Intro to Bayesian Methods

Rebecca C. Steorts

Bayesian Methods and Modern Statistics: STA 360/601

Lecture 1

- Course Webpage
- Syllabus
- LaTeX reference manual
- R markdown reference manual
- Please come to office hours for all questions.
  - Office hours are not a review period if you cannot come to class.
- Join Google group
- Graded on Labs/HWs, Exams.
  - Labs/HWs and Exams .R markdown format (it must compile).
  - Nothing late will be accepted.
  - You're lowest homework will be dropped.
- Announcements: Emails or in class.
- All your lab/homework assignments will be uploaded to Sakai.
- How to reach me and TAs – email or Google.

Expectations

- Class is optional but you are expected to know everything covered in lecture.
- Not everything will always be on the slides.
- 2 Exams: in class, timed. Closed book, closed notes. (Dates are on the syllabus).
- There are NO make up exams.
- Late assignments will not be accepted. Don't ask.
- Final exam: during finals week.
- You should be reading the book as we go through the material in class.

Expectations for Homework

- Your write ups should be clearly written.
- Proofs: show all details.
- Data analysis: clearly explain.
- For data analysis questions, don't just turn in code.
- Code must be well documented.
- Code style: https://google.github.io/styleguide/Rguide.xml
- For all homeworks, can use Markdown or LaTex. You must include all files that lead to your solutions (this includes code)!

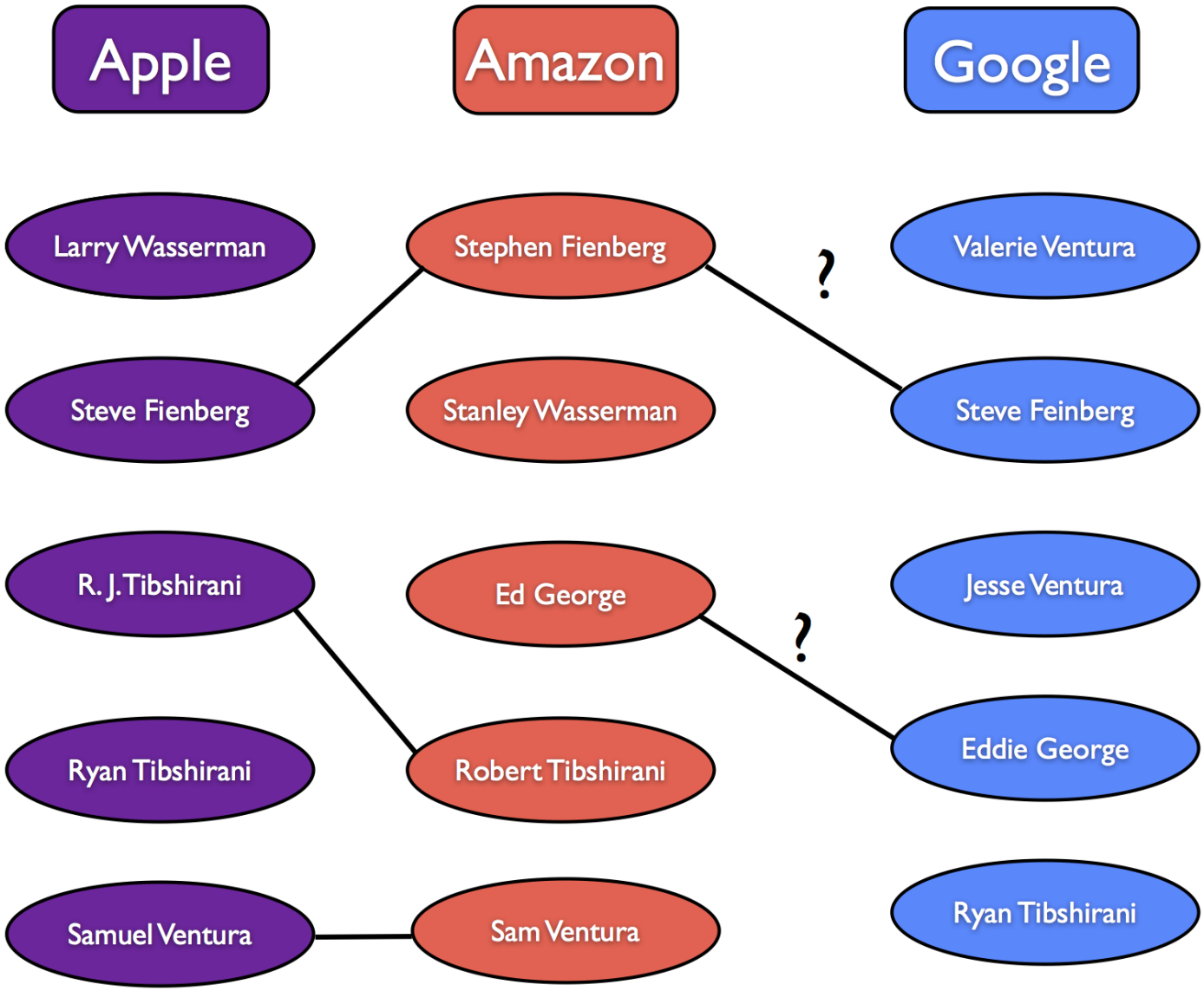
Things available to you!

