Testing Usability of Differentially Private Estimates

Caper Gooden*

August 15, 2015

1 Introduction

My project this summer has centered on the usability of differentially private estimates. To this end, I have had three main aspects of my project. First, I found various published data sets that would be useful for differentially private analysis and compiled them into one place. Second, I analyzed some univariate statistics to get an initial sense of how usable differential privacy is for social scientists. Finally, I designed a usability study that tests both the differentially private algorithms themselves and the actual user interface.

2 Dataset Curation

2.1 Background

In Summer 2014, several differentially private algorithms had been developed for various types of statistical analysis, such as histograms, means, and quantiles. These algorithms were then tested on synthetic data to make sure they functioned properly. Testing the algorithms on synthetic data was useful, but since the ultimate goal is for differential privacy to be used for many different types of data sets, the algorithms had to be tested on real data. Furthermore, it had to be clear that the analysis that was conducted with differentially private algorithms is the type of analysis that social scientists would actually be interested in. As such, the real data had to come from published papers and we would try to replicate the statistical analysis that had been done in those articles. Over the course of the summer, I looked through over 70 papers in various economics journals and in Dataverse in order to find appropriate data sets for differentially private statistical analysis. Thus far, I have found about 13 that I think would be good for testing, and am still working on adding more papers to the database.

2.2 Findings

Since I am most comfortable with economic papers, I predominantly used two economic academic journals to find data: the *Journal of Labor Economics* and the *Review of Economics and Statistics*. As a broad overview, the median N for the thirteen data sets is approximately 26,000. The smallest was 445 and the largest was 26 million. A spreadsheet containing all of the data sets and their corresponding papers can be found in the Privacy Tools github.

*REU Privacy Tools intern

One of the main papers that I used for analysis this summer was Antecol et.al from the *Journal of Labor Economics* that was published in 2015. This paper has an N of 1938 and analyzes the effect of teacher gender on student achievement in primary school. The data includes 32 variables, 11 of which are binary. Two of the variables are ID variables, and the remaining 19 are numerical. This paper was good for my univariate analysis because it provided summary statistics, but also contains multivariate regression, which will be useful for future analysis of differentially private statistics. Due to the inclusion of categorical data and binary variables, this paper also subsets very easily, making it useful for difference in means testing.

Another article that I used heavily in my work this summer was written by Gicheva, published in the the *Journal of Labor Economics* in 2013. This paper has an N of approximately 35,000 and analyzes the effect of working long hours on hourly wage growth annually for young professionals. The data includes 40 variables, seven of which are binary. Two of the variables are ID variables and the remaining 31 are numerical. Part of the data used for this paper comes from the 1979 National Longitudinal Survey of Youth which provides a lot of information about the individuals of the survey in addition to the wage growth data. This information is useful for differential privacy; since the data contains personal information such as race, marriage status, and income, it becomes increasingly important to protect the individual information. It has been de-identified, so the data itself is widely available. However, it may be interesting to analyze how well the de-identification protects the individual subjects.

The last data set that I used for my analysis was the Lalonde study. The Lalonde study, published in the *American Economic Review*, is an experimental study analyzing the affect of job training on income. There is a treatment group and a control group, making it ideal for difference-in-means testing. Although the actual study was performed on thousands of individuals, the data set that I had access to only had an N of 445. Regardless, this was useful for determining what parameters differentially private difference-of-means testing does not work.

I predict that the other ten papers that I found will be useful for differentially private analysis, but I did not have the opportunity to thoroughly analyze them as I did the Antecol, Gicheva, and Lalonde papers. I will discuss each here briefly. Their attributes can be found in the table below. A more comprehenisve list of attributes can be found on github`https://github.com/IQSS/PrivateZelig/blob/master/summer2015/means/Datasets%20for%20Differential%20Privacy.xlsx`

Author	Journal	N	Number of Variables	Type of Data	Type of Analysis
Bremzen et.al	Review of Economics and Statistics	4800	58	Numerical, Binary	Regression, quantiles
Dinkelman and Martinez A.	Review of Economics and Statistics	6233	37	Numerical, Binary	Regression, summary statistics
Gaule and Piacentini	Review of Economics and Statistics	16,080	14	Numerical, Binary, Categorical	Regression, summary statistics
Burgess and Greaves	Review of Economics and Statistics	17,000	approx 14	Numerical, Binary, Categorical	Regression, summary statistics, quantiles
Angelucci	Review of Economics and Statistics	26,532	39	Numerical, Binary	Regression
Scholz and Sincinski	Review of Economics and Statistics	60,441	126	Numerical, Binary	Regression
Mocan	Review of Economics and Statistics	116,000	112	Numerical, Binary	Regression, summary statistics, quantiles
Fernandes and Paunov	Review of Economics and Statistics	757,102	approx 27	Numerical, Binary	Regression
Bollinger and Hirsch	Review of Economics and Statistics	873,757	120	Numerical	Regression
Doleac and Sanders	Review of Economics and Statistics	26,000,000	65	Numerical, Binary	Regression

2.3 Conclusion

Although I did not have the chance to test differentially private algorithms on all of the data that I found, I anticipate that future researchers will find them useful for their own testing. With such a wide variety of sizes, variables, and analyses, these data sets should facilitate the development of general conclusions about what type of data differential privacy works well on.

3 Univariate Statistic Analysis

3.1 Introduction

One of my primary goals of this summer was to analyze the usability of differentially private statistics for social science research. Originally, I had intended to analyze differentially private univariate and bivariate statistics, but due to complications with the algorithms and code, I was only able to look at univariate statistics. The majority of my work analyzed differentially private means and identified the parameters across multiple data sets for which the algorithms perform the best.

