Model Based Inference for Political Science

Justin Grimmer

Professor
Department of Political Science
Stanford University

April 2nd, 2019

Your Training

0) Math Camp
1) Pol 450A: Regression
2) Pol 450B: Design-Based Inference
3) Pol 450C: Model-Based Inference
4) POL 450D: Topics in Quantitative Methods

   ⋮

Why Models?

Design-based inference:

- Randomness: comes from treatment assignment
- Examples: Experiments, Survey sampling, selection on observables
- Reliable, Clear, and Precise

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!

Why Models?

- Causal Inference$\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
  - What % of Latino voters supported Trump?

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
  - What % of Latino voters supported Trump?
  - Where do elected officials fall on ideological spectrum?

Why Models?

- Causal Inference$\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?

Why Models?

- Causal Inference$\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
  - What % of Latino voters supported Trump?
  - Where do elected officials fall on ideological spectrum?
  - What % of time do Women speak in community meetings?
  - When do diplomates discuss credibility concerns?

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference

Why Models?

- Causal Inference$\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference
    - When is war likely?

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference
    - When is war likely?
    - Who will win the election?

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
  - What % of Latino voters supported Trump?
  - Where do elected officials fall on ideological spectrum?
  - What % of time do Women speak in community meetings?
  - When do diplomates discuss credibility concerns?
- Predictive Inference
  - When is war likely?
  - Who will win the election?
  - How long will the regime last?

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference
    - When is war likely?
    - Who will win the election?
    - How long will the regime last?
- Exploratory Inference

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference
    - When is war likely?
    - Who will win the election?
    - How long will the regime last?
- Exploratory Inference
    - What issues do women raise in community meetings?

Why Models?

- Causal Inference$\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference
    - When is war likely?
    - Who will win the election?
    - How long will the regime last?
- Exploratory Inference
    - What issues do women raise in community meetings?
    - What do diplomats say in cables?

Why Models?

- Causal Inference $\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference
    - When is war likely?
    - Who will win the election?
    - How long will the regime last?
- Exploratory Inference
    - What issues do women raise in community meetings?
    - What do diplomats say in cables?
- Combination (heterogeneous treatment effects)

Why Models?

- Causal Inference$\rightsquigarrow$ Just One Goal!
- Descriptive Inference (Measurement)
    - What % of Latino voters supported Trump?
    - Where do elected officials fall on ideological spectrum?
    - What % of time do Women speak in community meetings?
    - When do diplomates discuss credibility concerns?
- Predictive Inference
    - When is war likely?
    - Who will win the election?
    - How long will the regime last?
- Exploratory Inference
    - What issues do women raise in community meetings?
    - What do diplomats say in cables?
- Combination (heterogeneous treatment effects)

All require models

Course Goals

Course Goals

1) Proficient user of likelihood based methods

Course Goals

1) Proficient user of likelihood based methods
2) An introduction to a sophisticated understanding of theoretical tools $\rightsquigarrow$ rest of your career

Course Goals

1) Proficient user of likelihood based methods
2) An introduction to a sophisticated understanding of theoretical tools $\rightsquigarrow$ rest of your career
3) An introduction to numerical optimization and sophisticated estimation procedures

Course Goals

1) Proficient user of likelihood based methods
2) An **introduction** to a sophisticated understanding of theoretical tools $\rightsquigarrow$ rest of your career
3) An introduction to numerical optimization and sophisticated estimation procedures
4) An introduction to basic machine learning methods

Course Goals

1) Proficient user of likelihood based methods
2) An **introduction** to a sophisticated understanding of theoretical tools $\rightsquigarrow$ rest of your career
3) An introduction to numerical optimization and sophisticated estimation procedures
4) An introduction to basic machine learning methods
5) Develop applied data analysis skills

Course Goals

1) Proficient user of likelihood based methods
2) An introduction to a sophisticated understanding of theoretical tools $\rightsquigarrow$ rest of your career
3) An introduction to numerical optimization and sophisticated estimation procedures
4) An introduction to basic machine learning methods
5) Develop applied data analysis skills
6) Continue improving programming skills

How to Accomplish those Goals?