- Come to office hours. We want to help you learn!
- Supplementary reading to go with the notes by yours truly. (Beware of typos).
  - Undergrad level notes
  - PhD level notes
  - Example form of write up in .Rmd on Sakai (Module 0).
  - You should have your homeworks graded and returned within one week by the TA's!

- Why should we learn about Bayesian concepts?
- Natural if thinking about unknown parameters as random.
- They naturally give a full distribution when we perform an update.
- We automatically get uncertainty quantification.
- Drawbacks: They can be slow and inconsistent.

Record linkage

Record linkage is the process of merging together noisy databases to remove duplicate entries.


Apple
Amazon
Google
Larry Wasserman
Stephen Fienberg
Valerie Ventura
?
Steve Fienberg
Stanley Wasserman
Steve Feinberg
R. J. Tibshirani
Ed George
Jesse Ventura
?
Ryan Tibshirani
Robert Tibshirani
Eddie George
Samuel Ventura
Sam Ventura
Ryan Tibshirani

240 Collins Dr
Pittsburgh PA 15235

50-54

412-793-3313

240 Collins Dr
Pittsburgh PA 15235

50-54

412-793-3313

537 N Neville St Apt 5d
Pittsburgh PA 15213

65+

412-683-5599

240 Collins Dr
Pittsburgh PA 15235

50-54

412-793-3313

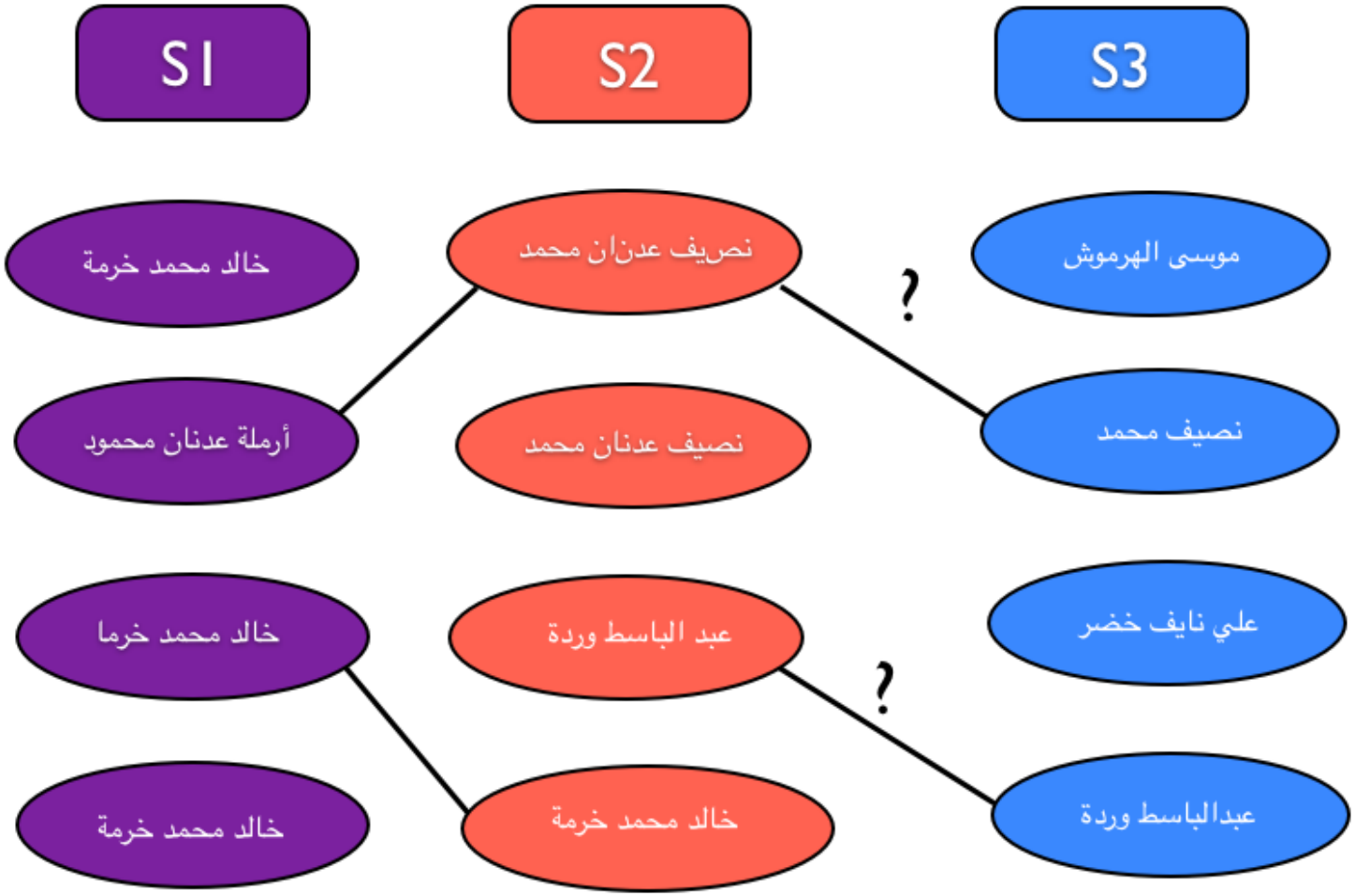
537 N Neville St Apt 5d
Pittsburgh PA 15213

65+

412-683-5599

These are clearly not the *same* Steve Fienberg!

Syrian Civil War