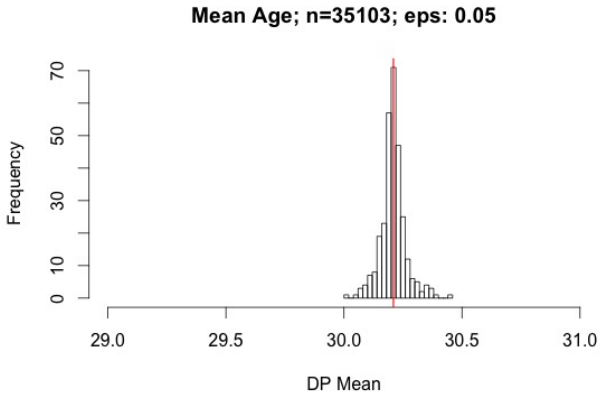
3.2 Differentially Private Mean Analysis

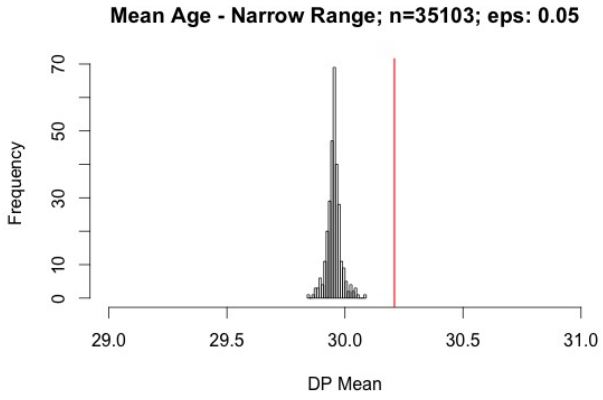
To begin, I identified and visualized the error in differentially private means for different values of epsilon, various truncated ranges, and different sizes of N. In order to accurately visualize the error and extrapolate meaningful conclusions, I ran a code with 300 repetitions of Laplace noise in order to get 300 differentially private estimate. This allowed me to understand what affects the error the most. The most interesting conclusion that I reached was that there is a trade off between variance and bias with differentially private mean estimates that depends on how narrow the truncated mean is. This trade-off can be seen for all levels of N and for all levels of epsilon.

In order to naively calculate a differentially private mean, the range of the variable must be truncated. When this happens, the values outside of the truncated range are recoded to the minimum or maximum value of the range. Part of the bias found in differentially private censored means stems from the truncated range, and if the truncated range is wide enough to encapsulate the true values of the variable, the differentially private censored mean is relatively unbiased.

However, when calculating differentially private means there is a trade-off between bias and variance. With wider ranges, the calculated means are unbiased but have a higher variance. Differentially private means with a narrower range have less variance but are more biased. This trade off can be seen for different levels of epsilon and various values of N.

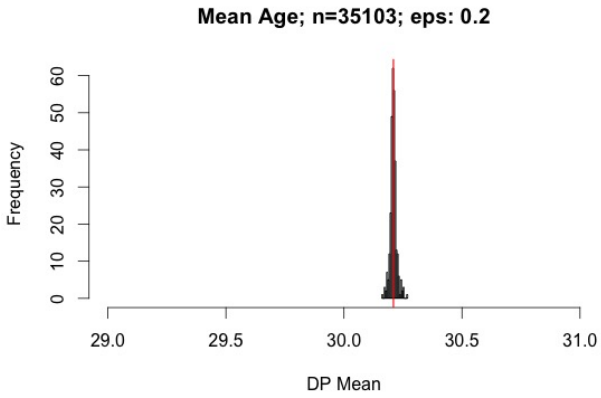
The graphs below show the relationship between bias and variance and how it depends on the range. The first set of graphs come from the Gicheva data with an N of 35,103. These graphs are looking at the variable of age. The real range was (20, 48); the wide range (shown in the figure on the left) was (1, 70) and the narrow range (shown in the figure on the right) was (1, 40).

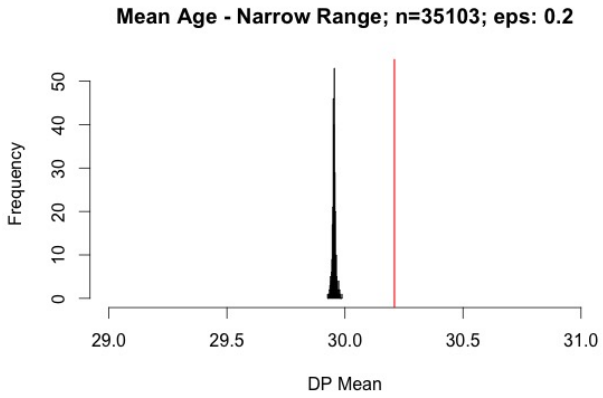




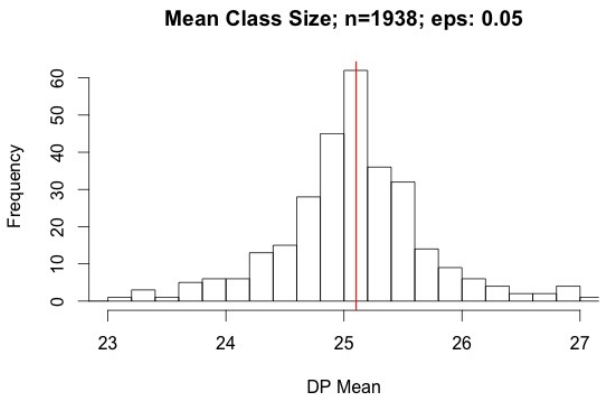
The following graphs are from the same data set and analyze the same variable. With a larger

