Chapter 4: Bayesian Inference

강의 목표

- 베이지안 추론의 이해
  - Likelihood Method
  - Bayesian Method

Statistical Models: Main Focus

- Inference about parameters, based on data.

Notations and Settings: Distributions

- Denote an unobserved parameter of interest as θ.
- Denote our data as $\mathbf{D}$.
- Our probability model for the data, given a value of θ, is denoted $P(\mathbf{D} \mid \theta)$.
- 가정: 주어진 데이터 D 혹은 X는 확률 변수 $\mathbf{X}$의 관측치이며 $\mathbf{X}$의 분포는 unknown parameter θ에 의존하는 density function $f(\cdot \mid \theta)$를 가진다.

  e.g.: Normal distribution $N(\mu, \sigma^2)$

Notations and Settings: Data

- Suppose we observe an iid sample of data $X = (X_1, ..., X_n)$.
- Now X is considered fixed and known.
- Denote our data as the $n \times k$ matrix X.
- We denote the parameter(s) of interest to be the vector θ.

Likelihood Theory

- The likelihood function: $L(\theta \mid X) = f(X \mid \theta)$.
- $L(\theta \mid X)$ is a function of θ that shows how "likely" are various parameter values θ to have produced the data X that **were observed**.

Likelihood Principle

- Mathematically, if the data X represent iid observations from probability distribution $p(X \mid \theta)$, then

$$L(\theta \mid X) = \prod_{i=1}^{n} P(X_i \mid \theta)$$

where $X_1, ..., X_n$ are the n data vectors.

Likelihood Theory

- 목표: Parameter θ
- 주어진 정보: 데이터 X
- 데이터 X가 θ정보를 가지고 있으므로 X를 통해 θ를 추측.
- 주의해야 할 점: θ가 X의 분포를 결정. X가 θ를 결정하는 것이 아님.

Maximum Likelihood Estimator (MLE)

- In classical statistics, the specific value of θ that maximizes $L(\theta \mid X)$ is the maximum likelihood estimator (MLE) of θ.
- e.g., 동전을 100회 던졌을때 앞면이 100회 연속 나왔다고 가정하자. 동전을 던졌을때 앞면이 나올 확률 p는 다음 중 어느 것이 더 가능성이 있을까?
  1. $p = 0$
  2. $p = 0.5$
  3. $p = 1$

Likelihood Limitations

In many common probability models, when the sample size n is large,

- $L(\theta \mid X)$ is unimodal in θ.
- $L(\theta \mid X)$ is strictly concave.
- Unlike $P(\theta \mid X)$, $L(\theta \mid X)$ does not necessarily obey the usual laws for probability distributions.
- In the classical framework, all the randomness within $L(\theta \mid X)$ is attached to X, not to θ

Likelihood Example

- 성공확률이 θ인 베르누이 시행을 10번 독립적으로 반복했을 때 성공횟수 **X**는 이항분포 $B(10, \theta)$를 따른다. **X**의 관측치로 $X = 3$을 얻었다면 θ의 Likelihoods는

$$L(\theta \mid X = 3) = f(3 \mid \theta) = \binom{10}{3} \theta^3 (1 - \theta)^7$$

Likelihood Example

- The first derivative of the log likelihood $\ell(\theta \mid X = 3)$ is as follows:

$$\frac{\partial \ell(\theta \mid X = 3)}{\partial d\theta} = 3\frac{1}{\theta} - 7\frac{1}{1-\theta}.$$

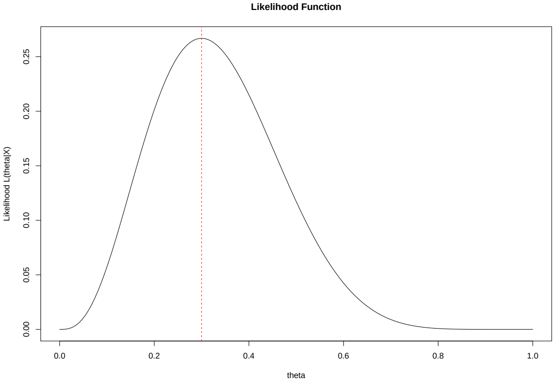
- From the first order optimality condition,

$$3(1-\theta) - 7\theta = 0$$

- Hence MLE: $\hat{\theta} = 0.3$.

