PSTAT 115: Bayesian Data Analysis

Professor Laura Baracaldo

Class Resources

Required Textbook

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

Course Pages

- Nectir

- Gradescope

Grades

- 35% - expect approximately 4 homeworks
- 20% - Midterm
- 15% - Quizzes
- 30% - Final exam

Homework

- There will be approximately 4 homeworks (35% of your grade total)

- Homework turned in within 24 hrs after the deadline without prior approval will receive a 20% deduction
- Homework will not be accepted more than 24 hrs late.

Homework submission format

- All code must be written to be reproducible in Rmarkdown
- 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*.
- 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 with eventual handwritten notes

Labs and Quizzes

- There will be a handful of quizzes throughout the quarter.
- The quizzes will be on Canvas.
- You will have 10 minutes to take the quiz any time within a window of 12 hours.
- The quizzes will be given on lecture days
- There are no makeups, but the lowest quiz grade will be dropped from your final score.
- Quizzes will be multiple choice and will test your comprehension of the basic concepts.

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

- Make sure you can write and compile an R markdown (`Rmd`) document online

- Text formatting is minimal but syntax is simple

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

Rmarkdown and Latex resources:

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

Other R resources

- Cheatsheets: https://www.rstudio.com/resources/cheatsheets/
- *An Introduction to R* - Venables and Smith

  http://cran.r-project.org/doc/manuals/R-intro.pdf
- *Using R for Introductory Statistics* - John Verzani

  http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf
- *R Markdown reference* - https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf
- Probability cheatsheet in resources folder of cloud environment

What is Bayesian statistics?

What is the version of statistics you already know?

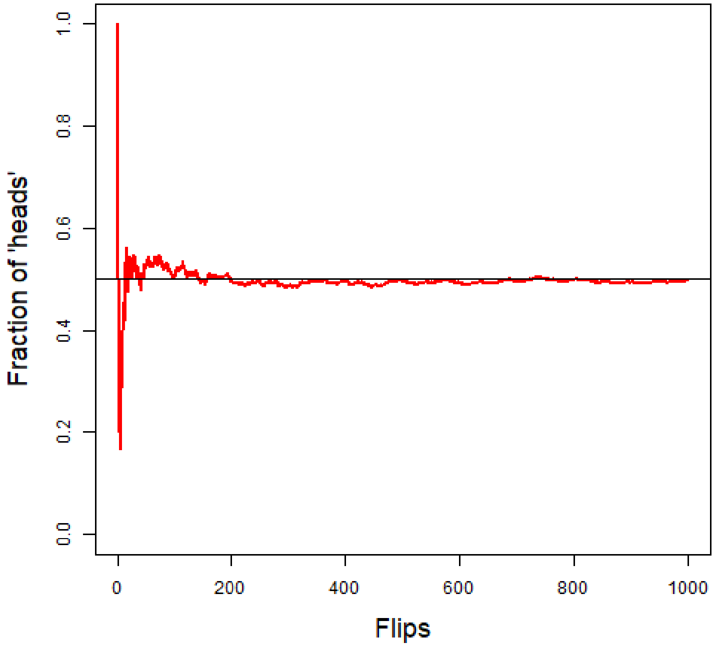
Frequentist statistics

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

Frequentist probability

The probability of a coin landing on heads is 50%



The long run fraction of heads is 50%

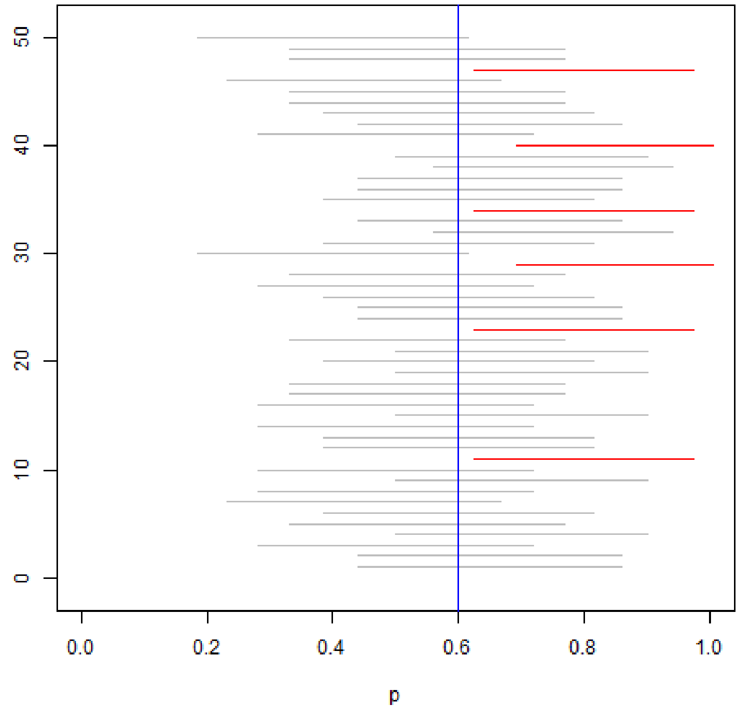
Frequentist statistics

What you learned in PSTAT 120B

- 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?

