Figure 12 – LDP dataset and SDP dataset correlation comparison to the original data for epsilon 5.99.

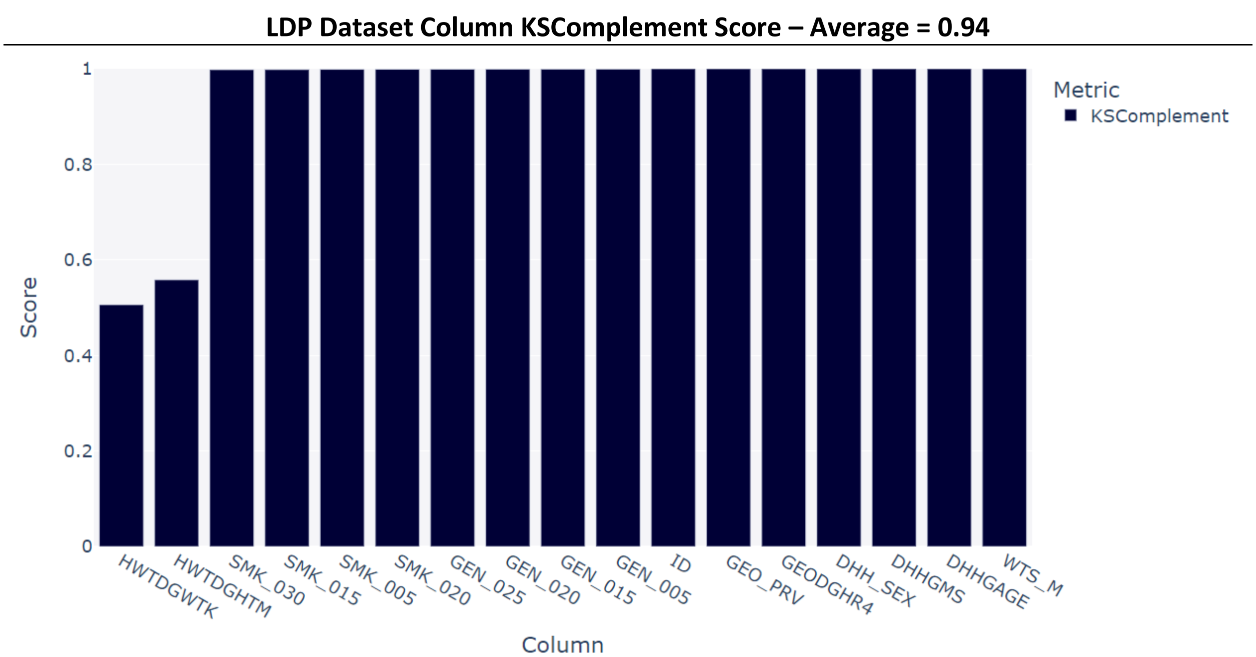


Figure 13 – LDP dataset column shape analysis and KSComplement score for epsilon 5.99.

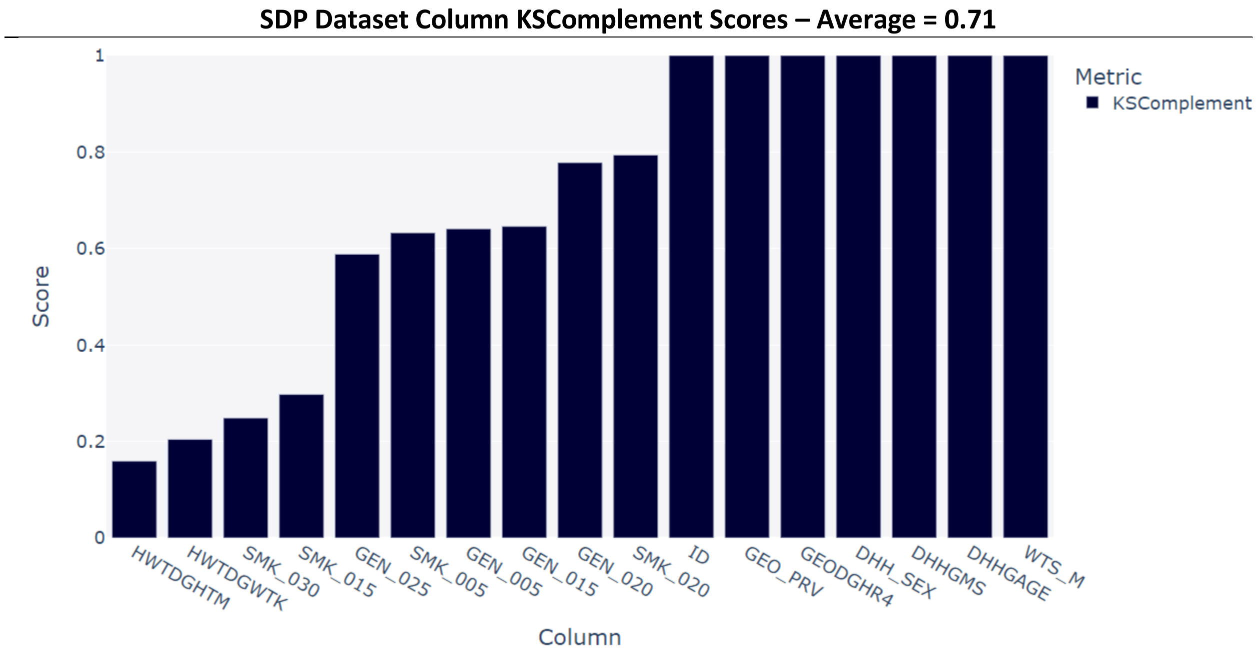


Figure 14 – SDP dataset column shape analysis and KSComplement score for epsilon 5.99.

From the correlation plots it is clear that LDP with little noise added is accurately matching the correlations for all except the height and weight values, which begin to deviate from the original data. This has also been observed in the query tests, where the means for these values brought more deviation in the general trends when compared to the counts of discrete values. However, it performs well overall, with a high 94% KSComplement score strengthening this claim. The column scores highlight that the height and weights are being changed too drastically, where bounding the values after adding noise may help in this scenario. SDP does not exhibit the same behaviours despite the higher epsilon, likely due to an issue with the chosen implementation. Overall, LDP's results highlight that adding little privacy results in a synthetic version of the dataset which adds plausible deniability to respondents while maintaining high utility in general. In practice we likely would prefer more noise to the discrete columns to ensure that the plausible deniability is higher. Also, note that any static variable, such as the ID, will always be the exact same since the DP process does not impact these features. This artificially boosts the scores assigned to the images.

Comparing LDP and SDP Datasets to the Original Dataset for Epsilon 1.0

Next, the LDP and SDP results with an epsilon of 1.0 are presented in the below correlation and shape plots (Figures 15, 16, 17).

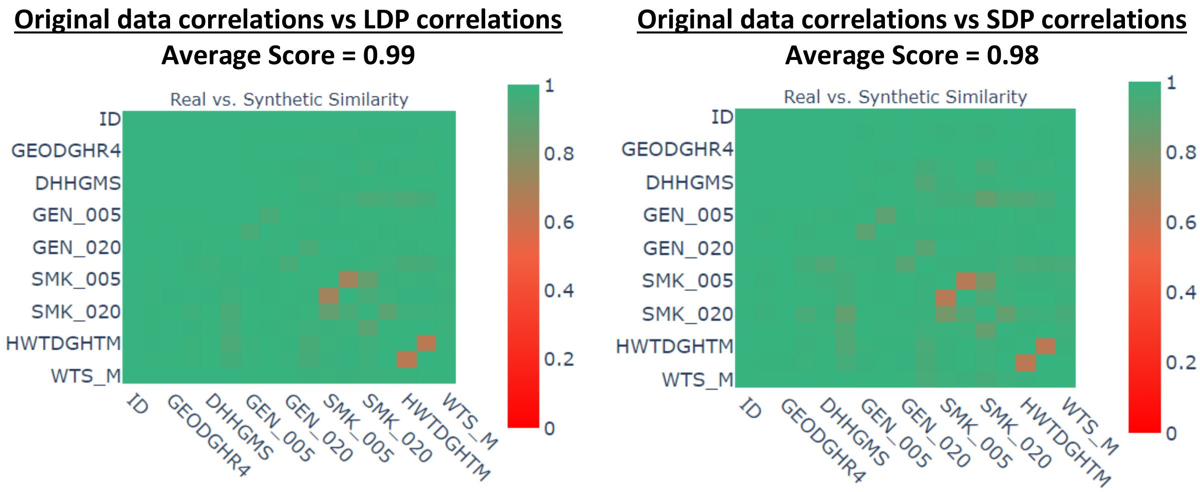


Figure 15 – LDP dataset and SDP dataset correlation comparison to the original data for epsilon 1.0.