1) Introduce lots of tools $\rightsquigarrow$ common logic unifying them
2) Use proofs + simulation to establish properties of estimators
3) Applied data in both problem sets and a replication assignment
4) Lots of hard work!!

Course Logistics

Lecture: Monday 900am-1050am GSL.

Course Logistics

Lecture: Monday 900am-1050am GSL.
TA: Matt Tyler

Course Logistics

Lecture: Monday 900am-1050am GSL.
TA: Matt Tyler
Attendance Policy:

Course Logistics

Lecture: Monday 900am-1050am GSL.
TA: Matt Tyler
Attendance Policy: Grad School is your job.

Course Logistics

Lecture: Monday 900am-1050am GSL.
TA: Matt Tyler
Attendance Policy: Grad School is your job. Show up to work.

Course Logistics

Lecture: Monday 900am-1050am GSL.
TA: Matt Tyler
Attendance Policy: Grad School is your job. Show up to work. (But you might have reasons to be absent too, which is fine just let me know so we can make sure you don't fall behind!)

Prerequisites:

1) Math Camp
2) POL 450A
3) POL 450B

OR My permission (if you haven't done 1-3, you shouldn't take the class)

Get some books!

Evaluation

Five components to evaluation:

1) Homework (30%): Weekly homework assignments to develop core intuition in class. Use R markdown to submit
2) Midterm Exam (15%): May 14th Closed Book, Pencil + Paper
3) Final Exam (25%): Closed Book, Computer Based
4) Replication Project (25%): (More on next slide)
5) Participation: ask lots of questions!

Replication Project

- Team of two: identify paper, replicate, and extend results

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

- 4/18: Discuss paper select + project with Brad and me

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

- 4/18: Discuss paper select + project with Brad and me
- 5/16: Confirm replication of original results

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

- 4/18: Discuss paper select + project with Brad and me
- 5/16: Confirm replication of original results
- 5/30: 20 page maximum replication and extension due

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

- 4/18: Discuss paper select + project with Brad and me
- 5/16: Confirm replication of original results
- 5/30: 20 page maximum replication and extension due
  - LaTeX document pdf

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

- 4/18: Discuss paper select + project with Brad and me
- 5/16: Confirm replication of original results
- 5/30: 20 page maximum replication and extension due
  - LaTeX document pdf
  - Includes references, figures, tables, etc

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

- 4/18: Discuss paper select + project with Brad and me
- 5/16: Confirm replication of original results
- 5/30: 20 page maximum replication and extension due
  - LaTeXdocument pdf
  - Includes references, figures, tables, etc
  - Replication code

Replication Project

- Team of two: identify paper, replicate, and extend results
- Science only proceeds if we know the strength of our findings. Procedural objections are almost never made in good faith (but you should be polite + nice to people)
- Scholars worked really hard on project, you'll learn a lot from them
- Please talk to Matt and me about project selection $\rightsquigarrow$ must be quantitative data
- Discuss communication with authors with me before writing them
- Focus on substance

Timeline

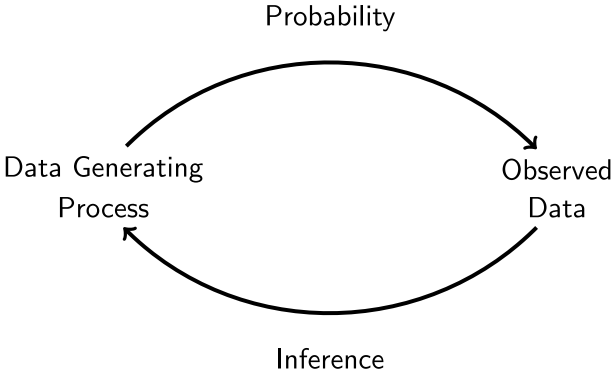
- 4/18: Discuss paper select + project with Brad and me
- 5/16: Confirm replication of original results
- 5/30: 20 page maximum replication and extension due
  - LaTeXdocument pdf
  - Includes references, figures, tables, etc
  - Replication code