Likelihood Example

```
> theta = seq(0,1, length = 1000)
> ltheta = choose(10,3)*theta^3*(1-theta)^7
> plot(theta, ltheta, type = "l", main ="Likelihood Function",
    ylab = "Likelihood L(theta|X)")
> abline(v = 0.3, lty = 2, col=2 )
```



Likelihood Example: Negative Binomial

- 성공확률이 θ인 베르누이 시행을 3번째 성공이 나올 때까지 실험을 계속하기로 한다면 실패횟수 **X**는 음이항 분포 $NB(3, \theta)$를 따르게 된다. 관측치 $X = 7$이라고 하자.

$$L(\theta \mid X = 7) = f(3 \mid \theta) = \binom{3+7-1}{7} \theta^3 (1-\theta)^7$$

- MLE: $\hat{\theta} = 0.3$.

Maximum Likelihood Estimator (MLE)

- MLE는 분포의 kernel에 의존한다.
- Binomial Dist: $\theta^3(1-\theta)^7$
- Negative Binomial Dist: $\theta^3(1-\theta)^7$

Likelihood Principle

- The Likelihood Principle of Birnbaum states that (given the data) **all** of the evidence about θ is contained in the likelihood function.
- 통계적 실험에서 데이터가 가지고 있는 θ의 추론에 관한 정보는 Likelihood function에 모두 포함되어 있다.
- Likelihood Principle implies: Two experiments that yield equal (or proportional) likelihoods should produce equivalent inference about θ.

Likelihood Ratio

- What if $L(\theta \mid X)$ is not differentiable?
- How to compare two values for θ?
- Likelihood Ratio:

$$f(X \mid \theta_a)/f(X \mid \theta_b) = L(\theta_a \mid X)/f(\theta_b \mid X)$$

Likelihood Ratio Example

$X = X_1, ..., X_n$ 이 $N(\theta, 1)$을 따를 θ_a와 θ_b의 Likelihood Ratio (LR)를 구하여라.

- 정의에 따르면 LR은 다음과 같다.

$$
\begin{aligned}
L(\theta_a \mid X)/L(\theta_b \mid X) &= \frac{(2\pi)^{n/2}\exp(-\sum_{i=1}^{n}(X_i - \theta_a)^2/2)}{(2\pi)^{n/2}\exp(-\sum_{i=1}^{n}(X_i - \theta_b)^2/2)} \\
&= \frac{\exp(-\sum_{i=1}^{n}(X_i - \theta_a)^2/2)}{\exp(-\sum_{i=1}^{n}(X_i - \theta_b)^2/2)} \\
\ell(\theta_a \mid X) - \ell(\theta_b \mid X) &\propto \sum(2\theta_a X_i - \theta_a^2) - \sum(2\theta_b X_i - \theta_b^2) \\
&= 2n(\theta_a - \theta_b)\bar{X} - n(\theta_a^2 - \theta_b^2)
\end{aligned}
$$

Likelihood Ratio Example

Consider $\theta_a = 0$ and $\theta_b = 1$. If $n = 10, \bar{x} = 0.1$

- 정의에 따르면 LR은 다음과 같다.

$$
\begin{aligned}
\ell(\theta_a \mid X) - \ell(\theta_b \mid X) &\propto 2n(\theta_a - \theta_b)\bar{X} - n(\theta_a^2 - \theta_b^2) \\
&= 2 \times 10(0 - 1)0.1 - 10(0 - 1) \\
&= -2 + 10 = 8
\end{aligned}
$$

Sufficient Statistics

- Sufficient Statistics (충분통계량):

  $\mathbf{X}$가 밀도함수 $f(X \mid \theta)$를 갖는다고 하자. $T(X)$가 주어졌을 때 $\mathbf{X}$의 조건부 분포가 θ에 의존하지 않으면 $T(X)$를 θ의 충분통계량이라고 한다.
- $T(X)$: θ의 모든 정보를 가진 통계량
- 위 Normal 분포의 예의 경우 충분통계량은 $\bar{X}$.
- 충분통계량 $T(X)$ 의 예: $\bar{X}, \sum X, \sum(X_i - \bar{X})^2, \max X, \min X$

Sufficient Statistics

- Likelihood Principle에 따르면, θ의 모든 정보는 X에 포함되어 있다.
- Sufficient Statistics에 따르면, θ의 모든 정보는 $T(X)$에 포함되어 있다.
- Hence, $T(X)$ is sufficient to estimate θ.
- 데이터가 가진 θ의 정보가 $T(X)$에 모두 포함되어 있으므로 $T(X)$만 알면 더 이상 데이터의 다른 내용은 몰라도 충분하다.