Confidence intervals

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



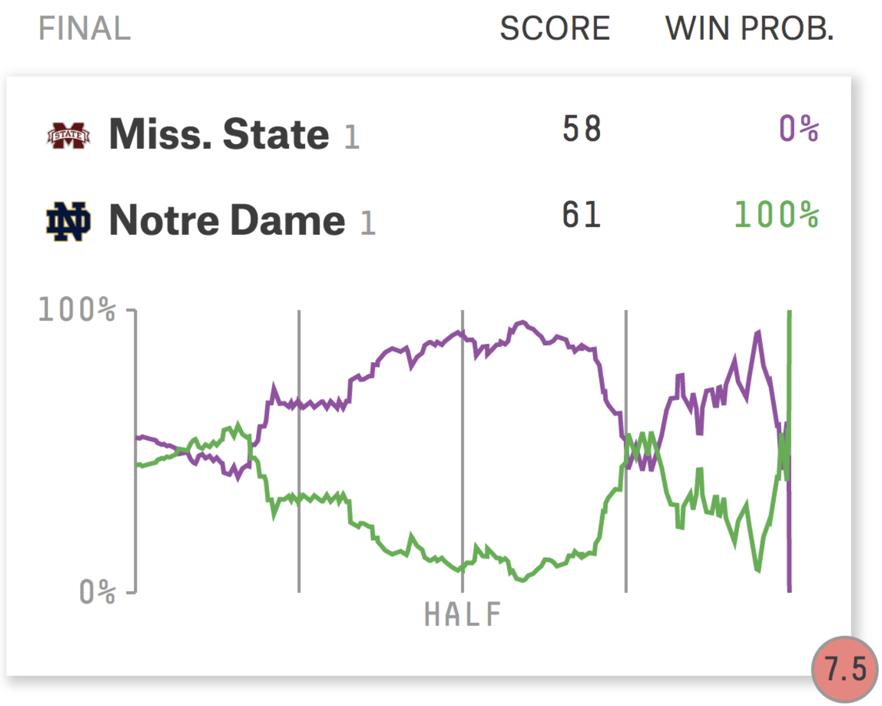
We expect $0.05 \times 50 = 2.5$ of the intervals to *not* cover the true parameter, $p = 0.6$, on average

Hypothesis Testing

H_0 : "All swans are white" vs H_A : "not all swans are white".

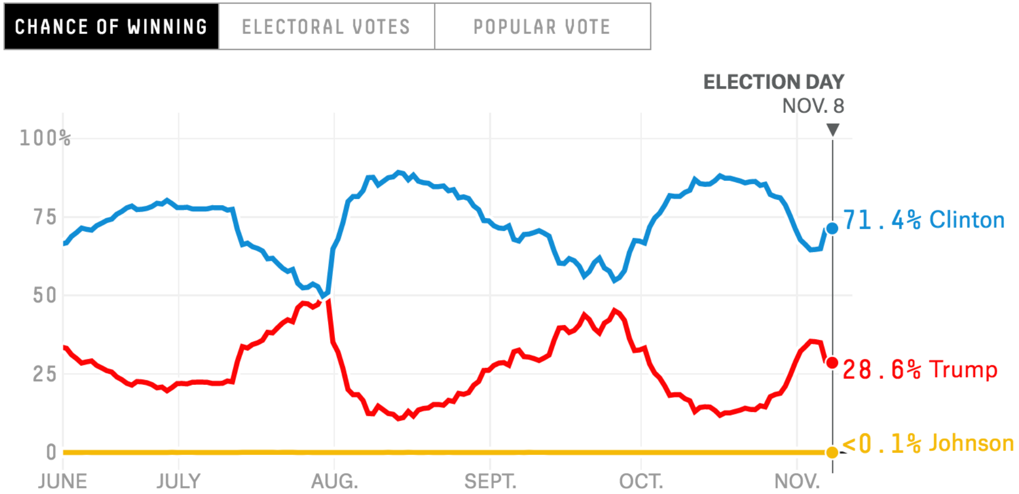
- What is the probability of rejecting the Null? -If we observe at least one back swan, we reject: Depends on how rare black swans are.
- A more complex null hypothesis: At least 50% of swans are black.

