Lecture 2: One Parameter Models

Professor Alexander Franks

2020-10-14

Announcements

- Reading: Chapter 3, Hoff
- Homework due: October 18, at midnight
  - + Submit homework1.Rmd and homework1.Pdf only
- Office Hours
- Quizzes

Bayesian Inference

- In frequentist inference, θ is treated as a fixed unknown constant
- In Bayesian inference, θ is treated as a random variable
- Need to specify a model for the joint distribution $p(y, \theta) = p(y \mid \theta)p(\theta)$

Bayesian Inference in a Nutshell

New

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

Same deal as before.

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!

Bayes' Rule

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

- $P(A \mid B)$ 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

likelihood / Sampling

Updated Belief

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

Prior distribution

- $P(\theta \mid y)$ 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

Hard to choose.

$$\boxed{P(\theta|y) \propto P(y|\theta)P(\theta)}$$

B/C $\frac{1}{P(y)}$ is constant in θ

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)

The posterior is proportional to the likelihood times the prior!

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

θ true population frac. that's infected.

$Y \sim Bin(20, \theta)$

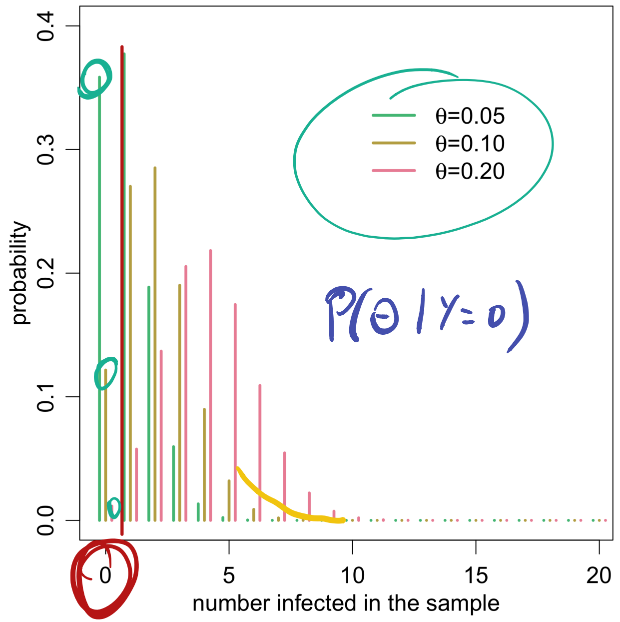
$\hat{\theta}_{MLE}: \frac{Y}{20}$. What about Bayes?

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

Sampling Model $P(Y|\theta)$



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

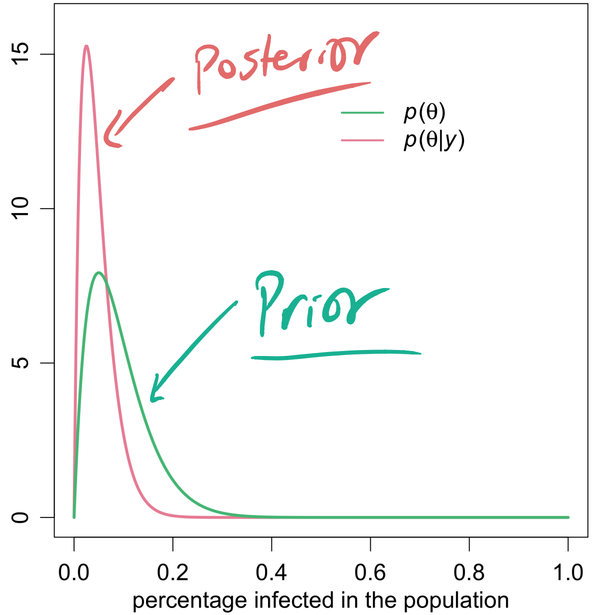


Table of Relevant Quantities

Bayesian vs Frequentist

- In frequentist inference, unknown parameters treated as constants
  - Estimators are random (due to sampling variability) Y
  - Asks: what would I expect to see if I repeated the experiment?"

