PSTAT 115: Bayesian Data Analysis

Class Resources

Required Textbook

- Bayes Rules: https://www.bayesrulesbook.com/

Course Pages

- Class website on Canvas: https://https://www.canvas.ucsb.edu/
- Gradescope: [https://www.gradescope.com/courses/1107727]https://www.gr
  - On Canvas site

Grades

- 25% - Homework
- 5% - Section Attendance
- 20% - Quizzes
- 20% - Midterm (October 29, in class)
- 30% - Final exam (December 9)

Homework

- There will be 5 homeworks (25% of your grade total)
- You will 2 weeks to complete the homeworks
- Every student *must* submit their own assignment on gradescope
- 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 written to be reproducible in Quarto
- 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 Rmarkdown pdf*.
- All files must be zipped together and submitted to Gradescope
- Ask a TA *early* if you have problems regarding submissions.

Software and Deliverables

Software

- R (R studio)

Homeworks submission format

- Electronic submission via Gradescope
- R markdown code
- Generated PDF file
- Any supplementary files (e.g. write up for math problems)

Section and Quizzes

- There are no makeups, but the lowest quiz grade will be dropped from your final score.
- Quizzes (20%) will test your comprehension of the basic concept. In Section.

Class Policies

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

RStudio Cloud Service

- Log on to pstat115.lsit.ucsb.edu
  - Cloud based rstudio service
  - Log in with your UCSB NetID
- Use tinyurl.com/pstat115) to sync new material (BOOKMARK THIS)
- Text formatting is minimal but syntax is simple

AI Policy

You may use large language models with any homework assignments in this course. All students will be asked to cite the tool you used, and include a brief reflection about the use of the tool.

AI Policy

> There are some cases where it makes sense to hire someone else to cook your food. There are essentially no cases where it makes sense to hire someone else to *eat* your food for you.
>
> Using AI inappropriately on an assignment is like having the chef eat the food for you: you don't get any of the nutrients yourself.

Markdown and mathematical formulas

The text inserted between two `$` signs will be interpreted as a Latex instruction, e.g. `$x$`

Code	Rendered math
`$x$`	x
`$\theta$`	θ
`$x_i^2$`	x_i^2
`$\frac{1}{n}\sum_{i=1}^n x_i$`	$\frac{1}{n}\sum_{i=1}^n x_i$

Code	Rendered math
`$\frac{1}{n}\sum_{i=1}^n (x_i-\bar x)^2$`	$\frac{1}{n}\sum_{i=1}^n (x_i-\bar x)^2$

Rmarkdown and Latex resources:

- Introduction to RMarkdown
- Latex cheat sheet
- Introduction to Latex

Other R resources

Conditional probability

Priors

Bayes Rule

What is Bayesian statistics?

What is the version of statistics you already know?

Frequentist Statistics

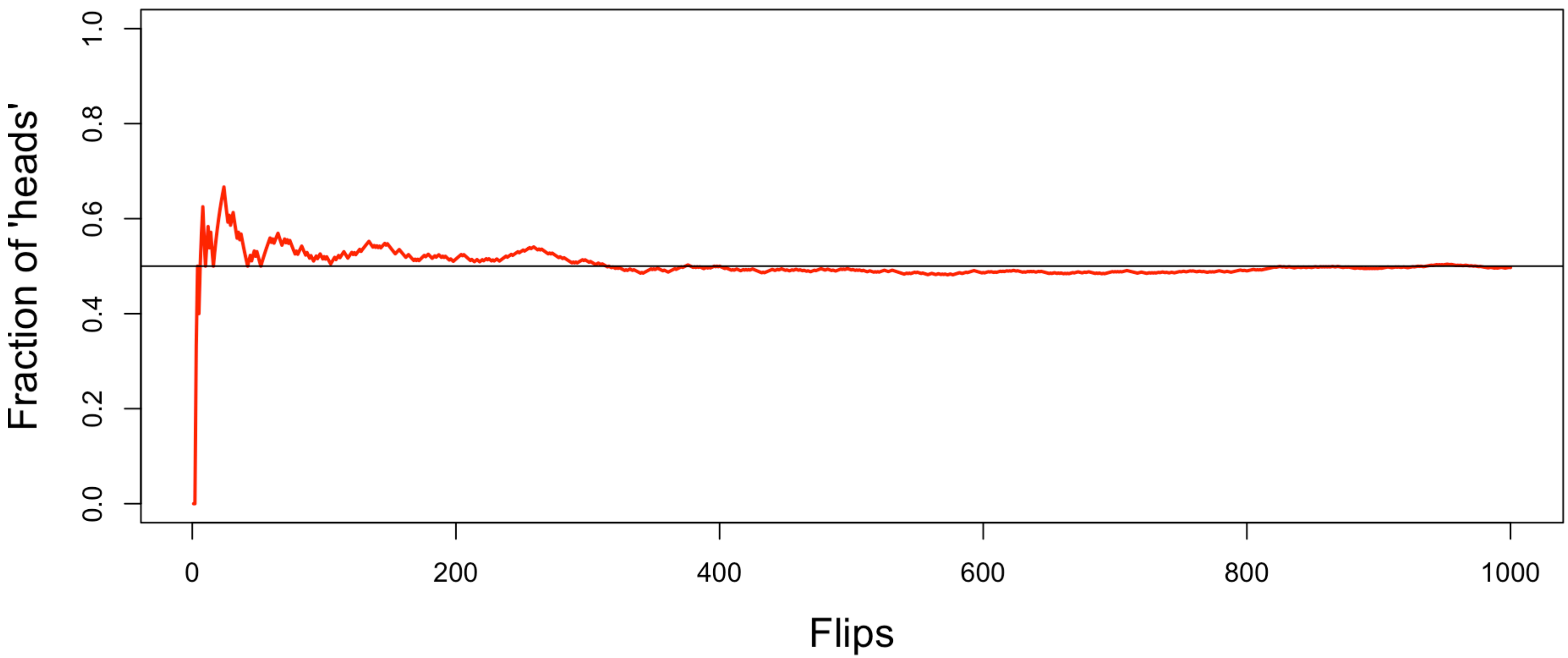
Frequentist statistics

What you learned in PSTAT 120B

- 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
- Null Hypothesis Significant Testing (NHST) and Confidence Intervals
  - Frequentist uncertainty premised on imaginary resampling of data
  - Example: If the null model is true, and I re-run the experiment many times, how often will I reject?

