# STAT201B Lecture 14 Bayesian Statistics

## 1 Bayesian Framework

**Logic: Overview**

Bayesian statistics 建立在对**概率的主观解释**之上。

换句话说，Bayesian statistician 使用概率语言来反映关于问题的两种不同的不确定性：

1. **偶然性不确定性 (aleatory uncertainty)**：
    - 由系统内在的随机性或对系统的随机观察引起。
    - 这种不确定性也被 frequentist statistics 所使用。
2. **认知性不确定性（epistemic uncertainty）**：
    - 来源于我们对系统理解的不完全
    - 科学研究的目的之一，就是减少这种认知性不确定性

**Logic**

关于 Bayesian framework 的更多论述, 见 STA3020 Lecture 25

### 1.1 Definition: Bayesian statistics 的基本概念

Bayesian statistics 基于 Bayes Theorem:

令 $x^n = (x_1, \dots, x_n)$ 表示 the observed data, 则有

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{f(x^n)}$$

我们由此定义以下概念:

- $f(\theta)$: the **prior** density – 在收集数据前对 $\theta$ 的先验认知
- $f(x^n|\theta)$: the **likelihood** – 给定特定的 $\theta$, data 的 joint density
- $f(\theta|x^n)$: the **posterior** density – 在收集数据后对 $\theta$ 的后验认知
- $f(x^n)$: the **normalizing constant** – data 的 marginal distribution (可能较难求出)

**Remark: STA3020 中的定义**

- **Prior distribution:** 任何定义在 parameter space $\Theta$ 上的 distribution $\Lambda(\theta|\lambda)$ 被称为 prior distribution with hyperparameter $\lambda$
- **Posterior distribution:** 对于一个 sample $X$ drawn from $f(x|\theta)$, posterior distribution $\pi(\theta|x,\lambda)$ 被定义为 $\theta$ conditional on observed $X = x$ 的概率, 即

$$\pi(\theta|x,\lambda) = \frac{f(x|\theta)\Lambda(\theta|\lambda)}{\int f(x|\theta)\Lambda(\theta|\lambda)d\theta} \propto f(x|\theta)\Lambda(\theta|\lambda) = K(\theta|x,\lambda)\cdot h(x,\lambda)$$

  其中 $K(\theta|x,\lambda)$ 被称为 **kernel of posterior distribution**

## 1.2 Definition: Kernel

$\theta$ 的 density 的 **kernel** 为 density 中与 $\theta$ 有关的部分 (剔除与 $\theta$ 无关的 constant terms)

**Example**

令 $\theta \sim \mathcal{N}(m, v)$, 其中 $v$ 已知, 求 $\theta$ 的 kernel

---

# 2 Conjugate prior

**Logic**

Prior 的选取往往很重要, 为了便于计算, 一种常见的选取方法是选择 conjugate priors

---

## 2.1 Definition: Conjugate prior

$\theta$ 的 **conjugate prior** 满足: $f(\theta)$ 和 $f(\theta|x^n)$ 属于相同的 parametric family

**Remark: STA3020 中的定义**

考虑:

- distribution family $\mathcal{F} = \{f(x|\theta), \theta \in \Theta\}$ (elements 定义在 $\mathcal{X}$ 上)
- distribution family $\mathcal{G} = \{\Lambda(\theta|\lambda), \lambda \in \tilde{\Lambda}\}$ (elements 定义在 $\Theta$ 上)

则:

- $\mathcal{F}$ 和 $\mathcal{G}$ 被称为 **conjugate distribution family**, 若对于 $\forall \Lambda(\theta|\lambda) \in \mathcal{G}$, 有

$$\pi(\theta|x, \lambda) = \frac{f(x|\theta)\Lambda(\theta|\lambda)}{\int f(x|\theta)\Lambda(\theta|\lambda)d\theta} \in \mathcal{G}$$

- 且 $\mathcal{G}$ 的任意 element 被称为 **conjugate prior**

换言之, conjugate prior 和 posterior 同属于一个分布族

**Remark**

若使用 conjugate prior, 则计算的重点在于找出 kernel, 并研究加入 likelihood 的影响前后的变化 (prior 和 posterior 的 kernels 的区别)

