PSTAT 115: Bavesian Data Analysis

Professor Alexander Franks

2020-10-05

Class Resources

Required Textbook

PDF on course site

- A First Course in Bayesian Statistical Methods, Peter Hoff

Optional Textbooks

- Bayesian Computation With R, Jim Albert
- Bayesian Data Analysis, Andrew Gelman et al ← More advanced

Course Pages

- Piazza for all course related questions: https://piazza.com/ucsb/fall2020/pstat115
- Gradescope: https://www.gradescope.com/courses/183592
  - Assignment submission
  - Quizes

Exams?

Grades

- 35% - expect approximately 6 homeworks
- 20% - Take home midterm (Due November 9)
- 10% - Quizzes
- 5% - Participation
- 30% - Final exam (Due December 16)

Homework

- There will be approximately 6 homeworks (30% of your grade total)
- You will typically have 1-2 weeks to complete the homeworks
- You are allowed and encouraged to work with a partner
  - Add partners name to your assignment
- 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 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*.
- 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 ~~GauchoSpace~~ Gradescope
- R markdown code
- Generated PDF file
- Any supplementary files (e.g. write up for math problems)

All should be zipped up and we should be able to run it to obtain identical PDF file

Labs and Quizzes

- There will be a handful of "pop" quizzes throughout the quarter.
    - The quizzes will be on Gradescope.
    - You will have 10 minutes to take the quiz any time within 24 hours of the announcement.
    - 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 (10%) will be multiple choice and will test your comprehension of the basic concept.
- Participation (5%). Includes lecture attendance, section attendance, and piazza posts.

Class Policies

- All questions should be posted on Piazza, *not by email* (unless they are personal or grade-related)
- You can use the "search for teammates" tool on Piazza to find homework partners.

RStudio Cloud Service

- Log on to pstat115.lsit.ucsb.edu
  - Cloud based rstudio service
  - Log in with your UCSB NetID
- Use https://bit.ly/3kZ2sVr to sync new material
- Make sure you can write and compile an R markdown (`Rmd`) document online
- Text formatting is minimal but syntax is simple

Resources/
- Textbook
- Probability Cheat Sheet
- Review Material

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

- Updating priors

- Decision trees?

It's not frequentist.

Predictions based on new info.

What is Bayesian statistics?

1203 (Math Stat)

What is the version of statistics you already know?

Frequentist

- Parameters are constants
- Frequentist interpretation of prob.

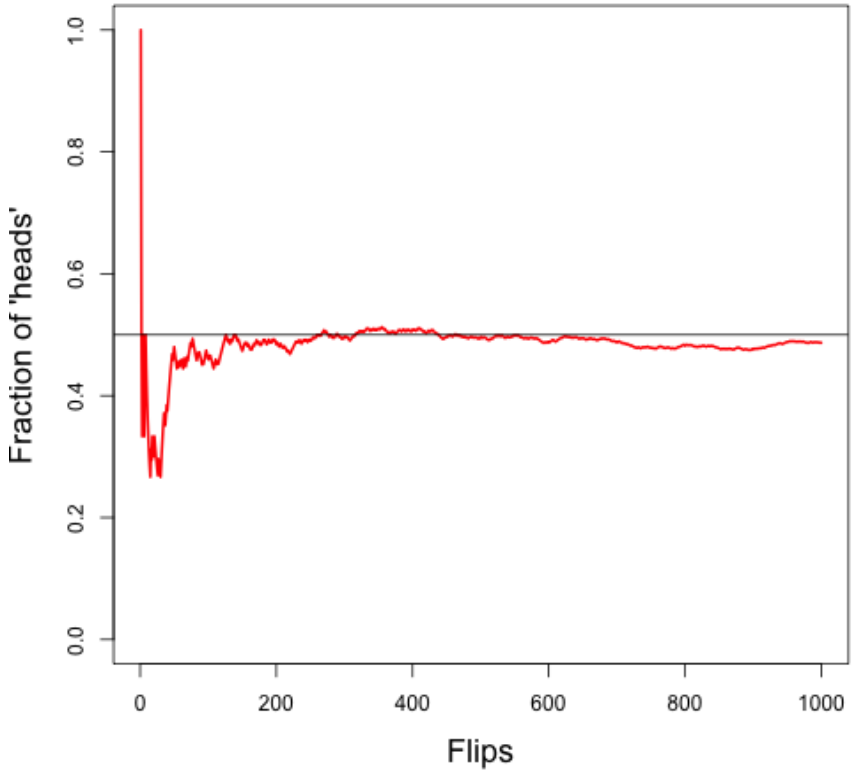
Frequentist statistics

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?

