Privacy Tools for Contingency Table Analysis

Project: Privacy Tools for Sharing Research Data

Hyun woo Lim

Mentors: Marco Gaboardi & Ryan Rogers

Table of Contents

- **Introduction**

- **Methods**
  - Application of Privacy Tools in a Contingency Table
  - Laplace Mechanism
  - Independence Testing: Chi-squared Test
    - The Framework of Hypothesis testing
    - Simulation Studies of Laplace Mechanism

- **Experiments**
  - Accuracy
    - Theory Bound Simulations
    - Comparison of Different Rounding Procedures
    - Accuracy of Chi-squared statistics with Privacy
    - The Comparison between the Size of Noise and Sampling Error
  - Utility
    - The Success Rate of a Private Table's Chi-squared Test
    - The Distribution of P-values from Chi-squared test with Privacy
    - Nature of Chi-squared Distribution and P-value
    - How Can the Variation of Private P-values Be Bound?
    - Chi-squared Value of a Private Table after Laplace mechanism
  - MWEM
  - Goodness of Fit Test
    - Transition from Chi-squared Test for Independence Testing to Goodness of Fit Test for Proportion Testing
    - Goodness of Fit Test without Privacy
    - New Goodness of Fit Test with Privacy

- **Further study**

- **Conclusions**

Introduction

Privacy tools for sharing research data is a multi-disciplinary study among computer science, statistics, law and social science for developing tools to resolve the paradox of utility and privacy. If we do not apply privacy tools in data, there are risks of identification. At the same time, if we change data too much for privacy, there is no utility of data left, so we cannot use the data for research because the data will be no more like the original data. We want to protect the individual's data but learn something about the global population, made up of individuals. Particularly, as a social scientist, I focused on developing, applying and understanding privacy tools for contingency table analysis. The idea of differential privacy is not new, but the application of differentially privacy to the social sciences, through statistical testing, has not been well studied. However, as the project title indicates, it is important to study differentially private mechanisms' practicality because researchers of social science want to know whether individual's privacy may be preserved while also being able to still perform statistics.

The main questions that we have asked in the summer include the following: can we implement (proven) differentially private algorithms to privately release contingency tables? How useful are differentially private mechanisms in statistical analysis from a contingency table? How can we further develop privacy tools for contingency table analysis for social science research? These questions are studied by a series of experiments on real data sets, as well as simulated ones.

At a high level, to meet the goal of differential privacy we require that if any adversary were to see a contingency table, he would not be able to conclude what any one particular individual's data was, regardless of what additional information the adversary may have about others' data. If we only had this goal in mind, we could easily output a private data set by simply outputting the same contingency table, regardless of the dataset. That is why we also include the goal of utility, where we want answers to any query on the private version of the contingency table to be close to the answer of the query on the actual (not private) contingency table. We focused on outputting a contingency table from a dataset as well as hypothesis testing that uses contingency tables (a table of counts) instead of the full dataset. Some statistical tests that we use a contingency table for include Goodness of Fit, where we test whether a vector of proportions indeed came from a known probability vector, or independence testing which tests to see whether two random variables are independent of one another or not. For instance, if a study was found to be statistically significant, we wanted our studies to also conclude the same while preserving privacy. These are two key goals that we want to achieve through differentially private tools.

The motivation of applying privacy tools in a contingency table is to protect sensitive information of a person or entity. One could question privacy tools' necessity in a contingency table because a contingency table is a summary of a data set and does not contain what may be termed "personally identifiable data." A contingency table consists of cells that count numbers, so people may understand a contingency table can provide privacy. However, as Latanya Sweeney's study of re-identification of "private" data by using public data, from anonymous data, we can identify private information (1996). There are still risks of revealing private information even if it is in a contingency table as an anonymous data. Therefore, we should apply privacy tools in a contingency table to ensure privacy.

Other than keeping privacy, privacy tools should provide good utility in terms of usefulness in social science research. Our intuition and previous privacy tool studies present that there is the trade –off between utility and privacy of data. For instance, if I change all numbers in a contingency table to 100, it is no more like the original table, so we have good privacy. However, now the table is independent from the original data, so there is no usefulness of the table. We introduce the parameter epsilon to be the dial between prefect privacy (epsilon = 0) and no privacy (epsilon = infinity).

Therefore, in this summer, I have focused on how to implement privacy tools for a contingency table and how privacy tools work for statistical tests on contingency table.

Methods

Application of Privacy Tools in a Contingency Table

There are many differentially private mechanisms developed for different goals and applications. In my study, I mainly focused on Laplace mechanism, but I also tried MWEM which is still in progress of study. I will briefly explain two mechanisms in here: Laplace mechanism adds carefully selected noise to each cell entry and MWEM computes a whole new synthetic dataset that is itself differentially private.

Laplace Mechanism

Laplace mechanism is the original differentially private mechanisms for computing counts on a dataset privately (Dwork, McSherry, Nissim, Smith, 2006). From the Laplace distribution I randomly select independent noise from a Laplace distribution, and add this noise to each cell in the contingency table. The new table (with noise) is now a private table. We have a parameter epsilon which is the level of privacy that we can guarantee which affects the noise we add to each cell. The Laplace mechanism is a good privacy tool in that it gives good utility guarantees and is easy to implement for people outside of computer science.

If an individual were to change his data in a dataset, the cell counts in two cells of the contingency table will change, because he will move from one cell (decrease count by one) and then move to another cell (increase count by 1). We then say that the global sensitivity of a contingency table is two. The global sensitivity is the difference in two "neighboring data sets" when we change an individual of a dataset. The Laplace mechanism then adds independent Laplace noise to each cell count in the contingency table, where the Laplace noise takes as a parameter the global sensitivity divided by epsilon, i.e. 2/epsilon. Therefore, for instance, if Epsilon is 0.1, we add or subtract roughly 20 people in each cell of a contingency table (note that it is not exactly this number because the noise is random). In the case of Chi-squared statistic that we will go over later of this paper, one could suggest adding noise directly to Chi-squared statistic. However, we cannot add noise to the Chi-squared statistic because the global sensitivity could be massive. If we check the Chi-squared statistic, on neighboring data sets, one player may be the only one in a row, so when that person moves, there could be nobody left in the row, which could cause Chi-squared statistic to divide by zero. Therefore, adding Laplace noise directly to Chi-squared statistic is not a good idea.

Independence Testing: Chi-squared test

Why do social scientists use a contingency table in their studies? What kind of statistical tests do social scientists apply with a contingency table? A contingency table is commonly used in statistics in social science because it is summarize vast information in a concise form. Moreover, social scientists present a contingency table when they want to check independence of categorical variables. Before social scientists do various statistical analyses such as Maximum likelihood estimate or regressions, they try Chi-squared test to make sure the variables are independent or not.

At first, I will briefly explain how to do Chi-squared test, and then what we expected to observe and what we have observed from simulation studies after applying Laplace mechanism.

Null hypothesis of Chi-squared test is X and Y are independent of each other, and the alternative hypothesis is that X and Y are correlated in some way. After calculating Chi-squared value of a table, social scientists read the P-value. Chi-squared values are not meaningful by itself, P-value tells us whether the variables are independent or not. However, P-value is not the fixed standard to follow, but it just implies that, the null hypothesis is actually right, but the test provides incorrect answer. That error is Type 1 error that the significance of the test is decided by the level of α. For instance, commonly used α in social science is 0.05. Although the null is actually true, there is α probability that Chi-squared test could say that the null should be rejected.

	Null is true	Null is false
Reject Null	Type 1 error	Correct rejection
Fail to reject Null	Correct acceptance	Type 2 error

< Table 1 >

The framework of hypothesis testing

Consider two random variables X and Y where each takes two possible outcomes X1, X2 and Y1, Y2, respectively. We summarize the number of times that each joint outcome occurred in the following table, e.g. the number of times that X=X1 and Y=Y1 is "a". The basic framework of hypothesis testing for Chi-square is the following:

Null hypothesis (H0): X and Y are independent of each other.

Alternative hypothesis (H1): X and Y are dependent.

	X1	X2	Total
Y1	a	b	a+b
Y2	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$\text{Chi square of the contingency table} = \sum_{n=1}^{N\ of\ cells} \left(\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \right)$$

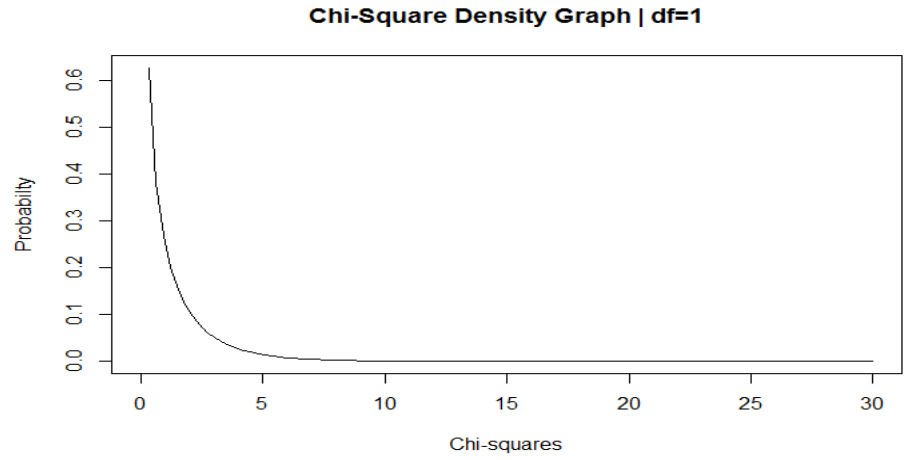
Expected value of each cell:

	X1	X2

Y1	$\frac{(a+b)*(a+c)}{(a+b+c+d)}$	$\frac{(a+b)*(b+d)}{(a+b+c+d)}$
Y2	$\frac{(a+c)*(c+d)}{(a+b+c+d)}$	$\frac{(b+d)*(c+d)}{(a+b+c+d)}$

Degree of freedom: (Number of rows – 1) * (Number of columns - 1)

To evaluate the Chi-square statistic calculated from the given table, we need to find the threshold value that we will fail to reject H0 as long as the statistic is smaller than the threshold and reject if the statistic is larger than the threshold. We find the threshold value by finding the critical point on the Chi-square distribution with1 degree of freedom where the Chi-square distribution at the critical points is 1-alpha.



