STAT 576 Bayesian Analysis

Lecture 2: Bayesian Inference I

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:

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

- 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:

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

- 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:

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

and

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

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 $X_1, X_2, \dots$ 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 $\mathrm{Bernoulli}(\theta)$.

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

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

- 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 $N \to \infty$)

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

Prior Elicitation

- In prevoius example, we used uniform prior for the binomial distribution parameter θ.
- Some bad choices:
  - $p(\theta \mid n) \propto \mathbb{I}_{[0,1/2]}$ (limited domain)
  - $p(\theta \mid n) \propto \sin(\pi\theta)$ (difficult to compute posterior)

Prior Elicitation

- In prevoius example, we used uniform prior for the binomial distribution parameter θ.
- Some bad choices:
  - $p(\theta \mid n) \propto \mathbb{I}_{[0,1/2]}$ (limited domain)
  - $p(\theta \mid n) \propto \sin(\pi\theta)$ (difficult to compute posterior)
- We desire the prior to be:
  - easy to compute posterior and to conduct inference
  - invariant under re-parametrization
  - least subjective

Conjugate Prior

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

- If we choose the prior in the form of

$$p(\theta \mid n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

  for some (α, β),

Conjugate Prior

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

- If we choose the prior in the form of

  $$p(\theta \mid n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

  for some (α, β),
- By Bayes' rule, the posterior is

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

Conjugate Prior

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

- If we choose the prior in the form of

$$p(\theta \mid n) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

  for some (α, β),
- By Bayes' rule, the posterior is

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

- The posterior has the same kernel format as in the prior with

$$\alpha \to \alpha + y, \quad \beta \to \beta + n - y$$

Conjugate Prior

- The prior is $\mathrm{Beta}(\alpha, \beta)$
- The sampling distribution is $\mathrm{Binom}(n, \theta)$
- The corresponding posterior is $\mathrm{Beta}(\alpha + y, \beta + n - y)$.

Conjugate Prior

- The prior is $\mathrm{Beta}(\alpha, \beta)$
- The sampling distribution is $\mathrm{Binom}(n, \theta)$
- The corresponding posterior is $\mathrm{Beta}(\alpha + y, \beta + n - y)$.
- The posterior and the prior belongs to the same distribution family.
- We call the Beta distribution is the **conjugate** prior for $\mathrm{Binom}(n, \theta)$ with fixed n.

Conjugate Prior

- The prior is $\mathrm{Beta}(\alpha, \beta)$
- The sampling distribution is $\mathrm{Binom}(n, \theta)$
- The corresponding posterior is $\mathrm{Beta}(\alpha + y, \beta + n - y)$.
- The posterior and the prior belongs to the same distribution family.
- We call the Beta distribution is the **conjugate** prior for $\mathrm{Binom}(n, \theta)$ with fixed n.
- α and β in the prior are called the **hyperparameters**.

Conjugate Prior

- The prior is $\text{Beta}(\alpha, \beta)$
- The sampling distribution is $\text{Binom}(n, \theta)$
- The corresponding posterior is $\text{Beta}(\alpha + y, \beta + n - y)$.
- The posterior and the prior belongs to the same distribution family.
- We call the Beta distribution is the **conjugate** prior for $\text{Binom}(n, \theta)$ with fixed n.
- α and β in the prior are called the **hyperparameters**.
- The $\text{Unif}[0, 1]$ is a special Beta distribution with $\alpha = \beta = 1$.
- List of common conjugate priors can be found at `https://en.wikipedia.org/wiki/Conjugate_prior#Table_of_conjugate_distributions`

Conjugate Prior — Example

- Exponential sampling distribution

$$p(x \mid \theta) \propto e^{-\theta x}$$

Conjugate Prior — Example

- Exponential sampling distribution

$$p(x \mid \theta) \propto e^{-\theta x}$$

  - We can set the prior to $p(\theta) \propto e^{-\kappa\theta}$, i.e. $\mathrm{Exp}(\kappa)$.
  - The corresponding prior is $\mathrm{Exp}(\kappa + x)$.