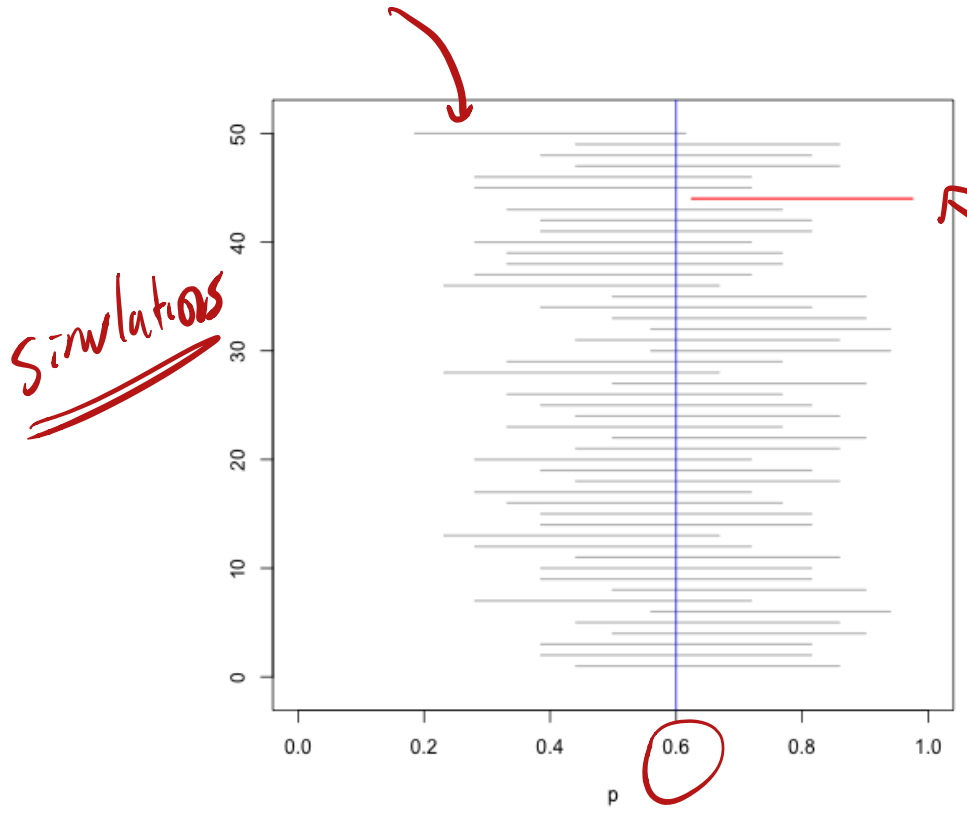
Frequentist probability

The probability of a coin landing on heads is 50%

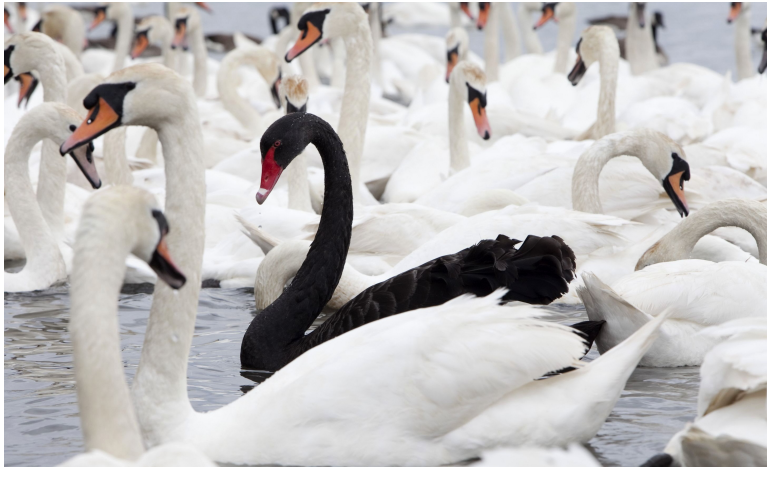


Confidence intervals

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



Falsification



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

Falsification

H_0 : "The Ivory-billed Woodpecker is extinct"

