Bayesian Methods

David S. Rosenberg

New York University

March 20, 2018

Contents

Classical Statistics

Parametric Family of Densities

- A **parametric family of densities** is a set

$$\{p(y \mid \theta) : \theta \in \Theta\},$$

  - where $p(y \mid \theta)$ is a density on a **sample space** $\mathcal{Y}$, and
  - θ is a **parameter** in a [finite dimensional] **parameter space** Θ.

- This is the common starting point for a treatment of classical or Bayesian statistics.

Density vs Mass Functions

- In this lecture, whenever we say "density", we could replace it with "mass function."
- Corresponding integrals would be replaced by summations.
- (In more advanced, measure-theoretic treatments, they are each considered densities w.r.t. different base measures.)

Frequentist or "Classical" Statistics

- Parametric family of densities

$$\{p(y \mid \theta) \mid \theta \in \Theta\}.$$

- Assume that $p(y \mid \theta)$ governs the world we are observing, for some $\theta \in \Theta$.
- If we knew the right $\theta \in \Theta$, there would be no need for statistics.
- Instead of θ, we have data $\mathcal{D}$: $y_1, \ldots, y_n$ sampled i.i.d. $p(y \mid \theta)$.
- Statistics is about how to get by with $\mathcal{D}$ in place of θ.

Point Estimation

- One type of statistical problem is **point estimation**.
- A **statistic** $s = s(\mathcal{D})$ is any function of the data.
- A statistic $\hat{\theta} = \hat{\theta}(\mathcal{D})$ taking values in Θ is a **point estimator of** θ.
- A good point estimator will have $\hat{\theta} \approx \theta$.

Desirable Properties of Point Estimators

- Desirable statistical properties of point estimators:
  - **Consistency:** As data size $n \to \infty$, we get $\hat{\theta}_n \to \theta$.
  - **Efficiency:** (Roughly speaking) $\hat{\theta}_n$ is as accurate as we can get from a sample of size n.
- **Maximum likelihood estimators** are consistent and efficient under reasonable conditions.

The Likelihood Function

- Consider parametric family $\{p(y \mid \theta) : \theta \in \Theta\}$ and i.i.d. sample $\mathcal{D} = (y_1, \ldots, y_n)$.
- The density for sample $\mathcal{D}$ for $\theta \in \Theta$ is

$$p(\mathcal{D} \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta).$$

- $p(\mathcal{D} \mid \theta)$ is a function of $\mathcal{D}$ and θ.
- For fixed θ, $p(\mathcal{D} \mid \theta)$ is a density function on $\mathcal{Y}^n$.
- For fixed $\mathcal{D}$, the function $\theta \mapsto p(\mathcal{D} \mid \theta)$ is called the **likelihood function:**

$$L_{\mathcal{D}}(\theta) := p(\mathcal{D} \mid \theta).$$

Maximum Likelihood Estimation

Definition

The **maximum likelihood estimator (MLE)** for θ in the model $\{p(y,\theta) \mid \theta \in \Theta\}$ is

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \Theta}{\arg\max}\, L_{\mathcal{D}}(\theta).$$

- Maximum likelihood is just one approach to getting a point estimator for θ.
- **Method of moments** is another general approach one learns about in statistics.
- Later we'll talk about **MAP** and **posterior mean** as approaches to point estimation.
  - These arise naturally in Bayesian settings.

Coin Flipping: Setup

- Parametric family of mass functions:

$$p(\text{Heads} \mid \theta) = \theta,$$

  for $\theta \in \Theta = (0, 1)$.

- Note that every $\theta \in \Theta$ gives us a different probability model for a coin.

Coin Flipping: Likelihood function

- Data $\mathcal{D} = (H, H, T, T, T, T, T, H, \ldots, T)$
  - n_h: number of heads
  - n_t: number of tails
- Assume these were i.i.d. flips.
- **Likelihood function** for data $\mathcal{D}$:

$$L_{\mathcal{D}}(\theta) = p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t}$$

- This is the probability of getting the flips in the order they were received.

Coin Flipping: MLE

- As usual, easier to maximize the log-likelihood function:

$$\begin{aligned}\hat{\theta}_{\text{MLE}} &= \underset{\theta\in\Theta}{\arg\max}\log L_{\mathcal{D}}(\theta) \\ &= \underset{\theta\in\Theta}{\arg\max}\left[n_h\log\theta + n_t\log(1-\theta)\right]\end{aligned}$$

- First order condition:

$$\begin{aligned}\frac{n_h}{\theta} - \frac{n_t}{1-\theta} &= 0 \\ \iff \theta &= \frac{n_h}{n_h+n_t}.\end{aligned}$$

- So $\hat{\theta}_{\text{MLE}}$ is the empirical fraction of heads.