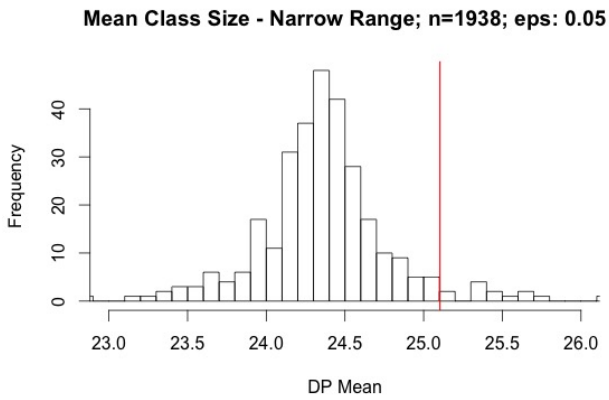
epsilon, the problem becomes significantly more pronounced. The wider range histogram is centered around the true mean, which is marked by the red line, while the narrow histogram has less variance but is clearly biased.



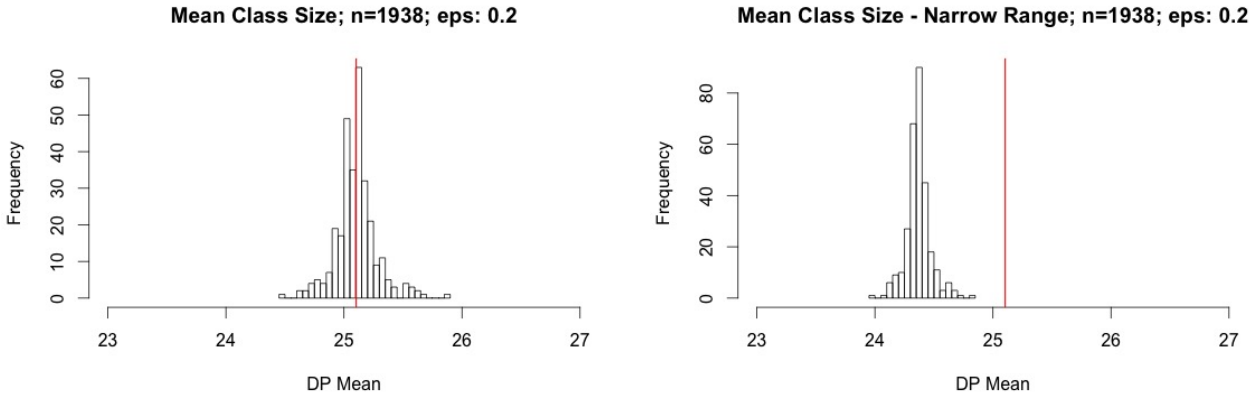


The graphs below use the Antecol, et.al. data set with the variable of class size. The N is 1938. The true range of the class sizes is (1, 37); the wide range was (1, 50) and the narrow range was (1, 30). Similar to the graphs with a larger N, the differentially private mean estimates with a wider range are generally unbiased and have a large variance. Conversely, the estimates with a narrow range have smaller variance but are consistently biased.





The following graphs are from the same data set and analyze the same variable. With a larger epsilon, the problem becomes significantly more pronounced. The wider range histogram is centered around the true mean, which is marked by the red line, while the narrow histogram has less variance but is clearly biased.



While the graphs in this paper only look at an epsilon of 0.05 and 0.2, my work has demonstrated that this trade off can be seen consistently for epsilon values as large as 0.9. This trade off is important to note for researchers using differentially private algorithms to calculate means. If they are unfamiliar with differential privacy, it is especially important to highlight the consequences of the width of the truncated range for mean estimation.

Fortunately, with both data sets, the biased estimated mean is not too far from the true mean. For example, with the age variable from the Gicheva data set, the differentially private mean estimate was consistently less than a year off from the true mean, so it is reasonable to expect that even a biased mean would not significantly affect statistical analysis. Even with the smaller Antecol data set, the biased differentially private estimate never exceeded 2 units away from the true mean. So while the trade off between bias and variance should be acknowledged and communicated to future researchers, it is unlikely to cause researchers to reach severely inaccurate conclusions about a data set.