< Figure 1 >

Based on this Chi-square distribution of the Figure 1, we need to compare the Chi-square value from the given table. If the value is larger than the Chi-square critical value for alpha = 0.05 probabilities, we reject the null hypothesis.

Simulation Studies of Laplace mechanism

The strengths of simulation studies in the application of differentially private mechanisms are the following: first, if the study starts from the empirical results, it is easy to find whether the mechanisms really work or not; second, the application of privacy tools in contingency table analysis is very new area of research that relates differentially private tools in the context of social science. I present research that relates differentially private tools in the context of social science. I present below what I have tested.

Experiments

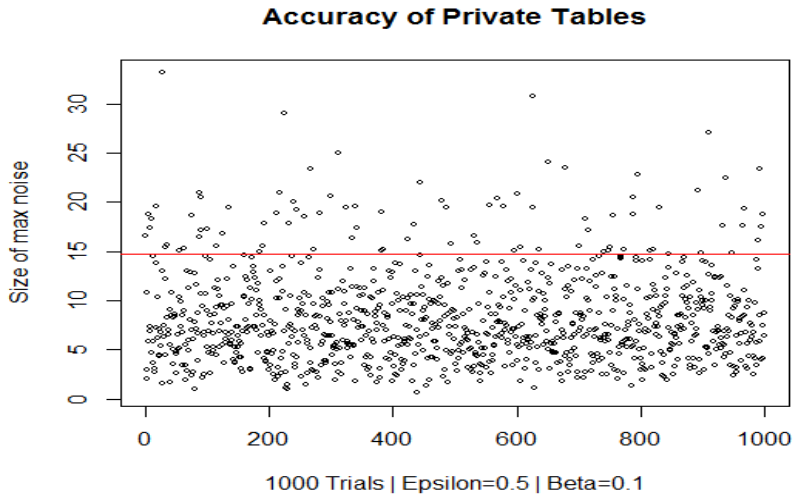
Accuracy:

Theory Bound Simulations

The accuracy of a private table means how similar a private table is to the original table. In theory, we know a bound for the accuracy of Laplace mechanism (Dwork and Roth, 34). Accuracy of a private table can be checked by using this theorem, but we would like to know whether this bound is actually realized in our experiments, or if the bound is just way too big. The parameters we have include: Epsilon which is a parameter of privacy and Beta which is the probability that the noise we add is too big and can be considered the probability that a "bad event" will occur.

Goal: Based on the theorem, $\frac{2}{Epsilon} \times \log\left(\frac{d^2}{Beta}\right)$ is the theory bound for a d by d table, that is with probability at least 1- beta, we can guarantee our noise to be smaller than the theory bound. We want to check whether this theoretical bound actually works in the application of Laplace mechanism in a contingency table.

Method: 1000 times of simulations to add Laplace noises in an original contingency table. In the experiment, Epsilon is 0.1 and Beta is 0.1.



< Figure 2 >

Conclusion: As the Figure 2 shows, because we set the Beta as 0.1, about 90 percent of the noise is under the theory bound. This result tells us that in terms of accuracy, Laplace mechanism works well because the outcomes in the plot is what we expected to see based on given conditions of Epsilon and Beta. In other words, it seems like our theoretical bound is tight in this simulation.

Comparison of Different Rounding Procedures

There are several problems if we use Laplace mechanism without anypost-processing on our contingency table. For example, the total number of an original table is changed after adding noise and a table's cell count can be negative as well as non-integers. Non-integers in a private table are minor issues because users could understand the table even if numbers are not integers as long as the users are properly told that the contingency table they are viewing is a noisy version of the actual table. However, a contingency table should not have negative value. The total counts of a private table should be adjusted to same as the actual table because the noise from Laplace mechanism could change the sum of counts in a private table. We test two rounding procedures: naïve rounding, we simply make a cell count zero if it is negative; MLE rounding which finds the most likely contingency table that could have been the result of adding

Laplace noise to each cell. Note that the naïve rounding may result in non integer values but all the cell counts will be nonnegative, and the MLE contingency table will have all integer values which all sum to the total number of individuals in the dataset.

Goal: we want to know the accuracy of a private table after different rounding procedures. We expected to see the MLE rounding work better than naïve rounding.

Method: There are two rounding procedures: Naïve and MLE rounding. Via simulation studies, we want to check the size of maximum noise after both procedures.

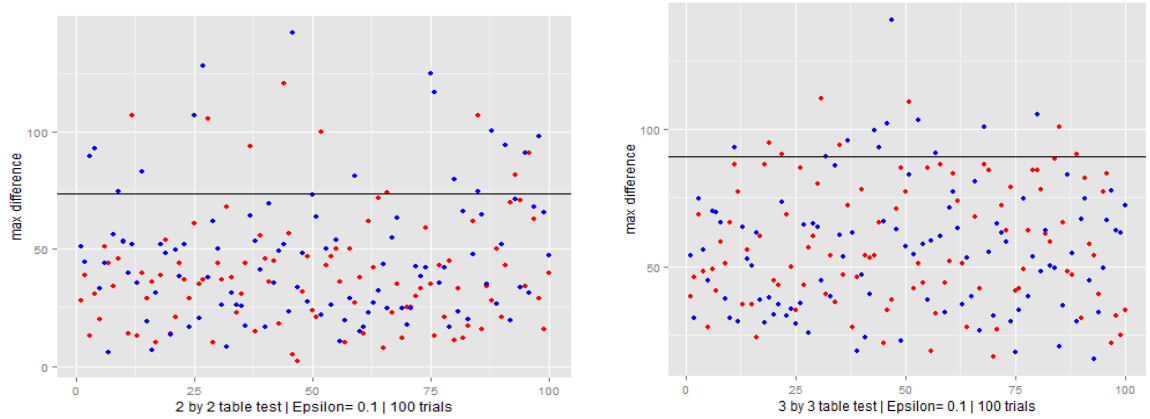
	Method	Strength	Weakness
Naïve rounding	Change negative values in a private table to zero	1. Short running time in R	1. Total number of counts is different from the original table 2. Still have non-integers in a table
MLE rounding	Using Maximum likelihood estimation, change a private table's components close to what the original table looks like	1. theoretically makes sense as an effective post-processing 2. all cells can have integers 3. Total counts of a private table is same as the original	1. Long running time in R (if we want to apply integer function) 2. Requires solving an integer linear program.

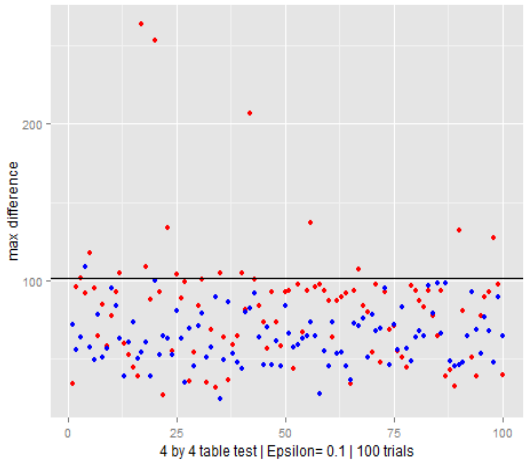
< Table 2 >

Results:

Red dots: MLE rounding

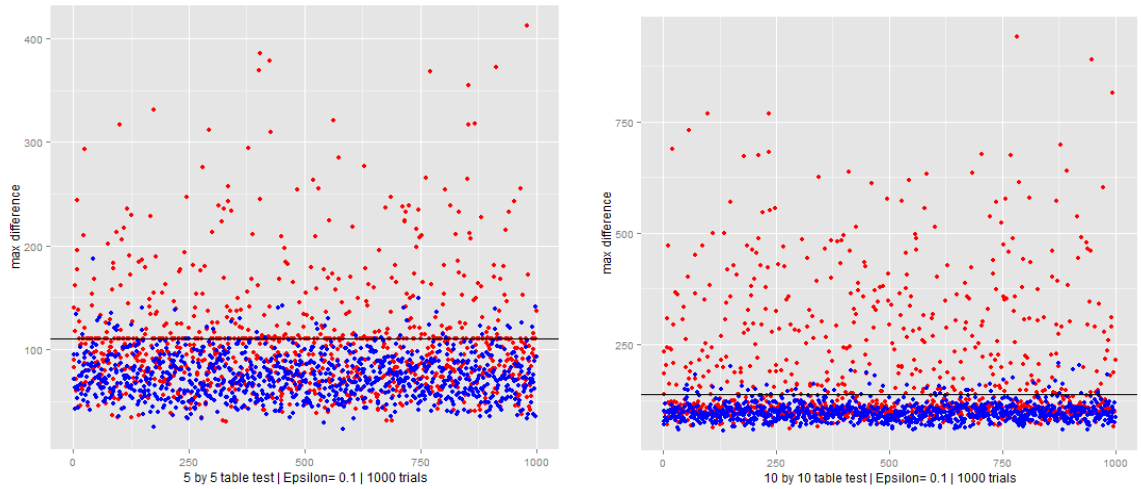
Blue dots: Naïve rounding





< Figure 3 >

These above three plots in the Figure 3 are when we wanted integer values in a table in the rounding function, so their running times are very long. I could run a 4 by 4 table, but our computers could not produce 5 by 5 table simulations even after 8 hours of computation. Based on the 100 simulations for each rounding process, it looks like there is no difference in the performance between MLE and naïve rounding process.