Ancillary Statistics

- Ancillary Statistics (보조 통계량): 통계량의 분포가 θ와 무관하여 θ에 대한 정보를 전혀 가지고 있지 않은 통계량
- Sufficient Statistics의 반대 개념

Ancillary Statistics

- Ancillary Statistics (보조 통계량): 통계량의 분포가 θ와 무관하여 θ에 대한 정보를 전혀 가지고 있지 않은 통계량
- Sufficient Statistics의 반대 개념
- i.e., 보조 통계량은 θ의 정보가 없으므로 θ의 추정에 아무런 도움이 안된다.

Ancillary Statistics Example

두 변수 X_1, X_2는 모두 $U(\theta - 1, \theta + 1)$의 Uniform 분포를 따른다.

이때 두 변수의 차이 $X_1 - X_2$는 θ의 정보를 전혀 갖지 않는다.

Ancillary Statistics Example

두 변수 $X_1, X_2, ..., X_n$는 모두 iid $N(\theta, 1)$의 분포를 따른다. 이때 sample variance (s^2)는 θ의 정보를 전혀 갖지 않는다.

$$s^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$$

Review

- The main difference between Bayesian and Frequentist is how to consider θ.
  - Constant: $P(\theta = 0) = 1$ or 0.
  - Variable: there exists a distribution for θ.

Bayesian Method

- Ultimate Goal: To make probability statements about θ, given some observed data: $p(\theta \mid \mathbf{D})$.
- Using Bayes' Law,

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

- There are two challenges
  1. How to find $p(\theta)$?
  2. How to find $p(\mathbf{D})$?

Prior Distribution

- How to find prior $P(\theta)$?

  In usual, we assume the prior distribution for θ.

  1. Informative Prior
  2. Non-informative Prior

- We must specify $P(\theta)$ based on any knowledge we have about θ before observing the data.
- This could be highly specific or quite vague, depending how uncertain we are about θ.

  e.g., Albert가 아빠일 확률

Data Distribution

- How to find $P(\mathbf{D})$?
  1. $P(\mathbf{D})$ does not depend on θ and thus carries no information about θ.
  2. It is simply a **normalizing constant** which makes $P(\theta \mid \mathbf{D})$ sum or integrate to 1.

Posterior Distribution

- For inference about θ, it is just as good to write.

$$p(\theta \mid \mathbf{D}) \propto p(\theta)p(\mathbf{D} \mid \theta)$$

- The LHS is called the posterior distribution of θ
- We can calculate the posterior distribution by
  1. Multiplying the prior by the likelihood.
  2. Normalizing the posterior at the last step.
- The posterior distribution represents a compromise between the prior information about θ in $p(\theta)$ and the information from the sample about θ in $p(\mathbf{D} \mid \theta)$.

Useful Statistics Using Bayes' Law

Once we obtain the posterior distribution we can use any summaries such as mean, median, variance and many others.

- Posterior mean

$$\mathbb{E}[\theta \mid \mathbf{D}] = \int \theta \cdot p(\theta \mid \mathbf{D}) d\theta.$$

For ease of notation,

- Posterior distribution: $\pi(\theta \mid x), p(\theta \mid x)$.
- Prior distribution: $\pi(\theta), p(\theta)$.

Statistics Using Bayes' Law

- The **posterior variance** is

$$\begin{aligned}
\mathrm{Var}(\theta \mid \mathbf{D}) &= E\left\{(\theta - E(\theta \mid \mathbf{D}))^2 \mid \mathbf{D}\right\} \\
&= \int (\theta - E(\theta \mid \mathbf{D}))^2 1 p(\theta \mid \mathbf{D}) d\theta \\
&= \int \theta^2 p(\theta \mid \mathbf{D}) d\theta - 2E(\theta \mid \mathbf{D}) \int \theta p(\theta \mid \mathbf{D}) d\theta \\
&+ E(\theta \mid \mathbf{D})^2 \int p(\theta \mid \mathbf{D}) d\theta \\
&= E(\theta^2 \mid \mathbf{D}) - E(\theta \mid \mathbf{D})^2
\end{aligned}$$

- If the values of θ are discrete, sums would replace the integrals.