Frequentist probability

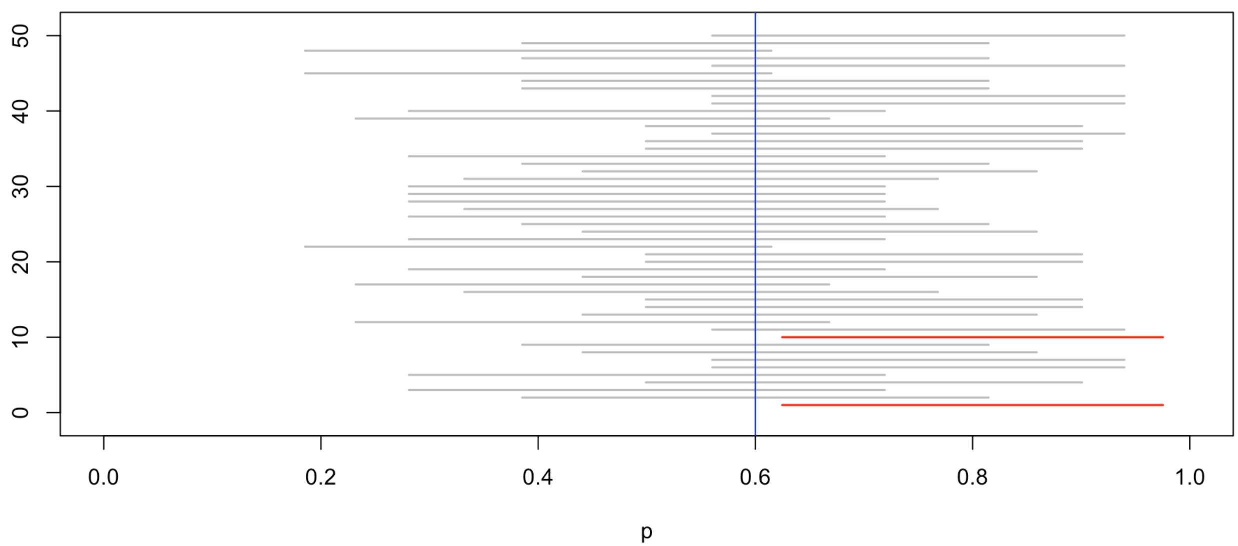
The probability of a coin landing on heads is 50%



The long run fraction of heads is 50%

Confidence intervals

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

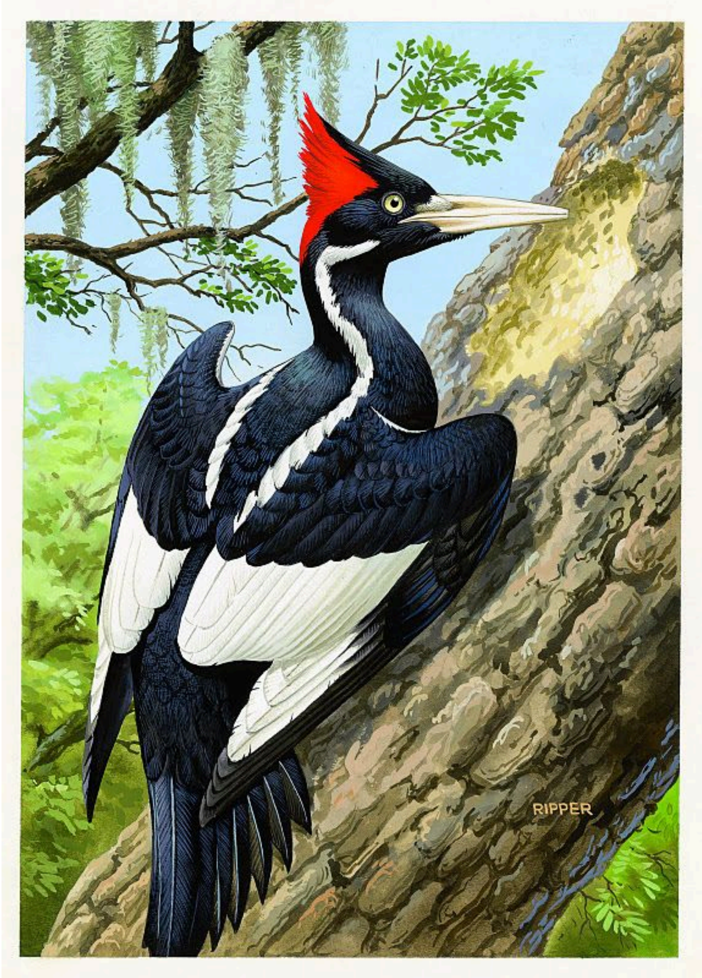


We expect $0.05 \times 50 = 2.5$ of the intervals to *not* cover the true parameter, $p = 0.6$, on average

Falsification

H_0 : “All swans are white” vs H_A : “not all swans are white”.

Falsification



H_0 : “The Ivory-billed Woodpecker is extinct”