Win probability



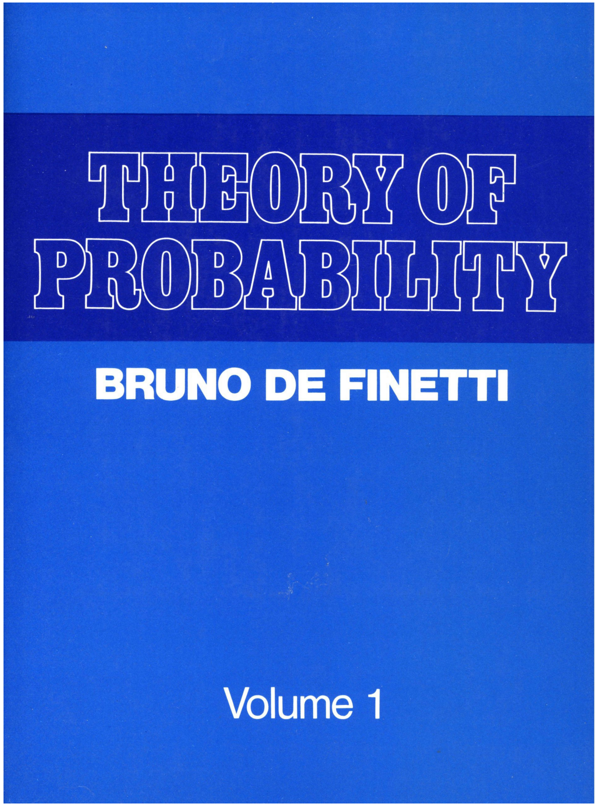
source: fivethirtyeight.com

Win probability



source: fivethirtyeight.com

Bayesian probability



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

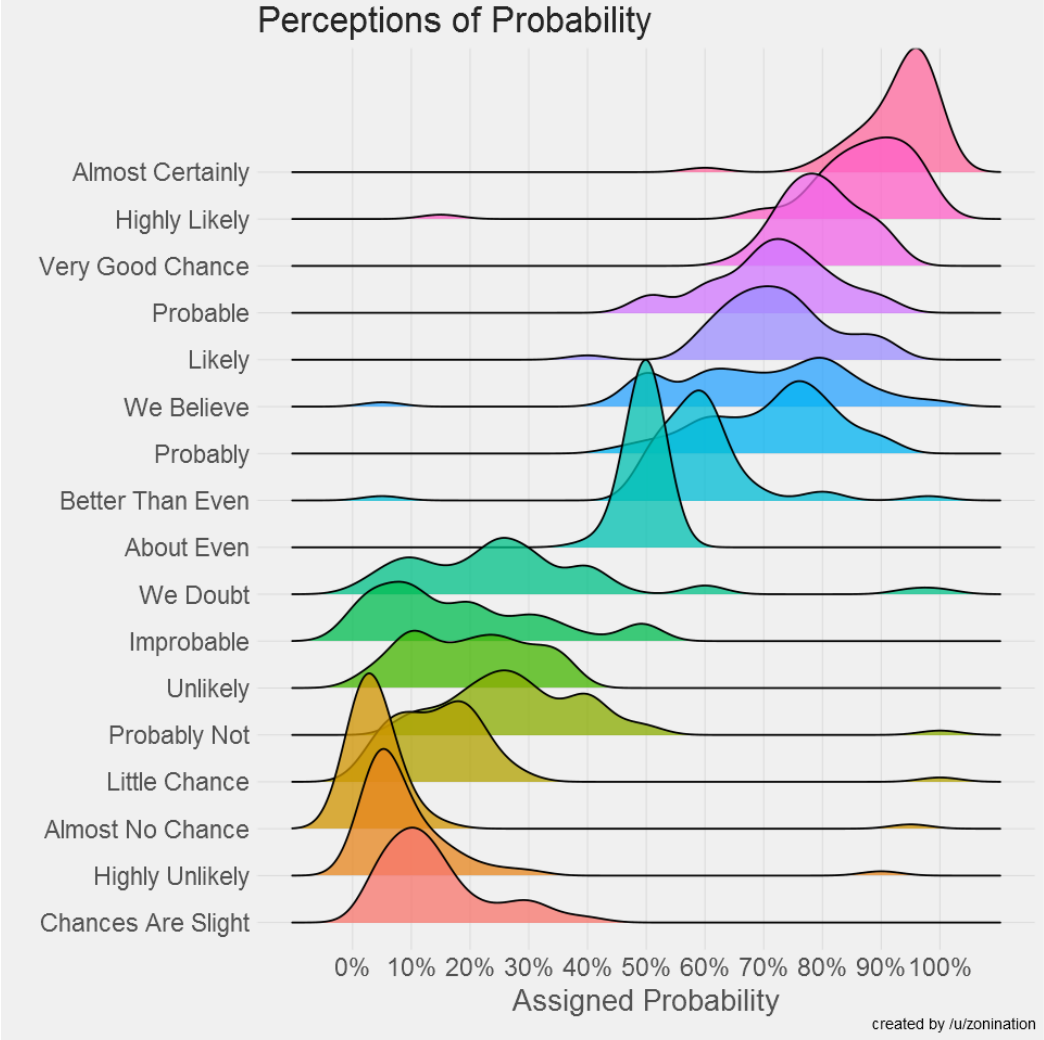
Bayesian probability

- de Finetti is arguing that probability is about *belief*
  - 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

Perceptions of Probability



source: https://github.com/zonination/perceptions

Why Bayesian statistics?