Bayesian Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.

Bayesian Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.
- G_ℓ: empirical distribution for field ℓ.

Bayesian Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.
- G_ℓ: empirical distribution for field ℓ.
- $W \sim F_\ell(w_0)$: $P(W = w) \propto \alpha_\ell(w) \exp[-c\, d(w, w_0)]$, where $d(\cdot, \cdot)$ is a string metric and $c > 0$.

Bayesian Model

- Define $\alpha_\ell(w) =$ relative frequency of w in data for field ℓ.
- G_ℓ: empirical distribution for field ℓ.
- $W \sim F_\ell(w_0)$: $P(W = w) \propto \alpha_\ell(w) \exp[-c\, d(w, w_0)]$, where $d(\cdot, \cdot)$ is a string metric and $c > 0$.

$$X_{ij\ell} \mid \lambda_{ij}, Y_{\lambda_{ij}\ell}, z_{ij\ell} \sim \begin{cases} \delta(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 0 \\ F_\ell(Y_{\lambda_{ij}\ell}) & \text{if } z_{ij\ell} = 1 \text{ and field } \ell \text{ is string-valued} \\ G_\ell & \text{if } z_{ij\ell} = 1 \text{ and field } \ell \text{ is categorical} \end{cases}$$

$$\begin{aligned} Y_{j'\ell} &\sim G_\ell \\ z_{ij\ell} \mid \beta_{i\ell} &\sim \text{Bernoulli}(\beta_{i\ell}) \\ \beta_{i\ell} &\sim \text{Beta}(a, b) \end{aligned}$$

$$\lambda_{ij} \sim \text{DiscreteUniform}(1, \ldots, N_{\max}), \quad \text{where } N_{\max} = \sum_{i=1}^{k} n_i$$

The model I showed you is very complicated.