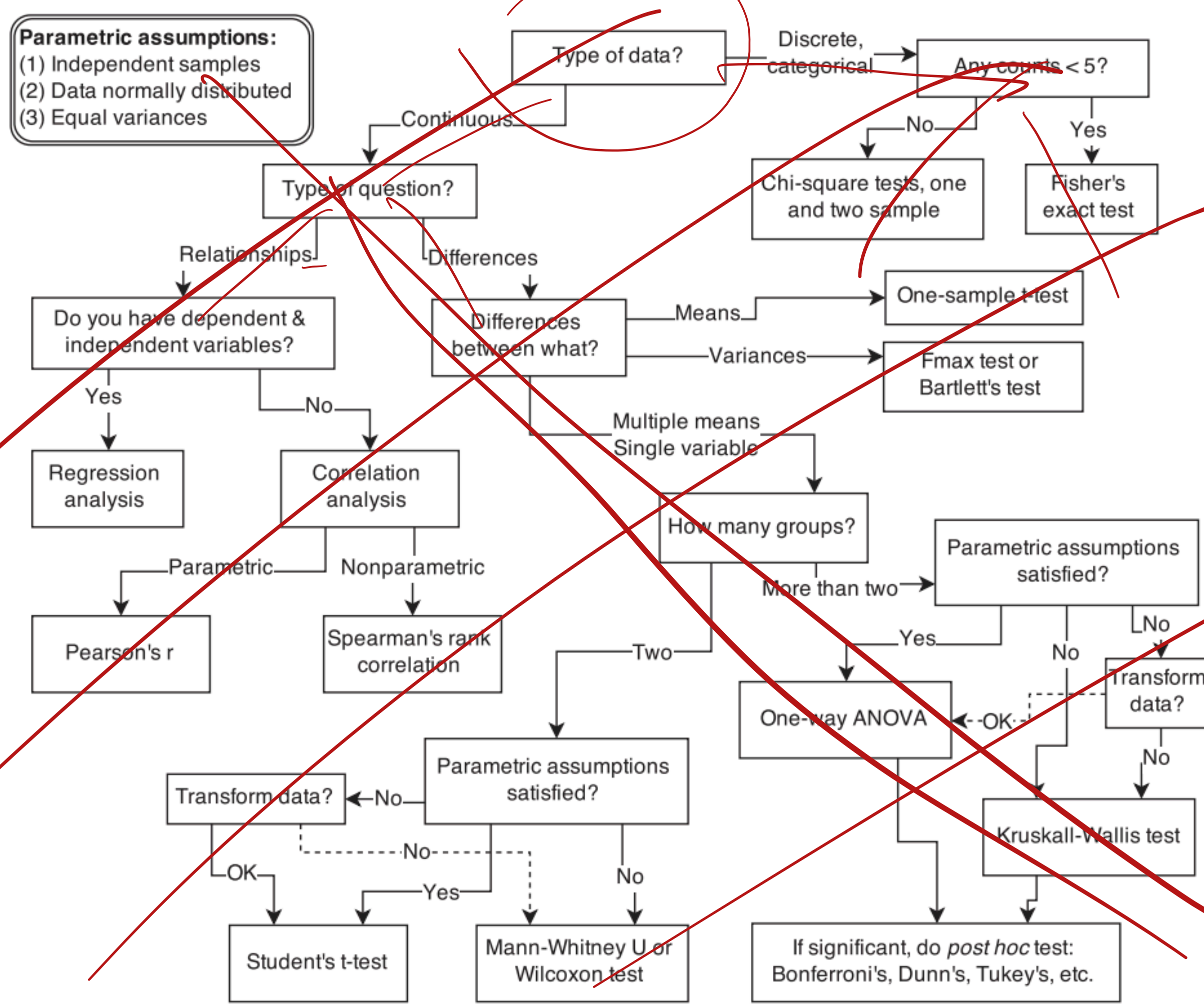
Falsification

H_0 : “Black swans are rare”

Falsification

- Is an observation real or spurious?
  - Importance of measurement error
  - Natural phenomena are usually continuous in nature
- Falsification requires consensus more than logic
  - Scientific communities argue toward consensus
  - Science is messy!

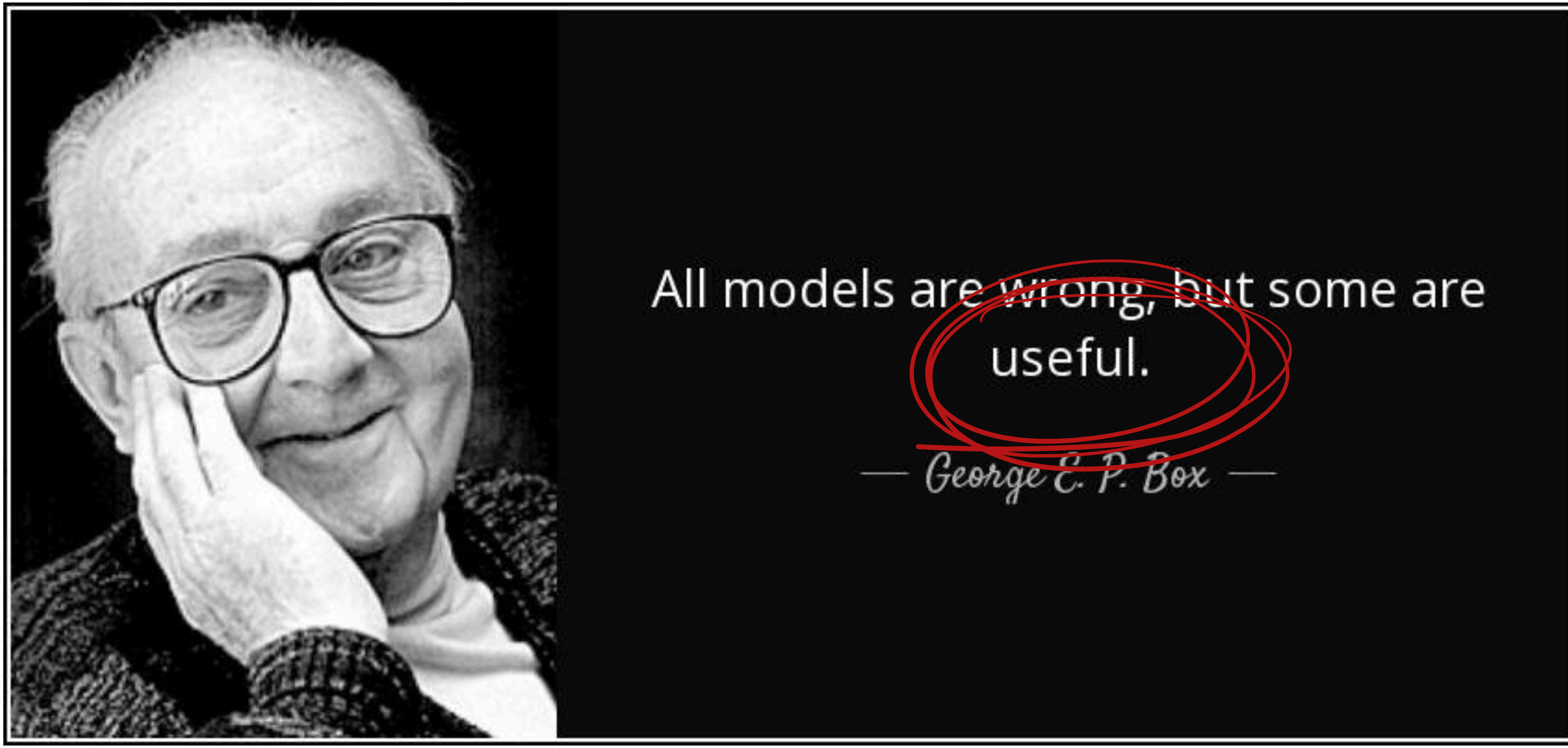
Significance Testing Flowchart



Alternative: focus on modeling!