---

## 2.2 Conjugate priors 的例子

**Logic**

关于 Conjugate priors 的例子, 见 STA2004 Lecture 21

**Example: Poisson-Gamma Model**

***问题:***

令 $X_1, \dots, X_n|\lambda \overset{i.i.d.}{\sim} Poisson(\lambda)$, prior 为 $\lambda \sim Gamma(a, b)$, 求 $\lambda$ 的 posterior distribution

***解答:***

首先用 kernel function 表示 posterior, 由于

$$
\begin{aligned}
f(\lambda) &= \frac{b^a}{\Gamma(a)} \cdot \lambda^{a-1} \cdot e^{-b\lambda} \propto \lambda^{a-1} \cdot e^{-b\lambda} \\
f(x_1, \dots, x_n|\lambda) &\propto \prod_{i=1}^{n} e^{-\lambda} \cdot \lambda^{x_i}
\end{aligned}
$$

因此 posterior 满足:

$$
\begin{aligned}
f(\lambda|x_1, \dots, x_n) &\propto e^{-n\lambda} \cdot \lambda^{\sum_{i=1}^n x_i} \cdot \lambda^{a-1} \cdot e^{-b\lambda} \\
&= \lambda^{(\sum_{i=1}^n x_i + a) - 1} \cdot e^{-(n+b)\lambda}
\end{aligned}
$$

注意到这是一个 $Gamma(a', b')$ 的 kernel, 其中 $a' = \sum_{i=1}^n x_i + a, b' = n + b$, 因此 posterior distribution 为 $Gamma(a', b')$

**⚠ Remark**

注意到对于 prior, 我们有:

$$
\begin{aligned}
\mathbb{E}[\lambda] &= \frac{a}{b} \\
Var[\lambda] &= \frac{a}{b^2}
\end{aligned}
$$

对于 posterior, 我们有:

$$
\mathbb{E}[\lambda|x_1, \dots, x_n] = \frac{\sum_{i=1}^n x_i + a}{n+b} = \frac{n}{n+b} \cdot \bar{X} + \frac{b}{n+b} \cdot \frac{a}{b}
$$

因此这里的 posterior mean 可以视作 sample mean 和 prior mean 的一个 weighted sum

当数据量足够大, 即 $n \to \infty$ 时, sample mean 占主导; 当数据量较小, 即 $n \to 0$ 时, prior mean 占主导

**⚠ Remark**

若 Gamma distribution 使用 scale parameter, 则 posterior 可以通过以下计算求得:

2. 常见的 conjugate model: Poisson - Gamma model

- 假设 $\mathbb{X} = \{x_1, \dots, x_n\}$ 为服从 Poisson$(\theta)$ 的 i.i.d. sample. 则 $\mathbb{X}$ 有 density (likelihood)

$$
p(\mathbb{X}|\theta) = \prod_{j=1}^{n} \frac{\theta^{x_j} e^{-\theta}}{x_j!}
$$

- 一个 conjugate prior 为 Gamma$(\alpha, \beta)$, $\alpha$ 与 $\beta$ 已知,

$$
\pi(\theta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \theta^{\alpha-1} e^{-\frac{\theta}{\beta}}
$$

- 则 posterior distribution 为

$$
\begin{aligned}
p(\theta|\mathbb{X}) &\propto P(\mathbb{X}|\theta) \cdot \pi(\theta) \\
&\propto \prod_{j=1}^{n} \theta^{x_j} e^{-\theta} \cdot \theta^{\alpha-1} e^{-\frac{\theta}{\beta}} \\
&= \theta^{\sum x_j + \alpha - 1} \cdot e^{-(n+\frac{1}{\beta})\theta}
\end{aligned}
$$

因此

$$
p(\theta|\mathbb{X}) = \frac{1}{\Gamma(\tilde{\alpha})\tilde{\beta}^{\tilde{\alpha}}} \theta^{\tilde{\alpha}-1} e^{-\frac{\theta}{\tilde{\beta}}}
$$

其中

$$
\tilde{\alpha} = \alpha + \sum_j x_j, \quad \tilde{\beta} = \frac{\beta}{1+n\beta}
$$

### Example: Normal-Normal Model

3、posterior distribution 的计算：data ~ N，prior ~ N

假设 mean $\mu \sim N(\nu, \tau^2)$，且 data $x_1, \cdots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$，其中 $\sigma^2$ 与 $\tau^2$ 已知.