3.3 Relative Percent Error

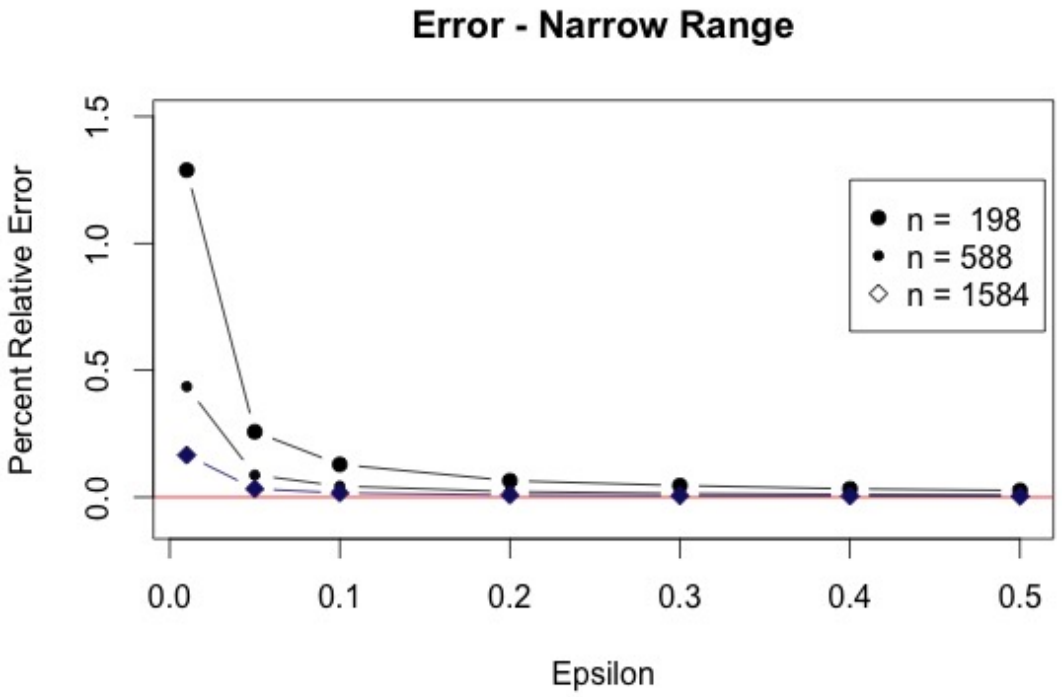
Even though it is helpful to analyze the bias of differentially private means through the distribution of multiple calculations, it is more useful to understand the average percent relative error of the differentially private means and their standard deviations to understand how significantly the differentially private estimates may deviate from the true mean.

I calculated relative percent error the following way:

$$\frac{\tilde{x} - \bar{x}}{\bar{x}}$$

where $\tilde{x}$ is the differentially private mean and $\bar{x}$ is the true mean of the data.

The following graph shows the effect of various N values and epsilon levels on relative percent error.



As shown in the graph, unsurprisingly small N and small epsilon values lead to larger percent relative error. For very small values of N and small epsilons, the relative percent error can be over 100 percent. However, regardless of the N, as epsilon increases, the relative percent error asymptotically goes to zero. This graph analyzes Antecol et.al. data and the maximum N is 1584. With a significantly larger N (e.g. 35,000), the percent relative error decreases drastically and goes to zero at a smaller epsilon.

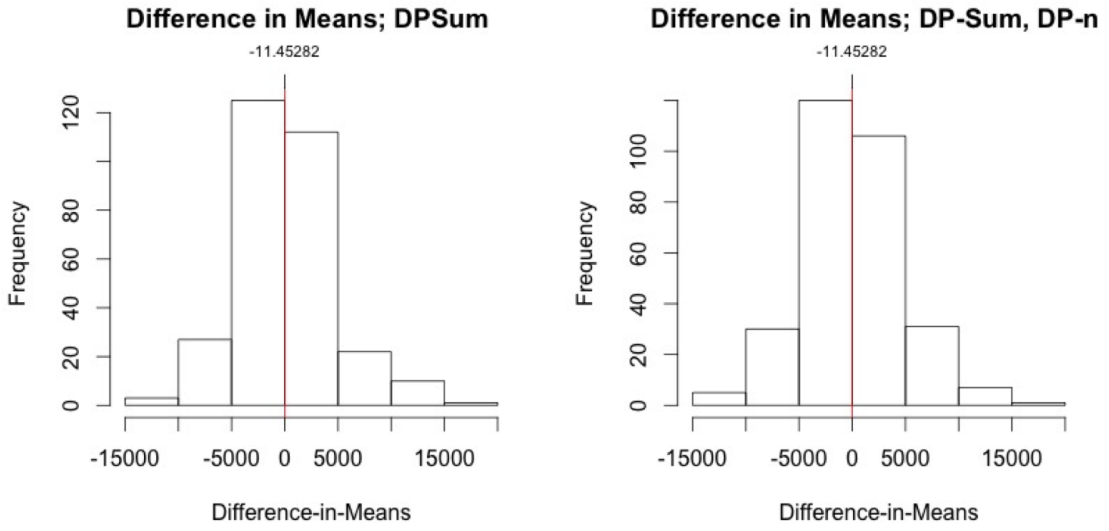
The potential for extremely large relative percent error is important to bear in mind when analyzing difference of means. For example, if the relative percent error is so large so as to obscure a statistically significant difference between two groups, the ability to accurately make inferences is impaired. Moreover, since this is only one statistic, it is unlikely that an epsilon of more than 0.05 would be allocated for its calculation. As such, for determining difference in means for subsets, I recommend using an $N > 2000$. Anything smaller may be too biased to be useful.