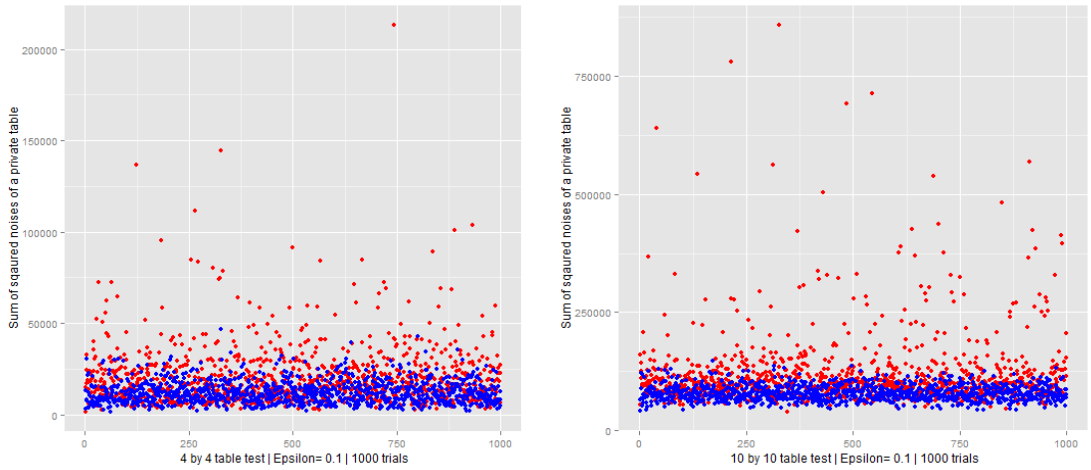


< Figure 4 >

The above plots in the Figure 4 are the comparison between Naïve and MLE rounding in terms of the maximum noise that we add across all the cells in the table. In these plots, we did not find an integer solution in MLE rounding procedure and it significantly decreases the running time of the rounding function. However, in the first plot, the difference between the actual contingency table and the one where we used the MLE rounding is much larger than we expected. It means that in more than a tenth of our simulations, we are seeing this difference bigger than our theory bound we had prior to any rounding. In this case, about 26 percent of red dots are above the theoretical bound. MLE rounding fails to meet our design's goal. However, blue dots that are produced by naïve rounding works well in terms of the given Beta of 0.1. About 5 percent of blue dots are above the theory bound.

Other than a 2 by 2 table which has small dimensions, we want to check how Naïve and MLE rounding behave in a table which has large dimensions such as a 10 by 10 table. The 10 by 10 table experiments also show unexpected results. About 36 percent of our contingency tables differ from the actual by more than the theoretical bound when we do not find an integer solution, thus relaxing our MLE rounding. Also, Naïve rounding shows that only 0.05 of dots are above the theory bound in case of applying naïve rounding.

Conclusion: In terms of accuracy, MLE rounding does not provide better post-processing that maximum noise of that we add in a table is not smaller than naïve rounding's noise. We have tested accuracy including sum of squared of all noises that we add in a private table. It means that if it is a 4by4 table, I sum up squared of 16 noises of the table. Through this test, we want to see in terms of chi-squared value, which a rounding procedure performs better.



< Figure 5 >

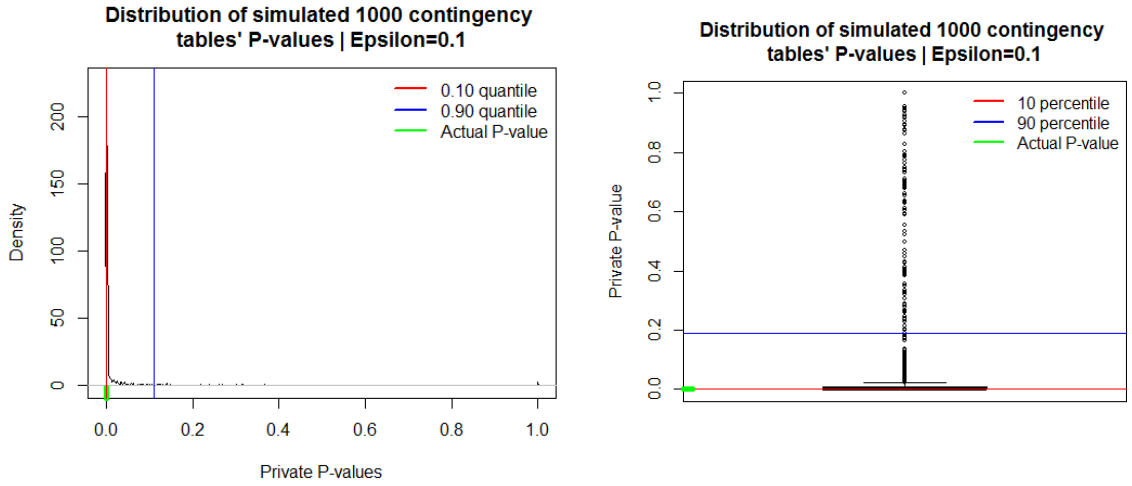
In the two plots of the Figure 5, the sum of squared of noises that we add in a private table after naïve rounding performs better empirically in light of smaller sum of squared noise compared to the MLE rounding.

Accuracy of Chi-squared Statistics with Privacy

Now I will present simulation studies of Chi-squared test of a private table.

Goal: we want to check how P-value of Chi-squared test is different after applying Laplace mechanism and ask if it is similar to the actual table's P-value.

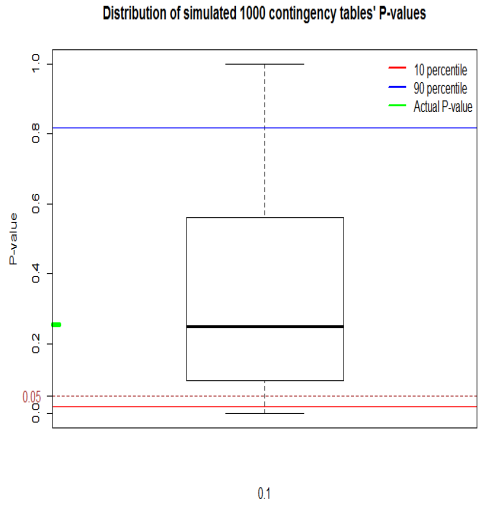
Method: we will empirically check how P-values vary by running 1000 simulations of our noise terms. We will use real data from a social science paper for Chi-squared test (War 5). We will fix the table and apply Laplace mechanism 1000 times and check the P-values.



< Figure 6 >

	Non-democracy	Democracy
No Civil war (years)	3002	1612
Civil war (years)	215	63
*Cold War Civil war years and Regime type Democracy and civil war: A note on the democratic peace proposition, Karin and Myers, 2008		

Conclusion: The Figure 6 has a density graph and a boxplot. Because simulated private P-values are highly concentrated, it is hard to read the boxplot, so I present density plot as well. This is the simulation results of Chi-squared test's P-values after applying Laplace mechanism in a 2by2 table. The actual table's P-value is very small and private tables' P-values are also very small and most of private P-values from simulations are concentrated around the actual P-value. Because the original table's P-value is extremely small, we decided to test another table which has large P-value.



< Figure 7 >

1457	634
1541	724
*Gender and Plan to go college Cite: Investing in Chile: the role of information about financial aid for higher education, Dinkelman and MartAnez, 2014 ***Data set from Dataverse** **https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26908**	

In this table, the original P-value is about 0.25 much larger than the actual 0.05 cutoff for significance. In this case, the private P-values from simulations spread. Median of private P-values are around the actual P-value, but the simulations show that private P-values are significantly different from the original table's P-value from Chi-squared test. These simulations of two tables cast questions that how does Laplace mechanism affect P-value of a private table. Therefore, we have tested various tables' P-value after applying Laplace mechanism.