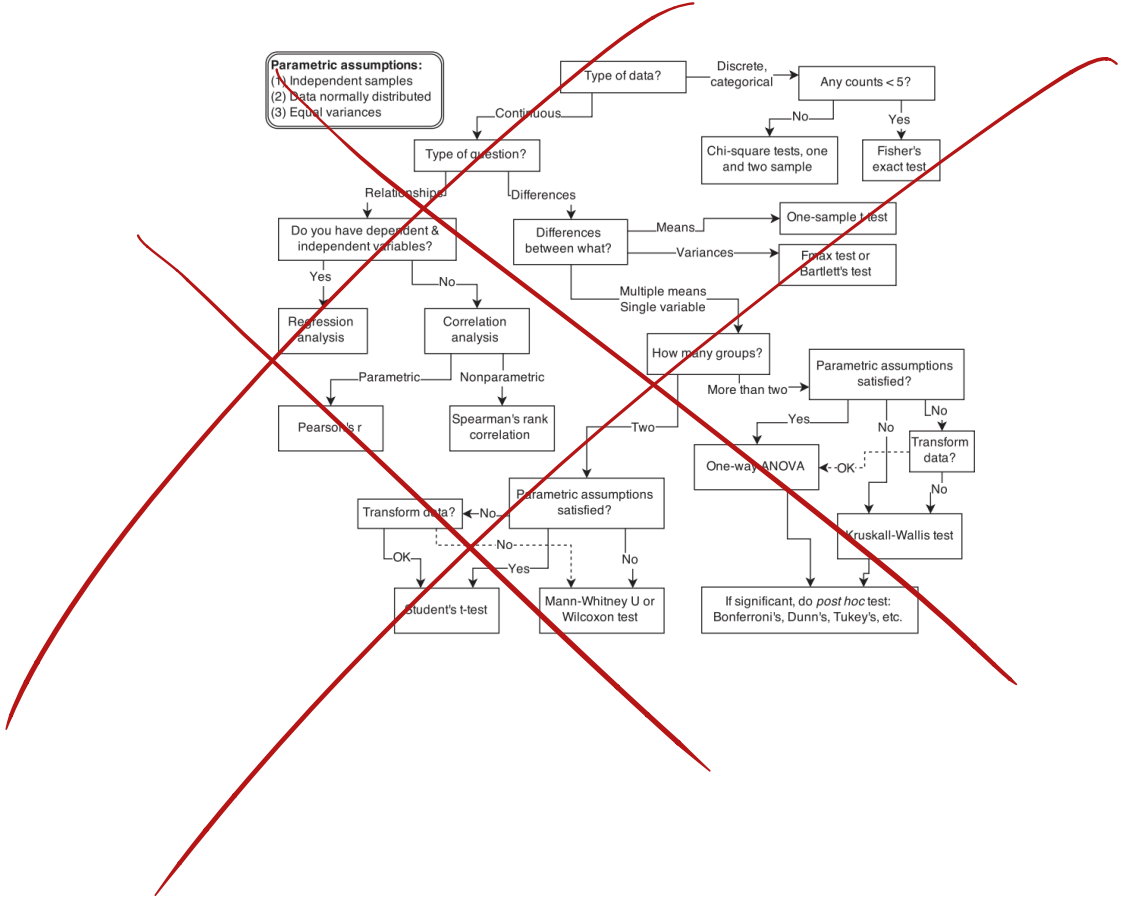
Falsification

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!

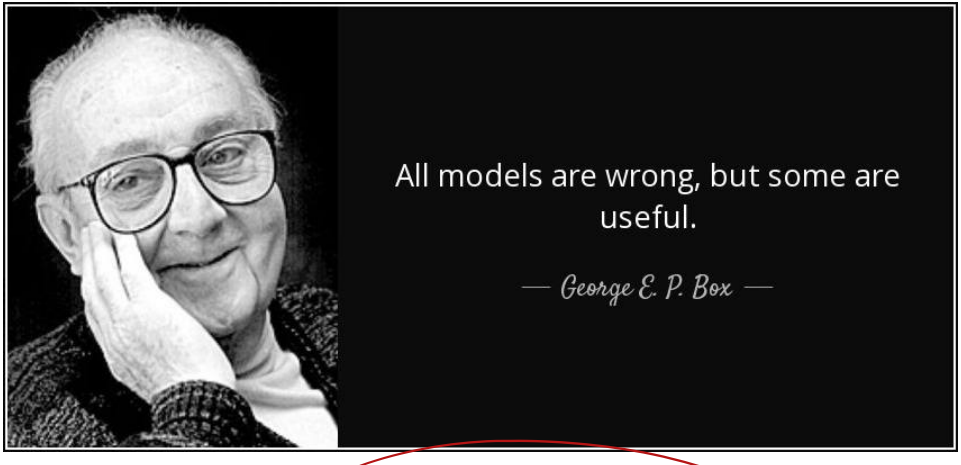
Significance Testing Flowchart



Alternative: focus on modeling!

