STAT 188: Variations, Information and Privacy

Fall 2024

Time: 3:00-5:45 PM EST Wednesdays, with 10 minute break in the middle

Location: Science Center Room 706

Format: lecture style, but questions and discussions encouraged.

Instructor: Xiao-Li Meng (meng@stat.harvard.edu),
Whipple V. N. Jones Professor of Statistics and Founding Editor-in-Chief of Harvard Data Science Review

Teaching Fellow: Kyla Chasalow (kyla_chasalow@g.harvard.edu)

Course Assistant: Warren Zhu (wzhu@college.harvard.edu)

Office Hours: *[Pending]* Mondays Afternoons at 4:30-6:00 in Science Center room 706.
If this time does not work with your schedule, a zoom meeting at another time may sometimes be possible, especially if requested in advance.

Course Overview

This course delves into the intriguing realms of variations, information, and privacy, with a attention to both their qualitative conceptualizations, such as contextual integrity, and their quantitative specifications, exemplified by differential privacy. Our primary goal is to examine these concepts through a foundational statistical lens, and study statistics from the dual perspectives of creating and limiting information from data. At the heart of our exploration is the concept of variations, serving as a unifying theme that intricately links information (revelatory variations) with uncertainty (obfuscatory variations). This nuanced approach enables us to recognize that the principles governing how we restrict the flow of information mirror those involved in generating information (the traditional focus of statistics).

A considerable portion of the course will focus on an in-depth study of differential privacy. First, we will dissect its mathematical framework through theory and examples, identify five key elements that define a general DP specification, and understand what DP guarantees – and what it does not. Second, we will delve into the intricacies of implementing DP via the case of the 2020 U.S. Census and the social and legal perspectives on privacy this raises. Third, we will learn about how to apply missing data methodologies to properly analyze differentially privatized data. Throughout the course, we will confront the challenge of meaningfully defining and quantifying individual privacy and information.

By the conclusion of this course, students are expected to have developed a deeper appreciation for the complex interplay among variations, information, and privacy. They will be equipped with foundational analytical tools and statistical insights, empowering them to navigate the theoretical and practical challenges associated with revealing and concealing information in data for statistical inference and learning.

Course Logistics

Prerequisites

This course is intended as an advanced level undergraduate course for students in statistics, computer science, computational sociology, digital humanities, science and technology studies, and similar data-adjacent fields. Graduate students are also welcome. There is a cap on enrollment to ensure effective classroom discussion. Enrolled students are expected to have foundational and theoretical proficiency at the level of STAT110 and STAT111. Some interest in philosophical and legal thinking is desirable. For students who wish to enroll but do not have the full STAT prerequisites, they should submit (via email) a maximum 400-word explanation for why they are a good fit for the course, or how they will contribute to the course without the STAT prerequisites. A justified motivation to study either statistical foundations, privacy or inter-disciplinary concepts more broadly can be a compensating factor for lacking the full prerequisites (e.g., having only taken one of STAT110 or STAT111).

Assignments

There are three components to your grade:

1. **Homework**: there will be **4 homeworks** across the semester with 2 weeks to complete each homework.
2. **Participation**: active classroom participation and discussion is an important part of the course and asking questions during lecture is welcome and encouraged. The course schedule is semi-flexible. If an interesting idea or connection arises, we will take time to explore it! It is important to attend regularly, and students should email the course TA if they will be absent. We understand that asking questions and sharing ideas in class can be daunting. If you find participating in class a challenge or are for any reason not getting the chance to contribute fully to class discussions, please talk to a member of course staff as soon as possible. Note that attending and asking questions in office hours can also contribute towards participation, though this is neither necessary nor sufficient to get a full participation score.
3. **Midterm**: working in pairs, students will pick a technical paper introducing an extension of differential privacy and provide a presentation teaching it to the class
4. **Final**: there will be a structured final project. More details will be released later in the semester.

Assignment	Assigned	Due	Grade Percentage
Participation	NA	NA	10 %
Homework 1	September 11	September 25	10 %
Homework 2	September 25	October 9	10 %
Midterm	NA	October 16	25 %
Homework 3	October 16	November 6	10 %
Homework 4	November 6	November 20	10 %
Final	NA	December 11	25 %

Information on the midterm and final will be released at least 4 weeks in advance of each due date.

Submission Logistics

All assignments should be submitted via Gradescope
https://www.gradescope.com/courses/794655

Late Policy

Students have **24 cumulative hours** of late time across all written assignments. This will be tracked via Gradescope, and you do not need to notify the TA that you are using your late hours. Additional lateness requires approval and documentation (e.g., medical absence).

Homework Guidelines

Homeworks will generally emphasize mathematical and statistical problems, but interpreting the significance of the question and results is also an important componant. We will also sometimes ask more conceptual questions for which there may not be one right answer. Each homework will be graded out of 100 points, **with partial credit possible for correct ideas and partial work.** Showing your work is crucial and part of the points allocation – correct answers without explanation might not receive full credit. Please also provide any code that you use to produce your answer.