求 posterior 分布的均值与方差

Step ①：写出 prior $\pi(\mu)$

$$\pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} exp(-\frac{(\mu-\nu)^2}{2\tau^2})$$

Step ②：写出 data distribution $f(x_i|\mu)$

$$f(x_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x_i-\mu)^2}{2\sigma^2})$$

Step ③：写出 $f(x_1, \cdots, x_n|\mu)$

$$f(x_1, \cdots, x_n|\mu) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2)$$

Step ④：写出 joint distribution $f(x_1, \cdots, x_n, \mu)$

$$f(x_1, \cdots, x_n, \mu) = \frac{1}{(\sqrt{2\pi\sigma^2})^n \times \sqrt{2\pi\tau^2}} exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{(\mu-\nu)^2}{2\tau^2})$$

注：接下来我们不必计算 $\int_\mu f(x_1, \cdots, x_n, \mu)\, d\mu$，因为对于关于 $\mu$ 的函数 $f(\mu|x_1, \cdots, x_n)$，这一项相当于常数（与 $\mu$ 无关），因此 $f(\mu|x_1, \cdots, x_n) = C\pi(\mu) f(x_1, \cdots, x_n|\mu) \propto \pi(\mu) f(x_1, \cdots, x_n|\mu)$.

通过观察 $\pi(\mu) f(x_1, \cdots, x_n|\mu)$，我们同样可以得到 posterior 分布.（C 被称为 normalizing constant）

如 $f(x) \propto exp(-\frac{1}{2\sigma^2}x^2)$，则可知 $x \sim N(0, \sigma^2)$

Step ⑤：写出与 posterior 成正比的项，整理并求出 posterior 分布

$$\begin{aligned}
f(\mu|x_1, \cdots, x_n) &\propto f(x_1, \cdots, x_n, \mu) \\
&\propto exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{(\mu-\nu)^2}{2\tau^2}) \quad \text{（消除与 } \mu \text{ 无关的项）} \\
&= exp[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2+\mu^2-2x_i\mu) - \frac{1}{2\tau^2}(\mu^2+\nu^2-2\nu\mu)] \\
&\propto exp[-\frac{1}{2\sigma^2}(n\mu^2-2n\bar{x}\mu) - \frac{1}{2\tau^2}(\mu^2-2\nu\mu)] \\
&= exp\{-\frac{1}{2}[(\frac{n}{\sigma^2}+\frac{1}{\tau^2})\mu^2 - (\frac{2n\bar{x}}{\sigma^2}+\frac{2\nu}{\tau^2})\mu]\} \\
&= exp\{-\frac{1}{2}(\frac{n}{\sigma^2}+\frac{1}{\tau^2})[\mu^2 - 2\frac{\frac{n\bar{x}}{\sigma^2}+\frac{\nu}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\mu]\} \\
&\propto exp[-\frac{1}{2}\frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}(\mu - \frac{n\bar{x}\tau^2+\sigma^2\nu}{n\tau^2+\sigma^2})^2]
\end{aligned}$$

$$\mu|x_1, \cdots, x_n \sim N(\nu_p, \tau_p^2)$$

其中 $\nu_p = \frac{n\bar{x}\tau^2+\sigma^2\nu}{n\tau^2+\sigma^2} = \nu \times \frac{\sigma^2}{n\tau^2+\sigma^2} + \bar{x} \times \frac{n\tau^2}{n\tau^2+\sigma^2}$

$$\tau_p^2 = \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}$$

注：① 由此可以构建出 $\mu$ 的 $1-\alpha$ CI（Bayesian confidence interval）：

$$A = [\nu_p - \sqrt{\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}} \cdot \alpha_{\alpha/2},\ \nu_p + \sqrt{\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}} \cdot \alpha_{\alpha/2}]$$

$$P(\mu \in A|x_1, \cdots, x_n) = 1-\alpha$$

② 当 sample size $n$ 为 0 时，$\tau_p^2 = \tau^2$，$\nu_p = \nu$（posterior = prior）

当 sample size $n \to \infty$ 时，$\tau_p^2 \approx \frac{\sigma^2}{n}$，$\nu_p \approx \bar{x}$（posterior ≈ frequentist point of view）

这被视作从 'experience' 到 'data' 的 transition.