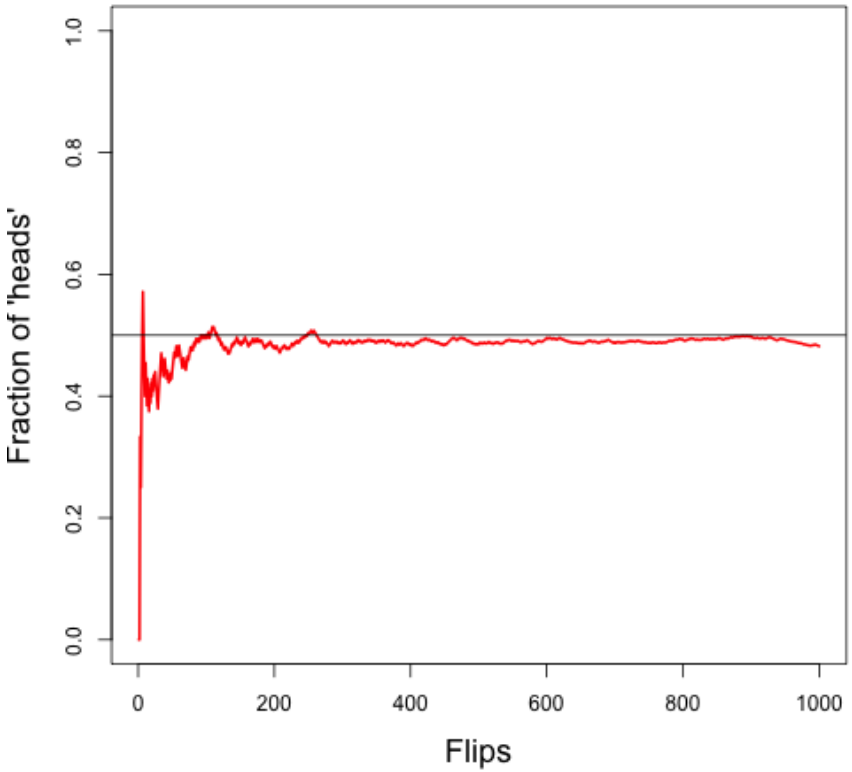
- A statistical model represents a set of assumption about how the data was generated.
- Models can still be used to develop statistical tests.
- Can also be used to make predictions or forecasts and describe sources of variability.
- Can (and should) be continuously refined and extended!