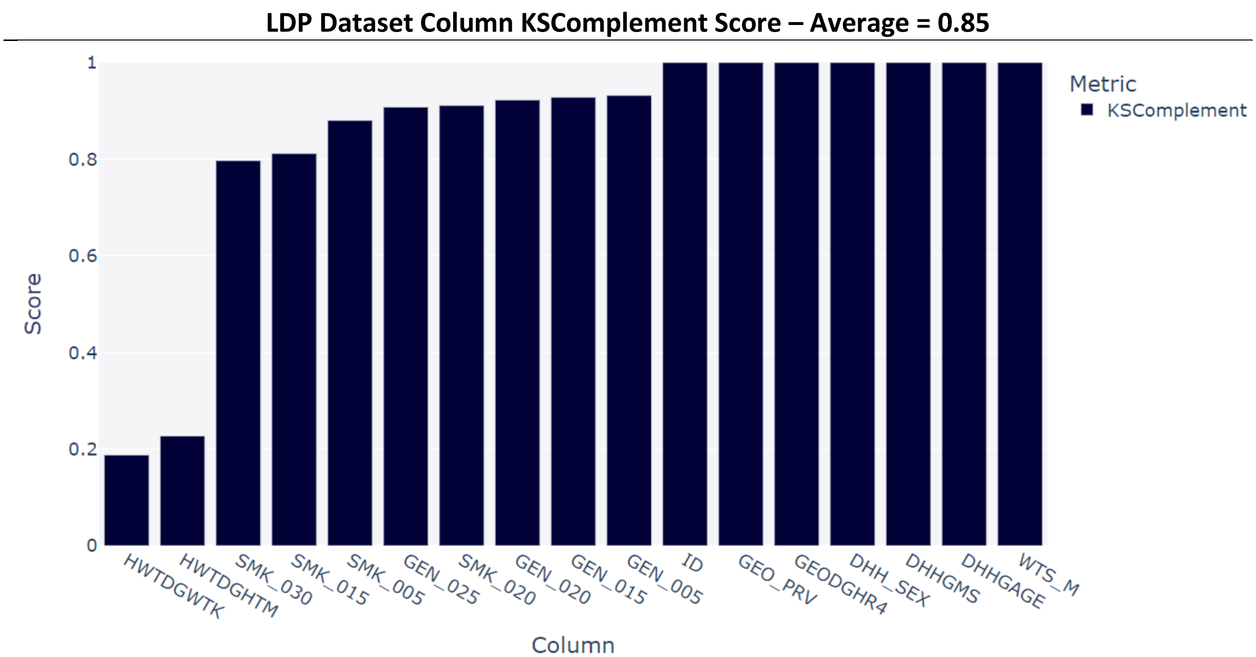


Figure 16 – LDP dataset column shape analysis and KSComplement score for epsilon 1.0.

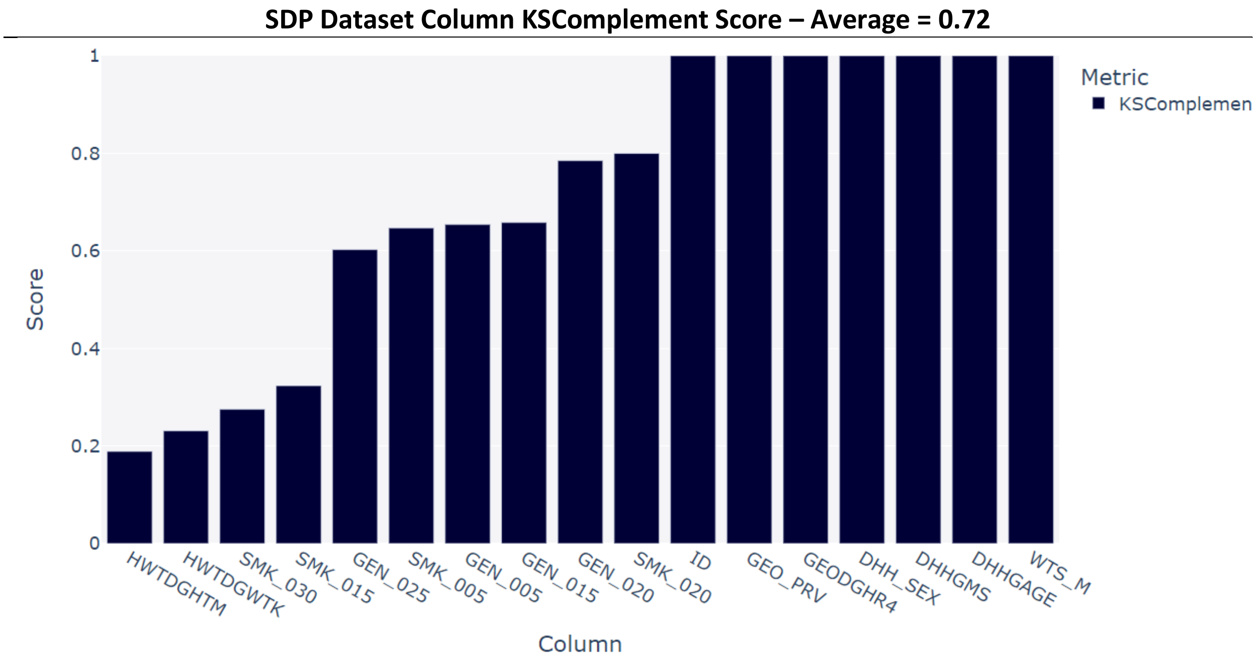


Figure 17 – SDP dataset column shape analysis and KSComplement score for epsilon 1.0.

Here, the reduced epsilon value has a clear impact for the LDP dataset. Unlike with an epsilon of 5.99, more of the discrete valued columns have a worse score and the correlation plots between the original data and the LDP data have gotten further apart. The height and weight values also lose more quality, which implies that the statistics generated from these columns will continue to get worse with the added privacy. They are becoming so low that they may be completely different in terms of their shape and correlations when compared to the original data. This tradeoff between privacy and utility is why it is crucial to properly evaluate which epsilon values to use based on the desired data and statistics.

<u>Comparing LDP and SDP Datasets to the Original Dataset for Epsilon 0.6</u>

Finally, the LDP and SDP results with an epsilon of 0.6 are presented in the below correlation and shape plots (Figures 18, 19, 20).

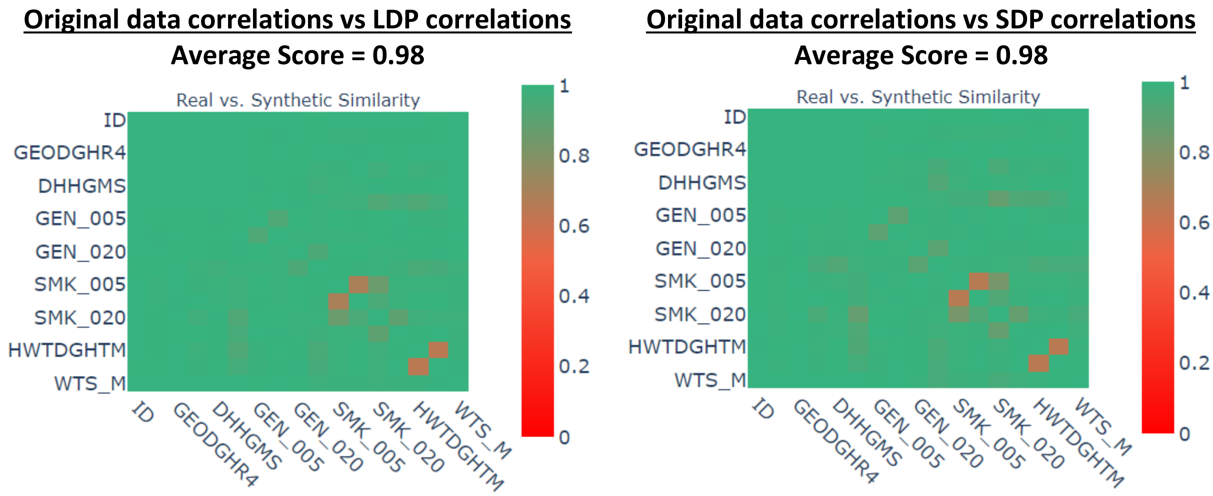


Figure 18 – LDP dataset and SDP dataset correlation comparison to the original data for epsilon 0.6.

