BTRY 6020: Introduction

Nayel Bettache

Spring 2025

Presentation

Who am I ?

- Visiting Assistant Professor since August 2024
- Fall 2024: Machine Learning (STSCI 5740) and Intro to R (STSCI 5120)
- PhD in Mathematical Statistics from Institut Polytechnique de Paris (2024)
- M.Sc in. DataScience from Ecole Polytechnique (2020)
- M.Sc in. Statistics from ENSAE Paris (2020)

Nice to meet you all

Welcome to BTRY 6020!

Lecture 1

1. Course logistics
2. Motivating examples
3. Group activity

BTRY 6020

Course goals: The goal of this course is to give you a basic foundation for applying statistical methods and reasoning to your research interests.

BTRY 6020

Course goals: The goal of this course is to give you a basic foundation for applying statistical methods and reasoning to your research interests.

- Focus on methods and applications
- Develop intuition for choosing methods
- Grow awareness of potential pitfalls
- Basic computation and visualization

Logistics

Instructor: Nayel Bettache
TA: Daniel Coulson and Tathagata Sadhukhan

Website: https://nayelbettache.github.io/STSCI6020.html

Course schedule: There are 28 total lectures and each TA will hold 14 labs.

Session	Time	Location	Instructor
Lecture	Mon/Wed 10:10 - 11:25 AM	Baker Lab 335	Bettache
Lab	Mon 2:55-4:10	Mann Library B30A	Coulson
Lab	Tu 1:25-2:40	Mann Library B30A	Sadhukhan
Office Hours	Mon 11:30-12:30	Surge 159	Bettache
Office Hours	TBD	TBD	Sadhukhan
Office Hours	Tues 1pm-2pm	Comstock Hall 1187	Coulson

Grading

- Module assessments: 80%
  - Roughly one every 1-2 weeks
  - Will mostly be walking through a data analysis with short answer and computational exercises
  - Can "consult" with other students in class, but write-up should be your own work
  - You should never have a copy of someone else's write-up. No copy/paste.
  - Lowest grade is dropped

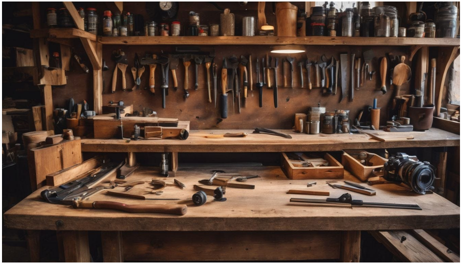
- Final Exam: 20%
  - Take-home test: Available on May 9. Answers to be sent by email on May 16.

Questions?

Why should you care about this class?

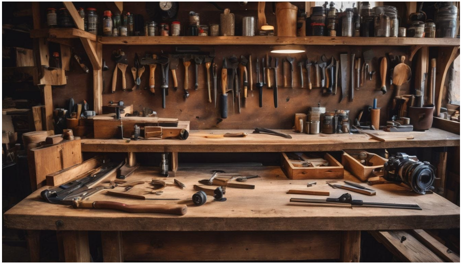
Statistical reasoning

Statistics is concerned with gathering and analyzing numerical data, and the interpretation and communication of subsequent results.



Statistical reasoning

Statistics is concerned with gathering and analyzing numerical data, and the interpretation and communication of subsequent results.



- "Apply statistical methods to your research" vs "Apply statistical reasoning to your research"
- Many different tools and procedures, and usually there is more than one "right answer" (also more than one wrong answer)
- How to think about reasonable procedures given your research problem

Trade-offs

Statistical reasoning involves thinking clearly about the trade-offs in data analysis

Trade-offs

Statistical reasoning involves thinking clearly about the trade-offs in data analysis

- Not enough samples
- Can't record all relevant variables
- Non-response
- May not generalize
- Not from an experiment
- No feasible experiment

Variance vs Bias

Sometimes complex models are better, sometimes simple models are better

Variance vs Bias

Sometimes complex models are better, sometimes simple models are better

(c) Betty Crocker Cake

(d) Stella Parks

Variance vs Bias

Fitting very complex models often fits our *observed* data well, but if the data is noisy, it may not represent the "truth" well