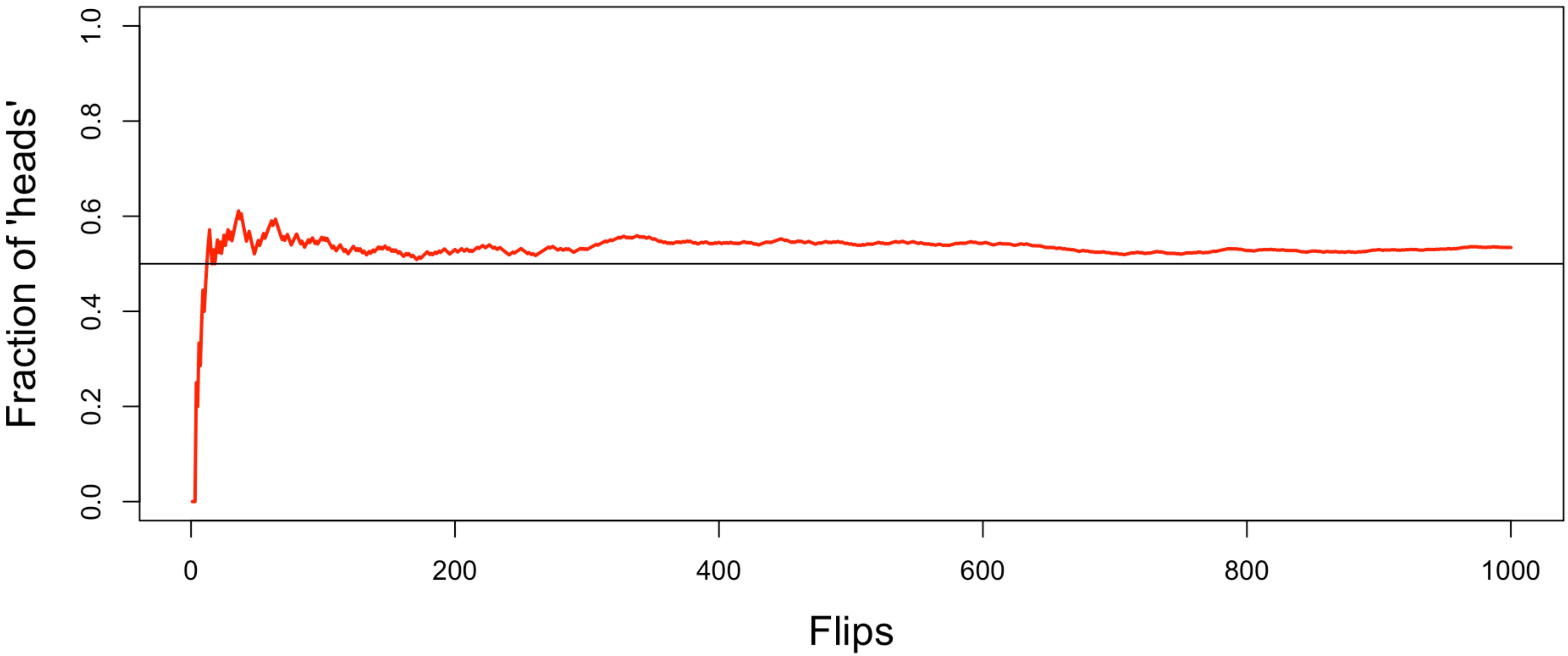
- A statistical model represents a set of assumption about how the data was generated.
- Models can still be used to develop statistical tests.
- Can also be used to make predictions or forecasts and describe sources of variability.
- Can (and should) be continuously refined and extended!