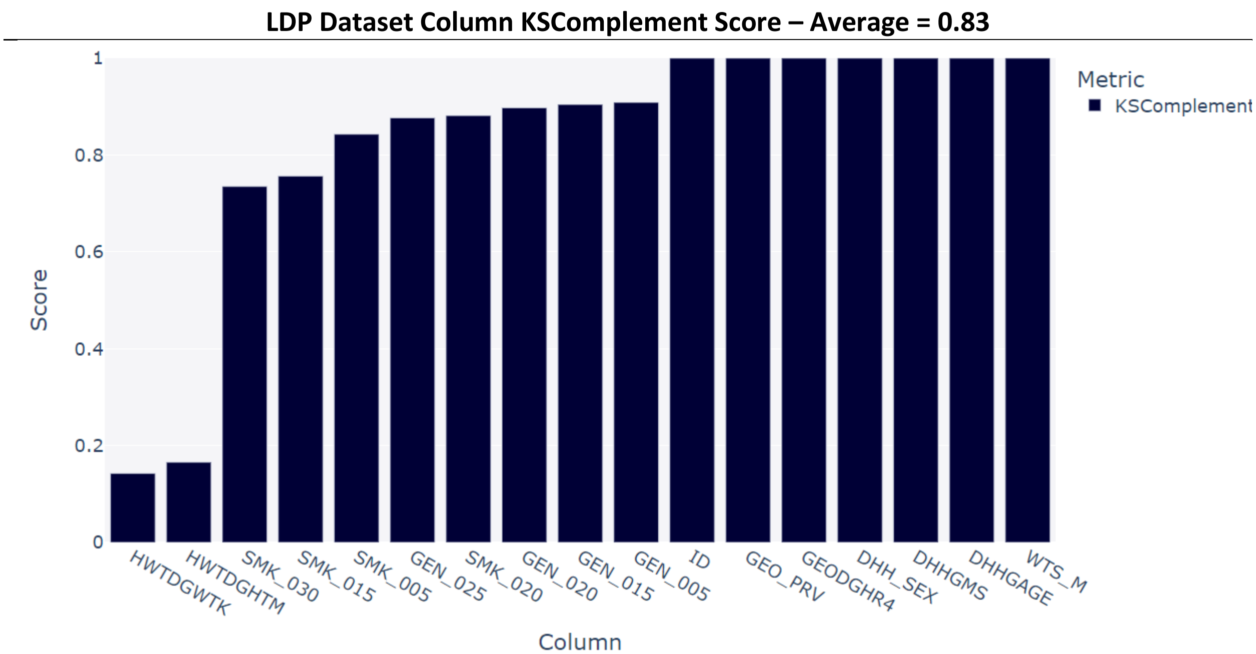


Figure 19 – LDP dataset column shape analysis and KSComplement score for epsilon 0.6.

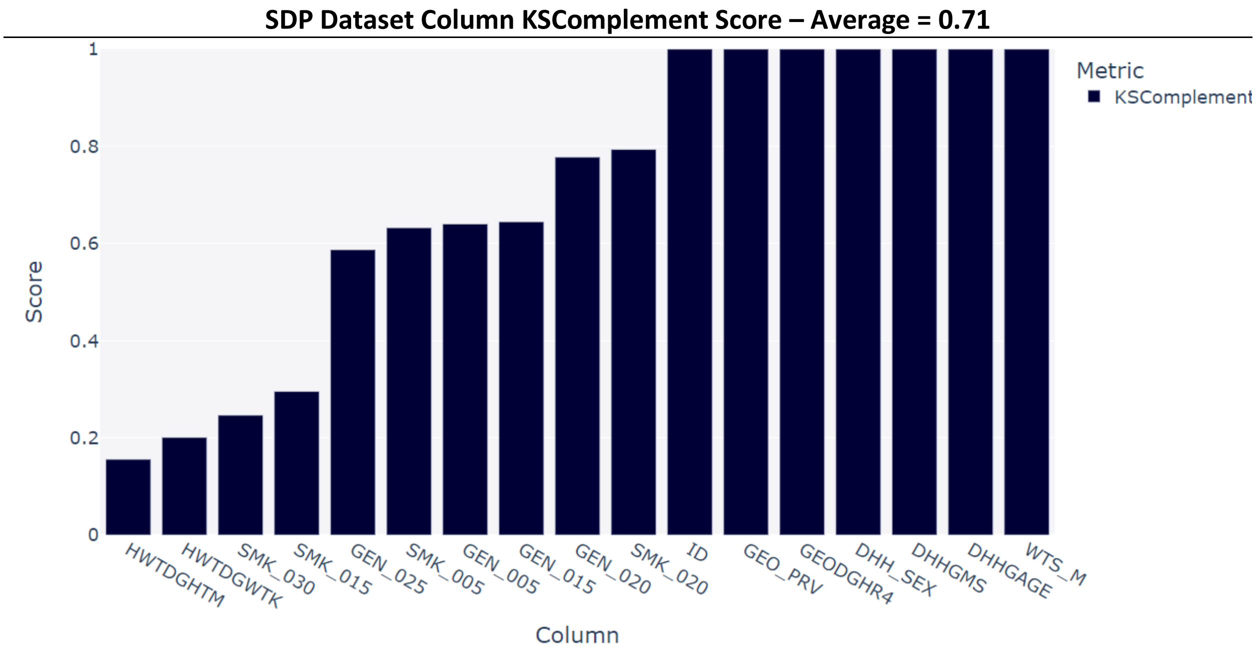


Figure 20 – LDP dataset column shape analysis and KSComplement score for epsilon 0.6.

Following the trends observed from the previous correlation and column shape plots, the resulting dataset from applying LDP with an epsilon of 0.6 further distances the data from the original dataset. All non-static columns see worse overall quality, where the height and weight columns degrade further. Overall, the plots analyzed clearly highlight that LDP with low epsilon values can result in a poor synthetic representation of the original for certain columns. A method which may improve these results for the heights and weights is to bound the minimum and maximum values, ensuring the added noise remains within these bounds. A formula exhibiting how this can be done will be presented in the following discussion subsection.

4.4 Discussion

Following both the query comparisons and correlation / column shape comparisons, we will discuss the key takeaways observed from the tests. This will help better understand how well each method works and the use cases which can benefit most from using each method. To start, the results output from the specified SDP implementation indicates that there is an issue with the implementation utilized. Unlike what is expected, there has been little variance observed within the outputs despite the variety of epsilon values used. This does not discount SDP as a viable solution but rather marks it as a challenging solution to correctly implement. Furthermore, due to the third-party shuffler required within this process, it is likely best to avoid starting with an SDP solution until after working and building trust with the techniques which we know can work well with out-of-the-box implementations. Speaking of, all approaches have been done in Python and require careful evaluation to ensure that the selected library is reliable and fixes vulnerabilities swiftly. JavaScript libraries are currently lacking for these techniques and thus custom implementations may be needed when applying LDP. Regardless, the key thing of note is that SDP still protects the privacy of the data but requires a trusted third party to process the data.

From the query results, it is clear that both GDP and LDP can be useful in protecting the privacy of the data while still allowing certain statistics to be inferred / computed. For GDP, the query outputs remained similar to those of the original non-privatized data, but the error slowly increased as epsilon shrunk. In fact, the outputs may even benefit from more noise than what has been used in the simulations to further increase the errors. However, these results prove that GDP being applied to the query outputs, rather than the data itself, reduces the utility lost in the process. The analytics derived can still be of a higher quality than other DP approaches. LDP can provide strong analytics as well but requires higher epsilon values to avoid degrading the utility of the data by too much. We have observed that adding little privacy in the process results in a much lower error when compared to the lower epsilon values. Reducing epsilon by too much may result in the general inferred statistics of a query being skewed (such as when looking for the province with the most daily smokers). At higher epsilon values the results themselves may be skewed, but the answers will remain similar to those from when no DP is applied. Thus, plausible deniability is added to the user responses with both LDP and GDP.

Within the LDP query tests, the counts received less of an observable impact than the means (with respect to the end distribution of query results). This is due to LDP arbitrarily adding noise to unbounded continuous values (i.e., the heights and weights). The correlation plots and shape evaluations further illustrate this since LDP is altering the height and weight distributions far more than the discrete columns. This may be mitigated by defining an expected range allowed for each categorical column and setting the value as the noise adjusted amount modulo the column maximum minus the column minimum. Then the column minimum can be added to ensure that the result will never be outside of the specified range (see the equation below).

$$x'_{col} = ((x_{col} + \text{noise})\ \%\ (max(col)\ -\ min(col))) + min(col)$$

Where,

- x_{col} is the original value of the column col for response x and x'_{col} is the updated value of x_{col}
- noise is the noise added by the LDP mechanism
- $max(col)$ is the maximum value possible for col and $min(col)$ is the lowest possible value for col

Despite the drastic changes to the query outputs with LDP, if the goal is to explore the overall results rather than individual outputs, then this succeeds in providing strong privacy to responses while allowing the general trends to be identified. Thus, there can still be benefit in leaving responses unbounded to ensure that they are privatized well and hold no meaning individually, only as a whole. While this may not always hold, if a significant number of results are received for each group, the general trends should be better maintained. Furthermore, since the values can be skewed heavily, this can be a defence against linkage attacks with datasets which may contain the same individual.