3.4 Difference of Means Testing

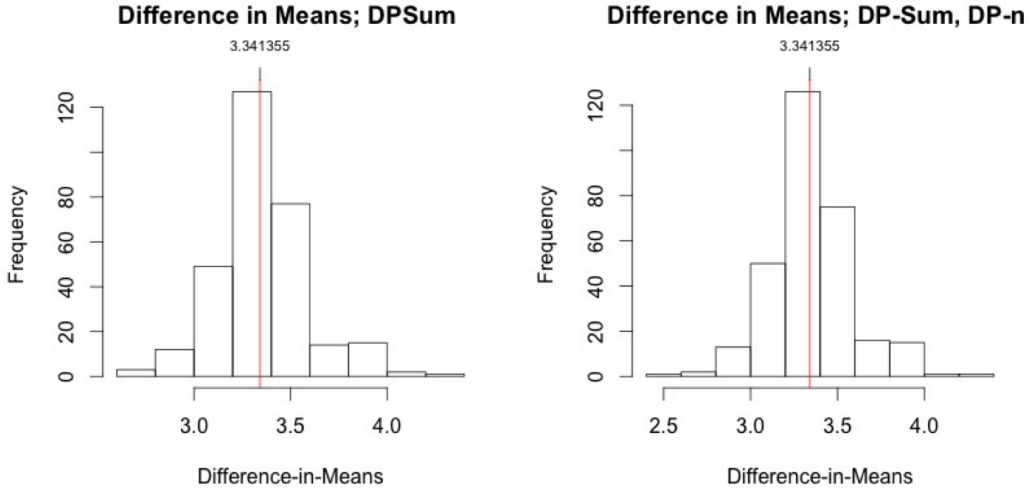
In addition to analyzing the bias and relative percent error of differentially private means, I also did some preliminary testing with difference of means. I looked at differentially-private absolute difference of means between subsets to understand how well the Laplace mechanism works for testing. Usually difference of means testing is a simple form of hypothesis testing, but in order to conduct hypothesis testing the standard errors must be known. Since we are still trying to calculate usable differentially private standard errors, we cannot do traditional difference in means testings. So as a preliminary step, we came up with two different ways to calculate differentially private means to find the absolute difference in means between subsets.

The first method (M1) uses a differentially private sum and the true data N. The second method (M2) uses a differentially private sum and differentially private N.

The following graph comes from the Lalonde study and looks at the difference in annual income between the treatment and control groups. The graph on the left uses the M1 way of calculating means and the graph on the right uses M2. With a global epsilon of 0.1 split evenly between the treatment group and control group and 300 iterations of mean calculation, this is the difference of means distribution. While both graphs appears to be centered around the true difference-in-means, -11, it varies widely, from -15000 to over 15000. However, the N is only 445 and since the variable is income, the range is large: (1, 20000).



As shown in the following graph from the Gicheva article, a large N and smaller range leads to better, more usable results. This graph is looking at the difference in hourly wage for married and unmarried men. The N is 35103 and the range is (0,100). Notably, both M1 and M2 give similar output, though M1 is slightly better.



3.5 Conclusion

With a few notable exceptions, my work with univariate statistics demonstrates that differentially private algorithms can be used on data of an N larger than 1000, even with an epsilon of 0.01. Since a lot of the data that social scientists use are larger than that, it is safe to say that differentially private means are usable for their research for univariate statistics. However, more work needs to be done to ensure that the differentially private algorithms do not obscure statistically significant differences between subsets, as seen with the relative percent error.

4 Drafting a Usability Study

4.1 Introduction

Although the development of differentially private tools is inherently interesting, one main goal of the project is to allow researchers to explore a sensitive data set before formally applying for access. Since a large portion of the researchers will be social scientists, it is vital that the tool facilitates the type of analysis that is standard in their fields. In order to understand how useful differentially private statistical analysis and the interactive query engine are for social scientists, a usability study will need to be conducted. As such, the final piece of my project this summer was drafting the design for the usability study so that it can be implemented in the future. My usability study work can be broken down into four main sections: creating personas, drafting questions for the subjects, determining the use cases for the algorithms, and writing functional requirement documents.

4.2 Personas

One of the first steps in usability study design is specifying your personas, or the target audience of your product. For the query engine and differentially private algorithms, I identified two main personas: the Uploader and the Analyst.

The Uploaders have a private data set and are uploading it to the interface. The Uploaders will then set the privacy budget, or the epsilon, for the entire data set, which will be distributed across various statistics. They can choose whether to split epsilon evenly among all the calculated statistics or to set a specific accuracy value for each statistic. If the Uploader chooses the latter, the corresponding epsilon value will automatically be allocated to the statistic when they select the accuracy level. The Uploader will also decide how much of the privacy budget to allocate for future users. Lastly, the Uploader can calculate some differentially private statistics, but they cannot recover the privacy budget that was spent on those statistics, even if they delete the calculations. Once the Uploader is satisfied with the differentially private statistics they calculated, they can make it available for other users, Analysts, to explore.

The Analysts are researchers who are interested in using sensitive data sets for their own analysis. As previously mentioned, the purpose of the interactive query engine is to give Analysts preliminary access to a data set in a privacy-protecting way so that they can decide whether to apply for full access to the data. Therefore, although Analysts would ideally be able to conduct a wide range of statistical analysis with the query engine, it is not designed to incorporate every statistical test that a researcher would be interested in. Analysts would be able to submit queries for differentially private statistics, generate new variables through transformation, and view the privacy budget for the data set. However, unlike the Uploaders, they would not be able to edit the privacy budget

of the data set and they would not be able to delete previously calculated differentially private statistics. Most importantly, the Analysts must understand that differentially private statistics are not precisely correct; they have some noise added to them.

4.3 Usability Questions

In order to ascertain the utility of the tool and better understand user's comfort with differential privacy, we need to ask them a series of questions both during and after testing. As part of my project, I drafted these questions. A document containing all of the questions can be found in the appendix.

A key component of the interactive query engine is that *users do not need to understand the math and computer science behind differential privacy.* As such, the first category of questions seeks to understand the extent to which users understand differential privacy and how it protects the sensitive data in the data set. If crafted correctly, any researcher should be able to use the interface easily without having to comprehend the exact algorithms protecting the sensitive data. However, in order to ensure that Uploaders feel comfortable with Analysts exploring sensitive data before formally applying, we will need to convey to them what differential privacy does on a basic level. These questions will reveal how well differential privacy was explained to Uploaders and inquire whether users would feel comfortable uploading their own data set given their understanding of differential privacy.

