STAT 576 Bayesian Analysis

Lecture 2: Bayesian Inference 1

Chencheng Cai

Washington State University

Binomial with Parameter θ

- Probability of “success” in trial: θ
- Probability of “failure” in trial: $1 - \theta$

Binomial with Parameter θ

- Probability of "success" in trial: θ
- Probability of "failure" in trial: $1 - \theta$
- If there are n independent trials, the probability of observing y "successes" is

$$p(y \mid \theta, n) = \mathrm{Bin}(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Binomial with Parameter θ

- Probability of "success" in trial: θ
- Probability of "failure" in trial: $1 - \theta$
- If there are n independent trials, the probability of observing y "successes" is

$$p(y \mid \theta, n) = \mathrm{Bin}(y \mid n, \theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$$

- The above probability is called **observation model** or **sampling distribution**.

Binomial with Parameter θ

- Probability of "success" in trial: θ
- Probability of "failure" in trial: $1 - \theta$
- If there are n independent trials, the probability of observing y "successes" is

$$p(y \mid \theta, n) = \mathrm{Bin}(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- The above probability is called **observation model** or **sampling distribution**.
- The likelihood function is a function of θ that

$$L(\theta; y) = p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

Proportional Notation

- Sometimes, we only care about a single variable in the formula.
- To reduce notational burden, we use $\propto$ to simplify equations.

Proportional Notation

- Sometimes, we only care about a single variable in the formula.
- To reduce notational burden, we use $\propto$ to simplify equations.
- The observation model is a function of y:

$$p(y \mid \theta, n) = \mathrm{Bin}(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Therefore, we may write

$$p(y \mid \theta, n) \propto \binom{n}{y} \left(\frac{\theta}{1-\theta}\right)^y$$

Proportional Notation

- Sometimes, we only care about a single variable in the formula.
- To reduce notational burden, we use $\propto$ to simplify equations.
- The observation model is a function of y:

$$p(y \mid \theta, n) = \mathrm{Bin}(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- Therefore, we may write

$$p(y \mid \theta, n) \propto \binom{n}{y} \left(\frac{\theta}{1-\theta}\right)^y$$

- The likelihood is a function of θ:

$$L(\theta; y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

- We may write

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

Bayes' Rule

- In order to conduct Bayesian inference, we need to assume a distribution for θ, which is known as the **prior** distribution, denoted by $p(\theta)$ here.

Bayes' Rule

- In order to conduct Bayesian inference, we need to assume a distribution for θ, which is known as the **prior** distribution, denoted by $p(\theta)$ here.
- Interpretation of the prior:
  - Populational/Marginal distribution for θ.
  - User's belief on the parameter θ **before** observing the data.
  - User's intention/preference over the parameter θ.

Bayes' Rule

- In order to conduct Bayesian inference, we need to assume a distribution for θ, which is known as the **prior** distribution, denoted by $p(\theta)$ here.
- Interpretation of the prior:
  - Populational/Marginal distribution for θ.
  - User's belief on the parameter θ **before** observing the data.
  - User's intention/preference over the parameter θ.
- **Bayes' Rule**:

$$p(\theta \mid y, n) = \frac{p(y \mid \theta, n) p(\theta \mid n)}{p(y \mid n)} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal}},$$

where $p(y \mid n) = \int p(y \mid \theta, n) p(\theta) d\mu(\theta)$.

Bayes' Rule

- In order to conduct Bayesian inference, we need to assume a distribution for θ, which is known as the **prior** distribution, denoted by $p(\theta)$ here.
- Interpretation of the prior:
  - Populational/Marginal distribution for θ.
  - User's belief on the parameter θ **before** observing the data.
  - User's intention/preference over the parameter θ.
- **Bayes' Rule**:

$$p(\theta \mid y, n) = \frac{p(y \mid \theta, n)p(\theta \mid n)}{p(y \mid n)} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal}},$$

  where $p(y \mid n) = \int p(y \mid \theta, n)p(\theta)d\mu(\theta)$.
- Proof:

$$p(\theta \mid y, n) = \frac{p(\theta, y \mid n)}{p(y \mid n)} = \frac{p(y \mid \theta, n)p(\theta \mid n)}{p(y \mid n)}$$

Bayes' Rule

- For now, we choose the prior as uniform on $[0, 1]$ such that

$$p(\theta \mid n) = 1$$

Bayes' Rule

- For now, we choose the prior as uniform on $[0,1]$ such that

$$p(\theta \mid n) = 1$$

- By Bayes' rule, we have the posterior:

$$p(\theta \mid y, n) = \frac{p(y \mid \theta, n)p(\theta \mid n)}{p(y \mid n)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y} \times 1}{\int \binom{n}{y}\theta^y(1-\theta)^{n-y}d\mu(\theta)} = \frac{\theta^y(1-\theta)^{n-y}}{\int \theta^y(1-\theta)^{n-y}d\mu(\theta)}$$