Overall, GDP and LDP are both feasible, but the goal of the analysis and what must be protected is important in guiding which solution to use and how to properly integrate the solution. GDP is performing the best overall but requires the full dataset itself. A good use case for GDP would be to share analytics about a dataset internally without requiring full access to be granted to the dataset itself. This can also be applied among different government agencies to help share analytics of sensitive data while providing privacy to the individuals in the dataset. LDP is good in a survey or crowdsourcing setting with many responses. Only having few responses can result in too much skew in the received data. Large-scale surveys can benefit from this approach if privacy is a key issue and if there is a lack of trust between PHAC and the respondents. LDP can also be applied to datasets from partner organizations being shared, but caution should be made on whether the noise will too drastically impact the analytics performed on the dataset.

LDP further exhibits strong correlations to the original data with high epsilon values and worse correlations as epsilon decreases. These correlation plots and shape comparisons help clearly highlight that LDP's naïve method of adding noise can result in certain columns not being as high quality as others. Therefore, it is important to understand the risks of applying LDP to continuously valued columns since there is a higher chance of it being altered substantially. The evaluations of the LDP dataset from the perspective of a synthetic dataset further highlight the delicate tradeoff between utility and privacy within its process. It will require careful tuning of epsilon to provide users with enough privacy while allowing PHAC to still gain the target statistics.

Although not tested within this research, one must also consider whether linkage attacks can be performed to violate the added privacy. By using existing data with similar columns or identifiers, it may be possible to link the privatized data with another dataset and derive some of the original values. Thus, it is important carefully select which columns need protection to ensure that privacy violations cannot occur. However, these DP mechanisms do add plausible deniability even in the case of a link since a respondent can argue that their data was altered significantly enough to match to an unrelated person.

Another approach which is not tested in this work but is interesting to consider is the hybrid use of LDP and no DP for data collection. By offering the users a choice, they will be able to offer either the full, quality data, or the altered version of that data. While this is a nice solution in theory, it can split the data into two sets and reduce the amount that can be used in both cases. Since LDP benefits from more responses, this can provide worse outputs. Similarly, if most people use LDP, the original datapoints may be insufficient for a proper analysis. The data can be put together if the LDP process bounds the results for each continuously valued column, but this work does not explore the impact of mixing LDP results with regular results in this scenario.

Additionally, there are several challenges which remain to be addressed when applying these techniques in practice. Although DP can be better audited due to the controlled privacy budget, establishing trust with a data holder such that they believe that PHAC will appropriately implement DP and not attempt to violate that added privacy is a challenging issue to tackle. Legal teams must also be aware of the techniques and understand that there will always be a tradeoff between privacy and utility. Understanding how this stands in the current laws and regulations to abide by is also important.

5. Web Application Prototype

In addition to the technical evaluations performed to understand the behaviours of different DP methods, a simple web application has been developed to help explain LDP to less technical audiences. Since DP can be a challenging topic to explain due to the paradox of it providing useful insights while reducing the quality of the data, this can help visualize what is happening in the LDP process. Doing so aims to help highlight how LDP techniques may enable collection efforts which better privatize the data before it arrives to PHAC. The web application is split into different sections, where it is presently available at dp-react-app-36sasy4jfa-pd.a.run.app, with the source code available within the project's GitHub repository. First, an overview of DP and LDP is provided with a clear way to visualize how it works at a high-level. Next, the user is guided through a three-step process on how LDP can be added to data which they are able to define.

The user will be presented with a sample form which is disconnected from any database and can be filled with corresponding health-related responses. Once filled, the user can proceed to a new page which presents the distribution being used to create noise, the epsilon and sensitivity values used, and the data before and after the noise is added (see Figure 21 below). By playing with sliders for the epsilon values, the noisy version of the user's input data is modified in real-time. This helps communicate precisely what is happening, that is, the data is being adjusted on their device before it is being sent anywhere. When the user is satisfied with the noise being added they can proceed to the final page where they can download the data which would have been sent to PHAC after the noise is added. This full process aims to demystify the LDP process and clearly present that it can feasibly be implemented within a real-world scenario. Accompanying descriptions are also provided to those who aim to learn more about the techniques rather than just interacting with it through the various sections.

Figure 21 – Screenshot of the LDP introductory website.

6. Conclusions and Future Work

To conclude, this report has provided a high-level overview of the project scope and the DP methods explored throughout the project. Supplemented by the additional presentations and references on these methods, this aims to help both executive / managerial and technical audiences understand LDP, GDP, and SDP. Following the background information, a discussion into privacy and trust has been provided to better outline the nuances that come with these methods. It is crucial to understand what privacy means within this context and to understand the corresponding trust that must be established between data curators and the data holders. The dataset utilized, simulation environment, and tests performed have been outlined to highlight how DP can be utilized, alongside its impacts to the statistics generated. Each approach has been compared and discussed. From the results, GDP performs the best whereas LDP performs well depending on the epsilon utilized and the target statistics being generated. SDP did not perform as expected and would require a better implementation alongside additional trust requirements in practice. Finally, a webpage guiding users through the LDP process has been discussed.

These results have indicated that GDP and LDP can both feasibly be applied for different use cases, but that careful considerations must be done based on their use. Future work should focus on identifying which use cases the methods can directly be applied to and any legal constraints which may limit its use at present. Furthermore, strategies must be derived to ensure that respondents and users whose data is held can trust the organization to responsibly use that data. These tasks can lead into potential deployments or mock deployments where the method(s) can be analyzed in a practical setting. Finally, relevant attacks, such as linkage attacks, should be considered for real use cases to understand mitigations that must occur. Although further testing can be done, this project has clearly outlined the feasibility of LDP and GDP with realistic health data in a survey context.

References

1. Dwork, Cynthia, and Aaron Roth. "The algorithmic foundations of differential privacy." *Foundations and Trends® in Theoretical Computer Science* 9, no. 3–4 (2014): 211-407.

2. Scott, Mary, Graham Cormode, and Carsten Maple. "Applying the shuffle model of differential privacy to vector aggregation." *arXiv preprint arXiv:2112.05464* (2021).

3. Le, Truong-Nhat, Shen-Ming Lee, Phuoc-Loc Tran, and Chin-Shang Li. "Randomized Response Techniques: A Systematic Review from the Pioneering Work of Warner (1965) to the Present." *Mathematics* 11, no. 7 (2023): 1718.

4. Holohan, Naoise, Spiros Antonatos, Stefano Braghin, and Pól Mac Aonghusa. "The bounded laplace mechanism in differential privacy." *arXiv preprint arXiv:1808.10410* (2018).

Appendix

1. Differential Privacy Definition

Let D be the set of all possible datasets, and $M : D \rightarrow R^d$ be a randomized mechanism that maps datasets to a set of query results in R^d.

Given a dataset $x \in D$, and a neighboring dataset $x' \in D$, differing by only one element.

A randomized algorithm $\mathcal{M}$ is ϵ-differentially private if

$$\forall \mathcal{S} \subset \mathcal{Y} \text{ and } \forall x, x' \in \mathcal{D} \text{ such that } x \simeq x'$$

$$\mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon} \mathbf{Pr}[\mathcal{M}(x') \in \mathcal{S}]$$

for all subsets $S \subseteq R^d$, where $\varepsilon > 0$ is the privacy parameter.

Remarks

1.1 Theoretical advantage of the definition

Strong definition which holds for all datasets and all outputs.

1.2 Practical disadvantage of the definition

Since it holds for all datasets (of a certain class) and all outputs, it may be hard to empirically verify (in practice).

2 Post-processing Theorem

A randomized algorithm $\mathcal{M}$ is ϵ-differentially private if

$$\forall \mathcal{S} \subset \mathcal{Y} \text{ and } \forall x, x' \in \mathcal{D} \text{ such that } x \simeq x'$$

$$\mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon} \mathbf{Pr}[\mathcal{M}(x') \in \mathcal{S}]$$

3 Composition Theorem

Suppose $\mathcal{M}_1, \ldots, \mathcal{M}_k$ is a sequence of ϵ-differentially private algorithms

then $\mathcal{M}_1 \circ \ldots \circ \mathcal{M}_k$ is $k\epsilon$-differentially private.

4 Approximate Differential Privacy

Approximate Differential Privacy is a relaxation of pure differential privacy where the guarantee needs to be satisfied only for events whose probability is at least δ. In practice, this relaxation can significantly reduce the complexity (sometimes the intractability) of applying pure differential privacy directly.

A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-differentially private if

$$\forall \mathcal{S} \subset \mathcal{Y} \text{ and } \forall x, x' \in \mathcal{D} \text{ such that } x \simeq x'$$

$$\mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon}\mathbf{Pr}[\mathcal{M}(x') \in \mathcal{S}] + \delta$$

Remark

Some algorithm that reliably expose data points can be made approximately differentially private under a large enough δ. For this reason, δ needs to be significantly smaller than the sample fraction of each unit in the dataset.

