Lecture 1: Introduction

Professor Alexander Franks

Expectations and prerequisites

- Comfortable with fundamentals of likelihood-based inference (e.g. 207A)

- Familiarity with (generalized) linear models (220A)

- Comfortable with R programming

Class Resources

Course Pages

- Class website: https://ucsb-pstat215.github.io/web/

- Nectir: Joint at https://app.nectir.io/invite/7hnEew

Required Textbook

- Bayesian Data Analysis (Third Edition), available for free as pdf

Grades

- 40% - weekly homework

- 20% midterm exam (Take home exam, November 3)

- 5% - Participation

- 35% - Final paper / project (Due final exam day, 12/6)

Homework

- There will be approximately 6 homeworks (40% of your grade total)

- You will typically have 1 week to complete the homework
  - In general: one theory problem, one computational problem and one applied problem

- Homework turned in within 24 hrs after the deadline without prior approval will receive a 10 pt deduction (out of 100)

- Homework will not be accepted more than 24 hrs late.

Homework submission format

- All code must be reproducible as a Quarto document

- All derivations can be done in any format of your choosing (latex, written by hand) but must be legible and _must be integrated into your Quarto document

- All files must be zipped together and submitted to Gradescope

Software and Deliverables

Software

- R studio. Latest R studio version needed for Quarto.

Homeworks submission format

- Electronic submission via Gauchospace

- Must turn in a pdf and the Quarto document used to generate it

- Any supplementary files

Final project

- At least 10 pages double spaced, including figures (no more than 20)

- Should cover an advanced topic of interest to you or relevant to your research

- Must include the analysis of simulated or real data (ideally both)

Class topics

- Conjugate priors for EF models (e.g binomial, poisson, and multivariate normal)

- Hierarchical modeling

- Monte Carlo Methods

- Decision Theory (push to 215B?)

- Advanced topics as time permits

Class Policies

- I will post textbook readings for each week. This is reading *required* not just recommended.

- All questions should be posted on nectir, *not by email* (unless they are personal or grade-related)

- https://app.nectir.io/group/ucsb/PSTAT215-F22

RStudio Cloud Service

- Log on to pstat215.lsit.ucsb.edu

  - Cloud based rstudio service

  - Log in with your UCSB NetID

- Use https://tinyurl.com/3bs8uu8d to sync new material to the cloud…

- … or pull content directly from https://github.com/ucsb-pstat215/fall22

- Make sure you can write and compile a Quarto document (qmd)

Other R resources

- Cheatsheets:
  https://www.rstudio.com/resources/cheatsheets/

- Most relevant cheat sheets are uploaded, in `resources` folder

Bayesian Statistics

Frequentist statistics

- Associated with the *frequentist* interpretation of probability

  - For any given event, only one of two possibilities may hold: it occurs or it does not.

  - The *frequency* of an event (in repeated experiments) is the *probability* of the event

  - Focus on finding estimators with well established properties (consistent, unbiased, low variance, coverage etc)

  - Premised on imaginary resampling of data

- Example: Null Hypothesis Significant Testing (NHST)

  - If the null model is true, and I re-run the experiment many times, how often will I reject?
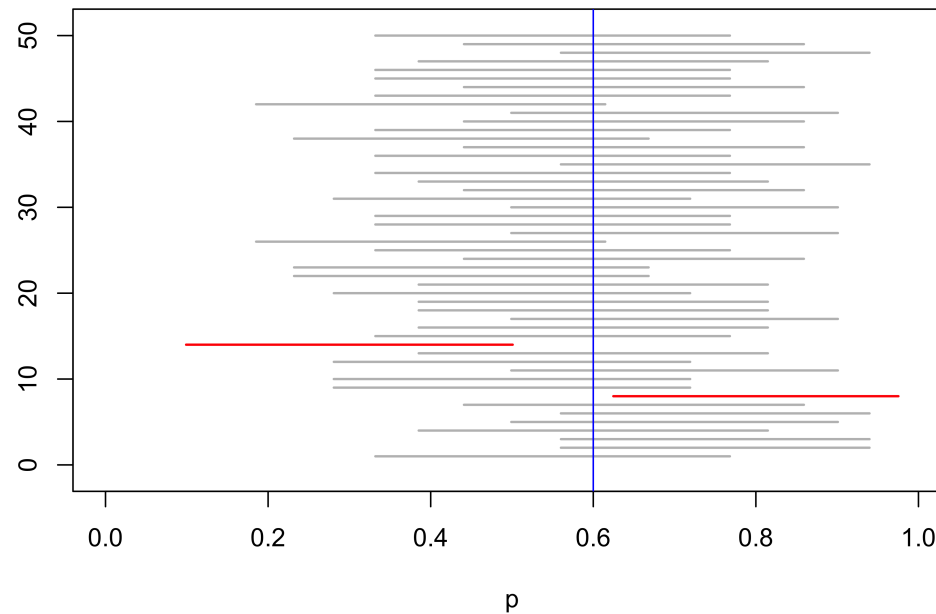
Frequentist probability