Probability

Data Generating Process

Observed Data

Data Generating Process $\xrightarrow{\text{Probability}}$ Observed Data


Probability
Data Generating Process
Observed Data
Inference

Probability Theory $\rightsquigarrow$ Refresher

1) Model of Probability, Axioms of Probability Function
2) Definition of Random Variable
3) Univariate ideas: Expectation, Variance
4) Multivariate ideas: Joint Distribution, Independence

We will review many random variables + properties as we develop data generating processes

Model of Probability

Model of Probability

1) Sample Space: set of all things that can occur

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space
3) Probability: chance of an event

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space
3) Probability: chance of an event
    a) For all events E

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space
3) Probability: chance of an event
    a) For all events E

$$0 \leq P(E) \leq 1$$

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space
3) Probability: chance of an event
   a) For all events E

$$0 \leq P(E) \leq 1$$

   b) P(Sample Space) $= 1$

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space
3) Probability: chance of an event
   a) For all events E

$$0 \leq P(E) \leq 1$$

   b) P(Sample Space) $= 1$
   c) For all sequences of mutually exclusive events (they share no outcomes) $E_1, E_2, \ldots, E_N$ (where $N \to \infty$)

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space
3) Probability: chance of an event
    a) For all events E

$$0 \leq P(E) \leq 1$$

    b) P(Sample Space) $= 1$
    c) For all sequences of mutually exclusive events (they share no outcomes) $E_1, E_2, \ldots, E_N$ (where $N \rightarrow \infty$)

$$P(\cup_{i=1}^{N} E_i) = \sum_{i=1}^{N} P(E_i)$$

Model of Probability

1) Sample Space: set of all things that can occur
2) Events: subsets of the sample space
3) Probability: chance of an event
   a) For all events E

$$0 \leq P(E) \leq 1$$

   b) P(Sample Space) $= 1$
   c) For all sequences of mutually exclusive events (they share no outcomes) $E_1, E_2, \ldots, E_N$ (where $N \rightarrow \infty$)

$$P(\cup_{i=1}^{N} E_i) = \sum_{i=1}^{N} P(E_i)$$

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)
- X's range is the Real line (or subset of it)

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)
- X's range is the Real line (or subset of it)
- If X is defined on outcomes that are discrete, we can write a probability mass function:

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)
- X's range is the Real line (or subset of it)
- If X is defined on outcomes that are discrete, we can write a probability mass function:

$$p(x) \quad = \quad P(X = x)$$

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)
- X's range is the Real line (or subset of it)
- If X is defined on outcomes that are discrete, we can write a probability mass function:

$$p(x) \quad = \quad P(X = x)$$

- If X is defined on outcomes that are continuous, we can write a probability density function, $f(x)$:

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X : \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)
- X's range is the Real line (or subset of it)
- If X is defined on outcomes that are discrete, we can write a probability mass function:

$$p(x) = P(X = x)$$

- If X is defined on outcomes that are continuous, we can write a probability density function, $f(x)$:

$$P(X \in [a, b]) = \int_a^b f(x)dx$$

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)
- X's range is the Real line (or subset of it)
- If X is defined on outcomes that are discrete, we can write a probability mass function:

$$p(x) \quad = \quad P(X = x)$$

- If X is defined on outcomes that are continuous, we can write a probability density function, $f(x)$:

$$P(X \in [a, b]) \quad = \quad \int_a^b f(x) dx$$

For all $[a, b] \in \Re$

Random Variable

A Random Variable X is a function from the sample space to real numbers. In notation:

$$X \quad : \quad \text{Sample Space} \rightarrow \Re$$

- X's domain are all outcomes (Sample Space)
- X's range is the Real line (or subset of it)
- If X is defined on outcomes that are discrete, we can write a probability mass function:

$$p(x) \quad = \quad P(X = x)$$

- If X is defined on outcomes that are continuous, we can write a probability density function, $f(x)$:

$$P(X \in [a, b]) \quad = \quad \int_a^b f(x)dx$$

For all $[a, b] \in \Re$

Expectation

Suppose X is a random variable, then:

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Suppose $g : \Re \rightarrow \Re$, then

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Suppose $g : \Re \rightarrow \Re$, then

- $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$ (discrete)

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Suppose $g : \Re \rightarrow \Re$, then

- $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$ (discrete)
- $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ (continuous)

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Suppose $g : \Re \to \Re$, then

- $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$ (discrete)
- $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ (continuous)

Variance of X is

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Suppose $g : \Re \rightarrow \Re$, then

- $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$ (discrete)
- $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ (continuous)

Variance of X is

$$\text{Var}(X) = E[(X - E[X])^2]$$

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Suppose $g : \Re \rightarrow \Re$, then

- $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$ (discrete)
- $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ (continuous)

Variance of X is

$$\begin{aligned} \text{Var}(X) &= E[(X - E[X])^2] \\ &= E[X^2] - 2E[XE[X]] + E[X]^2 \end{aligned}$$

Expectation

Suppose X is a random variable, then:

- $E[X] = \sum_{x:p(x)>0} xp(x)$ if X is discrete
- $E[X] = \int_{-\infty}^{\infty} xf(x)dx$ is continuous

Suppose $g : \Re \rightarrow \Re$, then

- $E[g(X)] = \sum_{x:p(x)>0} g(x)p(x)$ (discrete)
- $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$ (continuous)

Variance of X is

$$\begin{aligned}
\text{Var}(X) &= E[(X - E[X])^2] \\
&= E[X^2] - 2E[XE[X]] + E[X]^2 \\
&= E[X^2] - E[X]^2
\end{aligned}$$

Random variables $\rightsquigarrow$ Build Model

Participation

Model individual i's voter turnout decision, Y

Participation

Model individual i's voter turnout decision, Y

$$Y = 1, i \text{ turns out to vote}$$

Participation

Model individual i's voter turnout decision, Y

$$\begin{aligned} Y &= 1, \textit{ i} \text{ turns out to vote} \\ Y &= 0, \textit{ i} \text{ does not vote} \end{aligned}$$

Participation

Model individual i's voter turnout decision, Y

$$\begin{aligned} Y &= 1, \ i \text{ turns out to vote} \\ Y &= 0, \ i \text{ does not vote} \\ P(Y=1) &= \pi \end{aligned}$$

Participation

Model individual i's voter turnout decision, Y

$$\begin{aligned} Y &= 1, i \text{ turns out to vote} \\ Y &= 0, i \text{ does not vote} \\ P(Y=1) &= \pi \end{aligned}$$

$Y \sim \text{Bernoulli}(\pi)$

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \quad = \quad \pi^y (1 - \pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \quad = \quad \pi^y (1 - \pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

$$E[Y] \quad = \quad 1 \times P(Y = 1) + 0 \times P(Y = 0)$$

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \quad = \quad \pi^y (1-\pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

$$\begin{aligned} E[Y] &= 1 \times P(Y = 1) + 0 \times P(Y = 0) \\ &= \pi + 0(1-\pi) = \pi \end{aligned}$$

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \quad = \quad \pi^{y}(1-\pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

$$\begin{aligned} E[Y] &= 1 \times P(Y = 1) + 0 \times P(Y = 0) \\ &= \pi + 0(1-\pi) = \pi \\ \text{var}(Y) &= E[Y^2] - E[Y]^2 \end{aligned}$$