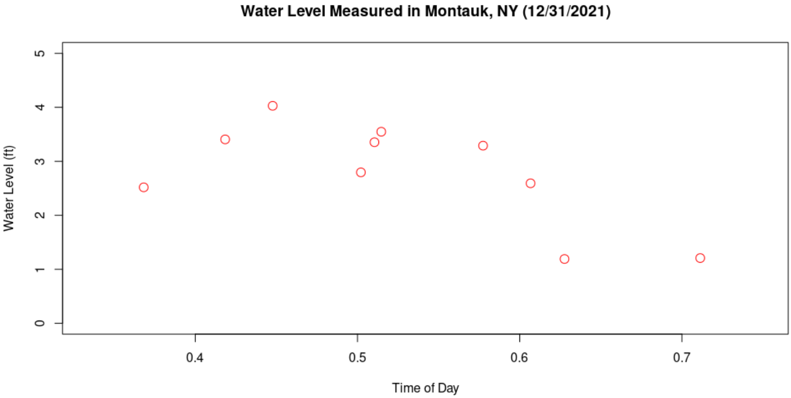


Figure: Water levels observed at Montauk NY on 12/31/2021.

Variance vs Bias

Fitting very complex models often fits our *observed* data well, but if the data is noisy, it may not represent the "truth" well

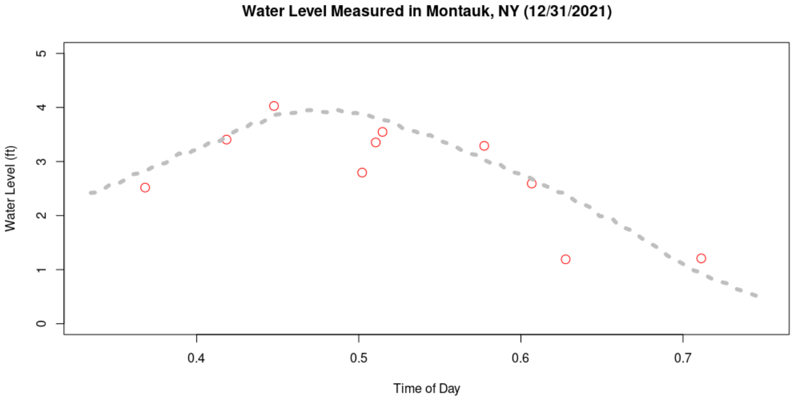


Figure: Water levels observed at Montauk NY on 12/31/2021.

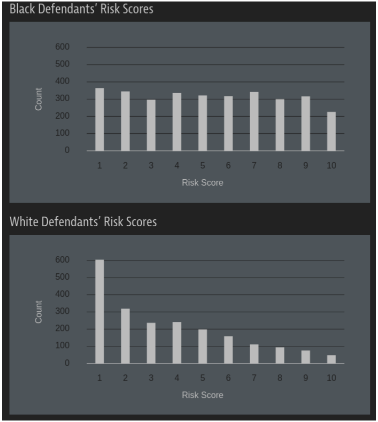
Complexity vs interpretation

Using a model which is more complex can sometimes lead to better fit or predictions. However, this often comes at the cost of being more difficult to interpret

[1] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Complexity vs interpretation

Using a model which is more complex can sometimes lead to better fit or predictions. However, this often comes at the cost of being more difficult to interpret



- In 2016, ProPublica analyzed software used by Broward County, FL to assess a defendant's risk of recidivism[1].
- Predictions were performed by Northpointe's Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm

[1] https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Complexity vs interpretation

Sometimes, simpler models with less predictive power may be preferred because of ease of interpretation

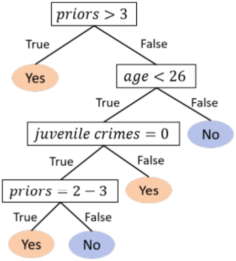


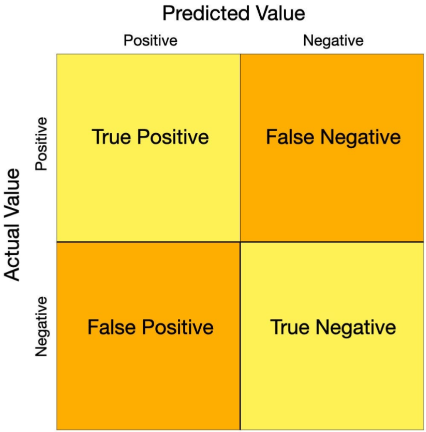
Figure: Potential decision tree from Rudin et al (2021)[2]

- The benefit of interpretability depends on setting
- Predictive power is more important in some context than others
- Choosing the appropriate model complexity is context specific
- Who is the audience? What is the end goal? What are the implications of a poor model?