Conjugate Prior — Example

- Exponential sampling distribution

$$p(x \mid \theta) \propto e^{-\theta x}$$

  - We can set the prior to $p(\theta) \propto e^{-\kappa\theta}$, i.e. $\mathrm{Exp}(\kappa)$.
  - The corresponding prior is $\mathrm{Exp}(\kappa + x)$.
- Poisson sampling distributon for n observations

$$p(x_1, \ldots, x_n \mid \theta) \propto \prod_{i=1}^{n} \theta^{x_i} e^{-\theta} \propto \theta^{S_n} e^{-n\theta}$$

Conjugate Prior — Example

- Exponential sampling distribution

$$p(x \mid \theta) \propto e^{-\theta x}$$

  - We can set the prior to $p(\theta) \propto e^{-\kappa\theta}$, i.e. $\mathrm{Exp}(\kappa)$.
  - The corresponding prior is $\mathrm{Exp}(\kappa + x)$.

- Poisson sampling distributon for n observations

$$p(x_1, \ldots, x_n \mid \theta) \propto \prod_{i=1}^{n} \theta^{x_i} e^{-\theta} \propto \theta^{S_n} e^{-n\theta}$$

  - We can set the prior to $p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$, i.e. $\mathrm{Gamma}(\alpha, \beta)$.
  - The posterior is $\mathrm{Gamma}(\alpha + S_n, \beta + n)$

Conjugate Prior for Exponential Family

- Suppose we have a sampling distribution from an exponential family:

$$p(y_i \mid \theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)} \propto g(\theta)e^{\phi(\theta)^T u(y_i)}$$

where both $\phi(\theta)$ and $u(y_i)$ are vector-valued.

Conjugate Prior for Exponential Family

- Suppose we have a sampling distribution from an exponential family:

$$p(y_i \mid \theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)} \propto g(\theta)e^{\phi(\theta)^T u(y_i)}$$

  where both $\phi(\theta)$ and $u(y_i)$ are vector-valued.
- Then

$$p(y_1, \ldots, y_n \mid \theta) \propto [g(\theta)]^n \, e^{\phi(\theta)^T t(y)},$$

  where $t(y) = \sum_{i=1}^n u(y_i)$ is the sufficient statistics.

Conjugate Prior for Exponential Family

- Suppose we have a sampling distribution from an exponential family:

$$p(y_i \mid \theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)} \propto g(\theta)e^{\phi(\theta)^T u(y_i)}$$

  where both $\phi(\theta)$ and $u(y_i)$ are vector-valued.
- Then

$$p(y_1, \ldots, y_n \mid \theta) \propto [g(\theta)]^n \, e^{\phi(\theta)^T t(y)},$$

  where $t(y) = \sum_{i=1}^n u(y_i)$ is the sufficient statistics.
- We can choose the prior to be

$$p(\theta) \propto [g(\theta)]^\alpha \, e^{\phi(\theta)^T \beta},$$

  for some real number α and vector β.

Conjugate Prior for Exponential Family

- Suppose we have a sampling distribution from an exponential family:

$$p(y_i \mid \theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)} \propto g(\theta)e^{\phi(\theta)^T u(y_i)}$$

  where both $\phi(\theta)$ and $u(y_i)$ are vector-valued.
- Then

$$p(y_1, \ldots, y_n \mid \theta) \propto [g(\theta)]^n \, e^{\phi(\theta)^T t(y)},$$

  where $t(y) = \sum_{i=1}^n u(y_i)$ is the sufficient statistics.
- We can choose the prior to be

$$p(\theta) \propto [g(\theta)]^\alpha \, e^{\phi(\theta)^T \beta},$$

  for some real number α and vector β.
- The posterior is

$$p(\theta) \propto [g(\theta)]^{\alpha+n} \, e^{\phi(\theta)^T (\beta + t(y))}$$

Conjugate Prior for Exponential Family

- For exponential sampling distribution, we have

$$g(\theta) \propto 1, \quad \phi(\theta) = -\theta, \quad u(y_i) = y_i$$

- Therefore, the conjugate prior is

$$p(\theta) \propto e^{-\theta\beta}$$

Conjugate Prior for Exponential Family

- For exponential sampling distribution, we have

$$g(\theta) \propto 1, \quad \phi(\theta) = -\theta, \quad u(y_i) = y_i$$

- Therefore, the conjugate prior is

$$p(\theta) \propto e^{-\theta\beta}$$

- For Poisson sampling distribution, we have

$$g(\theta) = e^{-\theta}, \quad \phi(\theta) = \log\theta, \quad u(y_i) = y_i$$

- Therefore, the conjugate prior is

$$p(\theta) \propto e^{-\alpha\theta} e^{\beta\log\theta} \propto e^{-\alpha\theta}\theta^{\beta}$$

Uninformative Priors

- Sometimes we want the prior to be less **subjective** or less **informative**.