Example

The following is $(0, 1/N)$-differentially private

$$\mathcal{M} : (x_1, \ldots, x_N) \rightarrow x_i \text{ where } i \sim \mathbf{Unif}\{1, \ldots, N\}$$

therefore we need $\delta \ll \frac{1}{N}$

5. Query Sensitivity

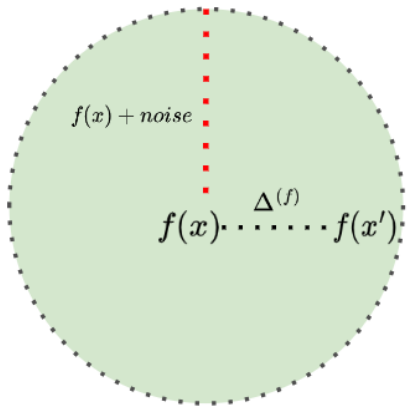
The sensitivity essentially captures how great a difference must be hidden by the additive noise generated by the curator.

Let $f : \mathcal{D} \rightarrow \mathbb{R}^k$ be a function of the data and consider a pair of neighbouring databases x and x'.

The l_p-sensitivity of f is

$$\Delta_p^{(f)} = \sup_{x, x' \in \mathcal{D}} ||f(x) - f(x')||_p$$

When f and p is clear from context, we simply denote the sensitivity by Δ



6. Examples of Differentially Private Mechanisms

In this section, we give examples of differentially private mechanisms. In particular, we describe the mechanism, give a formal mathematical proof that it is differentially private and then finally compare their accuracy bounds. In practice, the accuracy bounds would allow one to assess the utility-privacy trade-offs between different levels of privacy budget under specific statistical conditions such as sample complexity.

6.1 The Randomized Response Mechanism [Warner, 1965] is differentially private.

Mechanism Description

- N individuals answer a survey with one binary question
- The true answer for individual i is $x_i \in \{0,1\},\ i.e., \mathcal{D} = \{0,1\}^N$
- The false answer for individual i is z_i
- The answer collected for individual i is $y_i = \begin{cases} x_i & \text{with probability } \frac{e^\epsilon}{1+e^\epsilon} \\ z_i & \text{with probability } \frac{1}{1+e^\epsilon} \end{cases}$
- The mechanism is denoted by $\mathcal{M} : \mathcal{D} \to \mathcal{Y}, (x_1, \ldots, x_N) \to (y_1, \ldots, y_N)$

Proof:

For some set of outputs $\mathcal{S} \subset \mathcal{Y}$ and a database $x \in \{0,1\}^N$

$$\mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] = \sum_{b \in \mathcal{S}} \mathbf{Pr}[\mathcal{M}(x) = b]$$

$$= \sum_{b \in \mathcal{S}} \prod_{i=1}^{N} \mathbf{Pr}[y_i = b_i] \quad \text{remember} \quad \mathcal{M}(x_1, \ldots, x_N) = (y_1, \ldots, y_N)$$

We first rewrite the probability expression

$$\boxed{\mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] = \sum_{b \in \mathcal{S}} \prod_{i=1}^{N} \mathbf{Pr}[y_i = b_i]}$$

For some other neighbouring database x' we know that $\exists j \; s.t. \; x_j \neq x'_j$

$$\frac{\mathbf{Pr}[y_j = b_j]}{\mathbf{Pr}[y'_j = b_j]} = \begin{cases} e^{\epsilon} & \text{if } y_j \text{ is true} \quad y'_j \text{ is false} \\ e^{-\epsilon} & \text{if } y_j \text{ is false} \quad y'_j \text{ is true} \end{cases} \leq e^{\epsilon}$$

For some other neighbouring database x' we know that $\exists j \; s.t. \; x_j \neq x'_j$

$$\frac{\mathbf{Pr}[y_j = b_j]}{\mathbf{Pr}[y'_j = b_j]} = \begin{cases} e^{\epsilon} & \text{if } y_j \text{ is true} \quad y'_j \text{ is false} \\ e^{-\epsilon} & \text{if } y_j \text{ is false} \quad y'_j \text{ is true} \end{cases} \leq e^{\epsilon} \quad \Longrightarrow \quad \mathbf{Pr}[y_j = b_j] \leq e^{\epsilon} \mathbf{Pr}[y'_j = b_j]$$

$$\mathbf{Pr}[y_j = b_j] \leq e^{\epsilon} \mathbf{Pr}[y'_j = b_j] \quad \Longrightarrow \quad \sum_{b \in \mathcal{S}} \prod_{i=1}^{N} \mathbf{Pr}[y_i = b_i] \leq e^{\epsilon} \sum_{b \in \mathcal{S}} \prod_{i=1}^{N} \mathbf{Pr}[y'_i = b_i]$$

$$\sum_{b \in \mathcal{S}} \prod_{i=1}^{N} \mathbf{Pr}[y_i = b_i] \leq e^{\epsilon} \sum_{b \in \mathcal{S}} \prod_{i=1}^{N} \mathbf{Pr}[y'_i = b_i] \quad \Longleftrightarrow \quad \mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon} \mathbf{Pr}[\mathcal{M}(x') \in \mathcal{S}]$$

$$\mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] = \sum_{b \in \mathcal{S}} \prod_{i=1}^{N} \mathbf{Pr}[y_i = b_i]$$

Q.E.D.

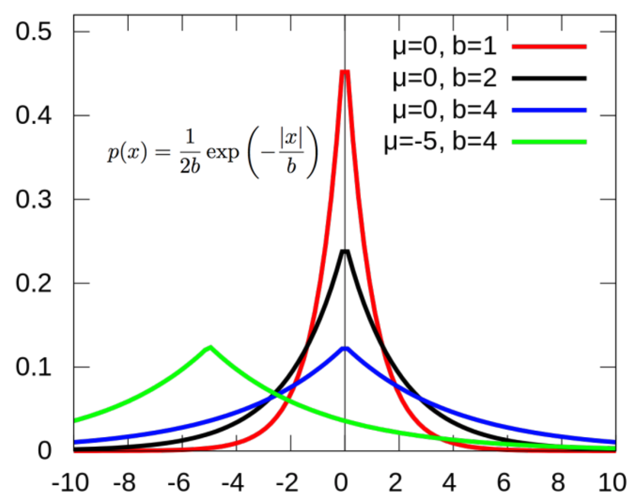
Accuracy, sample size and privacy trade-off of the Randomized Response Mechanism.

$$\left| \frac{1}{N} \sum_{i=1}^{N} x_i - \frac{1}{N} \sum_{i=1}^{N} \tilde{y}_i \right| \leq \mathcal{O}\left(\frac{1}{\epsilon\sqrt{N}}\right)$$

- If you fix ϵ and you keep increase N you get better accurary
- If you fix N and you keep decreasing ϵ you get less utility

6.2 The Laplace Mechanism [Naoise, 2018] is differentially **private.**

The Laplace Distribution



Mechanism Description

- A trusted curator holds a database $x \in \{0,1\}^N$
- 1 bit, $x_i \in \{0,1\}$, for each N individuals
- They then use the following mechanism, $\mathcal{M} : \{0,1\}^N \to \mathbb{R}$, to compute the mean
    1. Compute the mean: $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$
    2. Sample noise: $Z \sim \mathbf{Lap}(\frac{1}{\epsilon N})$
    3. Disseminate mean: $\tilde{\mu} = \mu + Z$

Proof:

For some set of outputs $\mathcal{S} \subset \mathcal{Y}$ and a database $x \in \{0,1\}^N$

$$\mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] = \int_{\mathcal{S}} \mathbf{Pr}[\mathcal{M}(x) = s]ds$$

Note that $s \sim \mathbf{Lap}(\mu, \frac{1}{\epsilon N})$ since $s = \mu + Z$ and $Z \sim \mathbf{Lap}(\frac{1}{\epsilon N})$

$$\mathbf{Pr}[\mathcal{M}(x) = s] = \frac{\epsilon N}{2} e^{-\epsilon N|s-\mu|}$$

For some other neighbouring database x' we know that $\exists j \; s.t. \; x_j \neq x'_j$

$$\begin{aligned}
\frac{\mathbf{Pr}[\mathcal{M}(x)=s]}{\mathbf{Pr}[\mathcal{M}(x')=s]} &= \frac{\frac{\epsilon N}{2} e^{-\epsilon N|s-\mu|}}{\frac{\epsilon N}{2} e^{-\epsilon N|s-\mu'|}} \\
&= e^{-\epsilon N(|s-\mu| - \epsilon N|s-\mu'|)} \\
&\leq e^{-\epsilon N|\mu-\mu'|} \quad \text{by reverse triangle inequality} \\
&= e^{-\epsilon|x_j - x'_j|} \\
&\leq e^{\epsilon}
\end{aligned}$$

Therefore $\mathbf{Pr}[\mathcal{M}(x) = s] \leq e^{\epsilon} \mathbf{Pr}[\mathcal{M}(x') = s]$

$$\mathbf{Pr}[\mathcal{M}(x) = s] \leq e^{\epsilon}\mathbf{Pr}[\mathcal{M}(x') = s] \implies \int_{\mathcal{S}} \mathbf{Pr}[\mathcal{M}(x) = s]ds \leq e^{\epsilon} \int_{\mathcal{S}} \mathbf{Pr}[\mathcal{M}(x') = s]ds$$

$$\int_{\mathcal{S}} \mathbf{Pr}[\mathcal{M}(x) = s]ds \leq e^{\epsilon} \int_{\mathcal{S}} \mathbf{Pr}[\mathcal{M}(x') = s]ds \Leftrightarrow \mathbf{Pr}[\mathcal{M}(x) \in \mathcal{S}] \leq e^{\epsilon}\mathbf{Pr}[\mathcal{M}(x') \in \mathcal{S}]$$

Q.E.D.