- Classic statistical toolbox may not be appropriate for all settings.
    - Inflexible and fragile
    - e.g. what if the assumptions of the test don't hold?
- Bayesian statistics provides a procedure for building our own tests / tools.
    - Design, build and refine procedures for you own models.
- A variety of powerful tools for inference with computer simulation
- Philosophy of science: quantifying degrees of belief often a more useful perspective than falsification

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

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

- $P(A \mid B)$ 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

Bayes' Rule for Bayesian Statistics

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

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

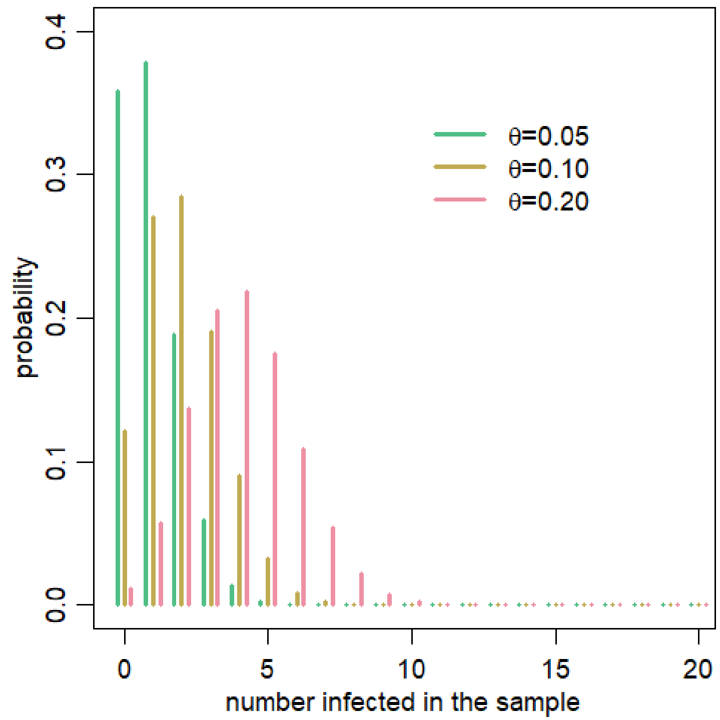
Example: Estimating COVID Infection Rates

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

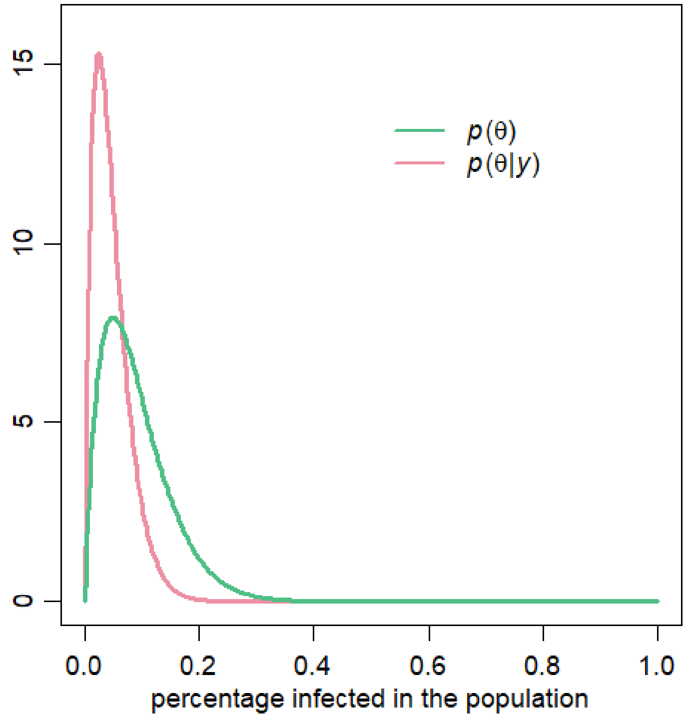
Example: Estimating Infection Rates



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?

Example: Estimating Infection Rates



Tentative syllabus

- One parameter models (binomial, poisson, and normal)
- Monte Carlo methods (i.e. simulation-based inference)
- Markov chain Monte Carlo (MCMC)
- Hierarchical modeling
- An introduction to probabilistic programming

Assignment

- Check Nectir, bookmark important links
- Start reviewing probability cheat sheet!
- Read chapters 1 and 2 of Bayes Rules