当样本容量较小时，'experience' 即 prior distribution dominates the posterior distribution

当样本容量较大时，'data' dominates the posterior distribution

### Example: Beta-Binomial Model

3、常见的 conjugate model：Beta-binomial model

- 假设 $\mathcal{X}=\{x_1,\cdots,x_n\}$ 为服从 Binomial $(m_j,\theta)$（$m_j$ 为 fixed constant）的独立 sample，则 $\mathcal{X}$ 有 density

$$p(\mathcal{X}|\theta)=\prod_{j=1}^{n}\binom{m_j}{x_j}\theta^{x_j}(1-\theta)^{m_j-x_j}$$

- 一个 conjugate prior 为 Beta$(\alpha,\beta)$，$\alpha$ 与 $\beta$ 已知，

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

- 则 posterior distribution 为

$$p(\theta|\mathcal{X})\propto p(\mathcal{X}|\theta)\cdot\pi(\theta)$$

$$\propto\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\Sigma_j x_j}(1-\theta)^{\Sigma_j(m_j-x_j)}$$

$$\propto\theta^{\Sigma_j x_j+\alpha-1}(1-\theta)^{\Sigma_j(m_j-x_j)+\beta-1}$$

因此

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

其中

$$\tilde{\alpha}=\alpha+\Sigma_j x_j\ ,\quad\tilde{\beta}=\beta+\Sigma_j(m_j-x_j)$$

### Example: Inverse Gamma-Normal Moel

4、常见的 conjugate model：Inverse Gamma - Normal model

- 假设 $\mathcal{X}=\{x_1,\cdots,x_n\}$ 为服从 $N(\mu,\sigma^2)$ 的 i.i.d. sample（其中 $\mu$ 已知），则 $\mathcal{X}$ 有 density

$$p(\mathcal{X}|\sigma^2)=\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_j-\mu)^2}{2\sigma^2}}$$

- 一个 conjugate prior 为 Inverse-Gamma distribution，$IG(\alpha,\beta)$，$\alpha$ 与 $\beta$ 已知，

$$\pi(\sigma^2)=\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}e^{-\frac{\beta}{\sigma^2}}$$

- 则 posterior distribution 为

$$p(\sigma^2|\mathcal{X})\propto p(\mathcal{X}|\sigma^2)\cdot\pi(\sigma^2)$$

$$\propto\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_j-\mu)^2}{2\sigma^2}}\cdot(\sigma^2)^{-\alpha-1}e^{-\frac{\beta}{\sigma^2}}$$

$$\propto(\sigma^2)^{-\frac{n}{2}-\alpha-1}e^{-\frac{\beta+\frac{1}{2}\Sigma_j(x_j-\mu)^2}{\sigma^2}}$$

因此

$$p(\theta|\mathcal{X})\sim IG\left(\alpha+\frac{n}{2},\ \beta+\frac{1}{2}\Sigma_j(x_j-\mu)^2\right)$$

注：关于 Inverse-Gamma distribution：

若 r.v. $X\sim IG(\alpha,\beta)$，$x>0$，则

$$\pi(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}(x)^{-\alpha-1}e^{-\frac{\beta}{x}}$$