Bayes' Rule

- For now, we choose the prior as uniform on $[0,1]$ such that

$$p(\theta \mid n) = 1$$

- By Bayes' rule, we have the posterior:

$$p(\theta \mid y, n) = \frac{p(y \mid \theta, n) p(\theta \mid n)}{p(y \mid n)} = \frac{\binom{n}{y} \theta^y (1-\theta)^{n-y} \times 1}{\int \binom{n}{y} \theta^y (1-\theta)^{n-y} d\mu(\theta)} = \frac{\theta^y (1-\theta)^{n-y}}{\int \theta^y (1-\theta)^{n-y} d\mu(\theta)}$$

- Notice that

$$\int \theta^y (1-\theta)^{n-y} d\mu(\theta) = B(y+1, n-y+1) = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

We know $p(\theta \mid y, n) = \mathrm{Beta}(\theta \mid y+1, n-y+1)$.

Bayes' Rule using Proportional Notation

- With proportional notation, the calculation can be speed up:

Bayes' Rule using Proportional Notation

- With proportional notation, the calculation can be speed up:
- We have
$$p(\theta \mid n) \propto 1, \quad p(y \mid \theta, n) \propto \theta^y (1-\theta)^{n-y}$$
- Therefore
$$p(\theta \mid y, n) \propto p(y \mid \theta, n) p(\theta \mid n) \propto \theta^y (1-\theta)^{n-y}$$

Bayes' Rule using Proportional Notation

- With proportional notation, the calculation can be speed up:
- We have

$$p(\theta \mid n) \propto 1, \quad p(y \mid \theta, n) \propto \theta^y (1-\theta)^{n-y}$$

- Therefore

$$p(\theta \mid y, n) \propto p(y \mid \theta, n) p(\theta \mid n) \propto \theta^y (1-\theta)^{n-y}$$

- It is immediate that $p(\theta \mid y, n)$ is $\mathrm{Beta}(y+1, n-y+1)$.

Bayes' Rule using Proportional Notation

- With proportional notation, the calculation can be speed up:
- We have

$$p(\theta \mid n) \propto 1, \quad p(y \mid \theta, n) \propto \theta^y (1-\theta)^{n-y}$$

- Therefore

$$p(\theta \mid y, n) \propto p(y \mid \theta, n) p(\theta \mid n) \propto \theta^y (1-\theta)^{n-y}$$

- It is immediate that $p(\theta \mid y, n)$ is $\mathrm{Beta}(y+1, n-y+1)$.
- Because the **kernel** of $\mathrm{Beta}(a, b)$ distribution is $\theta^{a-1}(1-\theta)^{b-1}$.

Kernel

- In Bayesian statistics, the **kernel** of a distribution family refers to the form of the pdf in which any factors that are not functions of any of the variables in the domain are omitted. (i.e. the proportional notation w.r.t. the parameter.)
- Common kernels:
  - Uniform: $p(x \mid \theta) \propto 1$
  - Gaussian: $p(x \mid \mu, \sigma) \propto \exp\{-(x-\mu)^2/(2\sigma^2)\} \propto \exp\{-(2\sigma^2)^{-1}x^2 + \mu\sigma^{-2}x\}$
  - Exponential: $p(x \mid \lambda) \propto \exp\{-\lambda x\}$
  - Gamma: $p(x \mid \alpha, \beta) \propto x^{\alpha-1}\exp\{-\beta x\}$
  - Beta: $p(x \mid \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}$
  - Binomial: $p(x \mid n, p) \propto p^x(1-p)^{n-x}$
  - Poisson: $p(x \mid \lambda) \propto \lambda^x/x!$
  - Geometric: $p(x \mid p) \propto (1-p)^x$

Point Estimation

- Now we have the posterior:

$$p(\theta \mid y, n) \sim \mathrm{Beta}(y+1, n-y+1)$$

- We can provide point estimators for θ based on the posterior:
  - Maximize a posteriori (MAP):

$$\hat{\theta} = \underset{\theta \in [0,1]}{\arg\max}\; p(\theta \mid y, n) = \underset{\theta \in [0,1]}{\arg\max}\; \theta^y (1-\theta)^{n-y} = \frac{y}{n}$$