Bayesian Statistics: Introduction

Bayesian Statistics

- Introduces a new ingredient: the **prior distribution.**
- A **prior distribution** $p(\theta)$ is a distribution on parameter space Θ.
- A prior reflects our belief about θ, **before seeing any data**..

A Bayesian Model

- A [parametric] Bayesian model consists of two pieces:
  1. A parametric family of densities
  $$\{p(\mathcal{D} \mid \theta) \mid \theta \in \Theta\}.$$
  2. A **prior distribution** $p(\theta)$ on parameter space Θ.
- Putting pieces together, we get a joint density on θ and $\mathcal{D}$:
$$p(\mathcal{D}, \theta) = p(\mathcal{D} \mid \theta)p(\theta).$$

The Posterior Distribution

- The **posterior distribution** for θ is $p(\theta \mid \mathcal{D})$.
- Prior represents belief about θ before observing data $\mathcal{D}$.
- Posterior represents the **rationally "updated" belief** about θ, after seeing $\mathcal{D}$.

Expressing the Posterior Distribution

- By Bayes rule, can write the posterior distribution as

$$p(\theta \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \theta)p(\theta)}{p(\mathcal{D})}.$$

- Let's consider both sides as functions of θ, for fixed $\mathcal{D}$.
- Then both sides are densities on Θ and we can write

$$\underbrace{p(\theta \mid \mathcal{D})}_{\text{posterior}} \propto \underbrace{p(\mathcal{D} \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}.$$

- Where $\propto$ means we've dropped factors independent of θ.

Coin Flipping: Bayesian Model

- Parametric family of mass functions:

$$p(\text{Heads} \mid \theta) = \theta,$$

for $\theta \in \Theta = (0, 1)$.

- Need a prior distribution $p(\theta)$ on $\Theta = (0, 1)$.
- A distribution from the Beta family will do the trick...

Coin Flipping: Beta Prior

- **Prior:**

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ p(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \end{aligned}$$

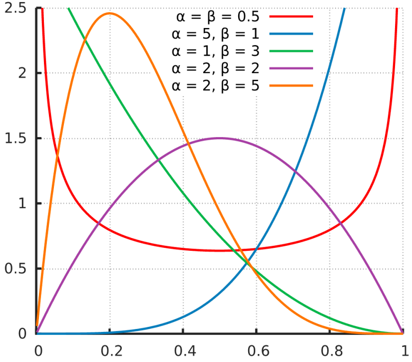


Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons http://commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg.

Coin Flipping: Beta Prior

- **Prior:**

$$\begin{aligned} \theta &\sim \text{Beta}(h,t) \\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1} \end{aligned}$$

- **Mean of Beta distribution:**

$$\mathbb{E}\theta = \frac{h}{h+t}$$

- **Mode of Beta distribution:**

$$\underset{\theta}{\arg\max}\, p(\theta) = \frac{h-1}{h+t-2}$$

for $h,t > 1$.

Coin Flipping: Posterior

- **Prior:**

$$
\begin{aligned}
\theta &\sim \text{Beta}(h,t) \\
p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}
\end{aligned}
$$

- **Likelihood function**

$$
L(\theta) = p(\mathcal{D} \mid \theta) = \theta^{n_h}(1-\theta)^{n_t}
$$

- **Posterior density:**

$$
\begin{aligned}
p(\theta \mid \mathcal{D}) &\propto p(\theta)p(\mathcal{D} \mid \theta) \\
&\propto \theta^{h-1}(1-\theta)^{t-1} \times \theta^{n_h}(1-\theta)^{n_t} \\
&= \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}
\end{aligned}
$$

Posterior is Beta

- **Prior:**

$$\begin{aligned}\theta &\sim \text{Beta}(h,t)\\ p(\theta) &\propto \theta^{h-1}(1-\theta)^{t-1}\end{aligned}$$

- **Posterior density:**

$$p(\theta \mid \mathcal{D}) \propto \theta^{h-1+n_h}(1-\theta)^{t-1+n_t}$$

- **Posterior is in the beta family**:

$$\theta \mid \mathcal{D} \sim \text{Beta}(h+n_h, t+n_t)$$

- **Interpretation**:
  - Prior initializes our counts with h heads and t tails.
  - Posterior increments counts by observed n_h and n_t.