This course will give you an intro to Bayesian models and methods.

Often Bayesian models are hard to work with, so we'll learn about approximations.

The above record linkage problem is one that needs such an approximation.

- "Bayesian" traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call "Bayes' Theorem".

- "Bayesian" traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call "Bayes' Theorem".

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

- "Bayesian" traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call "Bayes' Theorem".

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

- "Bayesian" traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call "Bayes' Theorem".

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

$p(\theta|x)$ posterior

- "Bayesian" traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call "Bayes' Theorem".

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

$p(\theta\|x)$	posterior
$p(x\|\theta)$	likelihood

- "Bayesian" traces its origin to the 18th century and English Reverend Thomas Bayes, who along with Pierre-Simon Laplace discovered what we now call "Bayes' Theorem".

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)p(\theta). \tag{1}$$

We can decompose Bayes' Theorem into three principal terms:

$p(\theta\|x)$	posterior
$p(x\|\theta)$	likelihood
$p(\theta)$	prior

Polling Example 2012

Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

Polling Example 2012

Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

- We take a random sample of 10 people in PA and find that 6 approve of President Obama.

Polling Example 2012

Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

- We take a random sample of 10 people in PA and find that 6 approve of President Obama.
- The national approval rating (Zogby poll) of President Obama in mid-December was 45%. We'll assume that in PA his approval rating is approximately 50%.

Polling Example 2012

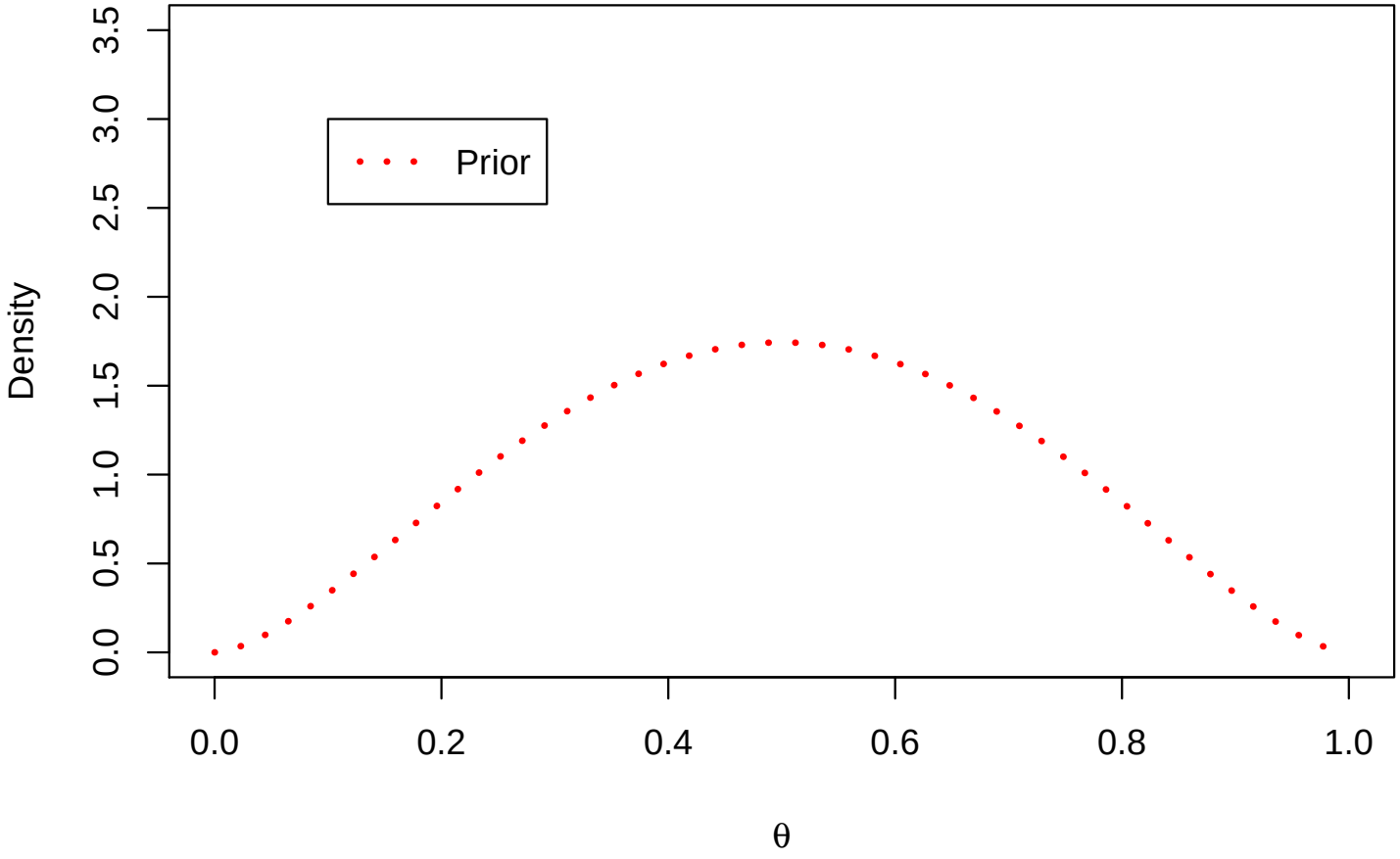
Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