  - Posterior mean:

$$\hat{\theta} = \mathbb{E}[\theta \mid y, n] = \frac{y+1}{n+2}$$

Point Estimation

- Now we have the posterior:

$$p(\theta \mid y, n) \sim \mathrm{Beta}(y+1, n-y+1)$$

- We can provide point estimators for θ based on the posterior:
  - Maximize a posteriori (MAP):

$$\hat{\theta} = \arg\max_{\theta \in [0,1]} \; p(\theta \mid y, n) = \arg\max_{\theta \in [0,1]} \; \theta^{y}(1-\theta)^{n-y} = \frac{y}{n}$$

  - Posterior mean:

$$\hat{\theta} = \mathbb{E}[\theta \mid y, n] = \frac{y+1}{n+2}$$

- Claim: MAP under uniform prior is the same as MLE.

Credible Interval

- An α-level **credible** interval $\mathcal{I} \subset \Omega$ is such that

$$\mathbb{P}(\theta \in \mathcal{I} \mid y, n) \geq \alpha$$

Credible Interval

- An α-level **credible** interval $\mathcal{I} \subset \Omega$ is such that

$$\mathbb{P}(\theta \in \mathcal{I} \mid y, n) \geq \alpha$$

- Quantile-baed interval (QBI): use quantiles of the posterior to construct $\mathcal{I} = [a, b]$:

$$a = q_{(1-\alpha)/2}(p(\theta \mid y, n)), \quad b = q_{(1+\alpha)/2}(p(\theta \mid y, n))$$

Credible Interval

- An α-level **credible** interval $\mathcal{I} \subset \Omega$ is such that

$$\mathbb{P}(\theta \in \mathcal{I} \mid y, n) \geq \alpha$$

- Quantile-baed interval (QBI): use quantiles of the posterior to construct $\mathcal{I} = [a, b]$:

$$a = q_{(1-\alpha)/2}(p(\theta \mid y, n)), \quad b = q_{(1+\alpha)/2}(p(\theta \mid y, n))$$

- Highest density region (HDI): use the superlevel set of the posterior:

$$\mathcal{I} = \{\theta \in \Omega : p(\theta \mid y, n) \geq c\}$$

with

$$c = \sup\{c : \mathbb{P}(\theta \geq c \mid y, n) \geq \alpha\}$$

Prediction

- Immagine $\tilde{y} \in \{0, 1\}$ is the outcome of another trial with the same parameter θ.
- $p(\tilde{y} \mid y, n)$ is the **predictive** distribution of $\tilde{y}$.

Prediction

- Immagine $\tilde{y} \in \{0, 1\}$ is the outcome of another trial with the same parameter θ.
- $p(\tilde{y} \mid y, n)$ is the **predictive** distribution of $\tilde{y}$.
- We claim

$$p(\tilde{y} \mid y, n) = \int p(\tilde{y} \mid \theta) p(\theta \mid y, n) d\mu(\theta)$$

Prediction

- Immagine $\tilde{y} \in \{0, 1\}$ is the outcome of another trial with the same parameter θ.
- $p(\tilde{y} \mid y, n)$ is the **predictive** distribution of $\tilde{y}$.
- We claim
$$p(\tilde{y} \mid y, n) = \int p(\tilde{y} \mid \theta) p(\theta \mid y, n) d\mu(\theta)$$
- Proof:
$$p(\tilde{y} \mid y, n) = \int p(\tilde{y}, \theta \mid y, n) d\mu(\theta) = \int p(\tilde{y} \mid \theta, y, n) p(\theta \mid y, n) d\mu(\theta).$$
The claim is immediate by observing $p(\tilde{y} \mid \theta, y, n) = p(\tilde{y} \mid \theta)$.

Prediction

- Immagine $\tilde{y} \in \{0, 1\}$ is the outcome of another trial with the same parameter θ.
- $p(\tilde{y} \mid y, n)$ is the **predictive** distribution of $\tilde{y}$.
- We claim
$$p(\tilde{y} \mid y, n) = \int p(\tilde{y} \mid \theta) p(\theta \mid y, n) d\mu(\theta)$$
- Proof:
$$p(\tilde{y} \mid y, n) = \int p(\tilde{y}, \theta \mid y, n) d\mu(\theta) = \int p(\tilde{y} \mid \theta, y, n) p(\theta \mid y, n) d\mu(\theta).$$
The claim is immediate by observing $p(\tilde{y} \mid \theta, y, n) = p(\tilde{y} \mid \theta)$.
- Therefore, we have
$$\mathbb{P}[\tilde{y} = 1 \mid y, n] = \int \theta p(\theta \mid y, n) d\mu(\theta) = \mathbb{E}[\theta \mid y, n] = \frac{y+1}{n+2}$$

De Finetti's Theorem

- The toy example with i.i.d. Bernoulli random trials with a common success probability θ from certain prior distribution is not trivial.