The Comparison between the Size of Noise and Sampling Error

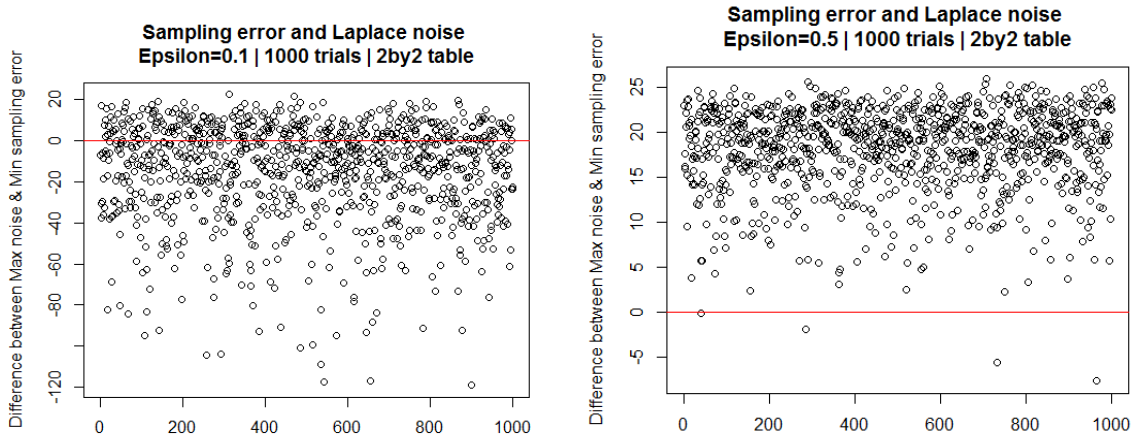
The idea that it could be safe to add Laplace noise for privacy and a private table will still be useful enough to provide meaningful statistical information is come from the concept of sampling error. Unless the study is about a specific small group, most of social science studies use a sample of a population. If a sample is suitable to a study's purposes that a sample is randomly selected or specifically designed depending on the characteristics of research, the sample should well represent the population. It means that even if we do not know the whole population's data, the sample should have very similar statistical information of the population. However, there are sampling errors. The sampling errors in a contingency table exist, but still the statistical analysis of a contingency table is still valid and remains robust to even these sampling errors

From this acknowledgment of sampling errors, we thought that it is safe to use differentially private mechanism if the noise that we add is smaller than sampling errors. It means that a private table's cells are inside of sampling errors, statistical analysis of a private table would be same as the original table because statistical analysis without privacy already has some room of sampling errors, but the statistical analysis is valid.

Goal: We want to check whether it is safe to use Laplace mechanism because the sampling error of a contingency table have already acknowledged for in traditional statistical testing. We also want to check what Epsilon will provide smaller Laplace noise than the sampling error of a cell.

Method: We run simulations of with different values of epilson for our privacy parameter. To test this idea, we need to subtract Maximum Laplace noise from Minimum Sampling error and the value should be positive because if it is negative, then the noise that we add is larger than the sampling error. In that case, either noise that we add is too large or Laplace mechanism could be not useful as a privacy tool for a contingency table.

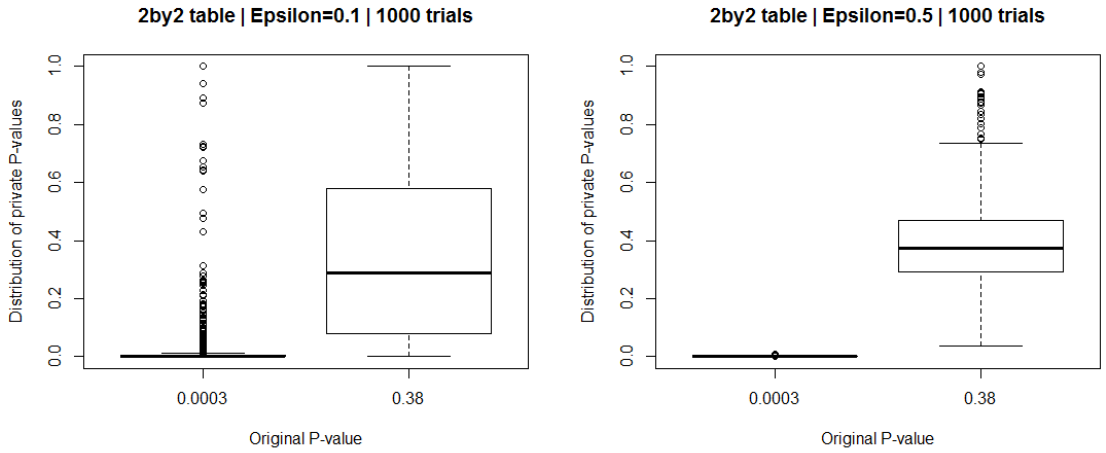
Conclusion:



< Figure 8 >

The left plot in the Figure 8 which has Epsilon of 0.1 shows that if we use Epsilon 0.1, the noise that we add is larger than the sampling error in most cases of simulations. If a subtraction from maximum noise to minimum sampling error is negative, it means the noise that the privacy mechanism adds is too large that statistical analysis from a private table would be incorrect. In other words, the original table's statistical test result will be different from a private table's result. It seems that Epsilon 0.1 could be too large. Therefore, I used Epsilon 0.5 for the simulation experiment. Simulation shows that in many times, the Laplace noises are small enough that the noise will not be larger than the sampling error, so we expect that statistical tests by using a private table after applying privacy mechanism will show similar statistical information.

However, P-value from Chi-squared test of a private table is significantly different from an original table's test result.



< Figure 9 >

The Figure 9's boxplots are simulations' results of private P-values in conditions of Epsilon 0.1 and 0.5. As we expected when Epsilon is large, the variation of private P-values are more concentrated around the original P-value. However, unlike our expectation that when Epsilon is 0.5, because the noise is smaller than the sampling error, statistics from a private table

is similar to the results from the original table, the private P-values vary. With Epsilon 0.5, although the original table's P-value from Chi-squared test is 0.38, more than 50 percent of private P-values are widely spread from about 0 to 0.8. These variations are huge that we cannot trust a private table's Chi-squared test result after applying Laplace mechanism.

The experiment results of checking the difference between sampling error and Laplace noise does not support our hypothesis that Laplace mechanism could provide good statistical test's outcomes if noises from Laplace mechanism is smaller than sampling error. One of the suspects that our experiment is not well-designed to check the hypothesis is the fact that the contingency table in the experiment starts from the expected value of each cell.

In the experiment, I framed a table in these conditions:

N: 2000

N*0.3 = p	N*0.4 = q
N*0.2 = r	N*0.1 = z

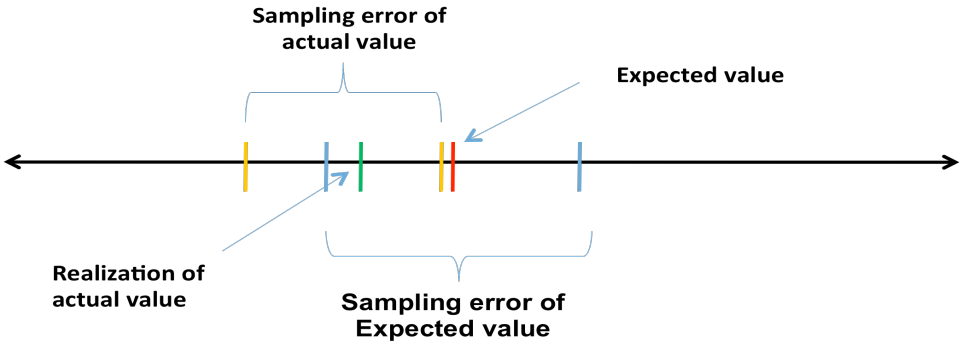
Because of this framework to calculate sampling error of each cell, the fake table has expected values in each cell. However, in the real world, there is very small chance that a table has expected values in all the cells.

These are the sampling errors with 95 percent coverage of each cell where we add or subtract the given amount in the table to the expected cell counts under the Null hypothesis:

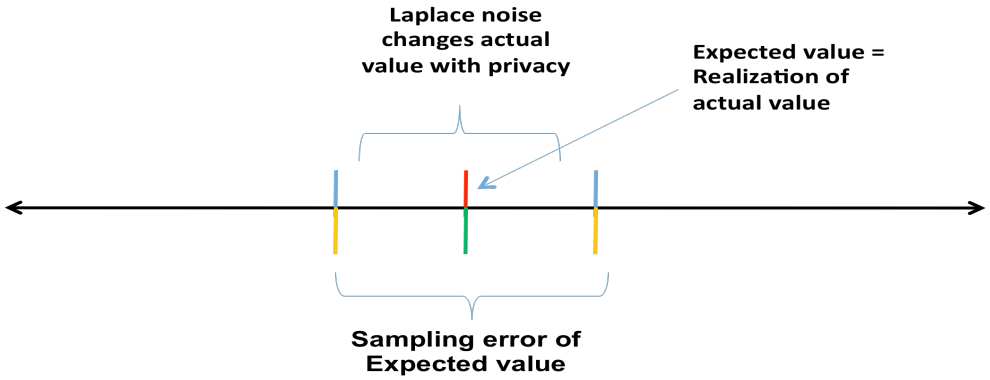
$\pm\sqrt{N * p * (1-p)} * 1.96$	$\pm\sqrt{N * q * (1-q)} * 1.96$
$\pm\sqrt{N * r * (1-r)} * 1.96$	$\pm\sqrt{N * z * (1-z)} * 1.96$

As I mentioned, the counts in each cell are starting from the expected values. In the experiment, we compare the noise from privacy mechanism to the sampling error of the expected value. However, in the real world, realization of each cell count in a real table could not be close to the actual sampling error from that table, and we add noises again with Laplace mechanism.