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \quad = \quad \pi^y(1-\pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

$$\begin{aligned} E[Y] &= 1 \times P(Y=1) + 0 \times P(Y=0) \\ &= \pi + 0(1-\pi) = \pi \\ \text{var}(Y) &= E[Y^2] - E[Y]^2 \\ E[Y^2] &= 1^2 P(Y=1) + 0^2 P(Y=0) \end{aligned}$$

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \;\; = \;\; \pi^y (1 - \pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

$$\begin{aligned} E[Y] &= 1 \times P(Y = 1) + 0 \times P(Y = 0) \\ &= \pi + 0(1 - \pi) = \pi \\ \text{var}(Y) &= E[Y^2] - E[Y]^2 \\ E[Y^2] &= 1^2 P(Y = 1) + 0^2 P(Y = 0) \\ &= P(Y = 1) = \pi \end{aligned}$$

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \quad = \quad \pi^y (1-\pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

$$\begin{aligned} E[Y] &= 1 \times P(Y=1) + 0 \times P(Y=0) \\ &= \pi + 0(1-\pi) = \pi \\ \text{var}(Y) &= E[Y^2] - E[Y]^2 \\ E[Y^2] &= 1^2 P(Y=1) + 0^2 P(Y=0) \\ &= P(Y=1) = \pi \\ \text{var}(Y) &= \pi - \pi^2 \end{aligned}$$

Participation

Y is a Bernoulli Random Variable
PMF:

$$P(Y = y) = p(y) \quad = \quad \pi^y (1-\pi)^{1-y}$$

for $y \in \{0, 1\}$ (0 otherwise)

$$\begin{aligned}
E[Y] &= 1 \times P(Y=1) + 0 \times P(Y=0) \\
&= \pi + 0(1-\pi) = \pi \\
\text{var}(Y) &= E[Y^2] - E[Y]^2 \\
E[Y^2] &= 1^2 P(Y=1) + 0^2 P(Y=0) \\
&= P(Y=1) = \pi \\
\text{var}(Y) &= \pi - \pi^2 \\
&= \pi(1-\pi)
\end{aligned}$$

Joint Random Variables: Many Turnout Decisions

- Interested in many individual's turnout decisions ($i = 1, \ldots, N$)
- Model joint decisions $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)$

Define:

$$p(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) \quad = \quad p(\boldsymbol{y})$$

This is very complicated $\rightsquigarrow$ Let's simplify

Joint Random Variables: Many Turnout Decisions

Independent Random Variables:
Random variables $Y_1, Y_2, \ldots, Y_N$ are independent if

Joint Random Variables: Many Turnout Decisions

Independent Random Variables:
Random variables $Y_1, Y_2, \ldots, Y_N$ are independent if

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) \quad =$$

Joint Random Variables: Many Turnout Decisions

Independent Random Variables:
Random variables $Y_1, Y_2, \ldots, Y_N$ are independent if

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) &= \\ P(Y_1 = y_1)P(Y_2 = y_2)\ldots P(Y_N = y_N) &= \end{aligned}$$

Joint Random Variables: Many Turnout Decisions

Independent Random Variables:
Random variables $Y_1, Y_2, \ldots, Y_N$ are independent if

$$\begin{aligned}
P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) &= \\
P(Y_1 = y_1)P(Y_2 = y_2)\ldots P(Y_N = y_N) &= \\
\prod_{i=1}^{N} P(Y_i = y_i)
\end{aligned}$$

Joint Random Variables: Many Turnout Decisions

Independent Random Variables:
Random variables $Y_1, Y_2, \ldots, Y_N$ are independent if

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) &= \\ P(Y_1 = y_1)P(Y_2 = y_2)\ldots P(Y_N = y_N) &= \\ \prod_{i=1}^{N} P(Y_i = y_i) \end{aligned}$$

$Y_i \sim \text{Bernoulli}(\pi)$

Joint Random Variables: Many Turnout Decisions

Independent Random Variables:
Random variables $Y_1, Y_2, \ldots, Y_N$ are independent if

