Bayesian Estimation

Foundations of Data Analysis

February 7, 2019

All models are wrong, but some are useful.

— George Box

Frequentist vs. Bayesian Statistics

Frequentist: θ is a parameter

$$L(\theta\,;\,x_1,\ldots,x_n) = \prod_{i=1}^{n} p(x_i\,;\theta)$$

Bayesian: θ is a random variable

$$p(\theta \mid x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n \mid \theta)p(\theta)}{p(x_1,\ldots,x_n)}$$

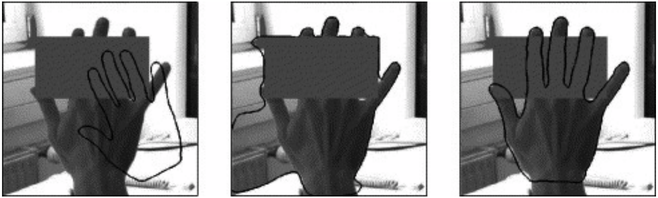
Why is Random θ Important?

- The prior, $p(\theta)$, let's us use our **beliefs, previous experience, or desires** in the model.
- We can make **probabilistic statements** about θ (e.g., mean, variance, quantiles, etc.).
- If θ is one of several competing **hypotheses**, we can assign it a probability.
- We can make **probabilistic predictions** of the next data point, $\hat{x}$, using

$$p(\hat{x} \mid x_1, \ldots, x_n) = \int p(\hat{x} \mid \theta)\, p(\theta \mid x_1, \ldots, x_n)\, d\theta$$

But Bayesian Analysis is *Subjective*, Right?

- Not necessarily (we'll cover noninformative priors)
- Frequentist models make assumptions, too!
- Whether using frequentist or Bayesian models, **always check the assumptions you make.**
- Sometimes prior knowledge is a good thing.



Cremers, et al., Pattern Recognition, 2003

MLE of Bernoulli Proportion

$$X_1, \ldots, X_n \sim \mathrm{Ber}(\theta)$$

$$L(\theta \,|\, x_1, \ldots, x_n) = \theta^k (1-\theta)^{n-k}, \quad \text{where } k = \sum_i x_i$$

$$\begin{aligned}
\frac{dL}{d\theta} &= k\theta^{k-1}(1-\theta)^{n-k} - (n-k)\theta^k(1-\theta)^{n-k-1} \\
&= \left(k(1-\theta) - (n-k)\theta\right)\theta^{k-1}(1-\theta)^{n-k-1} \\
&= (k - n\theta)\,\theta^{k-1}(1-\theta)^{n-k-1}
\end{aligned}$$

$$\frac{dL}{d\theta}\left(\hat{\theta}\right) = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{k}{n}$$

Bayesian Inference of a Bernoulli Proportion

Let's give θ a uniform prior: $\theta \sim \mathrm{Unif}(0, 1)$

$$p(\theta) = 1, \quad \text{for } \theta \in [0, 1]$$

Posterior:

$$\begin{aligned} p(\theta \,|\, x_1, \ldots, x_n) &= \frac{p(x_1, \ldots, x_n \,|\, \theta) p(\theta)}{p(x_1, \ldots, x_n)} \\ &= \frac{p(x_1, \ldots, x_n \,|\, \theta)}{p(x_1, \ldots, x_n)} \end{aligned}$$

Bayesian Inference of a Bernoulli Proportion

Just need the denominator (normalizing constant):

$$\begin{aligned} p(x_1, \ldots, x_n) &= \int_0^1 p(x_1, \ldots, x_n \,|\, \theta) p(\theta) d\theta \\ &= \int_0^1 \theta^k (1-\theta)^{n-k} d\theta \\ &= \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} \end{aligned}$$

Resulting posterior is:

$$p(\theta \,|\, x_1, \ldots, x_n) = \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} \theta^k (1-\theta)^{n-k}$$

Beta Distribution

$\theta \sim \mathrm{Beta}(\alpha, \beta)$ PDF:

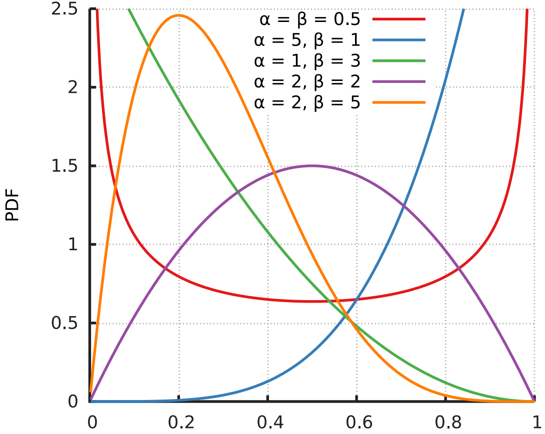
$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

So, posterior of Bernoulli with Uniform prior is $\theta \mid x \sim \mathrm{Beta}(k+1, n-k+1)$.

Also notice that $\mathrm{Beta}(1,1)$ is equivalent to $\mathrm{Unif}(0,1)$.

Mode: $\max p(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$

Beta pdf



See `https://en.wikipedia.org/wiki/Beta_distribution`

Bernoulli Likelihood with Beta Prior

$X_1, \ldots, X_n \sim \mathrm{Ber}(\theta)$
$\theta \sim \mathrm{Beta}(\alpha, \beta)$

Posterior:

$$\begin{aligned} p(\theta \mid x_1, \ldots, x_n) &\propto p(x_1, \ldots, x_n \mid \theta) p(\theta; \alpha, \beta) \\ &\propto \theta^k (1-\theta)^{n-k} \, \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &= \theta^{k+\alpha-1} (1-\theta)^{n-k+\beta-1} \end{aligned}$$

So, posterior dist. of θ is $\mathrm{Beta}(k+\alpha, n-k+\beta)$.

Conjugate Priors

Definition

Given a family (functional form) of likelihoods, $p(x \mid \theta)$, a **conjugate prior** $p(\theta; \alpha)$ is one in which the resulting posterior $p(\theta | x_1, \ldots, x_n; \alpha)$ has the same functional form as the prior.

- Conjugate priors result in closed-form posteriors.
- Often good approximation to what we want to model.
- Sometimes too simplistic, but provide building blocks for multivariate models.

Posterior Prediction for Bernoulli

Start with uniform prior: $\theta \sim \mathrm{Beta}(1,1)$

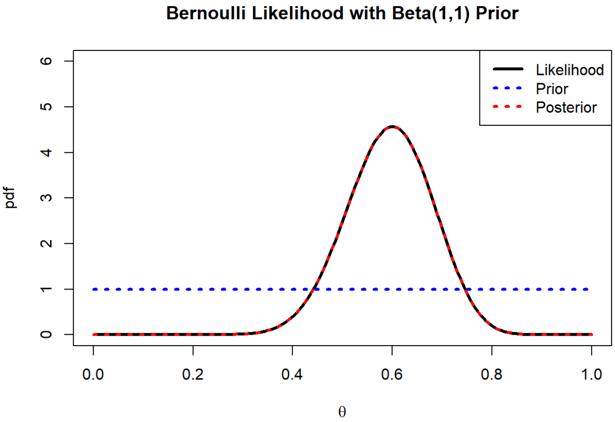
$$\begin{aligned}
p(\tilde{x} \mid k) &= \int_0^1 p(\tilde{x} \mid \theta, k)\, p(\theta \mid k)\, d\theta \\
&= \int_0^1 p(\tilde{x} \mid \theta)\, p(\theta \mid k)\, d\theta \\
&= \int_0^1 \theta\, p(\theta \mid k)\, d\theta \\
&= \mathrm{E}[\theta \mid k] \\
&= \frac{k+1}{n+2}
\end{aligned}$$

