MATH 5010 –Foundations of Statistical Theory and Probability

❖ Introduction

Instructor: He Wang
Department of Mathematics
Northeastern University

Probability vs. Statistical Inference

Probability: “What model explains the world?”

- "Probability" postulates a probability model and uses this to predict the behavior of observed data.

Statistics: “Given data, what can we infer about the model?”

- "Statistical inference" uses observed data to infer the probability model (distribution) from which the data was generated.

Probability + Statistics = A framework for uncertainty

Course Structure

Five Modules:

1. Probability Foundations
2. Expectation & Tools
3. Sampling & Limit Theorems
4. Estimation & Hypothesis Testing
5. Interval Estimation & Advanced Inference

Theoretical foundations

Probability → Statistics

Inference machinery

Learning Goals:

- Understand probability theory and random variables.
- Connect theory with real-world data through sampling.
- Apply methods of estimation, hypothesis testing, and intervals.
- Explore advanced inference (ANOVA, Bayesian, Monte Carlo).
- Build solid foundation for future courses.

- **Python Programing:**

This course requires Python, Jupyter Notebook Server and github, all of which are free and open source.

After today's class:

1. Install **Python** from here:

https://www.python.org/about/gettingstarted/

2. Install **Jupyter Notebook** from here: https://jupyter.org/

3. Try the tutorial examples using Jupyter-Notebook in the lab on Canvas.

4. Register a github account. https://github.com/

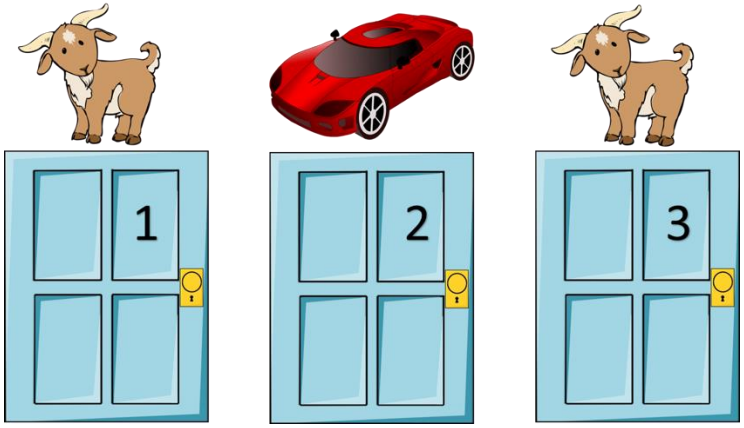
Paradox 1. Monty Hall Problem

In a game show, there are three doors with is a big prize behind only one door.

You choose one of them, then one of the left is opened and there is no prize behind the opened door.

You have a chance to switch your choice.

Will you switch?



(https://en.wikipedia.org/wiki/Monty_Hall_problem)

Paradox 2. Simpson's Paradox

Kidney stone treatment

Which treatment is better?

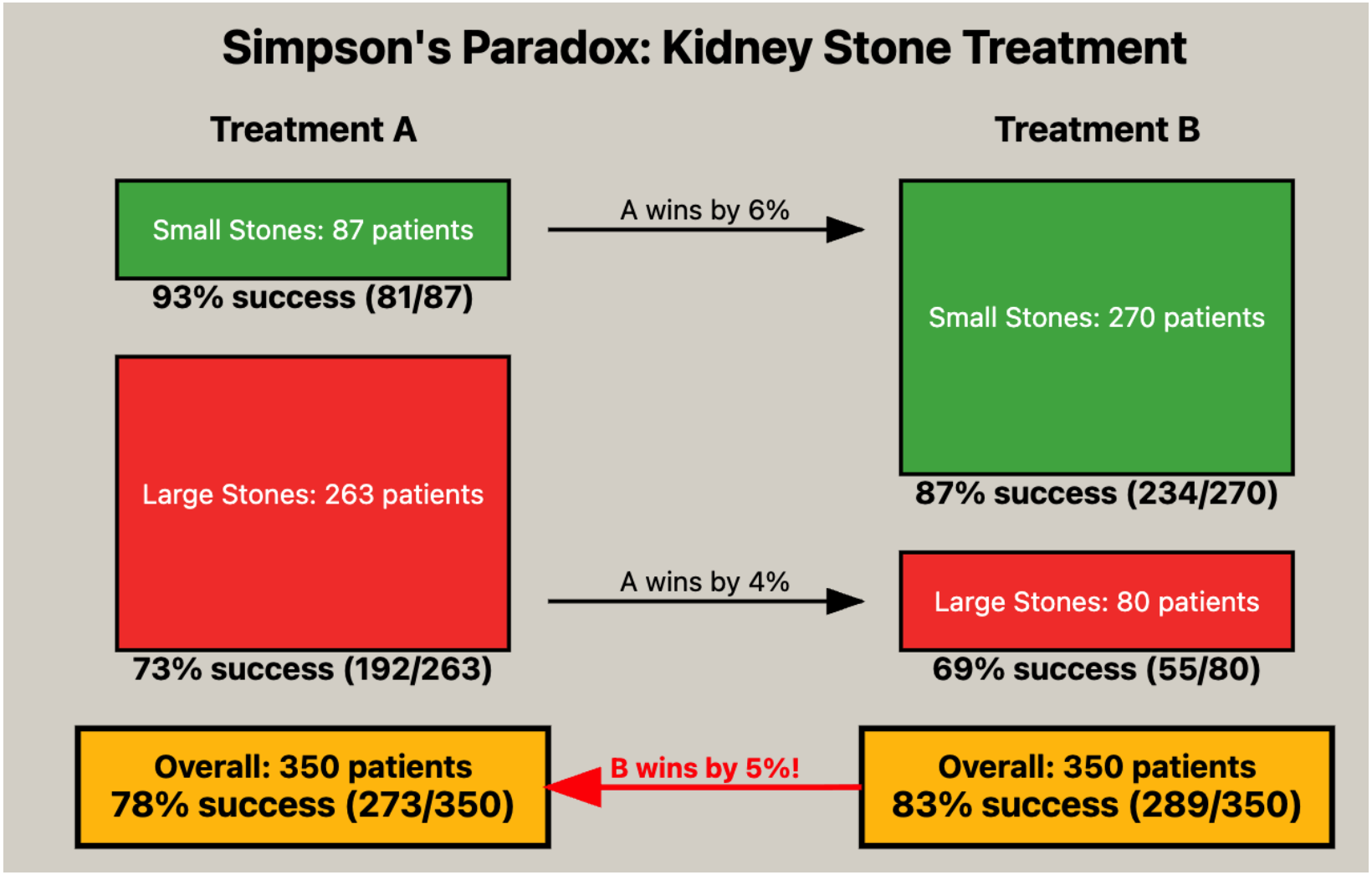
Treatment Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)