Another important thing we need to know is how clear the parameters of the statistics are for the users. In order to use the tool properly, users need to have some comprehension of the rationale behind setting certain parameters. For example, they should understand the implications of setting a certain privacy budget or truncating the range of value. Uploaders also need to know that some of the parameters can be changed prior to calculating the statistics, such as the granularity for histograms. Since Analysts can also submit queries for differentially private statistics, they will also need to comprehend the implications of setting certain parameters. While understanding the parameters is key to using the query engine properly, it is important to emphasize that neither Uploaders nor Analysts would need to understand the *mechanics* behind how the parameters affect the algorithms and, consequently, the output. The goal is for them to understand well enough to feel comfortable uploading their own sensitive data to the interface.

In addition to understanding the parameters of differentially private algorithms, we need to know how useful the tool is for researchers. Since they are the target audience for this tool, it is important for us to understand when and under what circumstances researchers would use it. For example, we want to know whether they would they find it beneficial to explore data sets with private information before applying for access. It would also be helpful to know how many times have they applied for access to a sensitive data set in the past only to find that it did not contain the information they thought it did. To help with this process, we will also be asking questions regrading some demographic information of the subjects, such as what field they are in and how long they have been working in that field. It has been suggested that users who are less familiar with quantitative research (e.g. sociologists) may be less willing to use a differentially private tool, so we want to know for sure what disciplines will get the most use out of this tool. Moreover, it has also been suggested that older researchers are likely to feel uncomfortable with uploading sensitive data to the interface. To get a sense of what researchers would find this tool most helpful, we hope to gather information on their general demographics and their prior research experiences.

Lastly, the questions will be asking for suggestions on improving the tool. Some of these questions will focus on the usability of the interface itself. For example, we need to know if the instructions are clear and if users feel they can navigate confidently and comfortably through the interface. Since the Javascript interface is still in the preliminary stages of development, specific questions regarding the interface have not yet been drafted. Moreover, we will also be asking for general feedback about how we can improve the tool itself in terms of additional features that the users would find useful. Since our ultimate goal is for this tool to integrate seamlessly into the research process, it is imperative to hear from the target audience exactly what capabilities they would like to see that we are not providing.

4.4 Use Cases and Functional Requirement Documents

Part of usability testing is specifying the use cases that we want the personas to be able to do. A use case is a specific action that a person should, or should not, be able to complete with the tool. All of the use cases have been drafted for the Uploaders, and can be found in the appendix. After developing clear Use Cases, Functional Requirement Documents (FRDs) can be drafted. Functional Requirement Documents outline the Use Cases, workflows, and requirements for each stage of the potential usability test. Not only do FRDs allow other people to implement the usability study, but they also help developers understand what type of interface we are hoping to have. Since the Use Cases have been specified for Uploaders, all of the FRDs are complete for that persona, and can be found in the Appendix.

As for the Analysts, composing Use Cases has been a bit more complicated. Although I generally understand what types of statistical analysis we would like them to be able to do with the tool, many of those methods are still being developed into workable differentially private algorithms. For example, testing is currently underway for the codes for simple and multivariate ordinary least squares regression. Moreover, another intern from this summer, Ally Kaminsky, developed a working command-line query engine that largely covers tasks for the Uploaders. This query engine has made drafting the Use Cases and FRDs for Uploaders significantly easier. Without a good understanding of the differentially private parameters of the statistical methods Analysts would use, developing Use Cases has been difficult. As such, I was unable to draft FRDs for the Analysts. However, my proposed Analyst Use Cases can be found in the Appendix.

4.5 Conclusion

As soon as the user interface is developed, the Uploader part of the usability study will be ready to implement. Testing usability for Analysts remains contingent on the functionality of differentially private algorithms for other statistical tests. Once the algorithms have been successfully tested on real data, the remainder of the usability study can be designed.

5 Conclusion

While I have made substantial progress on all three facets of my project, all of them can be continued. More entries can always be added to my curated repository of data for testing various differentially private algorithms. Further testing can be done, especially with differentially private standard errors and basic hypothesis testing. The Use Cases and Functional Requirement Documents for

Analysts still need to be drafted for the Usability Study. That being said, my work has laid a solid foundation for future work to be built upon. If I could continue working on the project, I would want to continue analyzing difference-of-means and help develop codes for differentially private hypothesis testing because that is a crucial element that needs to be included in the interactive query engine. I had a wonderful experience this summer and I look forward to keeping up with future progress of the Privacy Tools project.

A Appendix A

Persona Descriptions

Uploaders

Behaviors:

- Posts a data set that includes private information with an account
- May not have heard of differential privacy before.

Goals:

- Has a specific file to upload
- Wants to set privacy levels to allow other researchers to use data set.
- Should be able to easily understand implications of setting certain epsilon budget and delta.

Analyst

Behaviors:

- Knows how to perform statistical analysis.
- Most likely has never heard of differential privacy before.
- Understands that output has some noise added to it.

Goals:

- Has a specific dataset to analyze.
- Wants to get generally accurate output from standard statistical analysis (summary statistics, regression, t-tests)
- Should be able to easily perform this analysis without in-depth understanding of differentially private algorithms.