The Laplace Mechanism (GDP) has better accuracy bounds than the Randomized Response Mechanism (LDP). This is a common observation across GDP and LDP mechanisms for the same queries.

Laplace Mechanism
(Global DP)

$$|\mu - \tilde{\mu}| \leq \mathcal{O}\left(\frac{1}{\epsilon N}\right)$$

Randomized Response Mechanism
(Local DP)

$$\left|\frac{1}{N}\sum_{i=1}^{N} x_i - \frac{1}{N}\sum_{i=1}^{N} \tilde{y}_i\right| \leq \mathcal{O}\left(\frac{1}{\epsilon\sqrt{N}}\right)$$

7. Evaluation of Query Results

7.1 Table for Comparing the Absolute Error of the Means for the Heights in Provinces

Absolute Error for Means of HWTDGHTM (Epsilon = 5.99)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	8.914108673607758e-06	0.4444972202546407	3.37537676123071
BC	9.821078546945472e-06	0.42755463065307103	3.2591382426036075
MB	7.269382231278598e-06	0.4311928481256023	3.345780967864859
NB	1.0016031686443939e-05	0.43328913927907997	3.3060285693819567
NL	2.1229290336322038e-05	0.4228006573240273	3.3822744523895567
NT	1.910601690262581e-05	0.38872286209731366	3.3671010728107347
NS	5.803190891295884e-06	0.4457899363724761	3.443856103103111
NU	1.6835547320770415e-05	0.41577216379266246	2.842329678708855
ON	8.72231204755991e-06	0.43149054870367315	3.3226614899959896
PE	7.978923046003673e-06	0.4408484928176706	3.3719165413495262
QC	1.699607967999839e-05	0.4464415108477495	3.4278872962881275
SK	1.3586422190225988e-05	0.4169734165318624	3.486151808129274
YT	9.920148873021619e-06	0.45294140728142746	3.361748079595125

Absolute Error for Means of HWTDGHTM (Epsilon = 1.0)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.00010314562237971046	7.039419702393689	3.312821419379822
BC	2.801144757702545e-05	6.926023119351607	3.185379825972228
MB	9.67080916283306e-05	7.019717819321447	3.2449182631459457
NB	4.1604719258359066e-05	6.942585339537215	3.0733106019706544
NL	4.6378479663022885e-05	6.824893756021997	3.3124600831504267
NT	4.484214965314578e-05	6.818406632320662	3.3311384601881633
NS	4.075016981735782e-05	6.968297144725662	3.240508979704414
NU	5.26033002106588e-05	6.848637030971522	2.667208653712981
ON	0.00010010826969075284	6.9253665275093	3.1608549547066263
PE	8.70427344044342e-05	6.932551369264175	3.2265611839216275
QC	0.0001046656678796749	7.065652514092622	3.2917432916361813
SK	6.865998538828618e-05	6.955888543853345	3.2908020615863323
YT	8.256199019437436e-05	7.156865364594853	3.298611345832344

Absolute Error for Means of HWTDGHTM (Epsilon = 0.8)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	9.080051389022859e-05	9.160514791970503	3.4073632963035854
BC	0.00020044423520028002	9.094849442581337	3.256795720206257
MB	8.201608926278402e-05	8.994177178641966	3.360755574782909
NB	8.238326313524534e-05	9.081797324734382	3.151648458569241
NL	7.631207167511423e-05	8.97869486525423	3.446959254009678
NT	7.807362621573859e-05	9.197430143329758	3.3309295089483215
NS	0.00010078715250310525	9.098645576338733	3.4750414774474008
NU	5.411159227176878e-05	9.339978937402043	2.7502547464902776
ON	6.305942605315641e-05	9.100201276020988	3.3194367359428356
PE	7.005297806024656e-05	9.022140528380223	3.560428974658932
QC	0.0001788682777705411	9.079777500492053	3.4560052772111343
SK	0.00011022012138253759	9.117275713723327	3.5667709217657597
YT	0.00011240805420536914	9.222041904363	3.3618297006177107

Absolute Error for Means of HWTDGHTM (Epsilon = 0.6)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.00014536017355496636	12.686209771450995	3.4215138723933807
BC	0.00013586538878011925	12.510798111843425	3.3144409605988434
MB	7.659081190653652e-05	12.690695499453849	3.38293367089586
NB	0.000135044181711722	12.441390359654562	3.4125574394350293
NL	0.00010636289457061832	12.831560523965766	3.37858966395218
NT	3.9608561460191535e-05	12.774168759735414	3.224376767369539
NS	0.00013573415133431147	12.603441360043188	3.363714595352441
NU	0.00012811175537891259	11.90318949929011	2.6437444149109908
ON	0.00017772024990357868	12.627769724621961	3.3285588197366094
PE	9.660501505482968e-05	12.482650307727553	3.3403080602896305
QC	0.00016297359417377668	12.81584774525584	3.4453723061051624
SK	0.00012846235536824232	12.434620095783508	3.480491673479938
YT	0.0002372690850056029	12.731505154577112	3.4764156338115155

7.2 Table for Comparing the Absolute Error of Means for the Weights in Provinces

Absolute Error for Means of HWTDGWTK (Epsilon = 5.99)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.00174565883523308	96.59195627446593	364.276972137156
BC	0.00196670364882718	96.23188600231629	354.1481887025469
MB	0.0013396311291444048	95.48078593179221	351.623521454792
NB	0.0006471652984430161	96.96469052471666	347.8245686598534
NL	0.00250908281776668	91.27040329860077	348.6341108241246
NT	0.0012681425004217	90.10803111691365	374.1664940837381
NS	0.001129972729336232	94.71003981787898	360.7523093821659
NU	0.00130864356711408	97.22661328074123	301.4203691440189
ON	0.0029450755742629776	94.80476068272286	348.54391666095853
PE	0.0009330677101729	93.41993125023491	351.89704198199
QC	0.0005553339661531233	95.46629495979633	377.76477455883
SK	0.00254559790575962	95.46965924240624	362.9474728716704
YT	0.0019736115850036598	87.76373921266222	335.9477831910609

Absolute Error for Means of HWTDGWTK (Epsilon = 1.0)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.015230045402870419	862.2358208885337	362.7336598691059
BC	0.00254800715729852	859.6935637461844	346.31196671699684
MB	0.00601840369228622	867.4474835036078	346.75619087392
NB	0.004751956443283181	856.9288731073369	346.34734513022863
NL	0.007764670286968501	853.4443709199637	335.8169740611187
NT	0.00506740165664606	882.4094645281727	362.1986049600288
NS	0.00740085031348482	861.8494886558419	336.34471643774964
NU	0.00381134931305946	801.3891784309687	267.8067508286239
ON	0.00619764645788902	855.3585296752259	336.99805343453863
PE	0.0058867825598099	854.6704782869787	365.5923211496547
QC	0.00561199508331634	869.6979898039433	362.4485296466894
SK	0.00647236989783546	856.1104524602913	358.58369225123465
YT	0.01045436312884926	828.8951295701787	322.6004371258001

Absolute Error for Means of HWTDGWTK (Epsilon = 0.8)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.00599258475573374	1091.8510737064696	367.0419145228517
BC	0.012507646367401962	1093.729290718989	360.04289283302853
MB	0.010029842625004861	1097.4699933366842	359.6392184037235
NB	0.01357686696317722	1086.1506893337955	354.0625613929135
NL	0.0028038929475371196	1090.9476974560253	344.5272742649919
NT	0.005924517261126941	1097.9473579501614	361.4566501508305
NS	0.015350784273778081	1084.283877234872	361.5595707551571
NU	0.0068078178449411	1071.7500736847237	303.63911974744497
ON	0.0050786295936461195	1091.523000030324	347.6886538081885
PE	0.01008605661054954	1102.2000109521528	366.2889319398827
QC	0.00331076878045444	1114.5740857349617	377.9223022561536
SK	0.015379121437325638	1087.9486295805705	357.74158932255364
YT	0.015606694678990559	1047.2634044039253	360.8955090148744