Uninformative Priors

- Sometimes we want the prior to be less **subjective** or less **informative**.
- The idea:
  - We set the prior to be uniform on some symmetric parameter space.
  - We use change-of-variable to obtain the reasonable prior for other re-parametrization.

Uninformative Priors — Location Family

- The sampling distribution with the form

$$p(x \mid \theta) = f(x - \theta)$$

for some integrable f is called a **location family**.

Uninformative Priors — Location Family

- The sampling distribution with the form

$$p(x \mid \theta) = f(x - \theta)$$

  for some integrable f is called a **location family**.
- The Fisher's information is

$$I(\theta) = -\mathbb{E}_\theta[f''(x-\theta)] = -\mathbb{E}_0[f''(y)]$$

  with $y = x - \theta$ and $p(y \mid \theta) = f(y)$.
- The Fisher's information is irrelevant to θ.

Uninformative Priors — Location Family

- The sampling distribution with the form

$$p(x \mid \theta) = f(x - \theta)$$

  for some integrable f is called a **location family**.
- The Fisher's information is

$$I(\theta) = -\mathbb{E}_\theta[f''(x - \theta)] = -\mathbb{E}_0[f''(y)]$$

  with $y = x - \theta$ and $p(y \mid \theta) = f(y)$.
- The Fisher's information is irrelevant to θ.
- In this case, we naturally set the prior to

$$p(\theta) \propto 1$$

- Notice that $p(\theta) = 1$ is not a valid p.d.f.. It is called **improper prior distribution**.

Uninformative Priors — Scale Family

- The sampling distribution with the form

$$p(x \mid \theta) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$$

for some integrable f is called a **scale family**.

Uninformative Priors — Scale Family

- The sampling distribution with the form

$$p(x \mid \theta) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$$

for some integrable f is called a **scale family**.

- The Fisher's information is

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{1}{\theta^3} f''\left(\frac{x}{\theta}\right) \right] = \frac{1}{\theta^2} \mathbb{E}_1[f''(y)] \propto \frac{1}{\theta^2}$$

where $y = x/\theta$ and $p(y \mid \theta) = f(y)$.

Uninformative Priors — Scale Family

- The sampling distribution with the form

$$p(x \mid \theta) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$$

  for some integrable f is called a **scale family**.
- The Fisher's information is

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{1}{\theta^3} f''\left(\frac{x}{\theta}\right)\right] = \frac{1}{\theta^2}\mathbb{E}_1[f''(y)] \propto \frac{1}{\theta^2}$$

  where $y = x/\theta$ and $p(y \mid \theta) = f(y)$.
- The model is not uniform for all the θ.

Uninformative Priors — Scale Family

- Consider a re-parametrization with $\lambda = \log\theta$.

Uninformative Priors — Scale Family

- Consider a re-parametrization with $\lambda = \log \theta$.
- Then
$$p(x \mid \lambda) = e^{-\lambda} f(xe^{-\lambda})$$
- The Fisher's information is now
$$I(\lambda) = I(\theta) \left(\frac{d\theta}{d\lambda} \right)^2 \propto e^{-2\lambda} e^{2\lambda} \propto 1$$

Uninformative Priors — Scale Family

- Consider a re-parametrization with $\lambda = \log \theta$.
- Then

$$p(x \mid \lambda) = e^{-\lambda} f(xe^{-\lambda})$$

- The Fisher's information is now

$$I(\lambda) = I(\theta) \left(\frac{d\theta}{d\lambda} \right)^2 \propto e^{-2\lambda} e^{2\lambda} \propto 1$$

- The model is uniform in λ!
- We assign the prior in λ as $p(\lambda) \propto 1$.