Posterior Prediction for Bernoulli

For general Beta prior: $\theta \sim \mathrm{Beta}(\alpha, \beta)$

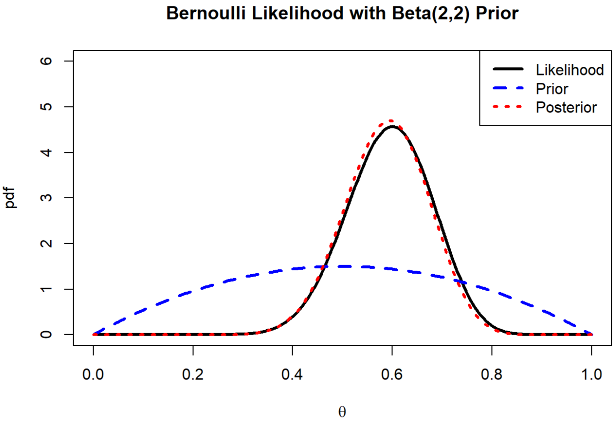
$$
\begin{aligned}
p(\tilde{x} \mid k) &= \mathrm{E}[\theta \mid k] \\
&= \frac{k + \alpha}{n + \alpha + \beta}
\end{aligned}
$$

Example

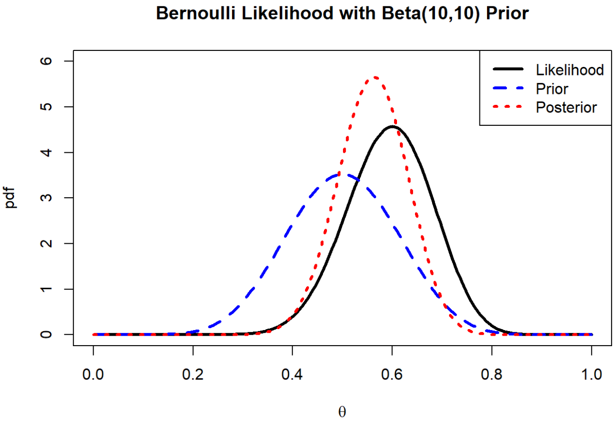


Bernoulli Likelihood with Beta(1,1) Prior
Likelihood
Prior
Posterior
pdf
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
θ

Example

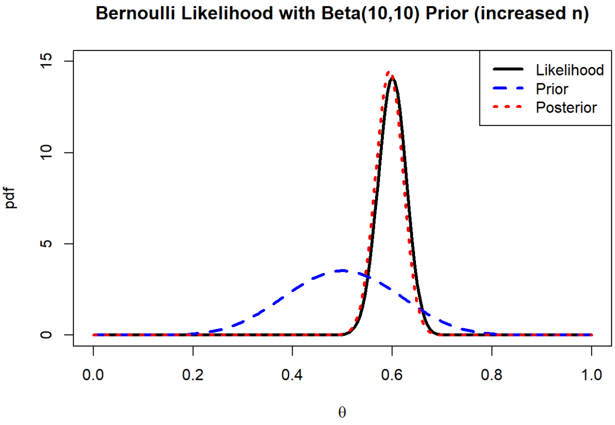

Bernoulli Likelihood with Beta(2,2) Prior
Likelihood
Prior
Posterior
pdf
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
θ

Example



Bernoulli Likelihood with Beta(10,10) Prior
Likelihood
Prior
Posterior
pdf
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
θ

Example


Bernoulli Likelihood with Beta(10,10) Prior (increased n)
Likelihood
Prior
Posterior
pdf
θ
0
5
10
15
0.0
0.2
0.4
0.6
0.8
1.0

Laplace's Analysis of Birth Rates

Mémoire sur les probabilités (1778)

`http://cerebro.xu.edu/math/Sources/Laplace/`

Problem: Boys were born at a consistently, but only slightly, higher rate than girls in Paris. Was this a real effect or just due to chance?

\# Boys: $k = 251527$ \# Girls: $n - k = 241945$

Solution: Model the proportion of boys as the posterior: $\theta \,|\, k \sim \mathrm{Beta}(251528, 241946)$. Then,

$$P(\theta \leq 0.5 \,|\, k) = F_{\theta|k}(0.5) = 1.15 \times 10^{-42}$$