令 $Y=\frac{1}{X}$，则 $X=\frac{1}{Y}=h(Y)$，$h'(y)=-\frac{1}{y^2}$，因此

$$\pi(y)=\frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{y}\right)^{-\alpha-1}e^{-\beta y}\cdot\left|-\frac{1}{y^2}\right|$$

$$=\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha+1-2}e^{-\beta y}$$

$$\sim Gamma\left(\alpha,\frac{1}{\beta}\right)$$

### Example: Beta-Geometric Model (补充)

5、常见的 conjugate model：Beta - Geometric model

- 假设 $X=\{x_1,\cdots,x_n\}$ 为服从 $Geo(\theta)$ 的 i.i.d. sample. 则 $X$ 有 density

$$p(X|\theta)=\prod_{j=1}^{n}(1-\theta)^{x_j-1}\theta$$

- 一个 conjugate prior 为 $Beta(\alpha,\beta)$，$\alpha$ 与 $\beta$ 已知，

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

- 则 posterior distribution 为

$$\begin{aligned} p(\theta|X) &\propto p(X|\theta)\cdot\pi(\theta) \\ &\propto \prod_{j=1}^{n}(1-\theta)^{x_j-1}\theta\cdot\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{n+\alpha-1}\cdot(1-\theta)^{\sum_j x_j+\beta-n-1} \end{aligned}$$

因此

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

其中

$$\tilde{\alpha}=\alpha+n,\quad \tilde{\beta}=\sum_j x_j+\beta-n$$

注：还有其他 pairs，如 Beta - Negative binomial model

---

♨ Logic

除了能使计算更加简单的 conjugate prior, prior 还有其他几种选择, 主要有以下几种学派:

1. **Subjective Bayesianism (主观贝叶斯学派)**
   - 观点: prior 应尽可能反映研究者对问题的uncertainty 的 prior knowledge
   - 实现方式: 通过 prior elicitation 方法, 从专家或研究者的知识中构建先验。
2. **Objective Bayesianism (客观贝叶斯学派)**
   - 观点: prior 应尽量减少主观信息, 使结果尽可能 objective (让数据在推断中起主导作用)
   - 实现方法: 使用 non-informative priors
3. **Robust Bayesianism (稳健贝叶斯学派)**
   - 观点: 不同的人可能持有不同的 priors, 且 priors 难以被精确表达, 因此, 应研究结果关于 prior 的选择的 sensitivity
   - 实现方法: 确保推断在合理的 change of prior 下仍然稳定可靠

♨ Logic

常见的 prior 包括 non-informative prior (Jeffrey's prior) 和 improper prior

关于 non-informative prior 和 improper prior 的详细论述, 见 STA3020 Lecture 25

## 3 Non-Informative Prior 和 Improper Prior

---

Logic

Non-informative prior 的目的是让数据本身 (而不是 prior assumption) 在推断中占主要作用, 常见的 non-informative prior 包括:

- uniform prior
- Jeffrey's prior

其中, Jeffrey's prior 的一个特殊性质是 **transformation invariant** (在 one-dimensional 情况下), 即在不同参数化下形式不变

## 3.1 Definition: Jeffrey's prior

若 sample $X \sim f(x|\theta)$, 其 information matrix 为 $I(\theta)$, 则 Jeffrey's prior 被定义为:

$$f(\theta) \propto (\det(I(\theta)))^{\frac{1}{2}}$$

Remark

在 $\theta$ 为 one-dimensional 的情况下, Jeffrey's prior 有以下性质:

- Jeffrey's prior 的形式为:

$$f(\theta) \propto (I(\theta))^{\frac{1}{2}}$$

- Jeffrey's prior is invariant under reparametrization, 即对于 reparametrization $\phi = h(\theta)$, 有

$$\begin{aligned} f_\phi(\phi) &= f_\theta(\theta) \cdot \left|\frac{\partial \theta}{\partial \phi}\right| \quad \text{(change of variable)} \\ &\propto I(\theta)^{\frac{1}{2}} \cdot \left|\frac{\partial \theta}{\partial \phi}\right| \\ &= I(\phi)^{\frac{1}{2}} \quad \text{(Fisher information 的性质)} \end{aligned}$$

- Jeffrey's prior 为 non-informative prior, 因为其 maximizes prior 和 posterior 之间的 KL 散度的期望值. 由于 $f(\theta)$ 和 $f(\theta|x)$ 之间的 gap 是 $x$ 的 information, 因此 Jeffrey's prior 可以被视作保留了最多的 $x$ 的 information, 即令 posterior 保留了最少的 $\theta$ 的 information