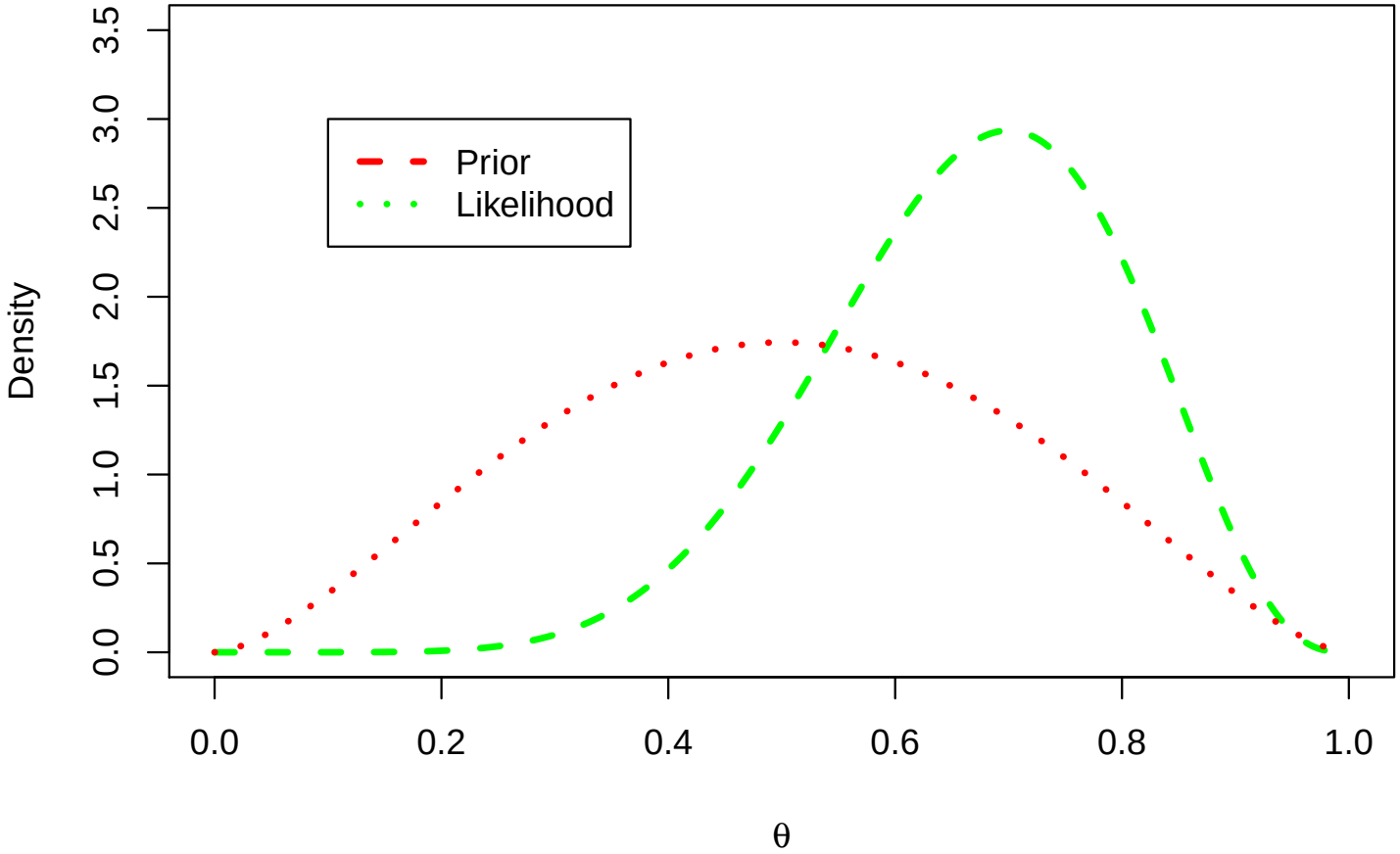
- We take a random sample of 10 people in PA and find that 6 approve of President Obama.
- The national approval rating (Zogby poll) of President Obama in mid-December was 45%. We'll assume that in PA his approval rating is approximately 50%.
- Based on this prior information, we'll use a Beta prior for θ and we'll choose a and b. (Won't get into this here).