Posterior Example

성공확률이 θ인 베르누이 시행을 10번 독립적으로 반복했을 때 성공횟수 **X**는 이항분포 $B(10, \theta)$를 따른다. **X**의 관측치로 $X = 3$을 얻었다면 θ의 Posteiror를 구하시오.

- Prior Distribution (사전 분포)에 대한 정보가 없으므로 θ의 분포를 $U(0, 1)$으로 가정한다.
- Posterior Distribution

$$\begin{aligned}\pi(\theta \mid X = 3) &= \frac{\binom{10}{3}\theta^3(1-\theta)^7}{\int_0^1 \binom{10}{3}\theta^3(1-\theta)^7 d\theta} \\ &= \frac{\Gamma(12)}{\Gamma(4)\Gamma(8)}\theta^3(1-\theta)^7.\end{aligned}$$

Posterior Example

Due the joint density function or kernel $\theta^3(1-\theta)^7$, the posterior distribution is Beta(4,8). Then,

- Posterior Mean: $1/3$
- Posteiror SD: 0.13

Principles of Bayesian Inference

- θ 에 대한 이론, 경험, 과거의 자료 등 가능한 정보로부터 사전분포 (Prior) $\pi(\theta)$를 구한다.
- 관측변수 X를 정하고 통계조사나 실험 등을 통하여 데이터를 얻는다. 적절한 통계 모형으로 부터 θ가 주어졌을 때 관측데이터의 조건부 밀도함수 $f(X \mid \theta)$를 구한다.
- 베이즈 정리를 이용하요 Posterior 분포 (사후분포)를 구하고 이를 추정에 사용한다.
- 즉, Posterior Distribution이 θ의 모든 정보를 가지고 있다.

Binomial Distribution

If the random variable X follows the binomial distribution with parameters $n \in \mathcal{N}$ and $p \in [0,1]$, we write $X \sim B(n,p)$.

$$\begin{aligned} \Pr(x; n, p) &= \Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}. \\ \mathbb{E}(X) &= np, \\ \mathrm{Var}(X) &= np(1-p). \end{aligned}$$

Negative Binomial Distribution

베르누이 시행을 미리 정한 성공횟수 r 회가 될 때까지 반복 시행할 때 확률변수 X (실패횟수 또는 시행횟수)가 나타내는 분포를 말한다. The probability mass function of the negative binomial distribution is

$$
\begin{aligned}
\Pr(X = x) &= \binom{k+r-1}{k} p^r (1-p)^x \quad \text{for} \quad k = 0, 1, 2, \ldots, \\
\mathbb{E}(X) &= \frac{pr}{1-p}, \\
\mathrm{Var}(X) &= \frac{pr}{(1-p)^2}.
\end{aligned}
$$

Poisson

A discrete random variable X is said to have a Poisson distribution with parameter $\lambda > 0$, if, for $x = 0, 1, 2, ...$, the probability mass function of X is given by

$$\begin{aligned} Pr(X = x) &= \frac{\lambda^x e^{-\lambda}}{x!} \\ \mathbb{E}(X) &= \lambda, \\ \mathrm{Var}(X) &= \lambda. \end{aligned}$$

Uniform

The probability density function of the continuous uniform distribution is

$$\begin{aligned} \Pr(x;\alpha,\beta) &= \Pr(X=x) = \frac{1}{\beta-\alpha} \quad \text{for} \quad \alpha \le x \le \beta \\ \mathbb{E}(X) &= \frac{\beta+\alpha}{2}, \\ \mathrm{Var}(X) &= \frac{(\beta-\alpha)^2}{12}. \end{aligned}$$

Gamma

The gamma distribution can be parameterized in terms of a shape parameter α and an inverse scale parameter β, called a rate parameter. The corresponding probability density function in the shape-rate parametrization is

$$
\begin{aligned}
\Pr(x;\alpha,\beta) &= \Pr(X=x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-\frac{x}{\beta}} \quad \text{for} \quad x > 0, \\
\mathbb{E}(X) &= \alpha\beta, \\
\mathrm{Var}(X) &= \alpha\beta^2.
\end{aligned}
$$