(counterfactual world)

Bayesian vs Frequentist

- In frequentist inference, unknown parameters treated as constants
  - Estimators are random (due to sampling variability)
  - Asks: 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)
  - Asks: "what do I *believe* are plausible values for the unknown parameters given the data?"
  - Who cares what might have happened, focus on what *did* happen by conditioning on observed data. $P(\theta | y)$

Example: estimating shooting skill in basketball

- On November 18, 2017, an NBA basketball player, Robert Covington, had made 49 out of 100 three point shot attempts.
- At that time, his three point field goal percentage, 0.49, was the best in the league and would have ranked in the top ten all time
- How can we estimate his true shooting skill?
  - Think of "true shooting skill" as the fraction he would make if he took infinitely many shots

$$\hat{\theta}_{MLE} = \frac{49}{100} = .49$$

Example: estimating shooting skill in basketball

- Assume every shot is independent (reasonable) and identically distributed (less reasonble?)
- Let $Y \sim \text{Bin}(\underset{}{\overset{=100}{n}}, \theta)$ where θ corresponds to his true skill
- Frequentist inference tells us that the maximum likelihood estimate is simply $\frac{y}{n} = 49/100 = 0.49$
- What would our estimates be if we use Bayesian inference?
    - What properties do we want for our prior distribution?

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

$$\theta \in [0, 1], \quad P(\theta) > 0 \text{ for all } \theta \in [0,1]$$

Cromwell's Rule

The use of priors placing a probability of 0 or 1 on events should be avoided except where those events are excluded by logical impossibility.

If a prior places probabilities of 0 or 1 on an event, then no amount of data can update that prior.

> I beseech you, in the bowels of Christ, think it possible that you may be mistaken.
>
> --- Oliver Cromwell

Cromwell's Rule

> Leave a little probability for the moon being made of green cheese; it can be as small as 1 in a million, but have it there since otherwise an army of astronauts returning with samples of the said cheese will leave you unmoved.
>
> --- Dennis Lindley (1991)

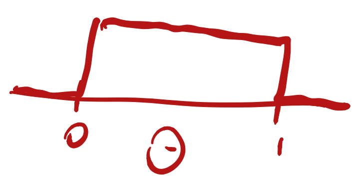
If $p(\theta = a) = 0$ for a value of a, then the posterior distribution is always zero, regardless of what the data says

$$p(\theta = a|y) \propto p(y|\theta = a)p(\theta = a) = 0$$

Example: estimating shooting skill in basketball

- Assume every shot is independent (reasonable) and identically distributed (less reasonble?)
- Let $Y \sim \text{Bin}(n, \theta)$ where θ corresponds to his true skill
- Frequentist inference tells us that the maximum likelihood estimate is simply $\frac{y}{n} = 49/100 = 0.49$
- What would our estimates be if we use Bayesian inference?
  - If our prior reflects "complete ignorance" about basketball?
  - What if we want to incorporate prior domain knowledge?

Uniform? $p(\theta) \propto \mathbb{1}[\theta \in [0,1]]$



$$P(\theta | y) \propto L(\theta) P(\theta)$$

Unif: $P(\theta) \propto \mathbb{1}[\theta \in [0,1]]$

$$L(\theta) \propto \theta^{y} (1-\theta)^{n-y}$$

$$P(\theta | y) \propto \theta^{y} (1-\theta)^{n-y} \mathbb{1}[\theta \in [0,1]]$$