All models are wrong

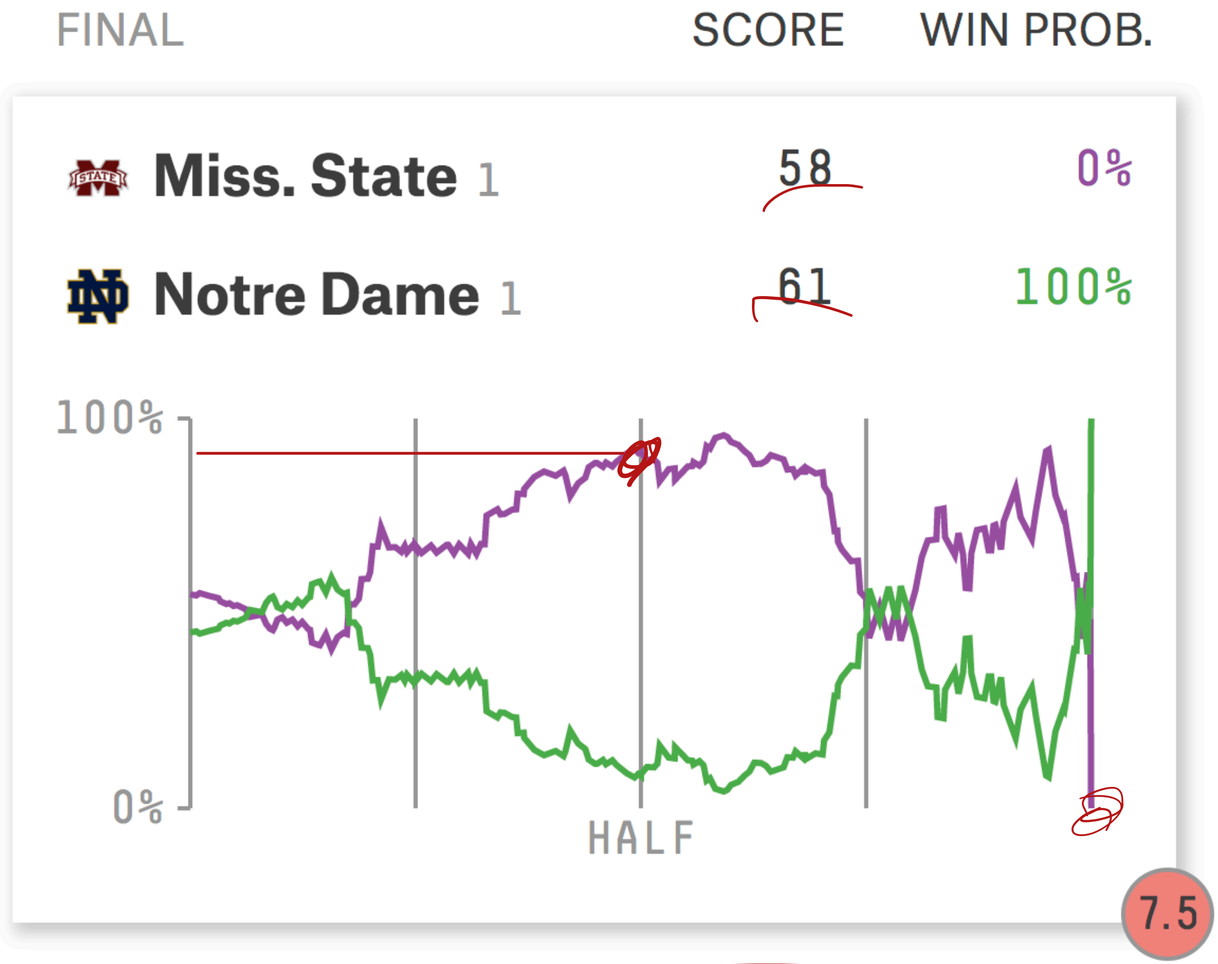


https://en.wikipedia.org/wiki/All_models_are_wrong

Frequentist probability

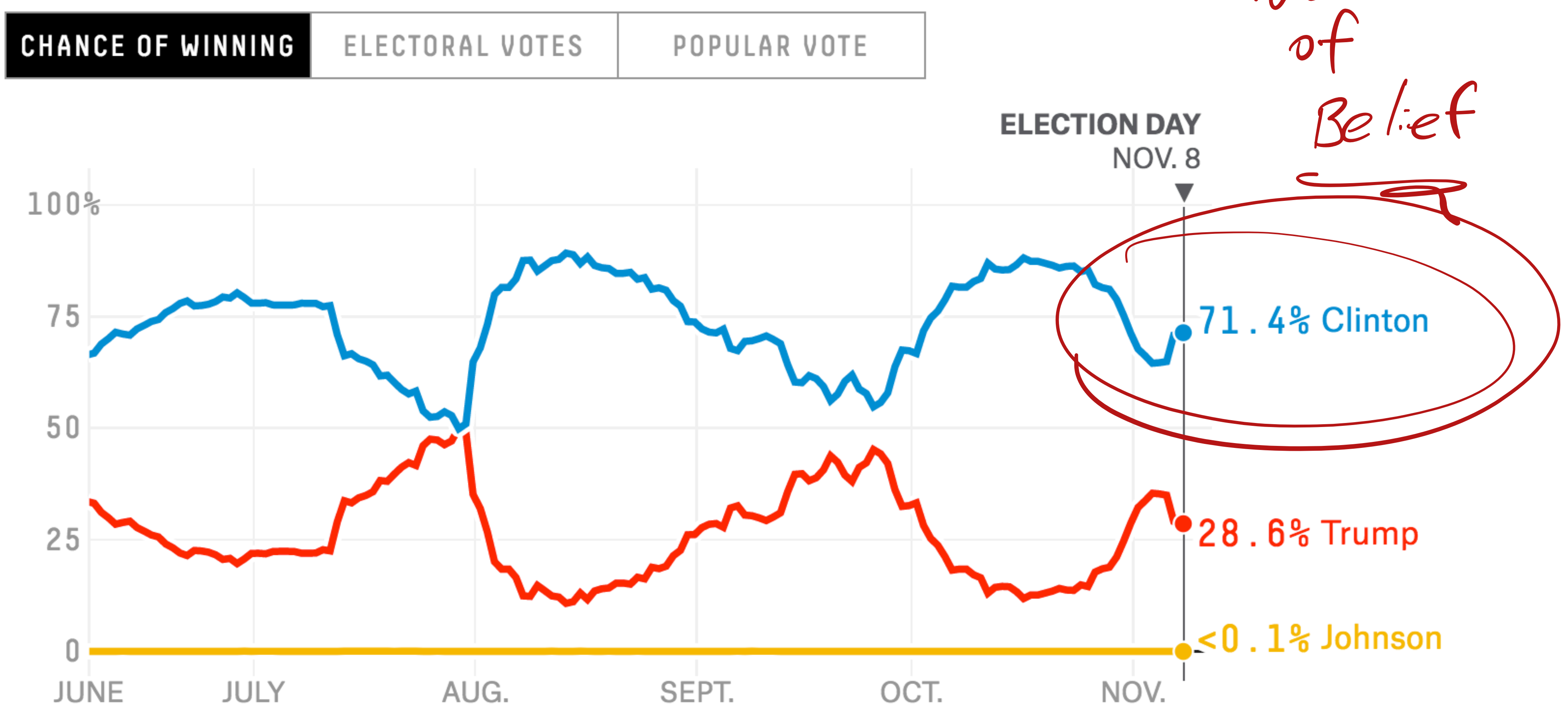


Win probability



source: fivethirtyeight.com

Win probability