< This is what real sample will be >



< The experiment assumes that expected value is equal to actual value >



< Figure 10 >

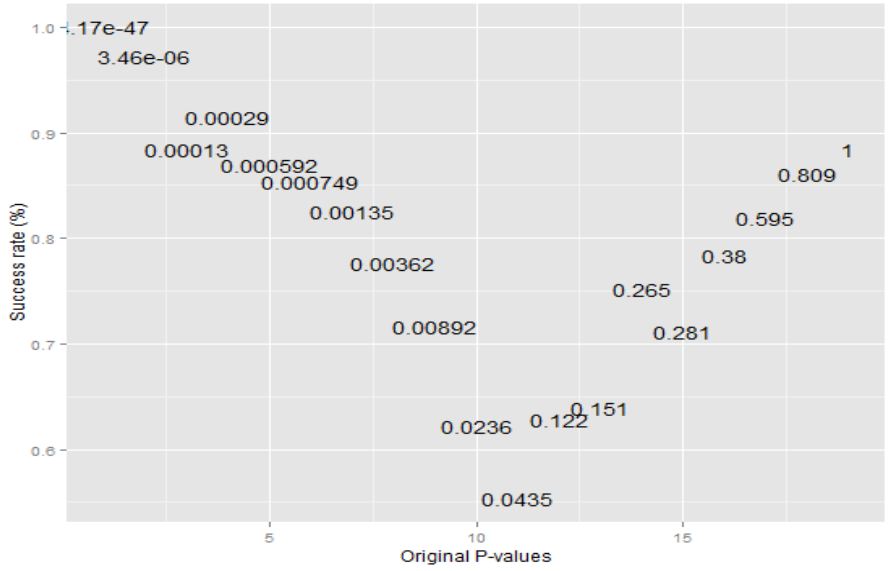
As the Figure 10 describes, my experiment cannot capture the realization of actual values and their sampling error because the realization of actual values is depending on a chance, and we cannot know what will be the realization of a sample. Our experiment assumes that Expected value is equal to the realization of actual value. However, the chance that actual sample has expected values are unrealistic. Therefore, our hypothesis regarding sampling error and Laplace mechanism in a contingency table cannot be well checked by this experiment.

Utility:

The success rate of a private table's Chi-squared test

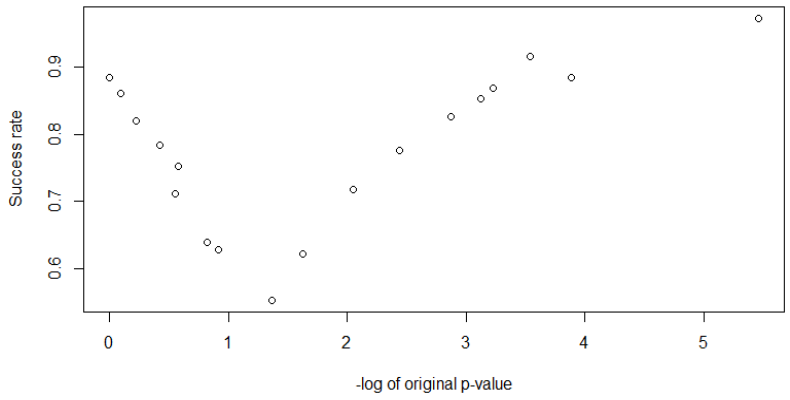
Goal: We want to know whether the Laplace mechanism is useful and whether we can trust a private table's Chi-squared test result. After applying Laplace mechanism, if the result of a private table's Chi-squared test is same as the original table, we could think that the Laplace mechanism is useful in real research.

Method: We ran simulations to check how reliable Laplace mechanism is for independence testing.



< Figure 11 >

Results: The plot in the Figure 11 is about the success rate of Laplace mechanism by testing P-values of different tables' Chi-squared test. The numbers in the plot are actual P-values of original tables. Success rate in the y-axis means that percentage from 1000 simulations of private tables that have the same Chi-squared test results of the original table. We called it "success" because when a private table's Chi-squared test result is same as the original table without privacy, it is the successful result consequentially. Therefore, if the original table's Chi-squared test rejects the null, then private table's Chi-squared test should also reject the null and vice versa. In this test, we fixed total counts (N) to 2000 and change the cells to have different P-values and Epsilon is 0.1.



< Figure 12 >

The plot in the Figure 12 is reorganized by using "–log of original P-value" in the x-axis. We changed the x-axis because we want to see how quickly the success rate drops depending on the original P-values. It seems that when original P-value is small the success rate slowly decrease

until the original P-value hits the threshold of 0.05, and after 0.05, when original P-value increases, the success rate increases rapidly.

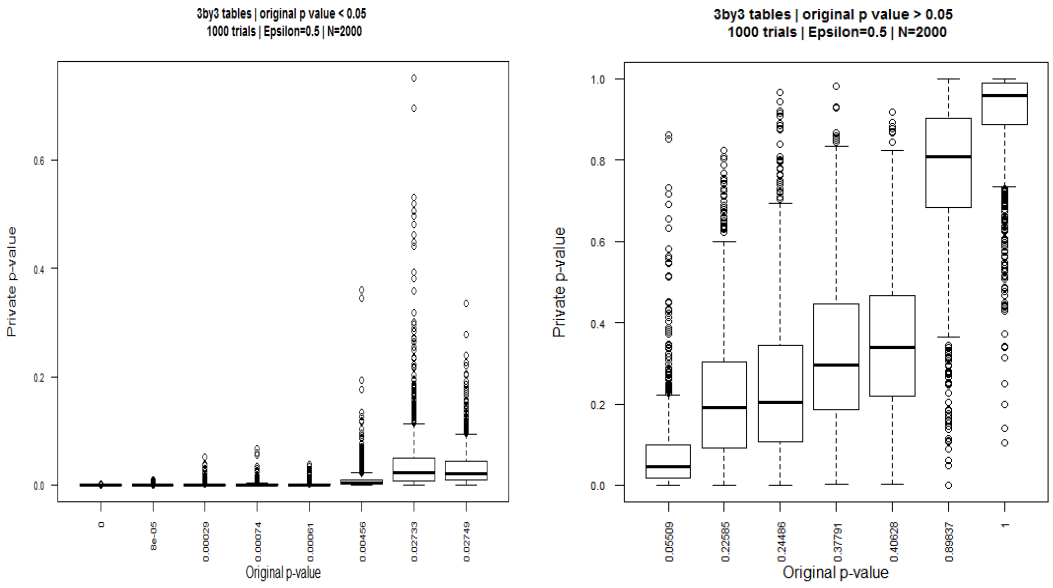
Conclusion: The plot implies several things: 1.When the original P-value is very small, private tables' P-values are also small and the Chi-squared test results are close to the original. 2. When the original table's P-value is around 0.05 which is the threshold to decide to reject or fail to reject the null, the success rate is about 50 percent. 3. When original table's P-value is large, the success rate is lower than the original tables that have low P-values. The third issue brings up some problem of applying Laplace mechanism in a table's Chi-squared test. Even if the original table has P-value of 1 which means even if the Chi-squared statistic is as low as possible, when we apply Laplace mechanism, there are about 15 percent chance that a private table could have P-value lower than 0.05, or a very high Chi-squared statistic.

These empirical findings from the experiment lead us to another experiment to check the utility of Laplace mechanism for Chi-squared test.

The distribution of P-values from Chi-squared test with privacy

Goal: We want to check what will be the P-value of independence testing after applying the Laplace mechanism in a table. Because we want to check more than the success rate of how our private test performs relative to the original conclusion on our dataset, we need to check the P-value distribution from many simulations.

Method: We ran 1000 simulations for several fake tables that have different P-values but fixed N = 2000. The fake table's original P-value will vary from 0 to 1 so that we can check how tables that have different P-value will have different private P-values.



< Figure 13 >

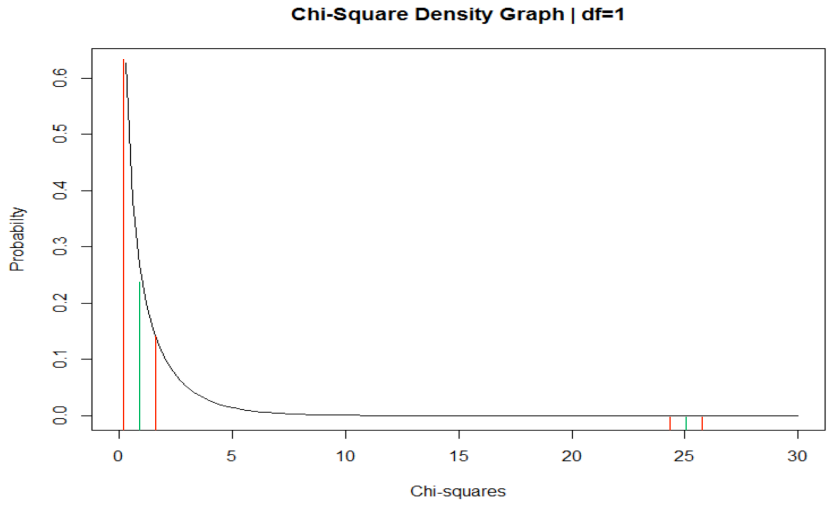
Results: These plots in the Figure 13 present the experiments of 1000 simulations of 15 different 3by3 tables that have different P-values. I have divided the experiments into two plots because their distributions are quite different that we cannot observe the plots well if all 15 boxplots are in one plot. These plot show that if the original table's P-value is small, the private P-values from

Chi-squared test are also small. The plot in the left describes that private P-values are concentrated when the original P-value is small. However, if the original table's P-value from Chi-squared test is large, then the distribution of private P-values are not concentrated.