## 3.2 Definition: Improper prior (补充)

令 sample $X \sim f(x|\theta)$, 则定义在 $\Theta$ 上的 prior function $f(\theta)$ 被称为 **improper prior**, 若

- $f(\theta) \geq 0$ (类似于 pdf, 非负)
- $\int_\Theta f(\theta)\, d\theta = \infty$ (区别于 pdf, 且不能被 normalize 成 pdf)
- $\int_\Theta f(x|\theta) f(\theta)\, d\theta < \infty$ (确保 marginal 是 pdf, 或能被 normalize 成 pdf)

Remark

尽管名字为 improper prior, 但它可以是一个 prior 的 proper choice

Example

1. 对于 location family with location parameter $\mu$ 且 $\Theta = \mathbb{R}$, 则一个常用的 improper prior 为 $f(\theta) \equiv 1$
2. 对于 scale family with scale parameter $\sigma$ 且 $\Theta = (0, \infty)$, 则一个常用的 improper prior 为 $f(\sigma) = \frac{1}{\sigma}$

## 3.3 Uniform prior

- **当 $\theta$ 的取值范围是有限区间时:** uniform prior 为 proper prior
- **当 $\theta$ 的取值范围是无限区间时:** uniform prior 为 improper prior

**Example**

若 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$, 则可以选择 uniform prior $f(\theta) \propto 1$, 此时 posterior 为

$$\begin{aligned} f(\theta|x^n) &\propto f(x^n|\theta)f(\theta) \\ &= \exp\left\{-\frac{1}{2}[n\theta^2 - 2n\theta\bar{X}_n]\right\} \end{aligned}$$

为 $\mathcal{N}(\bar{X}_n, 1/n)$ 的 kernel

---

# 4 Bayesian Estimation

**Logic**

在 Bayesian statistics 中, 所有的 inference 都是基于 posterior distribution, 类似于 frequentist statistics, 我们可以利用 posterior 进行 point estimation

关于 Bayesian estimator 的更多论述, 见 [STA3020 Lecture 26]

---

## 4.1 Definition: Average risk (补充)

**Logic**

在 point estimation中，我们将关于 parameter $\theta$ 和 estimator $\delta$ 的 loss function 和 risk function 分别定义为 $\mathcal{L}(\theta, \delta)$ 和 $R(\theta, \delta) = \mathbb{E}_\theta(\mathcal{L}(\theta, \delta)) = \int \mathcal{L}(\theta, \delta) dF(x|\theta)$, estimator $\delta$ 应 minimize risk

但在 Bayesian analysis 中， $\theta \in \mathbf{\Theta}$ 为 random variable. 考虑 $\theta = \theta_0$ 时的 $\mathcal{L}(\theta_0, \delta)$ 和 $R(\theta_0, \delta)$ 并不足够, 因此我们考虑 average risk

令:

1. loss function 为 $\mathcal{L}(\theta, \delta)$
2. risk function 为 $R(\theta, \delta) = \mathbb{E}_\theta[\mathcal{L}(\theta, \delta)]$
3. prior distribution 为 $\Lambda(\theta|\lambda)$

则 **average risk** 被定义为:

$$\begin{aligned} r(\Lambda, \delta) &= \mathbb{E}[R(\theta, \delta)] \\ &= \int_\Theta R(\theta, \delta) d\Lambda(\theta|\lambda) \\ &= \int_\Theta \int_{\mathcal{X}} \mathcal{L}(\theta, \delta)\, dF(x|\theta)\, d\Lambda(\theta|\lambda) \end{aligned}$$

**Remark**

average risk 实际上是关于 $\lambda$ 的函数, 但我们在 notation $r(\Lambda, \delta)$ 中省略了 $\lambda$

---

## 4.2 Definition: Bayesian estimator (补充)

**任意 minimize average risk $r(\Lambda, \delta)$** 的 estimator $\delta$ 被称为一个 **Bayesian estimator w.r.t prior $\Lambda$**

> **Remark**
>
> 对于 Bayesian estimator 的求解方法, 见 [STA3020 Lecture 26]