Please note also the following:

- **Collaboration Policy:** We encourage you to think deeply both on your own and in discussions with others. However, you must write up your solutions in your own words. Additionally, you must acknowledge people (e.g., other students you conversed with intellectually on the homework) and/or resources from which you received help – this is a well established norm in scholarly research of any kind. Copying someone else's solution (including ChatGPT), or just making cosmetic changes for the sake of not copying verbatim, is not acceptable, and can be taken as evidence of academic dishonesty.
- **ChatGPT:** With the arrival of ChatGPT, there has been a heated debate about the use of such technological advances in preparing essays and other educational assignments. We believe that you are taking this course to enhance your abilities as a data scientist and statistician or to gain a broader and deeper appreciation of these fields. With that in mind, we trust that you will exercise your best judgment in deciding when and how to take advantage of technologies to facilitate and enhance your learning. Our only requirement is that if you use any technology in a substantial way, you will include a note describing how you used it and in what way it helped your learning. **Making extensive use of ChatGPT or another LLM *without* attributing this is a violation of academic integrity.**
- **Marking:** For some homework problems there are multiple possible correct solutions. In marking these, we are looking for evidence that you have engaged with core issues underlying the problem at hand. Depth and clarity of thought and exposition are important. Please explain any choices you make and summarize their positive and negative consequences.

Lecture and Reading Schedule

The readings listed in each week are applicable to that week's lecture. It is up to you whether it is most helpful for you to read them after or before lecture. Some readings later in the semester are TBA. **We will re-upload the syllabus on Canvas as TBAs are filled in so please check it regularly**.

R = required reading. **O** = optional

	Date	Topic	Material
1	Sept. 4	Introduction, Data Science Life Cycle, Privacy and Variation	1. **O:** *The Data Science Life Cycle* (Wing, 2019)
2	Sept. 11	Contextual Integrity and Privacy in Statistics	1. **R:** *Contextual Integrity Up and Down the Food Chain Nissenbaum (2019).* **O**: For an example application, see Gerdon et al. (2020). 2. **R**: Warner on *randomized response* Warner (1965) 3. **O**: Re-identification attacks Dwork et al. (2017)
3	Sept. 18	Defining Differential Privacy (DP), Creating DP Mechanism	1. **R**: Chapter 1-2 of Dwork and Roth (2013). **O**: Chapter 3 2. **O**: Original papers defining DP Dwork et al. (2006) and Dwork et al. (2016)
4	Sept. 25	The 5 Building Blocks perspective; DP and Swapping **Guest Lect.** James Bailie	1. **R**: A Refreshment Stirred, Not Shaken (I): Five Building Blocks of Differential Privacy 2. **R**: A Refreshment Stirred, Not Shaken (II): Can Swapping Be Differentially Private?
5	Oct. 2	DP and Statistical Disclosure Limitation (SDL) **Guest Lect.** Aleksandra Slavković, James Seeman	1. **R**: Slavković and Seeman (2023)
6	Oct. 9	DP extensions	1. *TBA*

	Date	Topic	Material
7	Oct. 16	Student Presentations (**Midterm**)	See Midterm Assignment Sheet
8	Oct. 23	Case Study: The 2020 Census	1. *TBA*
9	Oct. 30	Case Study: The 2020 Census **Guest Lect.** Philip Leclerc	1. *TBA*
10	Nov. 6	Regression with DP Data, The EM Algorithm	1. *TBA*
11	Nov. 13	Further methods for analyzing DP data **Guest Lect.** Gary King	1. **R:** Evans et al. (2023) 2. **O:** Evans et al. (2022)
12	Nov. 20	Case Study: The impact of DP on the analysis of census data **Guest Lect.** David Van Riper	1. *TBA*
	Nov. 27	NA	No Class - Happy Thanksgiving Break!
13	Dec. 4	Wrap-up: Further Case Studies, Conclusions	1. *TBA*

References

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2016). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality 7*(3).

Dwork, C. and A. Roth (2013). *The Algorithmic Foundations of Differential Privacy*, Volume 9 of *Foundations and Trends® in Theoretical Computer Science*. Hanover, MA: Now Publishers Inc.

Dwork, C., A. Smith, T. Steinke, and J. Ullman (2017). Exposed! a survey of attacks on private data. *Annual Review of Statistics and Its Application 4* (Volume 4, 2017), 61–84.

Evans, G., G. King, M. Schwenzfeier, and A. Thakurta (2023). Statistically valid inferences from privacy-protected data. *American Political Science Review 117*(4), 1275–1290.

Evans, G., G. King, A. D. Smith, A. Thakurta, J. Katz, G. King, E. Rosenblatt, G. Evans, G. King, M. Schwenzfeier, et al. (2022). Differentially private survey research. *American Journal of Political Science 28*, 1–22.

Gerdon, F., H. Nissenbaum, R. L. Bach, F. Kreuter, and S. Zins (2020, May). Individual acceptance of using health data for private and public benefit: Changes during the covid-19 pandemic. *Harvard Data Science Review* (Special Issue 1).

Nissenbaum, H. (2019). Contextual integrity up and down the data food chain. *Theoretical Inquiries in Law 20*(1), 221–256.

Slavković, A. and J. Seeman (2023). Statistical data privacy: A song of privacy and utility. *Annual Review of Statistics and Its Application 10*(1), 189–218.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association 60*(309), 63–69.

Wing, J. M. (2019, 7). The data life cycle. *Harvard Data Science Review 1*(1).