Bayesian Methods

David Rosenberg

New York University

April 27, 2016

Classical Statistics

Frequentist or "Classical" Statistics

- Probability model with parameter $\theta \in \Theta$

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

  where $p(y;\theta)$ is either a PDF or a PMF.
- Assume that $p(y;\theta)$ governs the world we are observing.
- In **frequentist statistics**, the **parameter** θ is a
  - **fixed constant** (i.e. not random) and is
  - **unknown** to us.
- If we knew θ, there would be no need for statistics.
- Instead of θ, we have a **sample** $\mathcal{D} = \{y_1, \ldots, y_n\}$ i.i.d. $p(y;\theta)$.
- Statistics is about how to use $\mathcal{D}$ 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 $\hat{\theta} = \hat{\theta}(\mathcal{D})$ is a **point estimator** if $\hat{\theta} \approx \theta$.
- Desirable statistical properties of point estimators:
  - **Consistency:** As data size $n \to \infty$, we get $\hat{\theta} \to \theta$.
  - **Efficiency:** (Roughly speaking) $\hat{\theta}_n$ is as accurate as we can get from a sample of size n.
  - e.g. **maximum likelihood estimation** is consistent and efficient under reasonable conditions.
- In frequentist statistics, you can make up any estimator you want.
  - Justify its use by showing it has desirable properties.

Bayesian Statistics: Introduction

Bayesian Statistics

- Major viewpoint change in **Bayesian statistics**:
  - parameter $\theta \in \Theta$ is a **random variable**.
- New ingredient is the **prior distribution:**
  - It is a distribution on parameter space Θ.
  - Reflects our belief about θ.
  - Must be chosen before seeing any data.

The Bayesian Method

1. Define the model:
    - Choose a distribution $p(\theta)$, called the **prior distribution**.
    - Choose a probability model or "**likelihood model**", now written as:

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

2. After observing $\mathcal{D}$, compute the **posterior distribution** $p(\theta \mid \mathcal{D})$.
3. Choose **action** based on $p(\theta \mid \mathcal{D})$.

The Posterior Distribution

- By Bayes rule, can write the posterior distribution as

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

- **likelihood:** $p(\mathcal{D} \mid \theta)$
- **prior:** $p(\theta)$
- **marginal likelihood**: $p(\mathcal{D})$.
- Note: $p(\mathcal{D})$ is just a normalizing constant for $p(\theta \mid \mathcal{D})$. Can write

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

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 the **prior** and the **likelihood model**.
  - For decision making, need a **loss function**.
  - Everything after that is **computation**.

Coin Flipping: The Beta-Binomial Model

Coin Flipping: Setup

- **Parameter space** $\theta \in \Theta = [0,1]$:

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

- Data $\mathcal{D} = \{H, H, T, T, T, T, T, H, \ldots, T\}$
  - n_h: number of heads
  - n_t: number of tails
- **Likelihood model** (Bernoulli Distribution):

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

  - (probability of getting the flips in the order they were received)

Coin Flipping: Beta Prior

- **Prior:**

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

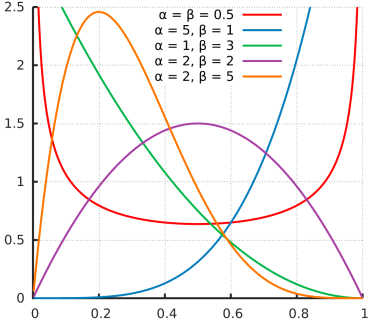


Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons `http://commons.wikimedia.org/wiki/File:Beta_distribution_pdf.svg`.

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:**

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

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 model**:

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

- **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**:

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

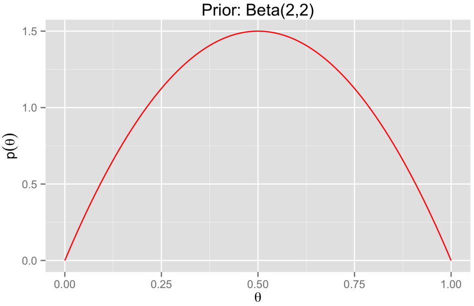
- **Interpretation**:
  - Prior initializes our counts with h heads and t tails.
  - Posterior increments counts by observed n_h and n_t.

Example: Coin Flipping

- Suppose we have a coin, possibly biased

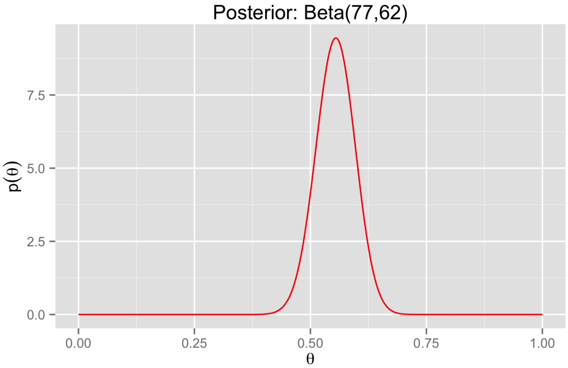
$$\mathbb{P}(\text{Heads} \mid \theta) = \theta.$$

- **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

- Suppose we have posterior $\theta \mid \mathcal{D}$...
- But we want a point estimate $\hat{\theta}$ or 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 (a Bayesian confidence interval).
    - e.g. Interval $[a, b]$ is a 95% **credible set** if

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

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