Absolute Error for Means of HWTDGWTK (Epsilon = 0.6)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.00663839637235238	1506.9895143908118	364.47965511702114
BC	0.025285012064426782	1487.3937580572965	360.5788586037996
MB	0.015302304989916101	1467.166539624236	351.7184536377579
NB	0.01212539297742406	1483.1018884272949	356.59197482455454
NL	0.0092229910879325	1516.11741286453	346.3130173199412
NT	0.01205047376413684	1541.0577746023223	385.45963161088486
NS	0.008430383299992121	1527.0783364916688	365.6243807818097
NU	0.014833357229213057	1455.0854707062813	311.4117643252938
ON	0.01821856622282777	1486.0517685682676	349.3204066735608
PE	0.01339494385270536	1518.2210790777635	338.44170072338545
QC	0.013704353572580841	1519.3036774171724	380.4120319933371
SK	0.00973408446391768	1467.3267705703079	365.13668750781153
YT	0.01966492576571564	1468.3059363551413	351.8233978621298

7.3 Table for Comparing the Absolute Error of the Counts of Daily Smokers in Provinces

Absolute Error for Sums of SMK_005 – Daily Smoker (Epsilon = 5.99)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.08892606445588166	1483.0660000000964	491319.8300000007
BC	0.281120446906425	2594.8679999999003	714424.3200000001
MB	0.14043984045274555	694.0199999999983	149355.33999999953
NB	0.09183360084134615	453.44600000000503	100928.31000000008
NL	0.16125844948110166	494.35800000000455	46288.98999999992
NT	0.06204539879618091	42.57399999999943	415.69000000000415
NS	0.2391611759579973	175.93399999999673	109588.56000000026
NU	0.08202110360762158	5.2979999999999565	9506.319999999983
ON	0.17905434723943472	3409.0180000003893	1759115.6400000013
PE	0.09169662852000324	102.15200000000115	15431.359999999986
QC	0.12543821106664832	2702.9539999999106	880882.2500000035
SK	0.3222201985830907	527.6120000000053	93364.64000000004
YT	0.16049213802871232	2.406000000000131	3286.989999999998

Absolute Error for Sums of SMK_005 – Daily Smoker (Epsilon = 1.0)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	1.4982668519485742	103837.77400000069	467089.46999999805
BC	0.9939643883146345	164467.3899999998	688514.0400000012
MB	1.2454090176936006	36433.3319999999	143469.90000000046
NB	0.5024410103593254	20086.127999999997	88589.98
NL	0.6202896010072436	6143.8779999999915	43076.59999999986
NT	0.5708251851301611	588.9900000000007	49.93000000000029
NS	0.7781306076591136	19203.692000000032	89510.42999999986
NU	0.8147325093770632	2958.257999999995	7114.449999999985
ON	0.8806329442653805	395895.61800000176	1590735.6900000083
PE	0.947646650619572	2782.2339999999995	14582.879999999986
QC	1.1538928777445108	190485.95400000009	879533.670000003
SK	2.7805952976807022	19648.62800000018	97560.8799999998
YT	0.862162410214296	712.0460000000006	3897.190000000001

Absolute Error for Sums of SMK_005 – Daily Smoker (Epsilon = 0.8)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.7331017173826695	110989.36200000081	471354.0599999997
BC	1.8597060116240756	184858.22999999972	691568.0000000026
MB	0.8217625267134281	32401.819999999854	156879.5300000004
NB	1.1538582060544287	24661.729999999974	100956.53999999983
NL	1.5475710286933464	4752.463999999978	28700.54000000006
NT	2.3800102285480533	828.9280000000006	720.1300000000065
NS	1.7350606409832836	21941.80999999999	90387.57999999975
NU	1.7319458149282583	4159.629999999995	8919.329999999978
ON	1.9851769138593227	458272.4340000011	1733296.3500000015
PE	1.0165886627786676	2290.3520000000017	16689.910000000003
QC	1.0533398873871191	210738.8840000003	920439.7699999961
SK	1.1591855243896134	26036.400000000267	95555.31999999998
YT	1.0514538020997861	1139.4560000000008	4852.990000000003

Absolute Error for Sums of SMK_005 – Daily Smoker (Epsilon = 0.6)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	1.9558088879100979	135155.45200000028	496772.97999999986
BC	1.1302557996008544	203439.7159999996	713222.7800000014
MB	2.600346182769863	40337.93199999987	172290.4400000001
NB	2.2190396255551605	27298.997999999985	97932.54000000034
NL	1.3664913257816806	6448.954000000036	37911.169999999925
NT	0.6354770600988559	876.280000000001	273.3300000000036
NS	1.486158767505549	23871.79400000003	89440.42999999983
NU	1.768678864033427	4746.517999999991	9822.259999999986
ON	2.8734803174156696	477312.9240000043	1702500.5800000117
PE	1.1630239520010945	3562.443999999997	15979.69999999995
QC	2.682100563752465	255752.34599999982	956257.1699999968
SK	2.524711318616755	21715.922000000184	95776.68000000012
YT	1.1631583974229216	1224.358000000001	3013.279999999997

7.4 Table for Comparing the Absolute Error of Counts of People with Good Perceived Health in Provinces

Absolute Error for Sums of GEN_005 – Good Perceived Health (Epsilon = 5.99)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.14496489616576577	529.7920000000042	346441.8699999993
BC	0.17871995742898433	1465.6179999999936	332779.5599999985
MB	0.11241449036460834	206.21200000000243	79180.84000000005
NB	0.21528721780632618	437.6959999999963	27081.090000000004
NL	0.17631251223501745	130.41800000000222	34683.27000000007
NT	0.3602191742035757	20.907999999999994	1925.5199999999973
NS	0.17166863670281599	220.3859999999986	45620.589999999895
NU	0.16122976105261838	6.888000000000011	1463.7299999999996
ON	0.2801923400489613	1690.1839999999852	1081365.810000001
PE	0.18983429085928946	155.5	8409.01999999999
QC	0.138165776617825	2180.1859999999406	641944.9299999982
SK	0.2716393701703055	376.21800000000223	76365.09000000007
YT	0.15443990324247348	8.40600000000013	2283.309999999999

Absolute Error for Sums of GEN_005 – Good Perceived Health (Epsilon = 1.0)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	1.0250389867112972	66495.84199999987	298899.72000000114
BC	0.4571990221971646	72618.58599999986	323561.67999999964
MB	1.171442723853397	22934.02999999998	62176.89000000007
NB	0.6797150787519058	6570.560000000003	30616.970000000125
NL	0.28279592029866757	10566.168000000042	35495.14000000001
NT	1.1932801695700619	569.3239999999994	1869.6899999999991
NS	0.5253024014789844	9796.533999999967	45129.27999999984
NU	2.6060191204121113	320.91199999999935	1692.64
ON	0.9892722014104948	244793.21200000183	1126208.440000001
PE	1.1441370840289893	1234.0119999999977	10580.689999999997
QC	1.0841843337053434	156039.91999999905	633968.8999999998
SK	0.914121098787291	17044.828000000045	71601.60000000014
YT	0.483836592427906	567.6519999999997	1268.8799999999987

Absolute Error for Sums of GEN_005 – Good Perceived Health (Epsilon = 0.8)			
Province	GDP Error	LDP Error	SDP Error
AB	1.0556450162199327	90781.7519999999	318097.86999999976
BC	1.540419123042375	80272.70199999916	339997.2500000009
MB	3.2173931158700726	22386.726000000002	81742.38999999981
NB	1.4283198543678737	4942.280000000002	28874.129999999976
NL	1.4148453215660992	7124.144000000017	39966.60000000004
NT	0.687435165829811	468.02800000000036	1753.8900000000008
NS	1.7606932617549318	12745.06599999998	51043.32000000008
NU	1.6414169664042675	216.8679999999992	1734.5099999999984
ON	1.0337596506346016	291913.57200000086	1019403.5399999938
PE	1.005099939694628	1908.9020000000012	9340.389999999974
QC	0.7521244637435303	171441.13199999923	699554.4199999947
SK	1.6032373369060224	19200.47000000001	73895.75000000001
YT	1.8091449035840923	323.39799999999985	1642.6999999999966