All models are wrong

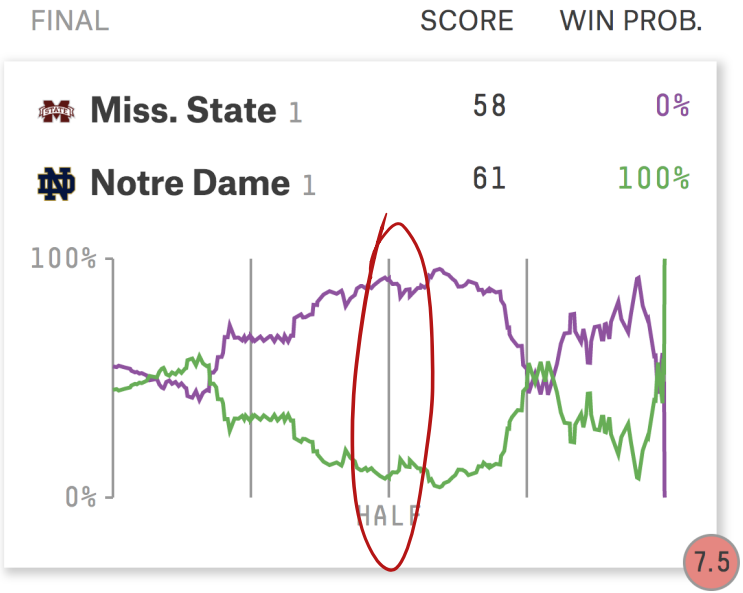


https://en.wikipedia.org/wiki/All_models_are_wrong

Frequentist probability



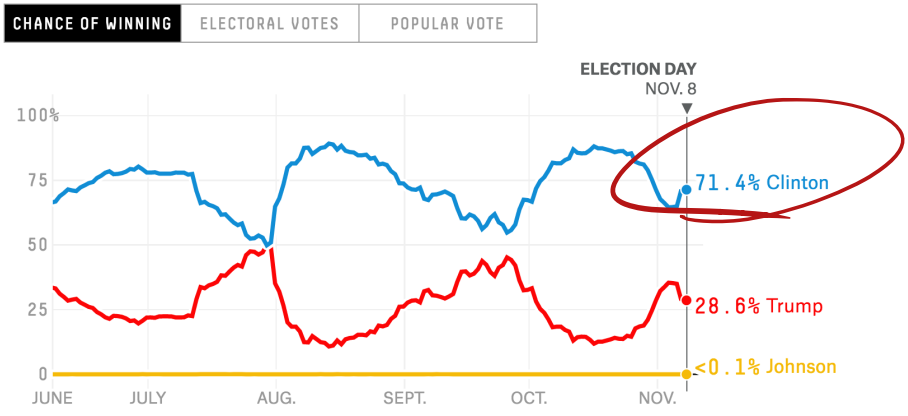
Win probability



source: fivethirtyeight.com

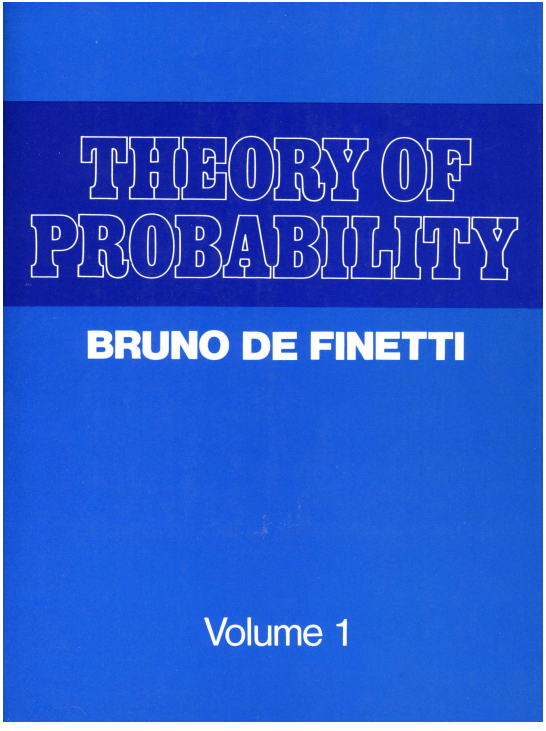
5% that Notre Dame wins (based on half time).

Win probability



source: fivethirtyeight.com

Bayesian probability



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

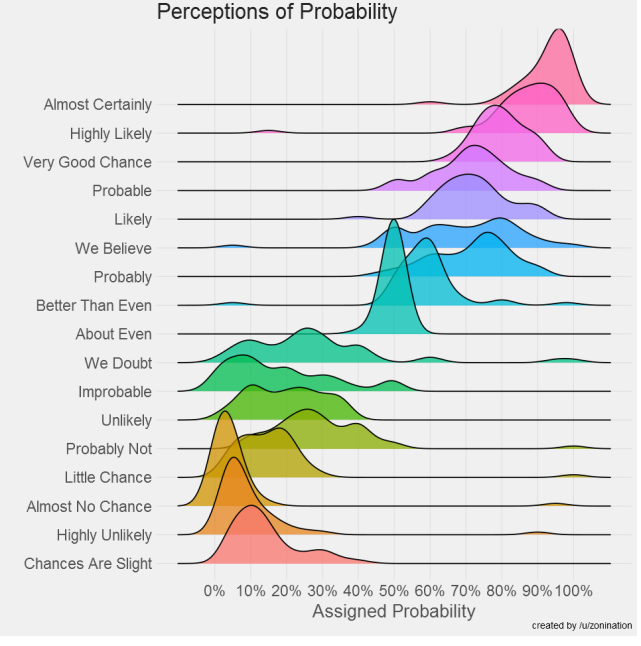
Bayesian probability

- 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



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!

n

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!

given data.

Key difference: θ is random!