$$\int_0^1 k\, \theta^{y} (1-\theta)^{n-y} \, d\theta = 1$$

Beta Distribution

$P(\theta | y) = \text{Beta}(\quad , \quad)$

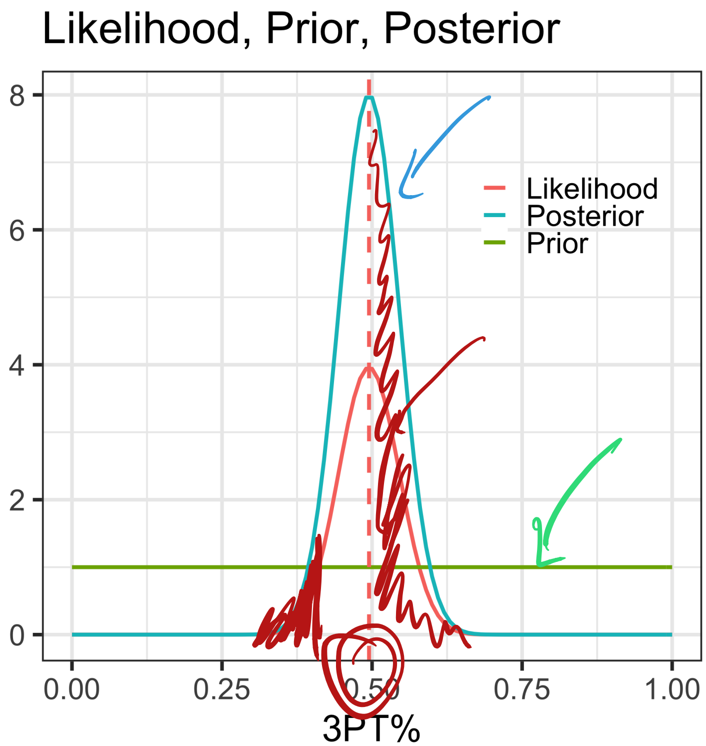
The Binomial Model

- The uniform prior: $p(\theta) = \text{Unif}(0, 1) = \mathbf{1}\{\theta \in [0, 1]\}$
  - A "non-informative" prior
- Posterior: $p(\theta \mid y) \propto \underbrace{\theta^y (1-\theta)^{n-y}}_{\text{likelihood}} \times \underbrace{\mathbf{1}\{\theta \in [0, 1]\}}_{\text{prior}}$
- The above posterior density is is a density over θ.

The Binomial Model

- The uniform prior: $p(\theta) = \text{Unif}(0,1) = \mathbf{1}\{\theta \in [0,1]\}$
  - A "non-informative" prior
- Posterior: $p(\theta \mid y) \propto \underbrace{\theta^y(1-\theta)^{n-y}}_{\text{likelihood}} \times \underbrace{\mathbf{1}\{\theta \in [0,1]\}}_{\text{prior}}$
- The above posterior density is is a density over θ.
- $p(\theta \mid y) \sim \text{Beta}(\underline{y+1}, \underline{n-y+1}) = \underbrace{\frac{\Gamma(n)}{\Gamma(n-y)\Gamma(y)}}_{= k} \underbrace{\theta^y(1-\theta)^{n-y}}_{}$

Example: estimating shooting skill in basketball



Posterior is proportional to the likelihood

Summarizing Posterior Results

- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

Summarizing Posterior Results

- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

- *Point estimates:* posterior mean or mode:

    - $E[\theta \mid y] = \int_{\Theta} \theta p(\theta \mid y) d\theta$ (the posterior mean)

    - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

      ∂

Summarizing Posterior Results

- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?

- *Point estimates:* posterior mean or mode:

    - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (the posterior mean)

    - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)

- Posterior variance: $\mathrm{Var}[\theta \mid y] = \int_\Theta (\theta - E[\theta \mid y])^2 p(\theta \mid y) d\theta$

Summarizing Posterior Results

- An entire distribution describes our beliefs about the value for θ. How can we summarize these beliefs?
- *Point estimates:* posterior mean or mode:
  - $E[\theta \mid y] = \int_\Theta \theta p(\theta \mid y) d\theta$ (the posterior mean)
  - $\arg\max p(\theta \mid y)$ (*maximum a posteriori* estimate)
- Posterior variance: $\mathrm{Var}[\theta \mid y] = \int_\Theta (\theta - E[\theta \mid y])^2 p(\theta \mid y) d\theta$
- Posterior credible intervals: for any region $R(y)$ of the parameter space compute the probability that θ is in that region: $p(\underbrace{\theta}_{\text{Random}} \in \underbrace{R(y)}_{\text{constant}})$