Inverse Gamma

The inverse gamma distribution's probability density function is defined over the support $x > 0$

$$
\begin{aligned}
\Pr(x;\alpha,\beta) &= \Pr(X=x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\frac{\beta}{x}} \quad \text{for} \quad x > 0 \\
\mathbb{E}(X) &= \frac{\beta}{\alpha-1}, \\
\mathrm{Var}(X) &= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}.
\end{aligned}
$$

Normal

The probability density of the normal distribution is

$$\begin{aligned}
\Pr(x;\mu,\sigma^2) &= \Pr(X=x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for} \quad -\infty < x < \infty, \\
\mathbb{E}(X) &= \mu, \\
\mathrm{Var}(X) &= \sigma^2.
\end{aligned}$$

Student T

Student's t-distribution has the probability density function given by

$$
\begin{aligned}
\Pr(x;\nu) &= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \text{for } -\infty < x < \infty \\
\mathbb{E}(X) &= 0 \\
\mathrm{Var}(X) &= \begin{cases} \frac{\nu}{\nu-2} & \text{if } \nu > 2 \\ \infty & \text{if } 1 < \nu \le 2 \end{cases}
\end{aligned}
$$

Posterior Intervals

- The ideal summary of θ is an interval (or region) with a certain probability of containing θ. For some positive α,

$$\Pr(L_\alpha \leq \theta \leq U_\alpha) = 1 - \alpha.$$

- Note that a classical (frequentist) confidence interval (CI) does not exactly have this interpretation.

Definitions of C.I. Coverage

- **Definition**: A random interval $L(\mathbf{X}), U(\mathbf{X})$ has $100(1-\alpha)\%$ frequentist coverage for θ if, **before** the data are gathered,

$$P[L(\mathbf{X}) < \theta < U(\mathbf{X}) \mid \theta] = 1 - \alpha.$$

(Pre-experimental $1-\alpha$ coverage)

- Note that if we observe $\mathbf{X} = x$ and plug x into our confidence interval formula,

$$P(L(x) < \theta < U(x) \mid \theta) = \begin{cases} 0 & \text{if } \theta \notin (L(x), U(x)) \\ 1 & \text{if } \theta \in (L(x), U(x)) \end{cases}$$

(Not Post-experimental $1-\alpha$ coverage)

Definitions of C.I. Coverage

- **Definition**: An interval $(L(x), U(x))$, **based on the observed data** $\mathbf{X} = x$, has $100(1-\alpha)\%$ Bayesian coverage for θ if

  $$P[L(\mathbf{X}) < \theta < U(\mathbf{X}) \mid \mathbf{X} = x] = 1 - \alpha.$$

  (Post-experimental $1-\alpha$ coverage)
- The Frequentist interpretation is less desirable if we are performing inference about θ based on a single interval.

Bayesian Credible Intervals

- A credible interval (or a credible set) is the Bayesian analogue of a confidence interval (C.I.)
- A $100(1-\alpha)\%$ credible set $\mathcal{C}$ is a subset of Θ such that

$$\int_{\mathcal{C}} \pi(\theta \mid X) d\theta = 1 - \alpha.$$

- This is equivalent to

$$\Pr(\theta \in C \mid x) = 1 - \alpha.$$

- If the parameter space Θ is discrete, a sum replaces the integral.

Quantile-Based Interval

- If θ_L^* is the $\alpha/2$ posterior quantile for θ, and θ_U^* is the $1-\alpha/2$ posterior quantile for θ, then (θ_L^*, θ_U^*) is a $100(1-\alpha)\%$ credible interval for θ.
- Note that $P(\theta < \theta_L^* \mid X) = \alpha/2$ and $P(\theta > \theta_U^* \mid X) = \alpha/2$.

$$\begin{aligned} P(\theta \in (\theta_L^*, \theta_U^*) \mid X) &= 1 - P(\theta \notin (\theta_L^*, \theta_U^*) \mid X) \\ &= 1 - (P(\theta < \theta_L^* \mid X) + P(\theta > \theta_U^* \mid X)) \\ &= 1 - \alpha. \end{aligned}$$