Absolute Error for Sums of GEN_005 – Good Perceived Health (Epsilon = 0.6)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	2.02447891284246	98251.1660000001	365807.3600000012
BC	1.195458721066825	91460.85600000001	342857.3400000006
MB	2.0198174065357306	28867.317999999963	91822.07000000012
NB	1.4083409181825117	3150.1759999999895	31354.040000000008
NL	2.7474721208971458	8216.964000000014	31202.880000000085
NT	1.5192521332026445	699.6659999999996	2276.029999999997
NS	2.2518210132111562	13349.361999999988	43691.70999999989
NU	1.0800063683223016	485.07799999999895	985.1399999999994
ON	2.127964633726515	311605.60000000143	1008189.1999999946
PE	0.7060646963349427	2603.978	9050.059999999976
QC	2.3760468965629116	201125.92799999946	678038.7499999992
SK	2.529763767219265	22727.354000000003	60832.80000000006
YT	1.4732561149344292	558.289999999999	2204.179999999996

7.5 Table for Comparing the Absolute Error of the Counts of People who Believe that Work is not Stressful at all in Provinces

Absolute Error for Sums of GEN_025 – Work Not at all Stressful (Epsilon = 5.99)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.04066863320767874	1807.5299999999813	623301.6799999988
BC	0.18866559681482606	4130.395999999996	587344.5199999986
MB	0.11845647556474428	902.6099999999976	168213.79000000044
NB	0.16778830757248214	319.17799999998533	81859.8300000001
NL	0.15445418703311584	194.77000000000407	59474.289999999935
NT	0.1066081804037821	25.88999999999978	6533.700000000005
NS	0.08257593399612229	419.76600000000326	126359.59
NU	0.2461033426447102	27.38000000000011	2565.6399999999994
ON	0.19928059671074153	4725.590000000409	1745304.8500000047
PE	0.20700206429319218	140.70800000000162	18734.30999999999
QC	0.1565972437150776	3330.797999999905	981473.3900000008
SK	0.19980984254507342	792.5400000000024	162376.7099999999
YT	0.312527450872949	22.022000000000116	6274.82

Absolute Error for Sums of GEN_025 – Work Not at all Stressful (Epsilon = 1.0)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.8222128683701158	208211.2859999983	594696.5699999984
BC	0.8209726714529098	190624.7739999974	579131.3999999978
MB	0.7880074182758108	53648.13000000028	161949.7800000004
NB	0.5702875012648292	26093.370000000075	74469.62000000011
NL	1.065173875459004	20811.89399999996	50249.559999999925
NT	0.8003388840596017	1920.1740000000052	5564.6700000000055
NS	0.7861976140295156	40609.18399999987	111275.81999999986
NU	0.9528438047778763	961.7079999999996	2257.77
ON	1.5714891436509788	588905.6820000021	1664010.190000002
PE	0.9456009572153562	6087.913999999995	19162.96999999999
QC	0.4888168912846595	321596.58600000007	969467.0399999965
SK	1.1604063885170035	50140.47399999985	153263.86999999985
YT	1.314775291638216	2417.9940000000047	5349.7300000000005

Absolute Error for Sums of GEN_025 – Work Not at all Stressful (Epsilon = 0.8)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.8427998959552496	239806.60399999865	613148.5899999985
BC	0.5426647666376084	232899.8059999977	591391.5699999982
MB	0.4479345416184515	62952.366000000446	174895.81000000035
NB	2.4113321383134463	33673.60000000011	93259.64
NL	1.0348136904154672	21288.735999999924	55757.69999999993
NT	2.521977726638943	2276.3340000000057	5214.210000000005
NS	0.7142531845369376	49236.35999999984	131594.07999999984
NU	0.8886401412773921	1112.5560000000007	2284.0000000000005
ON	1.1008482726290822	692014.7080000055	1768129.8000000012
PE	1.2646276411629516	7434.429999999993	18996.58999999998
QC	0.424901284230873	379497.8639999987	1016928.0199999992
SK	1.535428362514358	63798.75999999987	164878.46999999988
YT	1.190053897294638	2351.668000000003	6848.160000000001

Absolute Error for Sums of GEN_025 – Work Not at all Stressful (Epsilon = 0.6)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.7631062802392989	268933.1099999983	606626.0999999989
BC	4.074050439754501	258124.98199999673	604444.6299999983
MB	1.2637828749488107	73708.73000000046	175726.71000000025
NB	1.2153713585110382	38670.0240000001	89841.39000000014
NL	0.828346152897575	25851.375999999942	66437.94999999995
NT	2.3996370903572823	2225.330000000006	5491.720000000007
NS	2.022651542420499	52831.543999999856	127128.58999999982
NU	0.5424707590673279	1165.3900000000006	2884.6700000000005
ON	2.0547323202714325	762854.5400000049	1760753.9900000044
PE	2.187560693330306	8412.929999999988	17024.889999999985
QC	1.601262388844043	462593.35399999825	945921.4499999986
SK	1.2531290046754293	68546.21399999989	164109.68999999994
YT	3.3072761400751913	3137.6620000000025	6466.600000000001

7.6 Table for Comparing the Absolute Error of Counts of People who Believe that Work Quite a bit Stressful in Provinces

Absolute Error for Sums of GEN_025 – Work is Quite a bit Stressful (Epsilon = 5.99)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.11706522037857207	896.6059999999998	194164.7200000001
BC	0.32801863876520654	1126.2560000000055	280081.5800000003
MB	0.2264750506757991	444.07600000000383	49641.79000000005
NB	0.08892184365540737	181.33199999999925	30527.469999999965
NL	0.1937968182755867	199.2199999999968	10320.530000000012
NT	0.29595862057703926	58.705999999999946	337.7300000000014
NS	0.17794446873303965	361.2980000000054	47659.599999999904
NU	0.15851367184877746	12.38799999999992	2049.8899999999967
ON	0.1020467378897592	3305.4959999999965	730983.8999999998
PE	0.11963832471956262	55.81600000000035	7117.05
QC	0.19397423617774617	1850.9520000000252	394486.7600000005
SK	0.15852395192487162	470.3699999999997	67428.74000000008
YT	0.17236013207138964	13.607999999999993	1267.4600000000014

Absolute Error for Sums of GEN_025 – Work is Quite a bit Stressful (Epsilon = 1.0)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.4176708610786591	37261.09200000048	187051.9600000005
BC	1.2572838615858928	64362.42400000036	267404.3899999992
MB	1.4424364923906978	13356.473999999958	62598.849999999955
NB	0.7242570933100069	7372.822	31900.78999999989
NL	0.6917271144717233	934.6380000000048	5483.559999999998
NT	0.8771968206481688	455.9479999999995	1270.4399999999987
NS	1.216719347378239	11018.596000000012	50521.97
NU	1.0728310451499055	362.8499999999987	1206.7999999999984
ON	1.393656562641263	166591.44399999944	654033.8999999928
PE	1.1867147281674988	1182.2879999999982	7061.8299999999945
QC	1.323775988339912	94917.18600000013	413996.0099999999
SK	0.9504564474074868	12746.563999999957	56745.569999999934
YT	1.4333733885787296	429.7739999999999	1388.8900000000003

Absolute Error for Sums of GEN_025 – Work is Quite a bit Stressful (Epsilon = 0.8)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	2.1641124420042614	50976.44400000044	205008.5200000001
BC	0.561647824884858	74942.45800000032	298448.37999999983
MB	0.9471648105391068	17643.511999999962	57316.07000000008
NB	2.2087187818193343	8022.608	35274.49999999994
NL	2.1507342447090196	1407.074000000002	4849.819999999978
NT	1.218364462632053	98.46400000000031	175.48999999999978
NS	1.1813052588287973	11916.044000000013	46381.48000000001
NU	0.6520222020919391	620.4339999999983	1665.6999999999964
ON	1.2911702843150124	188403.36000000074	722378.5899999946
PE	2.0954705856755025	1950.7779999999973	6674.880000000002
QC	1.731612786243204	110642.90799999998	422069.780000001
SK	1.0117834260337986	14984.881999999949	61473.779999999955
YT	0.589565681106069	291.07599999999974	1684.7199999999998

Absolute Error for Sums of GEN_025 – Work is Quite a bit Stressful (Epsilon = 0.6)			
Province	**GDP Error**	**LDP Error**	**SDP Error**
AB	0.8181885236757808	57620.90200000034	180334.64000000036
BC	1.012877933710115	82017.1040000003	273396.7900000001
MB	0.6117325184808579	17556.88199999999	67012.25999999978
NB	0.9543485570524354	10511.26000000001	34455.79000000001
NL	0.682112206867896	3427.485999999984	6537.370000000075
NT	1.6420871918686317	482.85599999999977	264.34999999999945
NS	2.3291491938914985	10539.393999999998	46053.100000000006
NU	1.907363737135529	789.5019999999977	1399.5399999999972
ON	0.4916469380725175	230133.53400000068	751650.539999999
PE	2.049040236998553	1369.1559999999986	7026.200000000003
QC	2.0804377074702645	144338.68200000006	395783.1699999985
SK	2.158883521791722	19895.909999999945	57742.54000000008
YT	2.6360566504223244	683.2379999999998	2228.6600000000003