The probability of a coin landing on heads is 50%
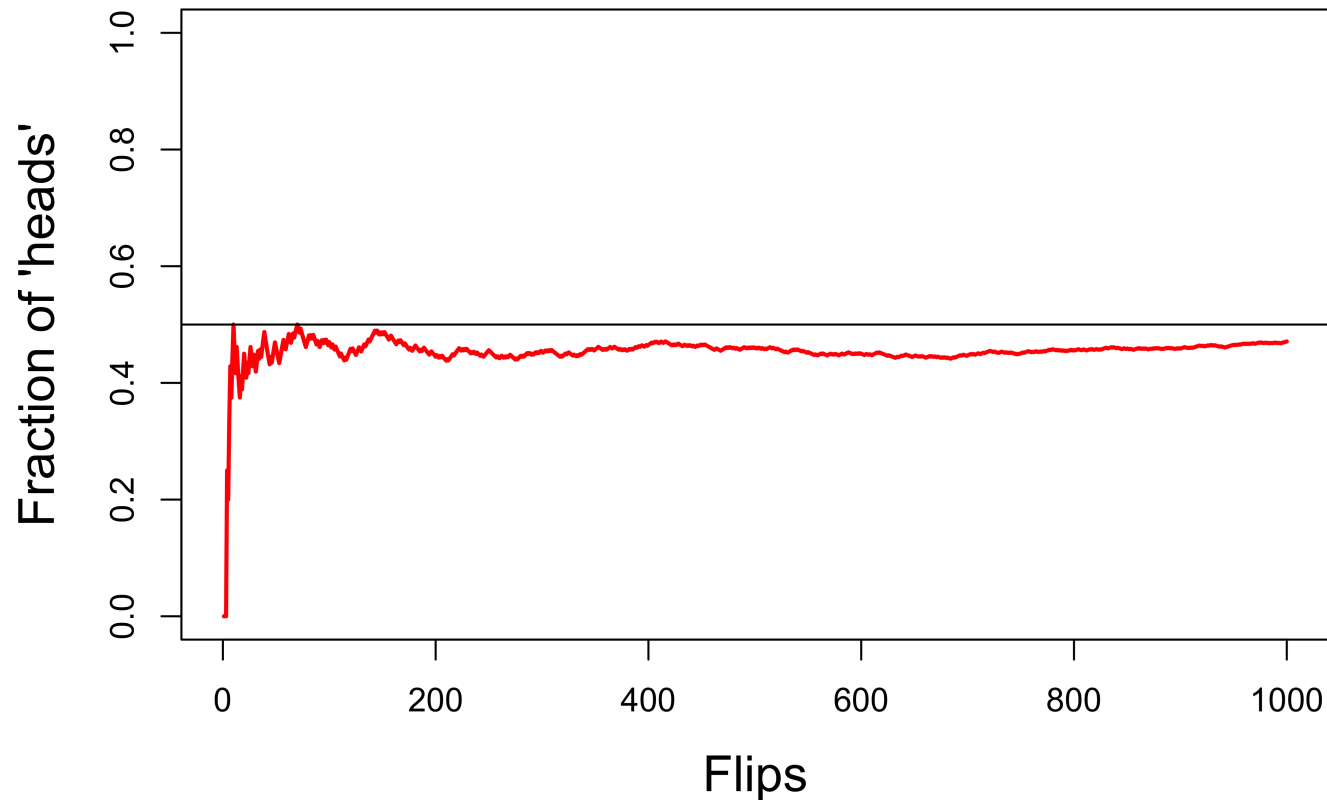
The long run fraction of heads is 50%

Confidence intervals

I have a 95% confidence interval for a parameter θ. What does this mean?