Conclusion: This experiment shows that Laplace mechanism does not make a private table more dependent. In other words, Laplace mechanism does not generate smaller P-value than the original table's P-value. However, simulation studies present that if an original table's P-value is larger, there is high chances that a private table's P-value varies a lot. The question is whether this empirical finding is because of Laplace mechanism's special characteristics or other factors of Chi-squared test affect this lager variation of P-values observed from simulations.

Nature of Chi-square distribution and P-value

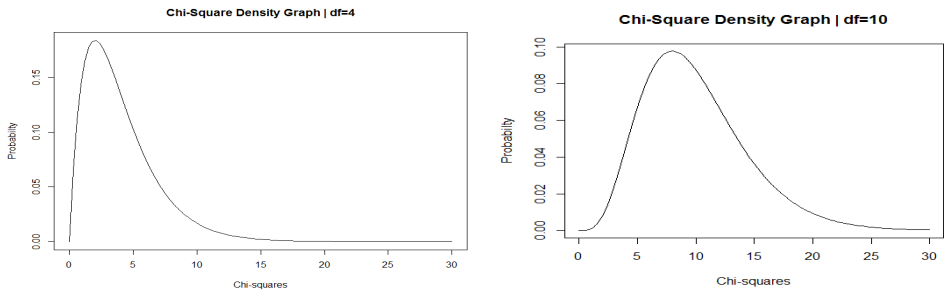
We found that there are larger variations if P-value from the original table's Chi-squared test is large. These large variations of P-value might be related to Chi-squared distribution because we read P-value from the Chi-squared distribution depending on the degree of freedom of a table.



< Figure 14 >

This is Chi-squared distribution with one degree of freedom. The P-value of a statistic is the area of under the density curve beyond the value of the statistic for the given statistic. We can see that if P-value is large, even if we change the Chi-squared value a little bit, the P-value can change a lot because the area around small Chi-squared values are much larger than for large Chi-squared values. It implies that it is not appropriate to check P-value of Chi-squared test to check how Laplace mechanism works for Chi-squared test. It is because the nature of Chi-squared distribution will change P-value a lot if the original Chi-squared value is small. On the other hand, if original Chi-squared value is large, regardless of the size of the noise, there is much higher chance that private P-value will not vary a lot.

Different tables will have different Chi-squared distributions as shown in the Figure 15 and the P-value will vary depending on the distribution. Therefore, checking P-value of Chi-squared test by using Laplace mechanisms as a privacy tool could not provide reliable outcomes.

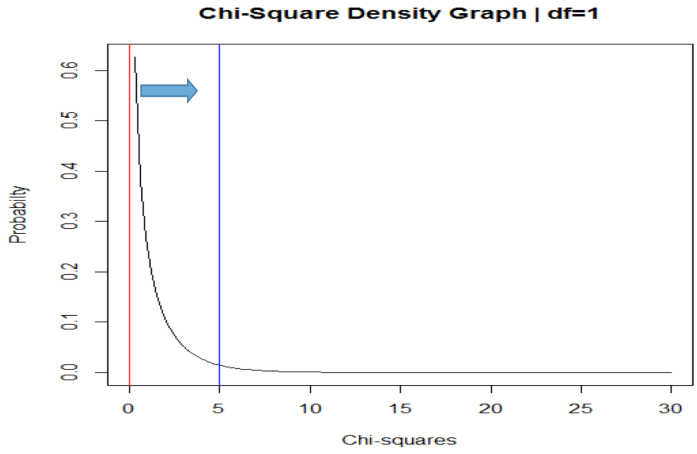


< Figure 15 >

How can variation of private P-values be bound?

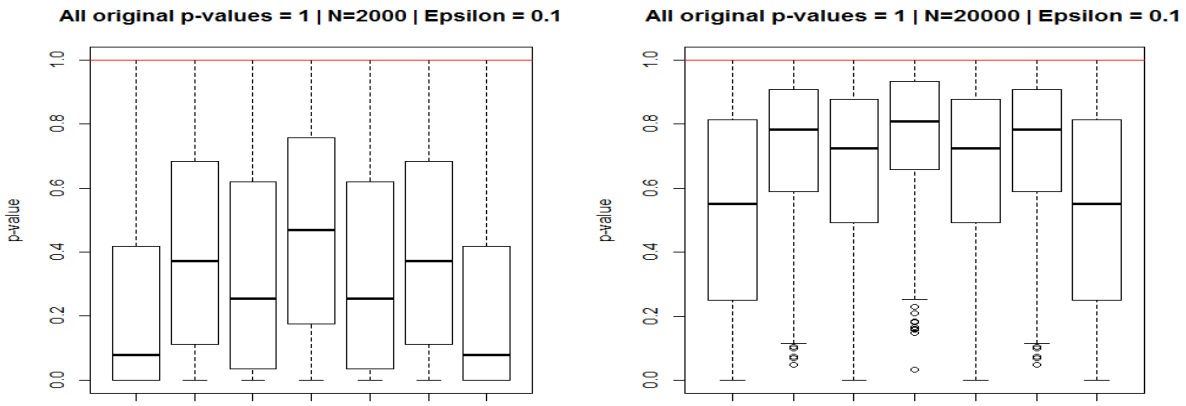
Goal: One of the problems that we have observed is that the variation of private p-values is too large, and it is hard to expect how they vary after applying Laplace noise. Therefore, we want to know whether it is possible to bound the variation of the P-value by increasing total numbers of a table (N).

Method: A possibility to understand the boundary of private p-value variation could be found from the study of original p-value = 1 condition. It is because as The Figure 16 shows, if an original table's p-value is one, after adding noise, there is only the possibility that the private table's p-value will be lower than or equal to the original table's P-value which is 1. What we want to check is how we can bind the variation of private P-values by changing the total number of a table (N). Therefore, I have tested various 2by2 tables' private P-value in condition that the original table's P-value is 1 through simulations. The only difference in these simulations is the total N.

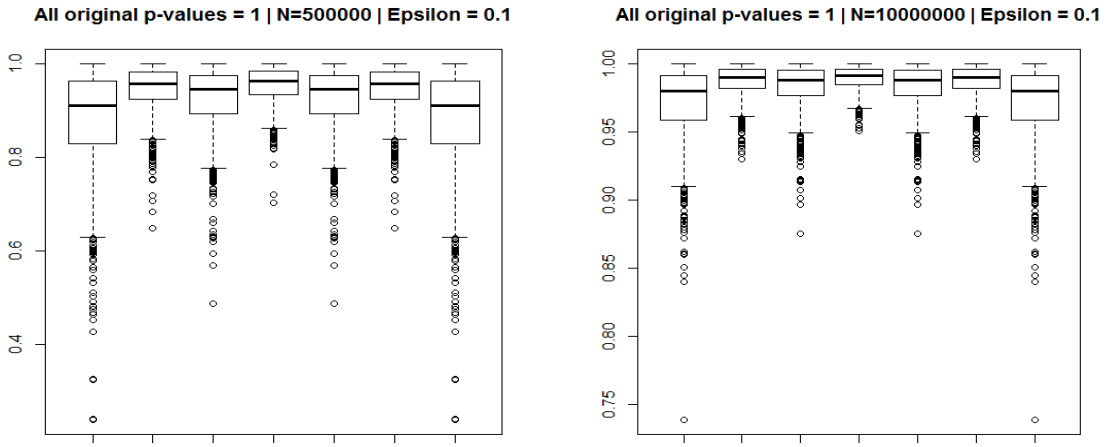


< Figure 16 >

Conclusion: The simulations in the Figure 17 and 18 show that when N is getting larger, the variation of private P-values are more concentrated to around 1 which is the original P-value. We tried to conduct experiments to find a "good" N to bind the variation of private P-vlaues, but the theoretical bound we obtained gives a sample size N that is much too large, so the development of the function that finds a not so large N to bound the variation of private P-value is till in progress.