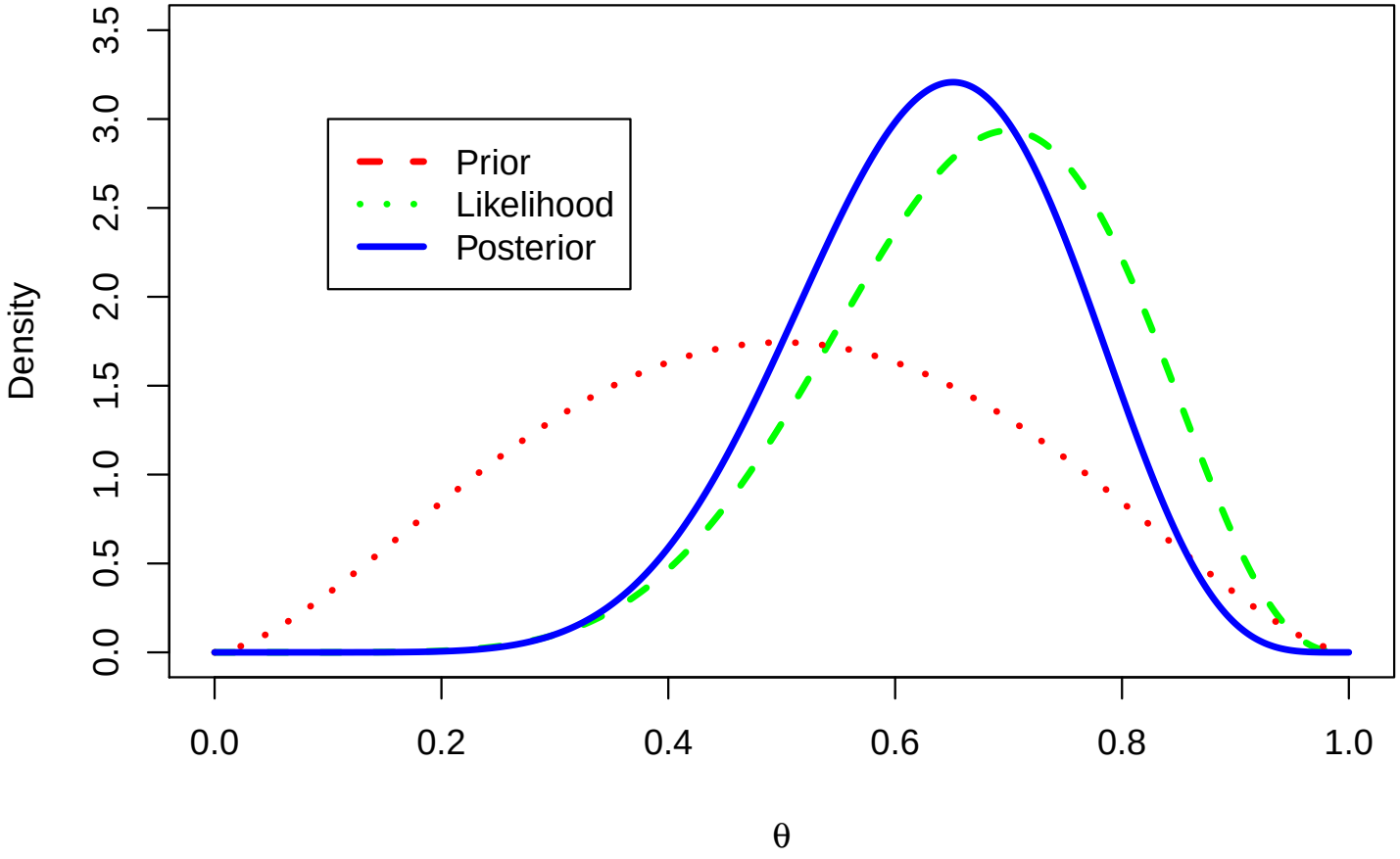
Polling Example 2012

Let's apply this to a real example! We're interested in the proportion of people that approve of President Obama in PA.

- We take a random sample of 10 people in PA and find that 6 approve of President Obama.
- The national approval rating (Zogby poll) of President Obama in mid-December was 45%. We'll assume that in PA his approval rating is approximately 50%.
- Based on this prior information, we'll use a Beta prior for θ and we'll choose a and b. (Won't get into this here).
- We can plot the prior and likelihood distributions in `R` and then see how the two mix to form the posterior distribution.


Prior
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ


Prior
Likelihood
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ


Prior
Likelihood
Posterior
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
θ

The basic philosophical difference between the frequentist and Bayesian paradigms is that

- Bayesians treat an unknown parameter θ as *random*.

The basic philosophical difference between the frequentist and Bayesian paradigms is that

- Bayesians treat an unknown parameter θ as *random*.
- Frequentists treat θ as unknown but *fixed*.

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

$$H_0 : \theta = 1/2, \qquad H_1 : \theta > 1/2$$

at a significance level of $\alpha = 0.05$. Suppose we observe the following sequence of flips:

heads, heads, heads, heads, heads, **tails** (5 heads, 1 tails)

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

$$H_0 : \theta = 1/2, \qquad H_1 : \theta > 1/2$$