Example: Quantile-Based Interval

- Suppose $X_1, \dots, X_n$ are the durations of cabinets for a sample of cabinets from Western European countries.
- We assume the X_i's follow an exponential distribution.

$$p(X_i \mid \theta) = \theta e^{-\theta X_i}, \quad X_i > 0,$$

$$L(\theta \mid X) = \theta^n e^{-\theta \sum_{i=1}^n x_i}.$$

- Suppose our prior distribution for θ is

$$p(\theta) \propto 1/\theta, \quad \theta > 0.$$

$\rightarrow$ Larger values of θ are less likely a **priori**.

Example: Quantile-Based Interval

- Then, we have

$$\pi(\theta \mid x) \quad \propto \quad p(\theta)L(\theta \mid x)$$

Example: Quantile-Based Interval

- Then, we have

$$\begin{aligned}\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\ &= \left(\frac{1}{\theta}\right)\theta^n e^{-\theta\sum_{i=1}^n x_i}\end{aligned}$$

Example: Quantile-Based Interval

- Then, we have

$$
\begin{aligned}
\pi(\theta \mid x) &\propto p(\theta) L(\theta \mid x) \\
&= \left(\frac{1}{\theta}\right) \theta^{n} e^{-\theta \sum_{i=1}^{n} x_i} \\
&= \theta^{n-1} e^{-\theta \sum_{i=1}^{n} x_i}.
\end{aligned}
$$

Example: Quantile-Based Interval

- Then, we have

$$\begin{aligned}\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x)\\ &= \left(\frac{1}{\theta}\right)\theta^n e^{-\theta\sum_{i=1}^n x_i}\\ &= \theta^{n-1}e^{-\theta\sum_{i=1}^n x_i}.\end{aligned}$$

- This is the kernel of a gamma distribution with "shape" parameter n and "rate" parameter $\sum_{i=1}^n x_i$.

Example: Quantile-Based Interval

- Then, we have

$$\begin{aligned}\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\ &= \left(\frac{1}{\theta}\right)\theta^n e^{-\theta \sum_{i=1}^n x_i} \\ &= \theta^{n-1} e^{-\theta \sum_{i=1}^n x_i}.\end{aligned}$$

- This is the kernel of a gamma distribution with "shape" parameter n and "rate" parameter $\sum_{i=1}^n x_i$.
- After including the normalizing constant,

Example: Quantile-Based Interval

- Then, we have

$$\begin{aligned}\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\ &= \left(\frac{1}{\theta}\right)\theta^n e^{-\theta\sum_{i=1}^n x_i} \\ &= \theta^{n-1}e^{-\theta\sum_{i=1}^n x_i}.\end{aligned}$$

- This is the kernel of a gamma distribution with "shape" parameter n and "rate" parameter $\sum_{i=1}^n x_i$.
- After including the normalizing constant,

$$\pi(\theta \mid X) = \frac{(\sum x_i)^n}{\Gamma(n)}\theta^{n-1}e^{-\theta\sum_{i=1}^n x_i}, \quad \theta > 0.$$

Example: Quantile-Based Interval

- Now, given the observed data $x_1, ..., x_n$, we can calculate any quantiles of this gamma distribution.

Example: Quantile-Based Interval

- Now, given the observed data $x_1, ..., x_n$, we can calculate any quantiles of this gamma distribution.
- The 0.05 and 0.95 quantiles will give us a 90% credible interval for θ.

Example: Quantile-Based Interval

- Suppose we feel $p(\theta) = 1/\theta$ is too subjective and favors small values of θ too much.

Example: Quantile-Based Interval

- Suppose we feel $p(\theta) = 1/\theta$ is too subjective and favors small values of θ too much.
- Instead, let's consider the **non-informative** prior

$$p(\theta) = 1, \quad \theta > 1$$

(favors all values of θ equally).

Example: Quantile-Based Interval

- Suppose we feel $p(\theta) = 1/\theta$ is too subjective and favors small values of θ too much.
- Instead, let's consider the **non-informative** prior

$$p(\theta) = 1, \quad \theta > 1$$

(favors all values of θ equally).