We expect $0.05 \times 50 = 2.5$ of the intervals to *not* cover the true parameter, $p = 0.6$, on average

This class

- Build statistical models representing a set of assumption about how the data was generated.

- Use models to develop statistical tests, predictions and forecasts

- Can (and should) be continuously refined and extended!

- Incorporate prior knowledge and condition on what we observe

- Can still consider frequentist properties of estimators derived from Bayesian principles
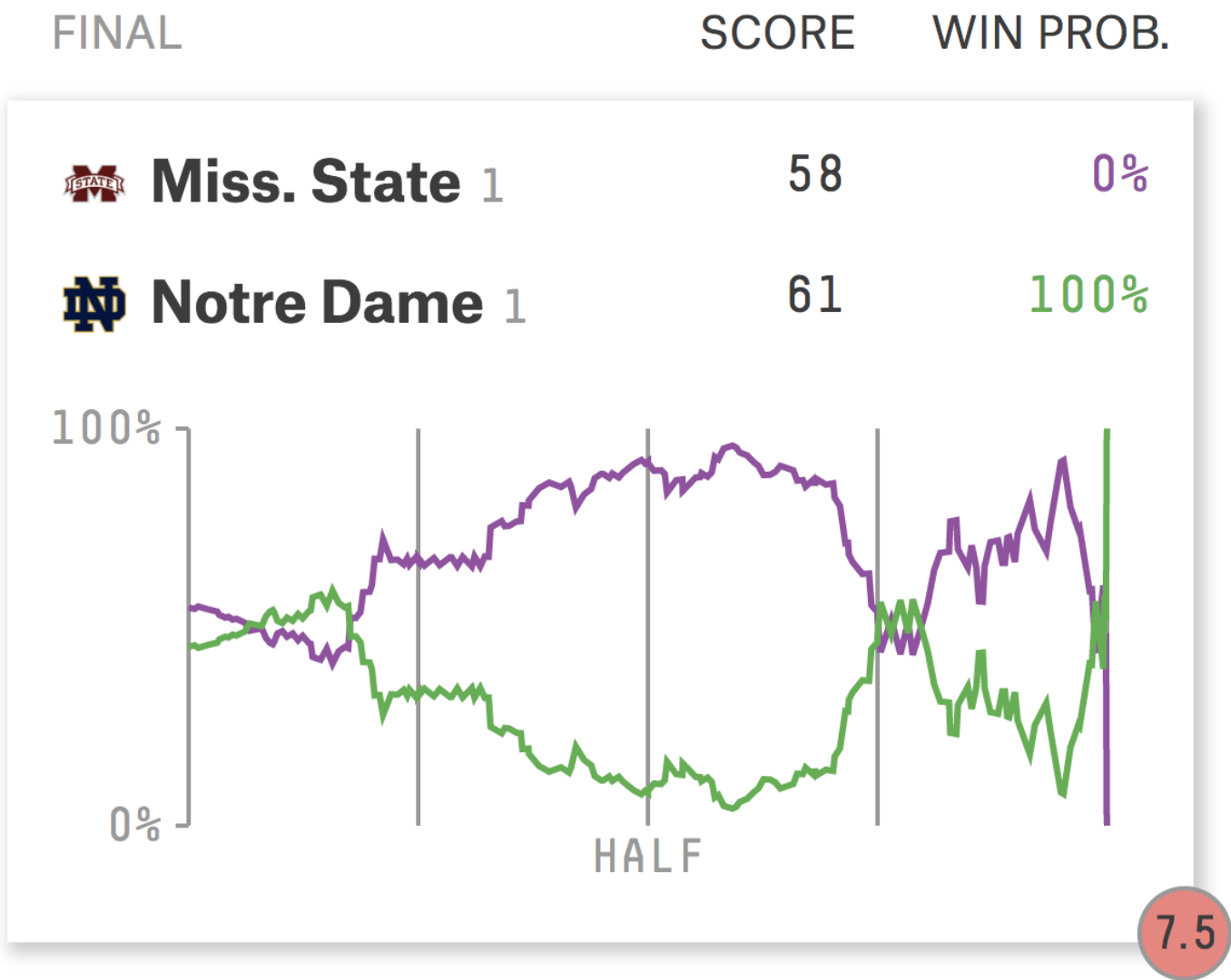
All models are wrong
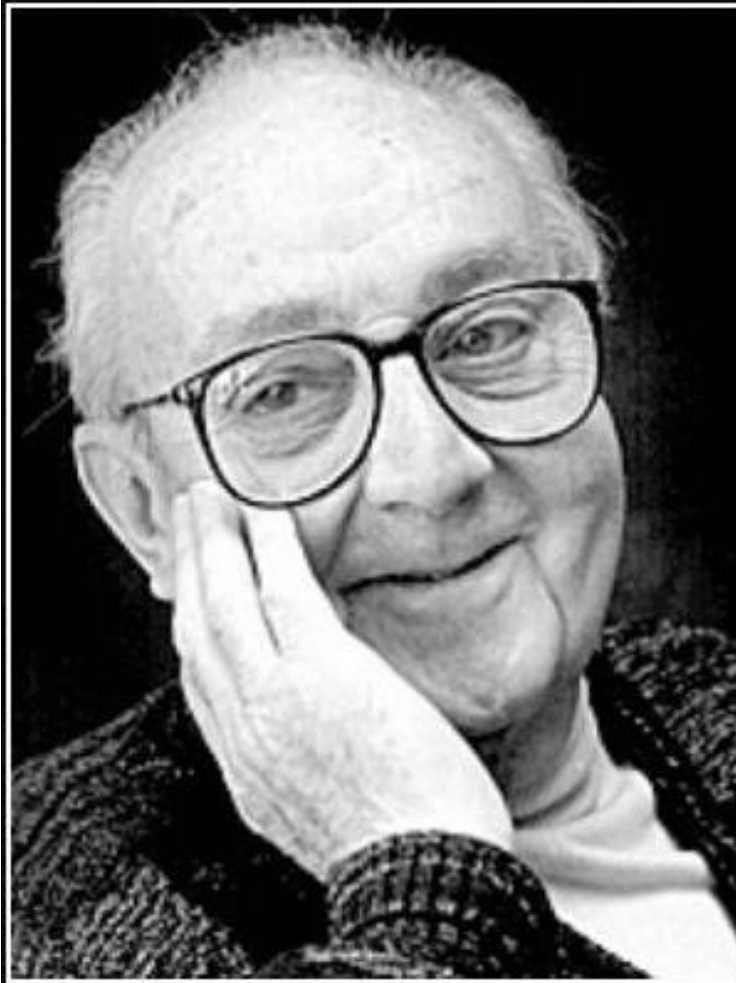
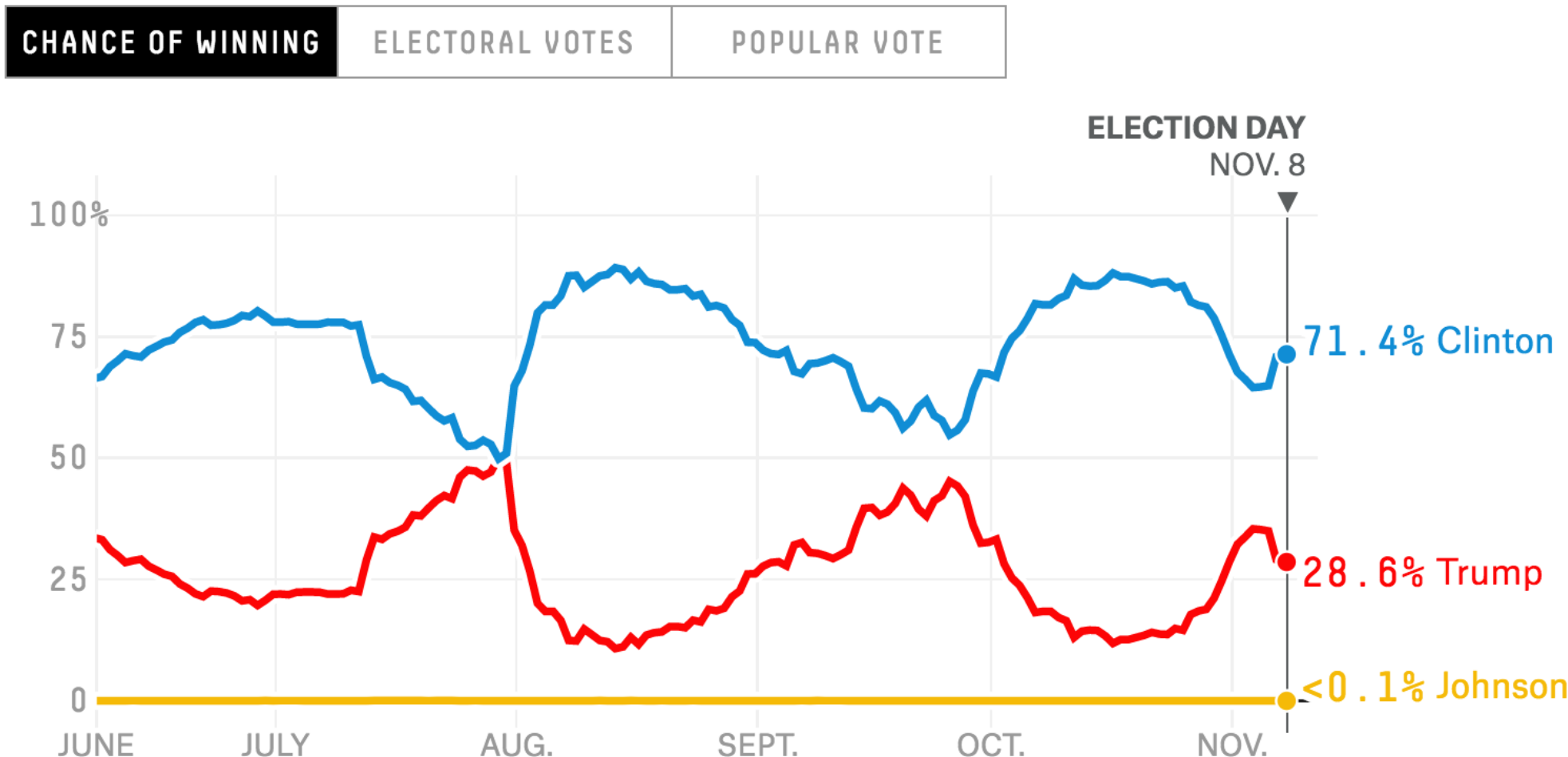All models are wrong, but some are useful.