at a significance level of $\alpha = 0.05$. Suppose we observe the following sequence of flips:

heads, heads, heads, heads, heads, **tails** (5 heads, 1 tails)

- To perform a frequentist hypothesis test, we must define a random variable to describe the data.

Stopping Rule

Let θ be the probability of a particular coin landing on heads, and suppose we want to test the hypotheses

$$H_0 : \theta = 1/2, \qquad H_1 : \theta > 1/2$$

at a significance level of $\alpha = 0.05$. Suppose we observe the following sequence of flips:

heads, heads, heads, heads, heads, **tails** (5 heads, 1 tails)

- To perform a frequentist hypothesis test, we must define a random variable to describe the data.
- The proper way to do this depends on exactly which of the following two experiments was actually performed:

- Suppose the experiment is "**Flip six times and record the results.**"

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 5) \\ &= P_{\theta=1/2}(X = 5) + P_{\theta=1/2}(X = 6) \end{aligned}$$

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 5) \\ &= P_{\theta=1/2}(X = 5) + P_{\theta=1/2}(X = 6) \\ &= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.109375 > 0.05. \end{aligned}$$

- Suppose the experiment is "**Flip six times and record the results.**"
  - X counts the number of heads, and $X \sim \text{Binomial}(6, \theta)$.
  - The observed data was $x = 5$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 5) \\ &= P_{\theta=1/2}(X = 5) + P_{\theta=1/2}(X = 6) \\ &= \frac{6}{64} + \frac{1}{64} = \frac{7}{64} = 0.109375 > 0.05. \end{aligned}$$

So we fail to reject H_0 at $\alpha = 0.05$.