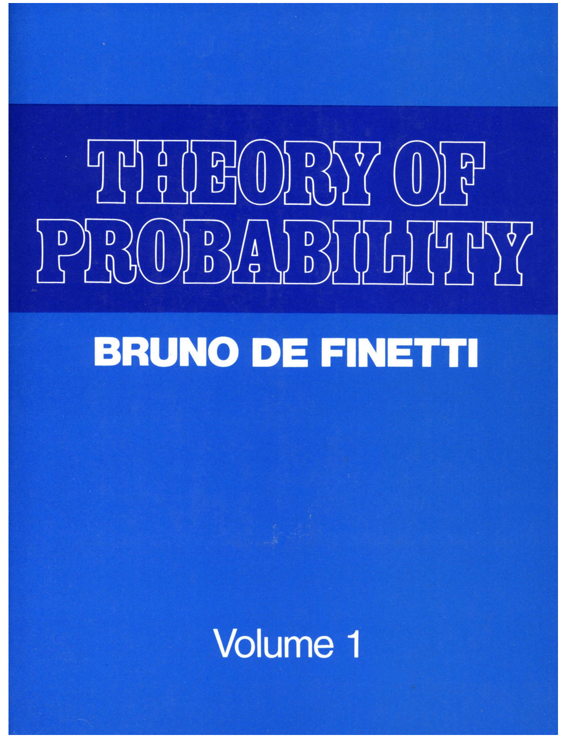


source: fivethirtyeight.com

Bayesian probability

Probability reflects Belief

What are fair odds? (gambling)



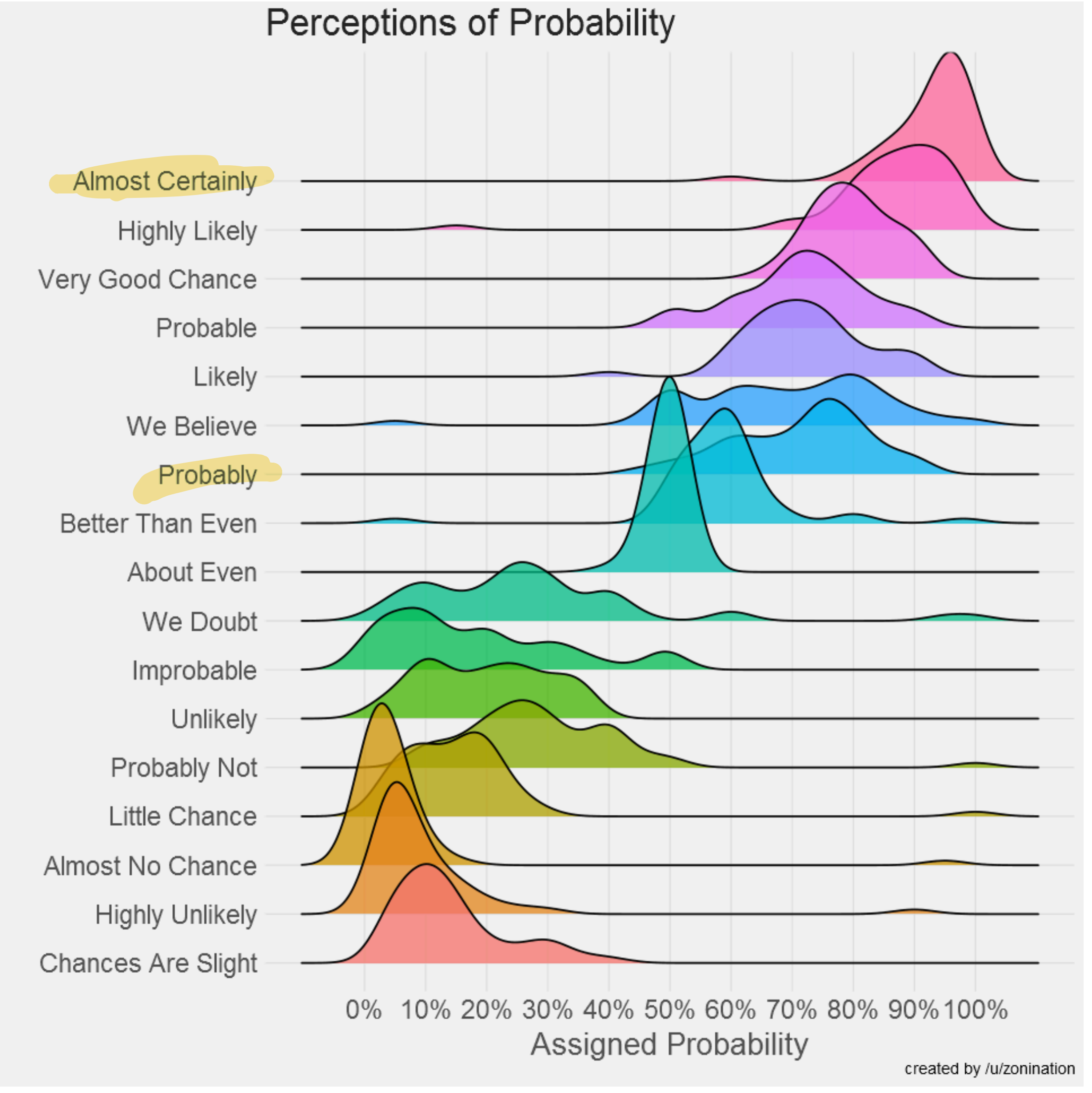
Bruno de Finetti begean his book on probability with:
"PROBABILITY DOES NOT EXIST"