Uninformative Priors — Scale Family

- Consider a re-parametrization with $\lambda = \log\theta$.
- Then
$$p(x \mid \lambda) = e^{-\lambda} f(xe^{-\lambda})$$
- The Fisher's information is now
$$I(\lambda) = I(\theta)\left(\frac{d\theta}{d\lambda}\right)^2 \propto e^{-2\lambda}e^{2\lambda} \propto 1$$
- The model is uniform in λ!
- We assign the prior in λ as $p(\lambda) \propto 1$.
- By change-of-variable, it corresponds to a prior for θ as
$$p(\theta) \propto p(\lambda)\frac{d\lambda}{d\theta} \propto \frac{1}{\theta}$$

(**improper prior distribution again**)

Uninformative Priors — Scale Family

- Consider a re-parametrization with $\lambda = \log\theta$.
- Then
$$p(x \mid \lambda) = e^{-\lambda} f(xe^{-\lambda})$$
- The Fisher's information is now
$$I(\lambda) = I(\theta)\left(\frac{d\theta}{d\lambda}\right)^2 \propto e^{-2\lambda}e^{2\lambda} \propto 1$$
- The model is uniform in λ!
- We assign the prior in λ as $p(\lambda) \propto 1$.
- By change-of-variable, it corresponds to a prior for θ as
$$p(\theta) \propto p(\lambda)\frac{d\lambda}{d\theta} \propto \frac{1}{\theta}$$
(**improper prior distribution again**)
- We observe that $p(\theta) \propto \sqrt{I(\theta)}$

Uninformative Priors — General Case

- Imageine a general samping distribution $p(x \mid \theta)$ with Fisher's information $I(\theta)$.
- Suppose there exists a bijective differentiable function g that re-parametrizes θ to $\lambda = g(\theta)$.

Uninformative Priors — General Case

- Imageine a general samping distribution $p(x \mid \theta)$ with Fisher's information $I(\theta)$.
- Suppose there exists a bijective differentiable function g that re-parametrizes θ to $\lambda = g(\theta)$.
- The Fisher's information for λ is

$$I(\lambda) = I(\theta) \left(\frac{d\theta}{d\lambda} \right)^2 = \frac{I(\theta)}{[g'(\theta)]^2}$$

Uninformative Priors — General Case

- Imageine a general samping distribution $p(x \mid \theta)$ with Fisher's information $I(\theta)$.
- Suppose there exists a bijective differentiable function g that re-parametrizes θ to $\lambda = g(\theta)$.
- The Fisher's information for λ is

$$I(\lambda) = I(\theta) \left(\frac{d\theta}{d\lambda} \right)^2 = \frac{I(\theta)}{[g'(\theta)]^2}$$

- If we choose g such that $g'(\theta) \propto \sqrt{I(\theta)}$, then we have

$$I(\lambda) \propto 1$$

Uninformative Priors — General Case

- Imageine a general samping distribution $p(x \mid \theta)$ with Fisher's information $I(\theta)$.
- Suppose there exists a bijective differentiable function g that re-parametrizes θ to $\lambda = g(\theta)$.
- The Fisher's information for λ is

$$I(\lambda) = I(\theta) \left(\frac{d\theta}{d\lambda} \right)^2 = \frac{I(\theta)}{[g'(\theta)]^2}$$

- If we choose g such that $g'(\theta) \propto \sqrt{I(\theta)}$, then we have

$$I(\lambda) \propto 1$$

- So we can assign a uniform prior for λ as $p(\lambda) \propto 1$.
- It corresponds to

$$p(\theta) \propto p(\lambda) \frac{d\lambda}{d\theta} \propto \sqrt{I(\theta)}$$

Jeffreys Prior

- The **Jeffreys Prior** is an uninformative prior defined by

$$p(\theta) \propto \sqrt{I(\theta)}$$

Jeffreys Prior

- The **Jeffreys Prior** is an uninformative prior defined by

$$p(\theta) \propto \sqrt{I(\theta)}$$

- The Jeffreys prior is invariance under re-parametrization in the sense that if $\lambda = g(\theta)$, then

$$p(\lambda) \propto \sqrt{I(\lambda)} = \sqrt{I(\theta)}\frac{d\theta}{d\lambda} \propto p(\theta)\frac{d\theta}{d\lambda}$$

Jeffreys Prior — Example

- Recall the binomial case with $y \mid \theta \sim \text{Binom}(n, \theta)$
- The conjugate prior is $\text{Beta}(\alpha, \beta)$ with $\alpha, \beta > 0$.