- Suppose now the experiment is "**Flip until we get tails.**"

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\text{p-value} \quad = \quad P_{\theta=1/2}(X \geq 6)$$

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\begin{aligned} \text{p-value} &= P_{\theta=1/2}(X \geq 6) \\ &= 1 - P_{\theta=1/2}(X < 6) \end{aligned}$$

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\begin{aligned}
\text{p-value} &= P_{\theta=1/2}(X \geq 6) \\
&= 1 - P_{\theta=1/2}(X < 6) \\
&= 1 - \sum_{x=1}^{5} P_{\theta=1/2}(X = x)
\end{aligned}$$

- Suppose now the experiment is "**Flip until we get tails.**"
  - X counts the number of the flip on which the first tails occurs, and $X \sim$ Geometric$(1-\theta)$.
  - The observed data was $x = 6$, and the p-value of our hypothesis test is

$$\begin{aligned}
\text{p-value} &= P_{\theta=1/2}(X \geq 6) \\
&= 1 - P_{\theta=1/2}(X < 6) \\
&= 1 - \sum_{x=1}^{5} P_{\theta=1/2}(X = x) \\
&= 1 - \left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32}\right) = \frac{1}{32} = 0.03125 < 0.05.
\end{aligned}$$

So we reject H_0 at $\alpha = 0.05$.

- The conclusions differ, which seems strikes *some people* as absurd.

- The conclusions differ, which seems strikes *some people* as absurd.
- P-values aren't close—one is 3.5 times as large as the other.

- The conclusions differ, which seems strikes *some people* as absurd.
- P-values aren't close—one is 3.5 times as large as the other.
- The result our hypothesis test depends on whether we would have stopped flipping if we had gotten a tails sooner.

- The conclusions differ, which seems strikes *some people* as absurd.
- P-values aren't close—one is 3.5 times as large as the other.
- The result our hypothesis test depends on whether we would have stopped flipping if we had gotten a tails sooner.
- The tests are dependent on what we call the *stopping rule*.

- The likelihood for the actual value of x that was observed is the same for both experiments (up to a constant):

$$p(x|\theta) \propto \theta^5(1-\theta).$$

- The likelihood for the actual value of x that was observed is the same for both experiments (up to a constant):

$$p(x|\theta) \propto \theta^5(1-\theta).$$

- A Bayesian approach would take the data into account only through this likelihood.

- The likelihood for the actual value of x that was observed is the same for both experiments (up to a constant):

$$p(x|\theta) \propto \theta^5(1-\theta).$$

- A Bayesian approach would take the data into account only through this likelihood.
- This would provide the same answers regardless of which experiment was being performed.

The Bayesian analysis is independent of the stopping rule since it only depends on the likelihood (show this at home!).

Hierarchical Bayesian Models

In a hierarchical Bayesian model, rather than specifying the prior distribution as a single function, we specify it as a hierarchy.

Hierarchical Bayesian Models

$$
\begin{aligned}
X|\theta &\sim f(x|\theta) \\
\Theta|\gamma &\sim \pi(\theta|\gamma) \\
\Gamma &\sim \phi(\gamma),
\end{aligned}
$$

where we assume that $\phi(\gamma)$ is known and not dependent on any other unknown *hyperparameters*.

Conjugate Distributions

Let F be the class of sampling distributions $p(y|\theta)$.

Conjugate Distributions

Let F be the class of sampling distributions $p(y|\theta)$.

- Then let P denote the class of prior distributions on θ.

Conjugate Distributions

Let F be the class of sampling distributions $p(y|\theta)$.

- Then let P denote the class of prior distributions on θ.
- Then P is said to be conjugate to F if for every $p(\theta) \in P$ and $p(y|\theta) \in F$, $p(\theta \mid y) \in P$.

Simple definition: A family of priors such that, upon being multiplied by the likelihood, yields a posterior in the same family.

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\pi(\theta|x) \quad \propto \quad p(x|\theta)p(\theta)$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned} \pi(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \end{aligned}$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)p(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1}
\end{aligned}$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)p(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \implies
\end{aligned}$$

Beta-Binomial

If $X|\theta$ is distributed as binomial(n, θ), then a conjugate prior is the beta family of distributions, where we can show that the posterior is

$$\begin{aligned}
\pi(\theta|x) &\propto p(x|\theta)p(\theta) \\
&\propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^x(1-\theta)^{n-x}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{x+a-1}(1-\theta)^{n-x+b-1} \implies
\end{aligned}$$

$$\theta|x \sim \text{Beta}(x+a, n-x+b).$$