STATS 401 and the future of undergraduate data science

- Data Science is **the use of modern technology for collecting and analyzing data.**
- This sounds a whole lot like **modern applied statistics** but data science is a more fashionable name.
- Either way, a modern course in applied statistics should be hard to distinguish from a course in data science.
- How to teach data science is a major open question, under discussion at the top levels of academia.
- It follows that how to teach applied statistics is worthy of the same attention.
- The new STATS 401 gives the topic the attention it deserves.

A 2018 report by the National Academies of Sciences, Engineering and Medicine on *Data Science for Undergraduates: Opportunities and Options*

- "A key goal is to give all students [with varied backgrounds and levels of preparation] the ability to make good judgments, use tools responsibly and effectively, and ultimately make good decisions using data. The committee defines this collection of abilities as 'data acumen.' To that end, students will need exposure to material from multiple disciplines—notably, mathematical, statistical, and computational foundations—and they will need training in data acquisition, modeling, management and curation, data visualization, workflow and reproducibility, communication and teamwork, domain-specific considerations, and ethical problem solving."
- We will consider how STATS 401 develops these topics, in the context of a second class in statistics.

National Academies report: Mathematical skills needed for data science

- *"Mathematics is essential for data science; however, how much and what types of mathematics are needed vary. Data scientists need to know how to test hypotheses and determine why they do or do not align to real-world problems. They need to be capable of assessing their data science models, determining when these models fail and how to make corrections that lead to scientific discovery. Tools [for us, R] can be utilized and combined to produce an outcome (e.g., simulation or visualization) that reinforces data scientists computational and statistical knowledge without demanding the study of calculus in full detail."*
- STATS 401 follows this approach. Mathematical thinking is required to apply computational and statistical skills creatively and correctly to a new dataset.

National Academies report: key mathematical concepts to engage in data science

- Working with sets and basic logic.
- Multivariate thinking via functions and graphical displays.
- Basic probability theory and randomness.
- Matrices and basic linear algebra.

STATS 401 makes some progress developing each of these skills.

National Academies report: statistical foundations for carrying out data science

- Variability, uncertainty, sampling error, and inference;
- Multivariate thinking;
- Nonsampling error, design, experiments, biases, confounding, and causal inference;
- Exploratory data analysis;
- Statistical modeling and model assessment;
- Simulations and experiments.

All these topice enter STATS 401.

National Academies report: Conclusions

Recommendation. Academic institutions should provide and evolve a range of educational pathways to prepare students for an array of data science roles in the workplace. Key concepts include:

- Mathematical foundations,
- Computational foundations,
- Statistical foundations,
- Data management and curation,
- Data description and visualization,
- Data modeling and assessment,
- Work flow and reproducibility,

All these topics enter STATS 401, some more than others.

How is a data science perspective different from tradional applied statistics?

- Applied statistics has been moving steadily toward the modern data science era.
- Traditionally, much emphasis was placed on learning specific statistical tests, how to carry them out, how to interpret them, how to use them wisely.
- Datasets of growing size and complexity put increasing emphasis on creativity. We end up making a statistical conclusion, but much of the work involves manipulating data and formalizing our questions to bring the two together.
- Modern computation lets us quickly apply many statistical methods. We must still understand what is going on inside the computer so we can guide the computer toward a sensible and correct analysis.

Discussion

Question 1. Should STATS 401 in future follow the data science perspective outlined above?

Question 2. Has this version of STATS 401 developed math/stats/computing topics at a suitable level (challenging but not unreasonable)?

Question 3. Has today's discussion helped to clarify the goals of this version of STATS 401?