---

## 4.3 特定 loss function 下的 Bayesian estimator (补充)

在特定 condition 下, 有以下结论:

1. 令 $\mathcal{L}(g(\theta), \delta) = (\delta - g(\theta))^2$ **(quadratic loss)**,
   则 $\delta_\Lambda(X) = \mathbb{E}[g(\theta)|X]$ **(posterior mean)**
2. 令 $\mathcal{L}(g(\theta), \delta) = \omega(\theta)(\delta - g(\theta))^2$ **(weighted quadratic loss)**,
   则 $\delta_\Lambda(X) = \frac{\int \omega(\theta) g(\theta) \cdot \pi(\theta|x) d\theta}{\int \omega(\theta) \cdot \pi(\theta|x) d\theta}$ $\left(\frac{\text{posterior weighted mean}}{\text{weight mean}}\right)$
3. 令 $\mathcal{L}(g(\theta), \delta) = |\delta - g(\theta)|$ **(absolute loss)**,
   则 $\delta_\Lambda(X) = \text{median}[g(\theta)|X]$ **(posterior median)**

---

## 4.4 一个常见的 Bayesian estimator: posterior mean

Posterior mean 通常被用作 point estimator:

$$\mathbb{E}[\theta|X_1, \ldots, X_n] = \int \theta \cdot f(\theta|X_1, \ldots, X_n) d\theta$$

> **Remark**
>
> Posterior mean 通常可被写作 prior mean 和 MLE (常为 MLE) 的 weighted sum, 这点在接下来的两个例子中有很好的展现

**Example: Binomial-Beta conjugacy (补充)**

例 2: (Binomial-Beta conjugacy)

令 random sample $X_1, \cdots, X_n \overset{i.i.d.}{\sim} Ber(p)$, 令 $p$ 的 prior 为 $Beta(\alpha, \beta)$.

求 quadratic loss 对应的 $p$ 的 Bayesian estimator

(Step 1: 求出 posterior distribution)

$$\begin{aligned} \pi(p|x) &\propto f(x|p) \cdot \Lambda(p|\alpha, \beta) \\ &\propto \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \cdot p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{\alpha + \Sigma x_i - 1}(1-p)^{\beta + n - \Sigma x_i - 1} \end{aligned}$$

$\Rightarrow p|x \sim Beta(\alpha + \Sigma x_i, \beta + n - \Sigma x_i)$

(Step 2: 求出 posterior mean)

$$\Rightarrow \delta_\Lambda(x) = E[p|x] = \frac{\alpha + \Sigma x_i}{n + \alpha + \beta} = \frac{\alpha+\beta}{n+\alpha+\beta} \cdot \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{mean of prior}} + \frac{n}{n+\alpha+\beta} \cdot \underbrace{\bar{x}}_{\text{mean of sample}}$$

注: Bayesian estimator 可以看成 mean of prior 和 mean of sample 的 weighted sum.

① 若 $n \to \infty$, 则 $\delta_\Lambda(X) \to \bar{X}$ (此时 $\alpha = \beta = 0$ 为一个 improper prior)

② 若 $\alpha \to \infty, \beta \to \infty, \alpha/\beta \to C < \infty$, 则 $\delta_\Lambda(X) \to \frac{C}{1+C}$

### Example: Normal mean Bayesian estimator

例 4: ( Normal mean Bayes estimator )

令 random sample $X_1, \dots, X_n \overset{i.i.d.}{\sim} N(\theta, \sigma^2)$, 其中 $\sigma$ 已知. 令 $\theta$ 的 prior 为 $N(\mu, \tau^2)$.