Sidebar: Conjugate Priors

- Interesting that posterior is in same distribution family as prior.
- Let π be a family of prior distributions on Θ.
- Let P parametric family of distributions with parameter space Θ.

Definition

A family of distributions π **is conjugate to** parametric model P if for any prior in π, the posterior is always in π.

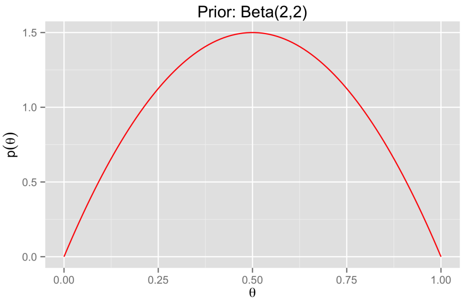
- The beta family is conjugate to the coin-flipping (i.e. Bernoulli) model.
- The family of all probability distributions is conjugate to any parametric model. [Trivially]

Example: Coin Flipping - Concrete Example

- Suppose we have a coin, possibly biased (**parametric probability model**):

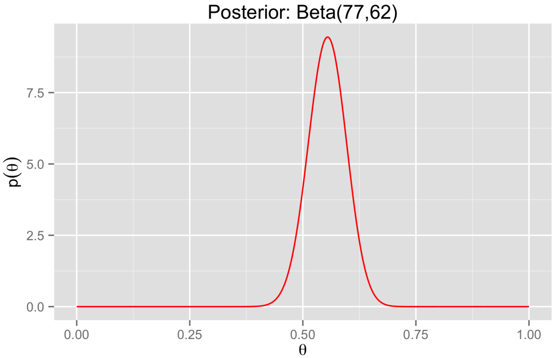
$$p(\text{Heads} \mid \theta) = \theta.$$

- **Parameter space** $\theta \in \Theta = [0,1]$.
- **Prior distribution:** $\theta \sim \text{Beta}(2,2)$.



Example: Coin Flipping

- Next, we gather some data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$:
- Heads: 75 Tails: 60
  - $\hat{\theta}_{\text{MLE}} = \frac{75}{75+60} \approx 0.556$
- **Posterior distribution:** $\theta \mid \mathcal{D} \sim \textbf{Beta}(77, 62)$:



Bayesian Point Estimates

- So we have posterior $\theta \mid \mathcal{D}$...
- But we want a point estimate $\hat{\theta}$ for θ.
- Common options:
    - **posterior mean** $\hat{\theta} = \mathbb{E}[\theta \mid \mathcal{D}]$
    - **maximum a posteriori (MAP) estimate** $\hat{\theta} = \arg\max_{\theta} p(\theta \mid \mathcal{D})$
        - Note: this is the **mode** of the posterior distribution

What else can we do with a posterior?

- Look at it.
- Extract "**credible set**" for θ (Bayesian version of a confidence interval).
  - e.g. Interval $[a, b]$ is a 95% **credible set** if

$$\mathbb{P}(\theta \in [a, b] \mid \mathcal{D}) \geqslant 0.95$$

- The most "Bayesian" approach is **Bayesian decision theory**:
  - Choose a loss function.
  - Find action **minimizing expected risk w.r.t. posterior**

Bayesian Decision Theory

Bayesian Decision Theory

- Ingredients:
  - **Parameter space** Θ.
  - **Prior**: Distribution $p(\theta)$ on Θ.
  - **Action space** $\mathcal{A}$.
  - **Loss function**: $\ell : \mathcal{A} \times \Theta \to \mathbf{R}$.
- The **posterior risk** of an action $a \in \mathcal{A}$ is

$$\begin{aligned} r(a) &:= \mathbb{E}\left[\ell(\theta, a) \mid \mathcal{D}\right] \\ &= \int \ell(\theta, a) p(\theta \mid \mathcal{D})\, d\theta. \end{aligned}$$