— George E. P. Box —

All models are wrong (wikipedia)
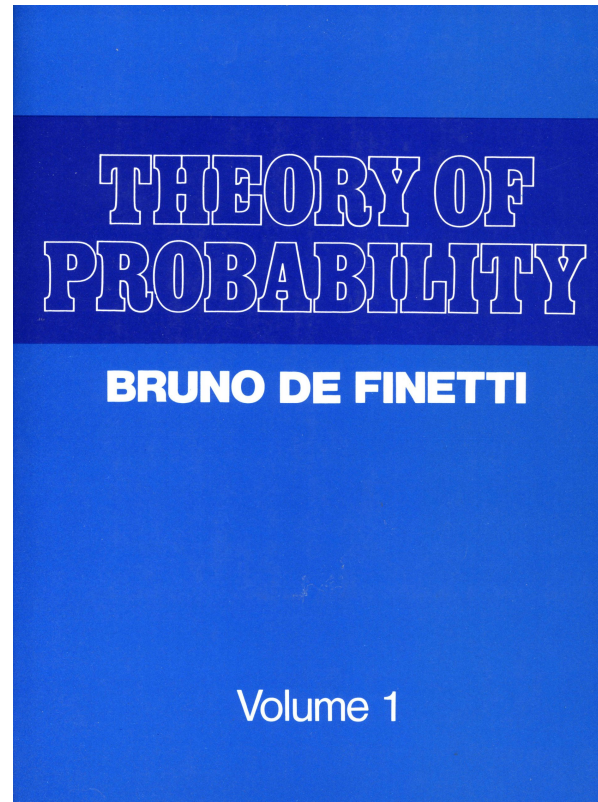
Win probability

source: fivethirtyeight.com

Win probability

source: fivethirtyeight.com

Bayesian probability

Bruno de Finetti began his book on probability with:
"PROBABILITY DOES NOT EXIST"