Bayesian probability

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

Perceptions of Probability

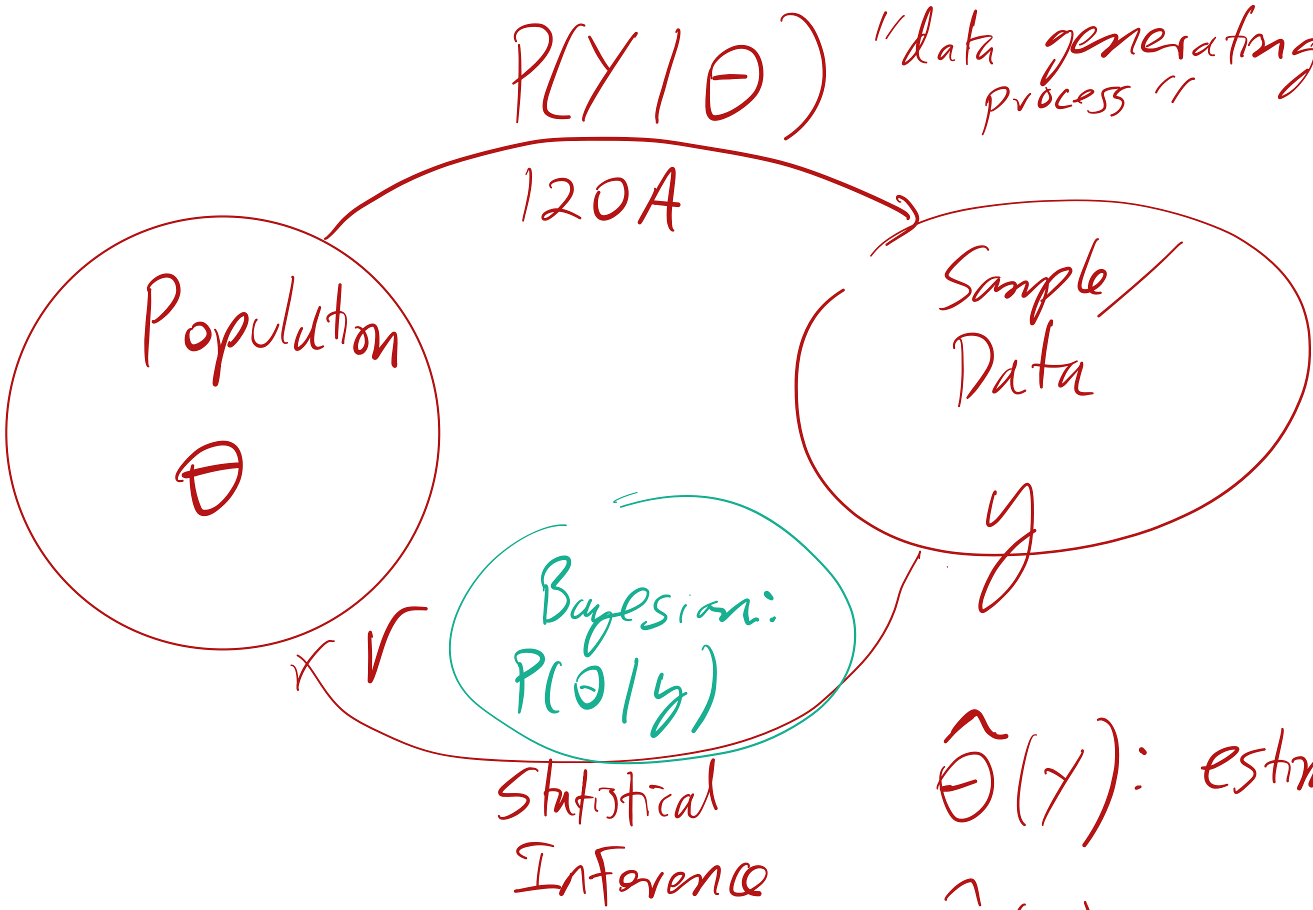


Why Bayesian statistics?

- Classical methods not always appropriate
  - Fragile / Inflexile
- Bayes provides procedures for building own tests / tools / inferences
- Powerful w/ computer simulation.
- Philosophy: degrees of belief.

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!



$\hat{\Theta}(Y)$: estimator

$\hat{\theta}(y)$: estimate

Three steps of Bayesian data analysis

1. Construct a plausible probability model governed by parameters θ
   $P(y|\theta)$, $P(\theta)$
   - 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
   $P(\theta|y)$
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

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

- $P(A \mid B)$ is the conditional probability of A given B
- $P(B \mid A)$ is the conditional probability of B given A
- $P(A)$ and $P(B)$ are called the marginal probability of A and B (unconditional)

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

$$P(\theta \mid y) \propto P(y \mid \theta) P(\theta)$$

Bayes' Rule for Bayesian Statistics

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

proportional to.

- Start with a subjective belief (prior)
- Update it with evidence from data (likelihood)
- Summarize what you learn (posterior)

Example: Estimating COVID Infection Rates

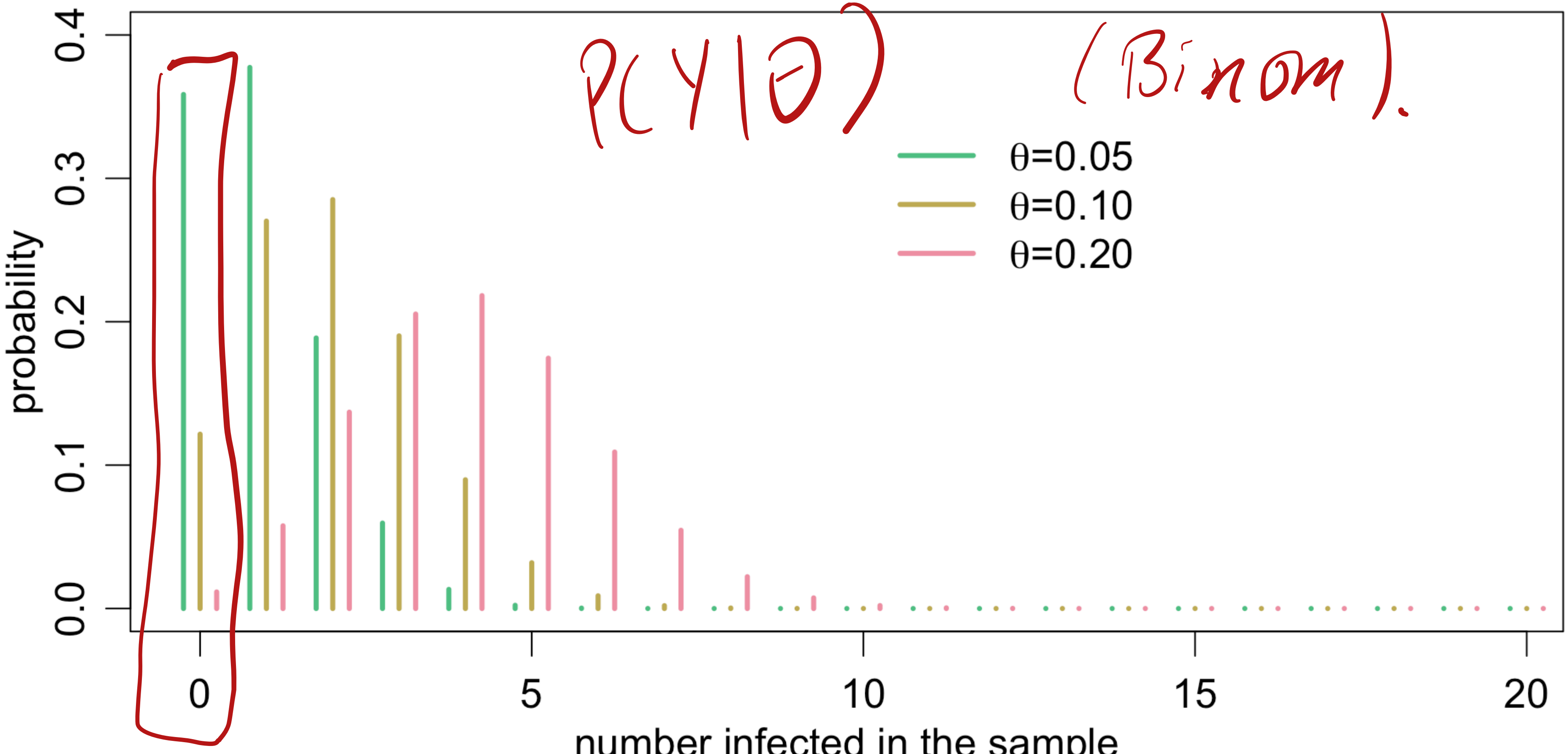
- We need to estimate the prevalence of a COVID in Isla Vista
- Get a small random sample of 20 individuals to check for infection