  - It's the **expected loss under the posterior.**
- A **Bayes action** a^* is an action that minimizes posterior risk:

$$r(a^*) = \min_{a \in \mathcal{A}} r(a)$$

Bayesian Point Estimation

- General Setup:
  - Data $\mathcal{D}$ generated by $p(y \mid \theta)$, for unknown $\theta \in \Theta$.
  - Want to produce a **point estimate** for θ.
- Choose the following:
  - **Prior** $p(\theta)$ on $\Theta = \mathbf{R}$.
  - **Loss** $\ell(\hat{\theta}, \theta) = \left(\theta - \hat{\theta}\right)^2$
- Find **action** $\hat{\theta} \in \Theta$ that minimizes **posterior risk:**

$$
\begin{aligned}
r(\hat{\theta}) &= \mathbb{E}\left[\left(\theta - \hat{\theta}\right)^2 \mid \mathcal{D}\right] \\
&= \int \left(\theta - \hat{\theta}\right)^2 p(\theta \mid \mathcal{D})\, d\theta
\end{aligned}
$$

Bayesian Point Estimation: Square Loss

- Find **action** $\hat{\theta} \in \Theta$ that minimizes **posterior risk**

$$r(\hat{\theta}) = \int \left(\theta - \hat{\theta}\right)^2 p(\theta \mid \mathcal{D})\, d\theta.$$

- Differentiate:

$$\begin{aligned}
\frac{dr(\hat{\theta})}{d\hat{\theta}} &= -\int 2\left(\theta - \hat{\theta}\right) p(\theta \mid \mathcal{D})\, d\theta \\
&= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta}\underbrace{\int p(\theta \mid \mathcal{D})\, d\theta}_{=1} \\
&= -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta}
\end{aligned}$$

Bayesian Point Estimation: Square Loss

- Derivative of posterior risk is

$$\frac{dr(\hat{\theta})}{d\hat{\theta}} = -2\int \theta p(\theta \mid \mathcal{D})\, d\theta + 2\hat{\theta}.$$

- First order condition $\frac{dr(\hat{\theta})}{d\hat{\theta}} = 0$ gives

$$\begin{aligned}\hat{\theta} &= \int \theta p(\theta \mid \mathcal{D})\, d\theta \\ &= \mathbb{E}[\theta \mid \mathcal{D}]\end{aligned}$$

- **Bayes action** for **square loss** is the posterior mean.

Bayesian Point Estimation: Absolute Loss

- **Loss:** $\ell(\theta, \hat{\theta}) = \left|\theta - \hat{\theta}\right|$
- **Bayes action** for **absolute loss** is the **posterior median.**
  - That is, the median of the distribution $p(\theta \mid \mathcal{D})$.
  - Show with approach similar to what was used in Homework #1.

Bayesian Point Estimation: Zero-One Loss

- Suppose Θ is discrete (e.g. $\Theta = \{\text{english}, \text{french}\}$)
- **Zero-one loss:** $\ell(\theta, \hat{\theta}) = 1(\theta \neq \hat{\theta})$
- **Posterior risk**:

$$
\begin{aligned}
r(\hat{\theta}) &= \mathbb{E}\left[1(\theta \neq \hat{\theta}) \mid \mathcal{D}\right] \\
&= \mathbb{P}\left(\theta \neq \hat{\theta} \mid \mathcal{D}\right) \\
&= 1 - \mathbb{P}\left(\theta = \hat{\theta} \mid \mathcal{D}\right) \\
&= 1 - p(\hat{\theta} \mid \mathcal{D})
\end{aligned}
$$

- **Bayes action** is

$$
\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\, p(\theta \mid \mathcal{D})
$$

- This $\hat{\theta}$ is called the **maximum a posteriori (MAP)** estimate.
- The MAP estimate is the **mode** of the posterior distribution.

Summary

Recap and Interpretation

- Prior represents belief about θ before observing data $\mathcal{D}$.
- Posterior represents the **rationally "updated" beliefs** after seeing $\mathcal{D}$.
- All inferences and action-taking are based on the posterior distribution.
- In the Bayesian approach,
  - No issue of "choosing a procedure" or justifying an estimator.
  - Only choices are
    - **family of distributions**, indexed by Θ, and the
    - **prior distribution** on Θ
  - For decision making, need a **loss function**.
  - Everything after that is **computation**.

The Bayesian Method

1. **Define the model**:
    - Choose a parametric family of densities:
    $$\{p(\mathcal{D} \mid \theta) \mid \theta \in \Theta\}.$$
    - Choose a distribution $p(\theta)$ on Θ, called the **prior distribution**.
2. After observing $\mathcal{D}$, compute the **posterior distribution** $p(\theta \mid \mathcal{D})$.
3. Choose **action** based on $p(\theta \mid \mathcal{D})$.