Bayesian probability

- de Finetti is implicitly arguing that probability is about *belief*
    - Claim: probability doesn't exist in an *objective* sense
    - "The coin is fair" means *I believe* that its equally likely to be heads or tails.
    - "Hillary Clinton has a 71% chance to win" reflects a belief, since the election happens only once
- Rarely, if ever, get *true* replications to estimate frequentist probabilities
- Bayesian idea: focus statistical practice around belief about parameters

Bayesian probability

"The terms *certain* and *probable* describe the various degrees of rational belief about a proposition which different amounts of knowledge authorise us to entertain. All propositions are true or false, but the knowledge we have of them depends on our circumstances

— John M Keynes

Why Bayesian statistics?

- Philosophy: quantify degrees of belief rather than reason about counterfactuals

    - Can easily "share information" across related observations

    - Ability to incorporate real prior knowledge

    - Particularly effective with small samples

- A variety of powerful tools for inference with computer simulation

- Can still characterize frequentist properties of Bayesian procedures

Setup

- The *sample space* $\mathcal{Y}$ is the set of all possible datasets.

    - Y is a random variable with support in $\mathcal{Y}$

    - We observe one dataset y from which we hope to learn about the world.

- The *parameter space* Θ is the set of all possible parameter values θ

- θ encodes the population characteristics that we want to learn about!

Three steps of Bayesian data analysis

1. Construct a plausible probability model governed by parameters θ

   - This includes specifying your belief about θ before seeing data (*the prior*)

2. Condition on the observed data and compute *the posterior* distribution for θ

3. Evaluate the model fit, revise and extend. Then repeat.

Bayesian Inference in a Nutshell

1. The *prior distribution $p(\theta)$* describes our belief about the true population characteristics, for each value of $\theta \in \Theta$.

2. Our *sampling model $p(y \mid \theta)$* describes our belief about what data we are likely to observe if θ is true.

3. Once we actually observe data, y, we update our beliefs about θ by computing *the posterior distribution $p(\theta \mid y)$*. We do this with Bayes' rule!

Key difference: θ is random!