Jeffreys Prior — Example

- Recall the binomial case with $y \mid \theta \sim \mathrm{Binom}(n, \theta)$
- The conjugate prior is $\mathrm{Beta}(\alpha, \beta)$ with $\alpha, \beta > 0$.
- The Jeffreys prior gives

$$p(\theta \mid n) \propto \sqrt{I(\theta)} \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \theta^{-1/2}(1-\theta)^{-1/2}$$

- The Jeffreys corresponds to $\mathrm{Beta}(1/2, 1/2)$ distribution.

Jeffreys Prior — Example

- Recall the binomial case with $y \mid \theta \sim \mathrm{Binom}(n, \theta)$
- The conjugate prior is $\mathrm{Beta}(\alpha, \beta)$ with $\alpha, \beta > 0$.
- The Jeffreys prior gives

$$p(\theta \mid n) \propto \sqrt{I(\theta)} \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \theta^{-1/2}(1-\theta)^{-1/2}$$

- The Jeffreys corresponds to $\mathrm{Beta}(1/2, 1/2)$ distribution.
- $\mathrm{Beta}(1/2, 1/2)$ is both **uninformative** and **conjugate** for the binomial case.

Case Study — Normal Distribution with Known Variance

- Consider an i.i.d. sequence of normal random variables:

$$x_1, \ldots, x_n \sim \mathcal{N}(\theta, \sigma^2)$$

  where μ is the unknown parameter and σ^2 is given.

Case Study — Normal Distribution with Known Variance

- Consider an i.i.d. sequence of normal random variables:

$$x_1, \ldots, x_n \sim \mathcal{N}(\theta, \sigma^2)$$

where μ is the unknown parameter and σ^2 is given.

- The likelihood function is

$$L(\theta; x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\} \propto \exp\left\{-\frac{n\theta^2}{2\sigma^2} + \frac{S_n}{\sigma^2}\theta\right\}$$

Case Study — Normal Distribution with Known Variance

- Consider an i.i.d. sequence of normal random variables:

$$x_1, \ldots, x_n \sim \mathcal{N}(\theta, \sigma^2)$$

where μ is the unknown parameter and σ^2 is given.

- The likelihood function is

$$L(\theta; x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right\} \propto \exp\left\{-\frac{n\theta^2}{2\sigma^2} + \frac{S_n}{\sigma^2}\theta\right\}$$

- The conjugate prior is Gaussian (with kernel $\exp\{-A\theta^2 + B\theta\}$)

$$p(\theta) \propto \exp\left\{-\frac{(\theta - \mu)^2}{2\tau^2}\right\} \propto \exp\left\{-\frac{\theta^2}{2\tau^2} + \frac{\mu}{\tau^2}\theta\right\}$$

Case Study — Normal Distribution with Known Variance

- The posterior is

$$p(\theta \mid x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\theta^2 + \left(\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}\right)\theta\right\}$$

Case Study — Normal Distribution with Known Variance

- The posterior is

$$p(\theta \mid x_1, \ldots, x_n) \propto \exp\left\{ -\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\theta^2 + \left(\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}\right)\theta \right\}$$

- The posterior follows

$$p(\theta \mid x_1, \ldots, x_n) \sim \mathcal{N}\left(\frac{\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

Case Study — Normal Distribution with Known Variance

- The posterior is

$$p(\theta \mid x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\theta^2 + \left(\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}\right)\theta\right\}$$

- The posterior follows

$$p(\theta \mid x_1, \ldots, x_n) \sim \mathcal{N}\left(\frac{\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

- MAP and posterior mean are both

$$\hat{\theta} = \frac{\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

Case Study — Normal Distribution with Known Variance

- The posterior is

$$p(\theta \mid x_1, \ldots, x_n) \propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\theta^2 + \left(\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}\right)\theta\right\}$$

- The posterior follows

$$p(\theta \mid x_1, \ldots, x_n) \sim \mathcal{N}\left(\frac{\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}\right)$$

- MAP and posterior mean are both

$$\hat{\theta} = \frac{\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

- If we generate a new observation $\tilde{x}$, then

$$\tilde{x} \sim \mathcal{N}\left(\frac{\frac{S_n}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} + \sigma^2\right)$$

Case Study — Normal Distribution with Known Variance

- Since the normal distribution with known variance is a location family of θ. The uninformative prior is
$$p(\theta) \propto 1$$
- The Jeffreys prior is (prove) $p(\theta) \propto 1$ as well.