B Appendix B

Usability Study Questions for Uploaders

- What is your career stage (e.g. Assistant Professor, post-doc, research assistant)?
- How many years have you been working in this position?
- What is your discipline? (Multiple choice)
- Did you understand how differential privacy protects your data sets? To what extent?
- What justifications of differential privacy do you find most convincing? (List different possibilities)
- How well was the privacy budget conveyed to you?
- How intuitive was the statistic selection process? (scale)
- Did you understand the implications of calculating the differentially private means/quantiles/histograms (as opposed to computing these statistics without privacy)?
  - If so, what are the implications?
- Did it seem possible to change the parameters of the statistics (e.g. the minimum/maximum values, privacy budget).
  - Was it clear to you how to change the parameters of your statistics? (scale)
  - Did you feel comfortable changing the parameters of your statistics? (scale)
  - Did you understand the implications of changing the parameters of your statistics? (scale) What were the implications?
  - Did you understand the purpose of setting a minimum and maximum value of some of the statistics? What was the purpose?
- Was it made clear that you cannot change the parameters after calculating a statistic?
- Would the statistical information you were able to release be useful for future researchers interested in your dataset (before they have access to the actual data)?
- Is there additional statistical information you would have liked to share, but the tool didn't support it?
- Did the privacy budget prevent you from releasing as much statistical information as you wanted?
- Under what circumstances can you imagine yourself using this tool? Would it integrate smoothly into the research process?

After each task

- Do you understand what you just did? Explain it to me in your own words.

  At end of testing

- How can we improve this tool? (free-form answer)

C Appendix C

Use Cases

Uploader

- Let the uploader decide what statistics to compute
- Let the uploader set the epsilon budget for a data set.
- Let the uploader save some of the privacy budget for future analysts and decide who should be able to use the privacy budget (e.g. registered users of dataverse, people affiliated with a certain organization, etc.). (they may be able to see what is left)
- Let the uploader set the accuracy for the metadata of each statistic.
- Let the uploader set the truncated ranges and granularity.
- Let the uploader edit the aforementioned values before calculating the statistic.
- Let the uploader understand that once the differentially private statistics are calculated, the parameters (epsilon, accuracy, etc) cannot be modified and the statistics cannot be deleted.
- Let the uploader view the calculated differentially private statistics.
- Let the uploader calculate differentially private statistics in batches.
- Do not let the uploader recover the privacy budget if a calculated differentially private statistic is deleted.
- Let the uploader save the metadata to memory.

Analyst *Feasible Currently*

- Let the analyst submit queries for differentially private means, histograms, quantiles, and standard deviations.
- Let analyst generate a new variable through transformations (e.g. log(GDP), age/income).
- Let the analyst select the data set they want to analyze.
- Let the analyst understand that the differentially private statistics are not precisely correct, but that they have noise added to them.
- Let the analyst view the epsilon and delta values for the data set.
- Do not let the analyst edit the privacy budget.
- Let the analyst view a graphical representation of the error bars.

D Appendix D

Functional Requirement Documents

Component: Uploader - Getting Started

Use Case 1: Uploader needs to be able to easily upload the data to the interface for further processing.
User story syntax: "As an uploader, I can upload a dataset so that I can set differentially private parameters to protect sensitive data."

Use Case 2: Uploader needs to understand how differential privacy protects their private datasets and who would have access to the datasets.
User story syntax: "As an uploader, I can understand how differentially private algorithms will protect my dataset so that I can feel comfortable uploading it to Dataverse"

Requirements:

- User should be able to view a clear definition of differential privacy and how it ensures the protection of sensitive data before being asked to upload anything.
- User should be given a quick overview of the upload process before she is directed to the upload link.
- User should be able to easily find additional information on differentially private algorithms if she so desires.
- Regardless of whether the user wants additional information on differentially private algorithms, the user should be able to easily find a link to upload her dataset.
- User should be able to select a data file from her computer to upload.
- User should be able to confirm that the dataset she's uploaded is the correct one. If the dataset is incorrect, the user should be able to delete the uploaded one and select the correct file from her computer.
- User should understand that solely uploading the sensitive dataset does not make it available for analysis for other users yet. She must understand that additional steps are required.
- User should be able to know what are the additional steps required.
- User should be able to easily tell who has access to the dataset they have uploaded.
- User should be able to exit at any time.
- User should be able to go back and complete additional steps after leaving the system.

UI Workflows:

- **Adding file:**
  - Log in→Access informational page about differential privacy (like a homepage)→Click upload data button→Access informational page about the additional steps after uploading→Select file to upload→Click save→Brought to landing page for file uploaded
- **Finding additional info on differential privacy:**
  - Log in→Access informational page about differential privacy (like a homepage)→Click "more information" button→Redirected in a new window to a page with information on differential privacy (recommendation: http://privacytools.seas.harvard.edu/differential-privacy)
- **Logging out and logging back in:**
  - Log in→Access user homepage→Click "log out" button→Access main homepage (Dataverse?)→Click "log in" button→Access user homepage→See uploaded dataset with"incomplete" identifier next to it→Click on the dataset→Continue where the user previously left off.
- **Determining access of file:**
  - Log in→Access user homepage→Click on the desired dataset→Click on the "status" button→Brought to landing page with information about the dataset, including who currently has access to the data
- **Deleting file:**
  - Log in→Access landing page for file uploaded→Option to delete file→Confirmation user wants to delete file→File deleted

Component: Uploader - Setting and Editing Parameters

Use Case 1: Uploader needs to be able to set the privacy budget (or epsilon) for her data set.
User story syntax: "As an uploader, I can set the privacy budget (or epsilon value) for the dataset and understand the relationship between the privacy budget and the accuracy of the calculated statistics."

Use Case 2: Uploader needs to be able to easily set and edit differentially private parameters for her uploaded dataset.
User story syntax: "As an uploader, I can set and edit specific differentially private parameters to protect sensitive data from my uploaded data set."