Bayes' Rule for Bayesian Statistics

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)}$$

- $P(\theta \mid y)$ is the posterior distribution
- $P(y \mid \theta)$ is the likelihood
- $P(\theta)$ is the prior distribution
- $P(y) = \int_{\Theta} p(y \mid \tilde{\theta})p(\tilde{\theta})d\tilde{\theta}$ is the model evidence
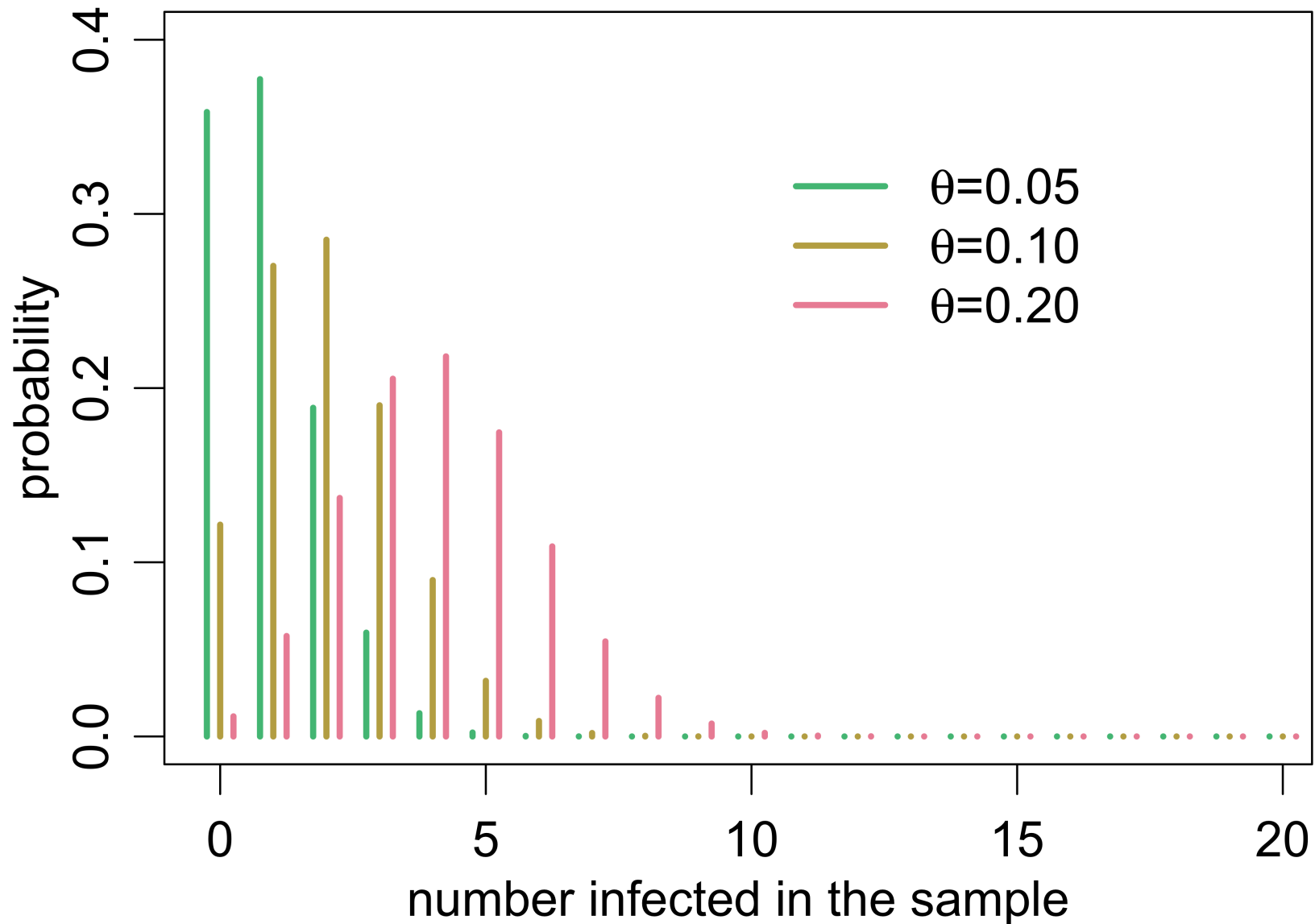
Example: Estimating COVID Infection Rates

- We need to estimate the prevalence of a COVID in Isla Vista

- Get a small random sample of 20 individuals to check for infection

Example: Estimating Infection Rates

- θ represents the population fraction of infected

- Y is a random variable reflecting the number of infected in the sample

- $\Theta = [0, 1] \quad \mathcal{Y} = \{0, 1, \ldots, 20\}$

- Sampling model: $Y \sim \text{Binom}(20, \theta)$

Example: Estimating Infection Rates

Example: Estimating Infection Rates

- Assume *a priori* that the population rate is low

  - The infection rate in comparable cities ranges from about 0.05 to 0.20

- Assume we observe $Y = 0$ infected in our sample

- What is our estimate of the true population fraction of infected individuals?

Bayesian vs Frequentist

- In frequentist inference, unknown parameters treated as constants

    - Estimators are random (due to sampling variability)

    - "What would I expect to see if I repeated the experiment?"

- In Bayesian inference, unknown parameters are random variables.

    - Need to specify a prior distribution for θ (not easy)

    - "What do I *believe* are plausible values for the unknown parameters **given the data**?"

Uncertainty Quantification: Bayes vs Frequentist

- $X_i \sim$ Rademacher

- $Y_i = \theta + X_i$.

- Observe Y_1 and Y_2 and construct the following confidence ''interval'':

$$\begin{cases} \{(Y_1 + Y_2)/2\} & \text{if} & Y_1 \neq Y_2 \\ \{Y_1\} & \text{else} \end{cases}$$

What fraction of time does this cover?

The age guessing game*

*Bayesian edition

Assignment

- Join Nectir

- Start reviewing probability cheat sheet!

- Read chapters 1 and 2 of BDA3