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

$p(Y|\theta)$

Example: Estimating Infection Rates



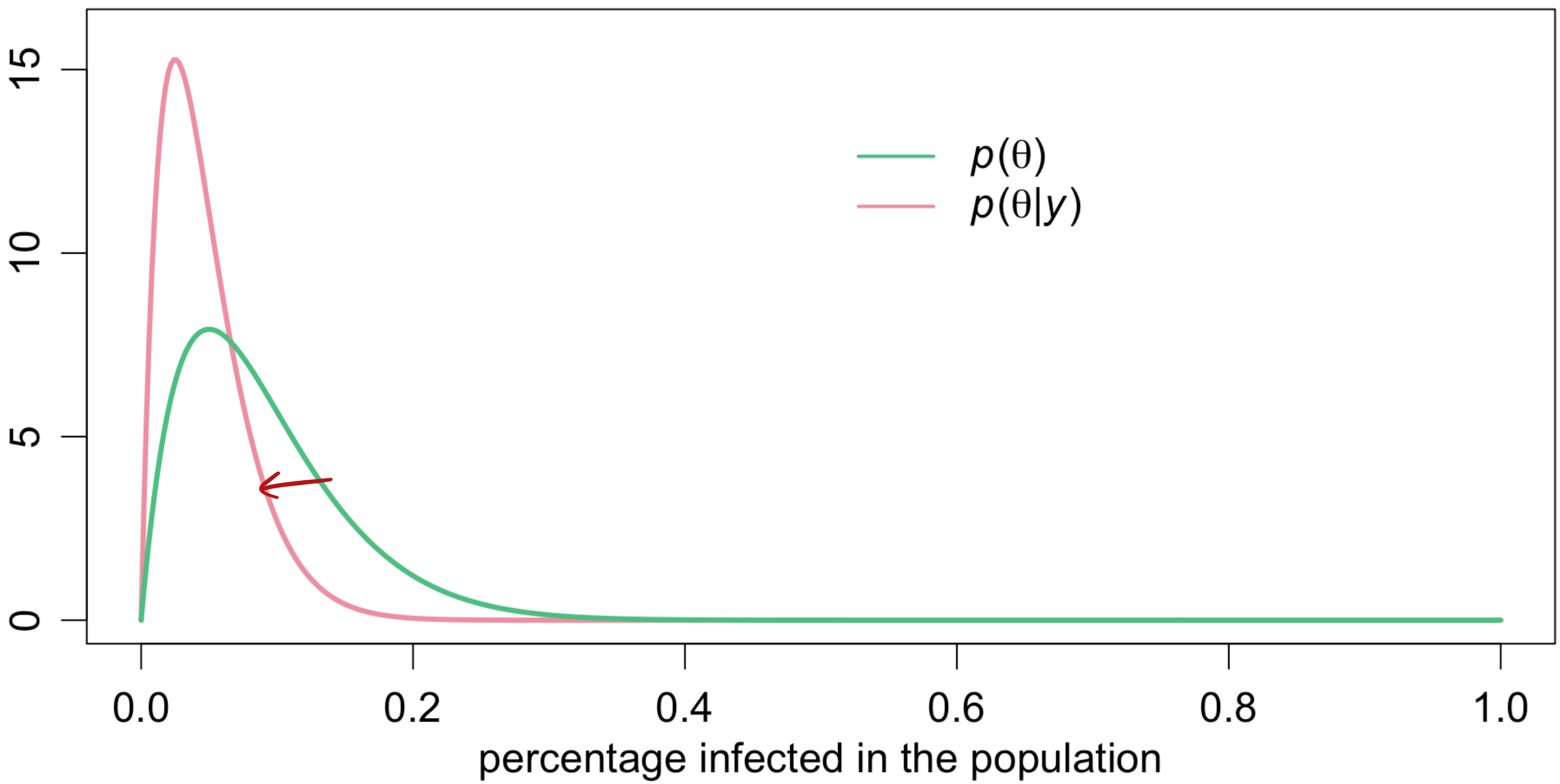
$P(Y|\theta)$ (Binom).

$P(Y=0|\theta) \longrightarrow P(\theta|y=0)$

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?

Example: Estimating Infection Rates



Content

- One parameter models (binomial, poisson, and normal) $N(\mu, \sigma^2)$ (2 param)
- Monte Carlo methods (i.e. simulation-based inference)
- Markov chain Monte Carlo (MCMC)
- Hierarchical modeling

Assignment

- Start reviewing probability cheat sheet!
- Read chapters 1 and 2 of Bayes Rules