< Figure 17 >



< Figure 18 >

Table 1		Table 2		Table 3		Table 4	
N*0.49	N*0.21	N*0.64	N*0.16	N*0.81	N*0.09	N*0.36	N*0.24
N*0.21	N*0.09	N*0.16	N*0.04	N*0.09	N*0.01	N*0.24	N*0.16

Table 5		Table 6		Table 7	
N*0.25	N*0.25	N*0.3025	N*0.2475	N*0.5625	N*0.1875
N*0.25	N*0.25	N*0.2475	N*0.2025	N*0.1875	N*0.0625

Chi-Squared value of a private table after Laplace mechanism

From the previous experiments' conclusions, we realize that although the P-value is the important measurement to read from the independence test, because of the nature of Chi-squared

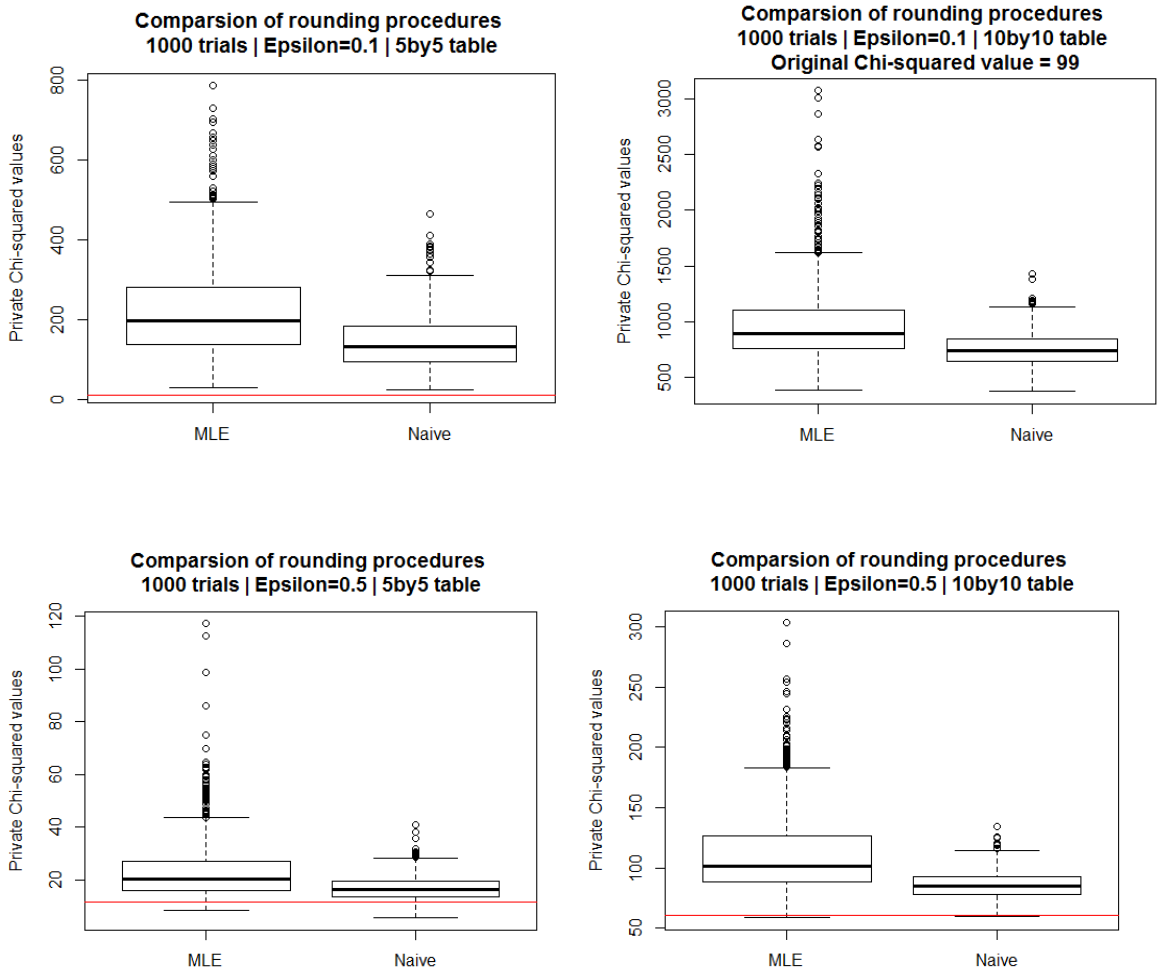
distrubution and P-value, we would rather focus on Chi-squared value itself rather than P-value to see how much the statistic value may change. This could be a better approach to understand Laplace mechanism applied in a cotingency table as well as Chi-sqaure statistics because Chi-squared value is what we get directly from a private contingency table.

Goal: We want to check private table's Chi-squared value not P-value.

Method: We can do simulations to check a private table's Chi-squared value and compare it to the original table's Chi-squared value.

Conclusion: The experiments' results of private Chi-squared value (x^2) are shown in the Figure 19. There is no significant difference in their performances between MLE rounding (without integer function) and naïve rounding. The red line in the left plot is the original Chi-squared value and even Naïve rounding is working better because Chi-squared value after naïve rounding is closer to the actual Chi-squared value. In the right table, Chi-squared value is far away from the original Chi-squared value (80) that it is not even captured in the plot.

The experiments that we have done so far suggest that Laplace mechanism of adding random Laplace noises in each cell of a table is not useful for Chi-squared test.



< Figure 19 >

MWEM

Other than Laplace mechanism, there are many differentially private algorithms for sensitive data. In the summer, we tried MWEM to check how it performed in independence testing of a contingency table. MWEM is a combination of the Multiplicative Weights approach and the Exponential mechanism (Hardt, Ligett, McSherry, 2012).The major difference between Laplace mechanism and MWEM is MWEM creates a synthetic dataset for privacy. Laplace mechanism that we have used adds noises in a given contingency table to keep privacy. However, MWEM generate a synthetic data set and then from this, we can create a contingency table.

To check the usefulness of MWEM for Chi-squared test, we have tried MWEM for independence testing as we did with our above tests using the Laplace mechanism.

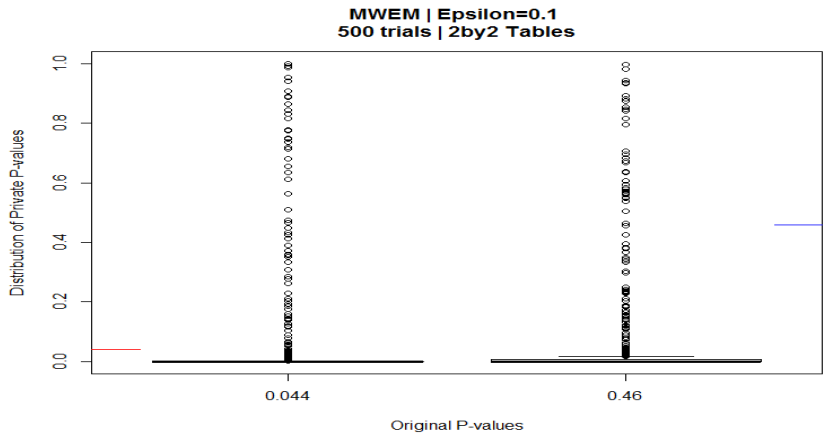
Goal: We want to check the utility of MWEM in Chi-squared test by checking P-values. We want to see if P-values from Chi-squared test on a synthetic data set generated by MWEM is similar to P-values from an actual data set by simulation studies.

Method: At first, we will go over how to execute MWEM in R for Chi-squared test. We need to import a data set in R and clean the data for application of MWEM. MWEM will not recognize missing variables, so we need to drop all data that have a missing variable. Then, we apply MWEM script written in the language Julia by Moritz Hardt. We can call the Julia file in R by using Window Command code in R, so we do not need to go back and forth from R to Julia and R again. When we run Julia file for MWEM, it creates a synthetic data set depending on the given Epsilon for privacy and what variables we want to make a contingency table for. On this synthetic data set, we can ask exponentially many queries (in the size of the dataset)but we must give all the queries in advance to MWEM so that the synthetic database it outputs will perform well on the queries it was given beforehand. For experiments of independence testing, we will run a query to generate a contingency table. In this table with privacy, we can test independence testing. Consequently, the script will save P-values of Chi-squared test.

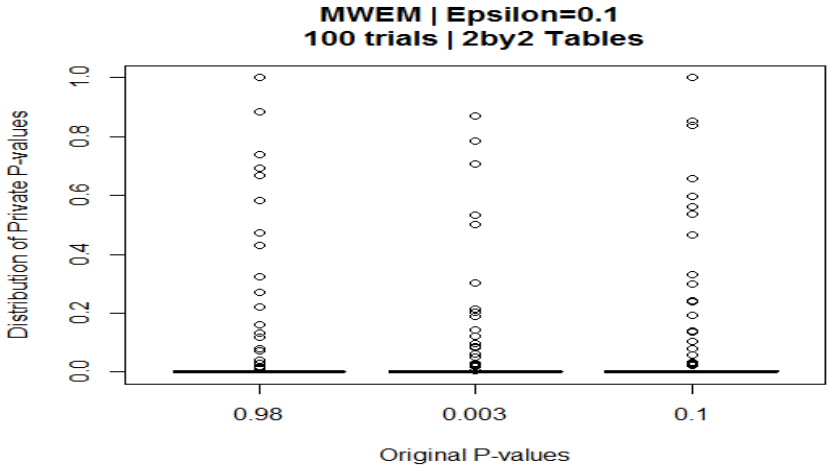
Running MWEM for creating a synthetic data and asking a query for a table takes about 20 seconds. For simulation studies to get empirical results, using MWEM requires much more time. It takes about 8 hours for running 500 simulations.

Result: At first, we have written R script for applying MWEM. R users can implement this privacy algorithm for general purpose by running a script just in R. As I have mentioned above, we use Moritz Hardt's MWEM script in our experiment (Hardt 2015). We found a bug from Moritz Hardt's script and fixed it.