求 quadratic loss 对应的 $\theta$ 的 Bayesian estimator

( Step 1: 求出 posterior distribution )

$$
\begin{aligned}
\pi(\theta|x) &\propto f(x|\theta) \cdot \Lambda(\theta|\mu, \tau) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\Sigma(x_i-\theta)^2\right) \cdot \exp\left(-\frac{1}{2\tau^2}(\theta-\mu)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}n\theta^2 + \frac{2\theta}{2\sigma^2}\Sigma x_i\right) \cdot \exp\left(-\frac{1}{2\tau^2}\theta^2 + \frac{2\theta}{2\tau^2}\cdot\mu\right) \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta - \frac{n\bar{x}/\sigma^2+\mu/\tau^2}{n/\sigma^2+1/\tau^2}\right)^2\right\}
\end{aligned}
$$

$$
\Rightarrow \theta|x \sim N\left(\frac{n\bar{x}/\sigma^2+\mu/\tau^2}{n/\sigma^2+1/\tau^2}, \frac{1}{n/\sigma^2+1/\tau^2}\right)
$$

( Step 2: 求出 posterior mean )

$$
\Rightarrow \delta_\Lambda(x) = E[\theta|x] = \frac{n\bar{x}/\sigma^2+\mu/\tau^2}{n/\sigma^2+1/\tau^2} = \frac{n/\sigma^2}{n/\sigma^2+1/\tau^2}\cdot\underbrace{\bar{x}}_{\text{mean of sample}} + \frac{1/\tau^2}{n/\sigma^2+1/\tau^2}\cdot\underbrace{\mu}_{\text{mean of prior}}
$$

---

## 5 Credible Interval (Posterior Interval)

### Logic

关于 credible interval 的更多论述, 见 [STA4042 Lecture Lecture 15]

---

### 5.1 Definition: Credible interval (Posterior interval)

$\theta$ 的一个 $1-\alpha$ credible interval $C_n$ 满足:

$$
\mathbb{P}(\theta \in C_n | X_1, \dots, X_n) = 1-\alpha
$$

#### Remark: Credible interval 和 confidence interval 的区别

相较于 confidence interval, credible interval 有以下不同:

- Credible interval 是一个关于 $\theta$ (而不是 $C_n$, 关于 $X_1, \dots, X_n$ 的一个 function) 的 statement
- Credible interval 是一个 equality statement, 这点有别于 frequentist interval (给 probability of coverage 设定了一个 lower bound), 这是因为 credible interval 只是提供了一个 summary of posterior distribution
- Credible interval 不一定有很好的 frequentist coverage rates

##### Example

令:

1. $X \sim \text{Bernoulli}(p)$
2. 只观测到一个样本 $(n=1)$
3. prior distribution 为 $p \sim \text{Uniform}(0,1)$

此时,

- 若观测到 $X=1$, 则 $p|X \sim \text{Beta}(2,1)$, $95\%$ posterior interval $\approx [0.0526, 0.997]$
- 若观测到 $X=0$, 则 $p|X \sim \text{Beta}(1,2)$, $95\%$ posterior interval $\approx [0.003, 0.947]$

若重复很多次实验 (每次只抽一个样本), 我们研究有多大比例的 credible interval 会覆盖 true value $p$:

- 若 $p=0.1$, 则无论观测值是 $0$ 还是 $1$, 都 100% 会覆盖 true value
- 若 $p=0.001$, 则无论观测值是 $0$ 还是 $1$, 都 100% 不会覆盖 true value
- 若 $p=0.999$, 则无论观测值是 $0$ 还是 $1$, 都 100% 不会覆盖 true value

换言之, 这种 credible interval 的 frequentist coverage 不是 95%, 而且取决于 true value $p$

⚠ **Remark**

相较于 report an interval, 直接作出 $f(\theta|x^n)$ 的图像往往更 informative

---

## 5.2 求 credible interval 的方法: equal-tail credible interval

The **$1-\alpha$ equal-tail credible interval** $(a,b)$ 满足:

$$\int_{-\infty}^{a} f(\theta|x^n)\,d\theta = \int_{b}^{\infty} f(\theta|x^n)\,d\theta = \alpha/2$$

即区间左侧的 cdf 和右侧的 tail probability 相等

---

## 5.3 求 credible interval 的方法: highest posterior density (HPD)

The **highest posterior density (HPD) region** $R_n$ 满足:

- $\mathbb{P}(\theta \in R_n|x^n) = 1-\alpha$
- $R_n = \{\theta : f(\theta|x^n) > k\}$ for some $k$

⚠ **Remark**

若 $f(\theta|x^n)$ 为 unimodal, 则 $R_n$ 为一个 interval