[1] Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges: https://arxiv.org/pdf/2103.11251.pdf

False Positives vs False Negatives

- False Positives: Incorrect identification of something as positive when it is not.
- False Negatives: Failing to identify something as positive when it is.



Practical Contexts

- **Medical Testing:**
  - False Positive: Diagnosing a healthy person with cancer.
  - False Negative: Missing cancer in a sick patient.
- **Spam Detection:**
  - False Positive: Marking an important email as spam.
  - False Negative: Missing a spam email in the inbox.
- **Security Systems:**
  - False Positive: Flagging a legitimate user as suspicious.
  - False Negative: Missing an actual security breach.

Causal model vs predictive model

Monday, January 24

12:00 am	**10°**	Partly Cloudy	8%	NW 4 mph	
1:00 am	**9°**	Partly Cloudy	8%	NW 3 mph	
2:00 am	**7°**	Partly Cloudy	8%	WNW 2 mph	
3:00 am	**7°**	Partly Cloudy	8%	SW 3 mph	
4:00 am	**8°**	Partly Cloudy	8%	S 3 mph	
5:00 am	**7°**	Partly Cloudy	8%	S 3 mph	
6:00 am	**7°**	Partly Cloudy	8%	SSE 4 mph	
7:00 am	**7°**	Partly Cloudy	8%	SSE 4 mph	

Causal model vs predictive model

- Causal model
  - Aim to **understand cause-and-effect relationships**.
  - Example: Does a new drug reduce blood pressure?
  - Require: - Strong assumptions (e.g., no confounders, correctly specified model). - Tools like randomized controlled trials (RCTs) or causal inference techniques.
  - Focus on **interventions**: What happens if we change X?
  - Example Tools: - Directed Acyclic Graphs (DAGs) - Structural Equation Models (SEMs)
- Predictive model
  - Aim to **forecast outcomes** based on patterns in data.
  - Example: Predict blood pressure levels based on patient data.
  - Focus on **accuracy of prediction**, not on understanding why the relationships exist.
  - Common in: - Machine Learning - Time Series Forecasting
  - Example Tools: - Linear Regression - Random Forests - Neural Networks

Group exercise

Introductions

- Your name
- Department
- Specific area of research, or problem you'd like to apply this class to
- Any broader goals you'd like to get out of class
- What are some data limitations or trade-offs in statistical analysis that you face in your field?
- A fun fact about yourself, or anything else you'd like the others to know
- Potentially get contact info

Wine vs Ratings

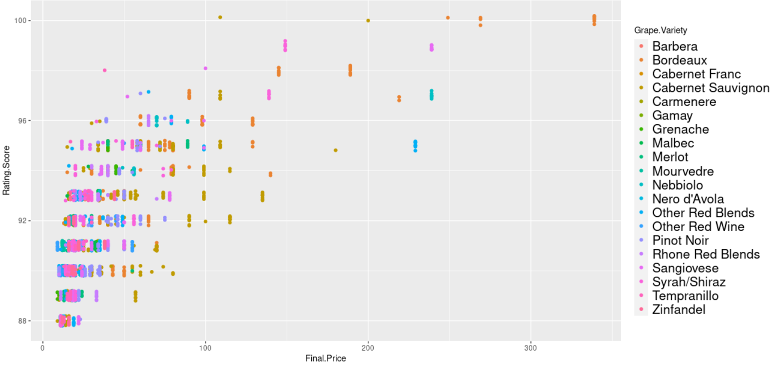


Figure: Data from Wine.com circa 2015

- How would you describe the relationship between wine price and wine rating?
- What are some hypotheses that you might have about the relationship?
- How would you test those hypotheses?
- What should you pay extra attention to as you examine these hypotheses? What additional data could you gather?