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) &= \\ P(Y_1 = y_1)P(Y_2 = y_2)\ldots P(Y_N = y_N) &= \\ \prod_{i=1}^{N} P(Y_i = y_i) & \end{aligned}$$

$Y_i \sim \text{Bernoulli}(\pi)$
Independent, Identically, Distributed (IID)

Joint Random Variables: Many Turnout Decisions

Independent Random Variables:
Random variables $Y_1, Y_2, \ldots, Y_N$ are independent if

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) &= \\ P(Y_1 = y_1)P(Y_2 = y_2)\ldots P(Y_N = y_N) &= \\ \prod_{i=1}^{N} P(Y_i = y_i) & \end{aligned}$$

$Y_i \sim \text{Bernoulli}(\pi)$
Independent, Identically, Distributed (IID)
Assumption$\rightsquigarrow$ similar to what assumption from design-based inference?

Joint Random Variables: Many Turnout Decisions

Suppose $Y_i \sim \text{Bernoulli}(\pi)$

$$\begin{aligned} p(\boldsymbol{y}) &= \prod_{i=1}^{N} P(Y_i = y_i) \\ &= \prod_{i=1}^{N} \pi^{y_i}(1-\pi)^{1-y_i} \\ &= \pi^{\sum_{i=1}^{N} y_i}(1-\pi)^{N-\sum_{i=1}^{N} y_i} \end{aligned}$$

Joint Random Variables: Many Turnout Decisions

Suppose $Y_i \sim \text{Bernoulli}(\pi_i)$
Independent

$$
\begin{aligned}
p(\boldsymbol{y}) &= \prod_{i=1}^{N} P(Y_i = y_i) \\
&= \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}
\end{aligned}
$$

Key insight:

- Given $\pi_i \rightsquigarrow$ probability of $\boldsymbol{y}$
- Given $\boldsymbol{y}$ we should be able to learn something about π_i

Modeling Incumbent Vote Share

Suppose we are interested in modeling an incumbent's vote share Y_i. Assume that Y_i is a Normal random variable. Specifically, we will assume:

$$Y_i \sim \text{Normal}(\mu, \sigma^2)$$

Equivalently, a normal random variable has pdf:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

Properties of normal random variables:

- $E[Y_i] = \mu$
- $\text{var}(Y_i) = \sigma^2$

Modeling Many Incumbents' Vote Share

Suppose we are interested in many Congressional incumbents' vote shares
Assume $Y_i \sim \text{Normal}(\mu_i, \sigma^2)$
independent (non-identical) samples $\rightsquigarrow \boldsymbol{Y}$

$$
\begin{aligned}
f(\boldsymbol{y}) &= \prod_{i=1}^{N} f(y_i) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\
&= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\sum_{i=1}^{N} \frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)
\end{aligned}
$$

How do we make inferences using the model?

Modeling Many Incumbents' Vote Share

Define $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)$

$$f(\boldsymbol{y}, \underbrace{\boldsymbol{\mu}, \sigma^2}_{parameters}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(\sum_{i=1}^{N} \frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)$$

Writing down model of $\boldsymbol{y}$ creates link between data, parameters

All inferences depend initial modeling assumption (that won't be reflected in other measures of uncertainty)

Probability Term Refresher

Probability Term Refresher

1) CDF

Probability Term Refresher

1) CDF
2) Marginal distribution, conditional density

Probability Term Refresher

1) CDF
2) Marginal distribution, conditional density
3) Scale/Location (Normal distribution is a ...)

Probability Term Refresher

1) CDF
2) Marginal distribution, conditional density
3) Scale/Location (Normal distribution is a ...)
4) X is a random variable, a and b are constants. $E[aX + b]$, var$[aX + b]$

Probability Term Refresher

1) CDF
2) Marginal distribution, conditional density
3) Scale/Location (Normal distribution is a ...)
4) X is a random variable, a and b are constants. $E[aX + b]$, var$[aX + b]$
5) Bayes' Rule

Wednesday: Introduction to Likelihood Theory of Inference (inversion)