For the experiment, we selected two real data sets from Harvard Dataverse: Citizenship, Involvement, Democracy (CID) Survey Data in 2006 and General Social Survey (GSS) in 2014. We applied MWEM and ran 500 simulations for GSS data set and 100 trials for CID data set. We created 2 tables for GSS data set and 3 tables for CID data set.



< Figure 20 >



< Figure 21 >

The Figure 20 for simulation results of GSS data set shows that MWEM is not effective to use for independence testing. Original tables have P-value of 0.44 and 0.46, but after we apply MWEM, P-value with privacy is significantly different from the original table's P-value from independence testing. The P-values with privacy are very small compared to the original table's actual P-value. Figure 21 also shows similar results from simulations. It seems that regardless of the actual table's P-value, tables from a synthetic data set have very small P-value. The empirical findings from simulation studies of MWEM imply that MWEM is not suitable to independence testing. Particularly, because a table from synthetic data set by using MWEM is highly likely to have very small P-value which means it will reject the null hypothesis regardless of an actual table, MWEM could be not useful in real research by using independence testing.

Goodness of Fit test

Transition from Chi-squared test for independence testing to Goodness of fit test for proportion testing

Laplace mechanism as a privacy tool for independence testing is not suitable to get reliable results of Chi-squared value and P-value. Traditional independence testing used in social science and other disciplines is useful because by testing it, we can learn about variables' independence with α level significance. Ultimately, what we want to accomplish is developing a new test by using Chi-squared statistics with privacy because the experiments that we have observed show that traditional independence testing with privacy is not effective.

We want to see that a new test has a few features:

1. The new test should satisfy differentially private mechanism's goal of keeping privacy and having good utility should be fulfilled.
2. The new test should use Chi-squared statistics that it could be an alternative of traditional Chi-squared statistics.
3. The new test should have P-value that it should be easy to interpret, and non-statistics or computer science researchers can use the test for actual research.
4. The new test should be reliable empirically.

We have developed a new test that preserves differential privacy in Goodness of Fit testing by using Gaussian mechanism and calculates p-values. This new test works effectively with privacy and satisfies aforementioned features. Gaussian mechanism is similar to the Laplace mechanism, but Gaussian noise distribution is easier to manipulate than Laplace noise distribution, so we used the Gaussian mechanism. Goodness of Fit test is different from Chi-squared test, but the Goodness of Fit can be used in various conditions.

Goodness of Fit test without privacy

Goodness of Fit test can test whether a distribution of a variable follows a particular hypothesized distribution. This test uses a Chi-squared statistic and a categorical variable like what we formed in independence testing.

This is the basic framework of Goodness of Fit test:

Null hypothesis (H0): The data follows a multinomial distribution with n trials and probability vector p0. (Observed distribution = Expected distribution)

Alternative hypothesis (H1): The probability vector is not p0. (Observed distribution ≠ Expected distribution)

$$\text{Chi squared statistics} = \sum_{n=1}^{N} \left(\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \right)$$

The formula for Chi-squared statistics is nearly the same as the independence test, but we are given the probability vector in the null hypothesis, so we use that to determine the expected cell counts.

$$\text{Chi squared statistics for Goodness of Fit test} = \sum_{n=1}^{N\ of\ P} \left(\frac{(\text{Observed} - \text{N}\times\text{P}_\text{i})^2}{\text{N}\times\text{P}_\text{i}} \right)$$

Goodness of Fit test has different Expected value calculation compared to Chi-squared test. This is an example of Goodness of Fit test's Expected value calculation.

N=1000

$P_1 = 0.3$ $P_2 = 0.2$ $P_3 = 0.1$ $P_4 = 0.4$

$E_1 = 1000 * 0.3$ $E_2 = 1000 * 0.2$ $E_3 = 1000 * 0.1$ $E_4 = 1000 * 0.4$

Goodness of fit test also has the significance level and 0.05 is used as a standard cutoff for making a decision to reject or fail to reject the null. Goodness of Fit test does not tell us about the independence of variables, but it can test the proportion of values. It could be applied to a contingency table as well as a vector of length d of counts.

New Goodness of Fit test with privacy

Goal: We want to check new Goodness of Fit test with privacy has similar test results as that on actual data.

Method: With various probabilities and sizes of N, we have tested various new Goodness of Fit test with privacy. The experiments are mainly focused on checking significance of the new test.

Result: The new Goodness of Fit test has good utility that new test can find a new threshold for each case and the threshold contains 1-α percent of data with privacy. When we set the significance level as 0.05, a new threshold for Goodness of Fit test should hold 95 percent of data and only 5 percent of data should be over the threshold that the test finds. The test results are very promising. Although the new test empirically finds a threshold by numerical calculation, the test finds new significance with privacy accurately.

The tables are from Ryan Roger's notes to describe the new Goodness of Fit test's experiment. The tables show that the new test's result is almost equivalent to the traditional Goodness of Fit test without privacy. When we set the significance level as 0.05, the new test provides a test result correspondingly. The significance level of the new test is a bit off that it should be 0.95, but there are some cases that the new significance is about 0.94 or 0.96. However, still the test result is very reliable with privacy.

Table 1: Proportion Testing for Multinomial data with different $\mathbf{p^0}$ and $d = 4$

$\mathbf{p}^0$				n	$\chi^2_{d-1,1-\alpha}$	Significance	τ^α_ϵ	New Significance
0.25	0.25	0.25	0.25	100	7.81	0.00	2172.11	0.94
0.25	0.25	0.25	0.25	1,000	7.81	0.018	223.63	0.95
0.25	0.25	0.25	0.25	10,000	7.81	0.39	28.95	0.95
0.25	0.25	0.25	0.25	100,000	7.81	0.90	9.85	0.95
0.1	0.4	0.2	0.3	10,000	7.81	0.30	37.63	0.96
0.1	0.4	0.2	0.3	100,000	7.81	0.86	10.62	0.94
0.05	0.25	0.1	0.6	10,000	7.81	0.19	63.34	0.95
0.05	0.25	0.1	0.6	100,000	7.81	0.80	13.02	0.94
0.01	0.29	0.1	0.6	100,000	7.81	0.57	29.05	0.95

Table 2: Proportion Testing for Multinomial data with different size n with $d = 100$

$\mathbf{p}^0$			n	$\chi^2_{d-1,1-\alpha}$	Significance	τ^α_ϵ	New Significance
0.01	$\cdots$	0.01	10000	123.23	0.00	7216.47	0.96
0.01	$\cdots$	0.01	100000	123.23	0.00	832.46	0.95
0.01	$\cdots$	0.01	1000000	123.23	0.06	194.1151	0.95

Detailed descriptions and experiment' results are explained in Ryan Rogers's Note. This new Goodness of Fit test could be an alternative of Chi-squared test. Because we have observed the necessity of new test with privacy, the new test will meet the needs of new testing.

Further Study

The development and improvement of new Goodness of Fit test with privacy is still in progress. We also want to apply new Goodness of Fit test into the real data set to make sure its application in real studies. Moreover, we need to check integer function applied in MLE rounding and examine MLE rounding more closely. It is because we think that our hypothesis that MLE rounding would perform better than Naïve rounding could be true. If there are some errors in integer function in MLE rounding, there is a chance that our experiment result could be different. Lastly, Dual query as another differentially private mechanism is also in progress of development.

Conclusions

Via various experiments, we have observed how to implement differentially private mechanism in a contingency table and checked statistical tests with privacy. Laplace mechanism can create a contingency table with theoretically tight bound by given level of epsilon. However, testing Chi-squared test in a private table cannot provide useful information that the Chi-squared test result can be significantly different from the actual table's test result. One of the reasons why Laplace mechanism is not suitable to Chi-squared test is because the nature of Chi-squared distribution and following P-value.

Other than Laplace mechanism, we applied MWEM and created a synthetic data set. From the data set, we run as many queries as possible so that we can make a contingency table.

However, a Chi-squared test result on a private data set created by MWEM does not provide useful statistical test result as well. P-values from a private table are significantly smaller than the actual table's P-value from Chi-squared test.

To satisfy the needs of new statistical test using Chi-squared statistics with privacy, we created new Goodness of Fit test with privacy. This new test provides accurate statistical outcomes from a private data. By using this new test, we can find correct significance threshold for Goodness of Fit test with α level significance. Although the technique that we uses for new Goodness of Fit test with privacy only has empirical strength, but the development of analytic and theoretical explanations are in progress.

References

Cynthia Dwork, Frank McSherry,Kobbi Nissim, Adam Smith, Calibrating Noise to Sensitivity in Private Data Analysis, Theory of Cryptography, 2006: 265-284.

Cynthia Dwrok, Aaron Roth, The Algorithmic Foundations of Differential Privacy. Now Publishers Inc, 2014

Latanya Sweeney. Replacing Personally-Identifying Information in Medical Records, the Scrub system. In: Cimino, JJ, ed. Proceedings, Journal of the American Medical Informatics Association (AMIA). Washington, DC: Hanley & Belfus, Inc, 1996:333-337.

Moritz Hradt, Katrina Ligett, Frank McSherry. A Simple and Practical Algorithm for Differentially Private Data Release. Advances in Neural Information Processing Systems. 2012

Moritz Hardt, Private Multiplicative Weights, GitHub repository, https://github.com/mrtzh/PrivateMultiplicativeWeights.jl. 2015