De Finetti's Theorem

- The toy example with i.i.d. Bernoulli random trials with a common success probability θ from certain prior distribution is not trivial.
- An infinite sequence $X_1, X_2, \ldots$ is said to be **exchangeable** if for any finite sequence $i_1, \ldots, i_n$ and any permutation of them $\pi : \{i_1, \ldots, i_n\} \to \{i_1, \ldots, i_n\}$, we have

$$(X_{i_1}, \ldots, X_{i_n}) \sim (X_{\pi(i_1)}, \ldots, X_{\pi(i_n)}).$$

De Finetti's Theorem

- The toy example with i.i.d. Bernoulli random trials with a common success probability θ from certain prior distribution is not trivial.
- An infinite sequence $X_1, X_2, \ldots$ is said to be **exchangeable** if for any finite sequence $i_1, \ldots, i_n$ and any permutation of them $\pi : \{i_1, \ldots, i_n\} \to \{i_1, \ldots, i_n\}$, we have
$$(X_{i_1}, \ldots, X_{i_n}) \sim (X_{\pi(i_1)}, \ldots, X_{\pi(i_n)}).$$
- **De Finetti's Theorem**:
If $X_1, X_2, \ldots$ is an infinite exchangeable Bernoulli random variables, then there exists a probability measure Π on $[0, 1]$ such that
  - $\theta \sim \Pi$;
  - $X_1, X_2, \ldots$ are conditionally independent given θ;
  - The conditional distribution of X_i given θ is $\mathrm{Bernoulli}(\theta)$.

De Finetti's Theorem

- The toy example with i.i.d. Bernoulli random trials with a common success probability θ from certain prior distribution is not trivial.
- An infinite sequence $X_1, X_2, \dots$ is said to be **exchangeable** if for any finite sequence $i_1, \dots, i_n$ and any permutation of them $\pi : \{i_1, \dots, i_n\} \to \{i_1, \dots, i_n\}$, we have

$$(X_{i_1}, \dots, X_{i_n}) \sim (X_{\pi(i_1)}, \dots, X_{\pi(i_n)}).$$

- **De Finetti's Theorem**:
  If $X_1, X_2, \dots$ is an infinite exchangeable Bernoulli random variables, then there exists a probability measure Π on $[0, 1]$ such that
  - $\theta \sim \Pi$;
  - $X_1, X_2, \dots$ are conditionally independent given θ;
  - The conditional distribution of X_i given θ is $\mathrm{Bernoulli}(\theta)$.
- In summary, if $(X_1, \dots, X_n)$ are exchangeable random variables, then

$$p(X_1, \dots, X_n) = \int \theta^S (1-\theta)^{n-S} d\Pi(\theta)$$

with $S = \sum_{i=1}^n X_i$ and Π some probability on $[0, 1]$.

Sketch of Proof

- Let $S_n = \sum_{i=1}^{n} X_i$.
- By exchangeablility, we have

$$p(X_1, \ldots, X_n) = \binom{n}{y}^{-1} p(S_n = y) = \binom{n}{y} \sum_{Y=y}^{N-(n-y)} \frac{\binom{Y}{y}\binom{N-Y}{n-y}}{\binom{N}{n}} p(S_N = Y)$$

- Define probability measure Π_N by

$$\Pi_N([0, \theta]) = p(S_N \le \theta N)$$

- Then we have

$$p(X_1, \ldots, X_n) = \int \frac{(\theta N)^{\downarrow y}((1-\theta)N)^{\downarrow n-y}}{N^{\downarrow n}} d\Pi_N(\theta)$$

Sketch of Proof

$$p(X_1, \ldots, X_n) = \int \frac{(\theta N)^{\downarrow y}((1-\theta)N)^{\downarrow n-y}}{N^{\downarrow n}} d\Pi_N(\theta)$$

- On the one hand,

$$\frac{(\theta N)^{\downarrow y}((1-\theta)N)^{\downarrow n-y}}{N^{\downarrow n}} \to \theta^y(1-\theta)^{n-y}$$

  uniformly.
- On the other hand, Π_N has a convergent subsequence by Helly's selection theorem. Denote the limit by Π.
- So we have (by taking $N \to \infty$)

$$p(X_1, \ldots, X_n) = \int \theta^y(1-\theta)^{n-y} d\Pi$$