Use Case 3: Uploader needs to be able to decide what differentially private statistics to compute.
User story syntax: "As an uploader, I can decide what differentially private statistics I want to calculate prior to making the dataset available for analysts to explore"

Requirements:

- User should be able to set the privacy budget (epsilon value) for the entire dataset User should be able to set the percentage of the privacy budget to save for future analysts.
- User should be able to decide who should be able to use the privacy budget for her dataset (e.g. registered users of Dataverse, people affiliated with a certain organization, etc).
- User should be able to decide whether to divide the privacy budget evenly amongst all the statistics OR to each set the accuracy for the metadata of each statistic, which would automatically populate the corresponding epsilon value.
- User should be able to select which variable she wants to calculate a differentially private statistic for.
- When calculating a differentially private mean, user should be able to easily set a minimum and maximum value for the truncated range.
- User should be given a brief explanation of the effects of setting a minimum and maximum value for mean computation. This explanation should not appear on a separate page.
- When calculating a differentially private histogram, user should be able to easily set the granularity.
- User should be given a brief explanation of the effects of setting a particular granularity. This explanation should not appear on a separate page.
- User should be able to easily edit the granularity and truncated range before calculating the differentially private statistic.

UI Workflows:

- **Setting a privacy budget for a dataset that's already uploaded**
  - Log in→Access user homepage→Click on the desired dataset→Brought to landing page for the file uploaded→Click "next step" button→Brought to page asking for privacy budget for dataset with brief explanation of what the privacy budget is→Type in a number between 0.01 and 1→Click "save" →Brought to landing page for privacy budget set
- **Allocating part of the privacy budget for future users:**
  - Log in→Access user homepage→Click on the desired dataset→Brought to landing page for privacy budget set→Click "next step"→Brought to page asking what percentage of the privacy budget to save for future analysts→Type in a number between 1 percent and 99 percent→Click "save"→Brought to landing page for privacy budget shared
- **Deciding what users can use the privacy budget**
  - Log in→Access user homepage→Click on the desired dataset→Brought to landing page for privacy budget shared→Click "next step"→Brought to page asking what users should be able to use the privacy budget with notification that this can be chosen later→Select a group of users from a drop down menu or click "skip"→Click "save"
- **Deciding how to split the privacy budget**
  - Log in→Access user homepage→Click on the desired dataset→Brought to page asking how to split the privacy budget among the calculated statistics with brief explanation of each option→Select "split evenly" or"set accuracy individually" from a drop down menu→Click "confirm"→Shown warning that this action cannot be undone→Click "save"
- **Setting parameters for a differentially private mean**
  - Log in→Access user homepage→Click on the desired dataset→Click on "calculate statistic"→Brought to page for setting parameters for a statistic→Select the desired variable from the first column drop down menu →Select "mean" from the second column drop down menu→Type minimum value for truncated range in the third column→Type maximum value for truncated range in the fourth column→Type accuracy in fifth column (if applicable)→Click "save" at end of row
- **Setting parameters for a differentially private histogram**
  - Log in→Access user homepage→Click on the desired dataset→Click on "calculate statistic"→Brought to page for setting parameters for a statistic→Select the desired variable from the first column drop down menu→Select "histogram" from the second column drop down menu→Type granularity in third column→Type accuracy in fourth column (if applicable)→Click "save" at end of row

Component: Uploader - Calculating Statistics

Use Case 1: Uploader needs to be able to calculate and view differentially private statistics.
User story syntax: "As an uploader, I can calculate differentially private statistics for my uploaded dataset for other users to view, but view them myself before they are published"

Use Case 2: Uploader needs to be able to delete calculated statistics.
User story syntax: "As an uploader, I can delete differentially private statistics if IâĂŹd like, but I **cannot** recover the privacy budget that was spent on calculating them."

Use Case 3: Uploader needs to be able to save the metadata to memory
User story syntax: "As an uploader, I can save the differentially private statistics to the metadata of my dataset for other users to view."

Requirements:

- User should be able to calculate multiple differentially private statistics at one time.
- User should be able to calculate differentially private statistics in several batches.
- User should understand that once the differentially private statistics are calculated the parameter cannot be modified and the privacy budget cannot be recovered.
- User should be able to view the differentially private statistics
- User should be able to delete a differentially private statistics
- User should be able to save the metadata to memory

UI Workflows:

- **Calculating differentially private statistics**
  - Log in→Access user homepage→Click on the desired dataset→Set parameters for the desired statistics→Click on "calculate statistics"→Brought to confirmation page showing how much privacy budget will be left over if statistics are calculated and prompt if want to proceed→Click "proceed"→Brought to warning pop-up stating that once statistics are calculated the privacy budget cannot be recovered in any way→Click "calculate"→Brought to confirmation page showing the differentially private statistics
- **Deleting differentially private statistics**
  - Log in→Access user homepage→Click on the desired dataset→Click on "view statistics"→Brought to page showing all calculated differentially private statistics→Click on "delete statistic" button next to the desired statistic→Pop up message confirming that this will delete the statistic and will not recover the privacy budget used to calculate it→Click "delete"→Brought to confirmation page stating that the statistic was deleted→Redirect to page showing all calculated differentially private statistics
- **Saving differentially private statistics to metadata**

- Log in→Access user homepage→Click on the desired dataset→Click on “publish dataset”→Brought to warning pop-up stating that once the dataset is published, all calculated differentially private statistics will be shown in the metadata, which everyone can view→Click “confirm and publish”→Brought to confirmation page stating that the dataset has been published