Paradox 2. Simpson's Paradox

Kidney stone treatment

Which treatment is better?

Treatment Stone size	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

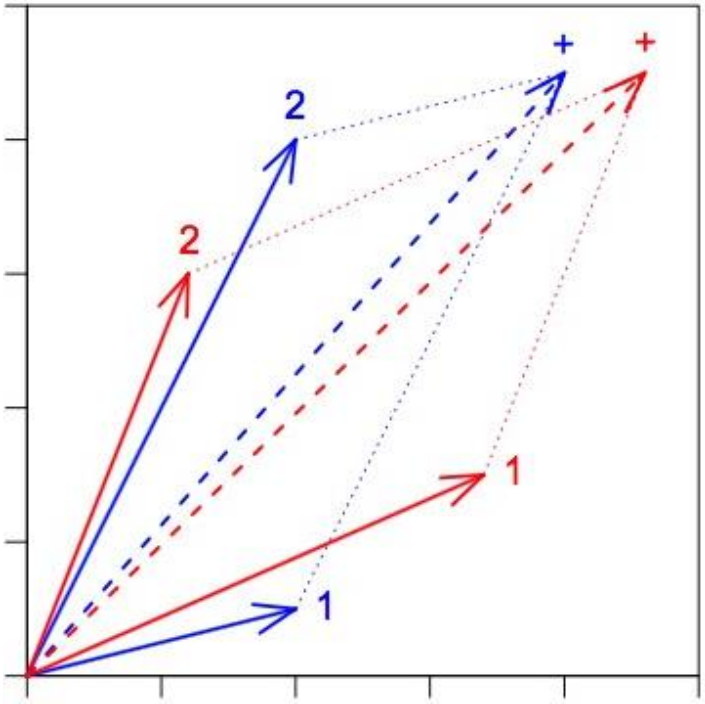


Generated by https://claude.ai/

Vector interpretation

Red: Treatment A

Blue: Treatment B



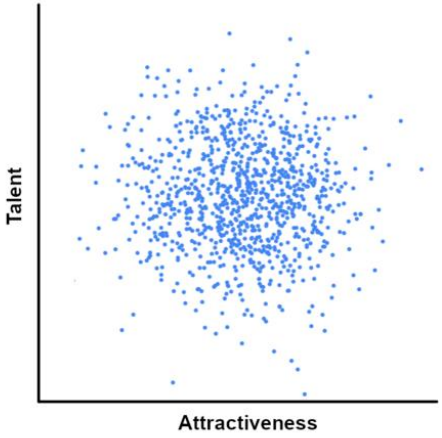
Paradox 2. Simpson's Paradox

Batting averages of players in professional baseball

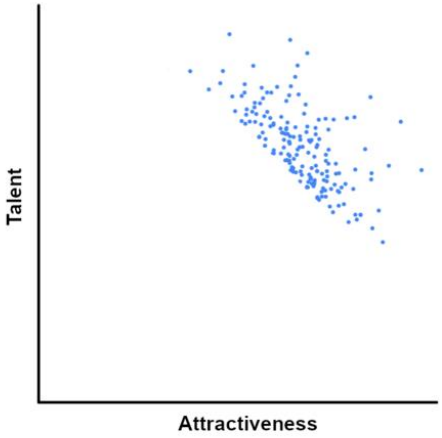
Year / Batter	1995		1996		Combined	
Derek Jeter	12/48	.250	183/582	.314	195/630	**.310**
David Justice	104/411	**.253**	45/140	**.321**	149/551	.270

https://en.wikipedia.org/wiki/Simpson%27s_paradox

Paradox 3. Berkson's paradox



Data show that talent and attractiveness are **uncorrelated** in the population.



The same graph truncated to only include celebrities (where a person must be both talented and attractive, in some combination, to have become a celebrity). Someone sampling this population may wrongly infer that talent is **negatively correlated** with attractiveness.

https://en.wikipedia.org/wiki/Berkson%27s_paradox

https://medium.com/@penguinpress/why-are-handsome-men-such-jerks-f385a46d314f

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- **Book 3. Introduction to Probability**. C.M. Grinstead and J.L. Snell. American Mathematical Society, 2012
- **Book 4. Introduction to Probability Models**, S. Ross, 12th edition (published by Academic Press).

Online books:

- https://www.probabilitycourse.com/
- https://online.stat.psu.edu/stat415/
- https://stat110.hsites.harvard.edu/
- https://bookdown.org/egarpor/inference/