Case Study — Normal Distribution with Known Variance

- Since the normal distribution with known variance is a location family of θ. The uninformative prior is

$$p(\theta) \propto 1$$

- The Jeffreys prior is (prove) $p(\theta) \propto 1$ as well.
- Both correspond to the conjugate prior with $\tau^2 \to \infty$ and mu some constant.

Case Study — Normal Distribution with Known Variance

- Since the normal distribution with known variance is a location family of θ. The uninformative prior is

$$p(\theta) \propto 1$$

- The Jeffreys prior is (prove) $p(\theta) \propto 1$ as well.
- Both correspond to the conjugate prior with $\tau^2 \to \infty$ and mu some constant.
- Using uninformative prior, the posterior is

$$p(\theta \mid x_1, \dots, x_n) \sim \mathcal{N}\left(\frac{S_n}{n}, \frac{\sigma^2}{n}\right)$$

Case Study — Normal Distribution with Known Mean

- Suppose we have i.i.d. sequence $x_1, \ldots, x_n$ from $\mathcal{N}(0, \sigma^2)$.

Case Study — Normal Distribution with Known Mean

- Suppose we have i.i.d. sequence $x_1, \ldots, x_n$ from $\mathcal{N}(0, \sigma^2)$.
- The likelihood

$$L(\sigma^2) \propto \prod_{i=1}^{n} (\sigma^2)^{-1/2} \exp\left\{-\frac{x_i^2}{2\sigma^2}\right\} \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{S_{xx}}{2\sigma^2}\right\}$$

Case Study — Normal Distribution with Known Mean

- Suppose we have i.i.d. sequence $x_1, \ldots, x_n$ from $\mathcal{N}(0, \sigma^2)$.
- The likelihood

$$L(\sigma^2) \propto \prod_{i=1}^{n} (\sigma^2)^{-1/2} \exp\left\{-\frac{x_i^2}{2\sigma^2}\right\} \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{S_{xx}}{2\sigma^2}\right\}$$

- The conjugate prior is the inverse-Gamma distribution:

$$p(\sigma^2) \propto (\sigma^2)^{-\alpha-1} \exp\left\{-\frac{\beta}{\sigma^2}\right\}$$

Case Study — Normal Distribution with Known Mean

- Suppose we have i.i.d. sequence $x_1, \ldots, x_n$ from $\mathcal{N}(0, \sigma^2)$.
- The likelihood

$$L(\sigma^2) \propto \prod_{i=1}^{n} (\sigma^2)^{-1/2} \exp\left\{-\frac{x_i^2}{2\sigma^2}\right\} \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{S_{xx}}{2\sigma^2}\right\}$$

- The conjugate prior is the inverse-Gamma distribution:

$$p(\sigma^2) \propto (\sigma^2)^{-\alpha-1} \exp\left\{-\frac{\beta}{\sigma^2}\right\}$$

- The posterior is $\mathrm{InvGamma}(\alpha + n/2, \beta + S_{xx}/2)$.

Case Study — Normal Distribution with Known Mean

- Suppose we have i.i.d. sequence $x_1, \ldots, x_n$ from $\mathcal{N}(0, \sigma^2)$.
- The likelihood

$$L(\sigma^2) \propto \prod_{i=1}^{n} (\sigma^2)^{-1/2} \exp\left\{-\frac{x_i^2}{2\sigma^2}\right\} \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{S_{xx}}{2\sigma^2}\right\}$$

- The conjugate prior is the inverse-Gamma distribution:

$$p(\sigma^2) \propto (\sigma^2)^{-\alpha-1} \exp\left\{-\frac{\beta}{\sigma^2}\right\}$$

- The posterior is $\mathrm{InvGamma}(\alpha + n/2, \beta + S_{xx}/2)$.
- The Jeffreys prior is $p(\sigma^2) \propto \sigma^{-2}$, or $\mathrm{InvGamma}(0, 0)$