- Then our posterior is

$$\begin{aligned} \pi(\theta \mid x) &\propto p(\theta) L(\theta \mid x) \\ &= (1)\theta^n e^{-\theta \sum_{i=1}^n x_i} \\ &= \theta^n e^{-\theta \sum_{i=1}^n x_i}. \end{aligned}$$

Example: Quantile-Based Interval

- This is the kernel of a gamma distribution with "shape" parameter $n+1$ and "rate" parameter $\sum_{i=1}^{n} x_i$.

Example: Quantile-Based Interval

- This is the kernel of a gamma distribution with "shape" parameter $n+1$ and "rate" parameter $\sum_{i=1}^{n} x_i$.
- We can similarly find the equal-tail credible interval.

Example: Quantile-Based Interval

- First Case: $\mathbb{E}(\theta \mid X_1, ..., X_n) = \frac{n-1}{\sum X_i}$.
- Second Case: $\mathbb{E}(\theta \mid X_1, ..., X_n) = \frac{n}{\sum X_i}$.

Example: Quantile-Based Interval

- First Case: $\mathbb{E}(\theta \mid X_1, ..., X_n) = \frac{n-1}{\sum X_i}$.
- Second Case: $\mathbb{E}(\theta \mid X_1, ..., X_n) = \frac{n}{\sum X_i}$.
- As $n \to \infty$ the both becomes similar.

Example: Quantile-Based Interval

- First Case: $\mathbb{E}(\theta \mid X_1, ..., X_n) = \frac{n-1}{\sum X_i}$.
- Second Case: $\mathbb{E}(\theta \mid X_1, ..., X_n) = \frac{n}{\sum X_i}$.
- As $n \to \infty$ the both becomes similar.
- Although the priors different, the posterior distributions are similar when n is sufficiently large enough.

Example 2: Quantile-Based Interval

- Consider 10 flips of a coin having $\Pr(\mathit{Heads}) = \theta$.

Example 2: Quantile-Based Interval

- Consider 10 flips of a coin having $\Pr(\textit{Heads}) = \theta$.
- Suppose we observe 2 "heads".

Example 2: Quantile-Based Interval

- Consider 10 flips of a coin having $\Pr(\mathit{Heads}) = \theta$.
- Suppose we observe 2 "heads".
- We model the count of heads as binomial:

$$p(X = x \mid \theta) = \binom{10}{x} \theta^x (1-\theta)^{10-x}, \quad x = 0, 1, ..., 10.$$

- Let's use a uniform prior for θ:

$$p(\theta) = 1, \quad 0 \leq \theta \leq 1.$$

Example 2: Quantile-Based Interval

- Then the posterior is:

Example 2: Quantile-Based Interval

- Then the posterior is:

$$\pi(\theta \mid x) \quad \propto \quad p(\theta) L(\theta \mid x)$$

Example 2: Quantile-Based Interval

- Then the posterior is:

$$
\begin{aligned}
\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\
&= \binom{10}{x}\theta^{x}(1-\theta)^{10-x}
\end{aligned}
$$

Example 2: Quantile-Based Interval

- Then the posterior is:

$$\begin{aligned}\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\ &= \binom{10}{x}\theta^x(1-\theta)^{10-x} \\ &\propto \theta^x(1-\theta)^{10-x}, \quad 0 \le \theta \le 1.\end{aligned}$$

Example 2: Quantile-Based Interval

- Then the posterior is:

$$\begin{aligned}\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\ &= \binom{10}{x}\theta^{x}(1-\theta)^{10-x} \\ &\propto \theta^{x}(1-\theta)^{10-x}, \quad 0 \leq \theta \leq 1.\end{aligned}$$

- This is a beta distribution for θ with parameters $x+1$ and $10-x+1$.

Example 2: Quantile-Based Interval

- Then the posterior is:

$$\begin{aligned}\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\ &= \binom{10}{x}\theta^x(1-\theta)^{10-x} \\ &\propto \theta^x(1-\theta)^{10-x}, \quad 0 \le \theta \le 1.\end{aligned}$$

- This is a beta distribution for θ with parameters $x+1$ and $10-x+1$.
- Since $x = 2$ here, $\pi(\theta \mid x = 2)$ is beta $(3, 9)$.

Example 2: Quantile-Based Interval

- Then the posterior is:

$$\begin{aligned}
\pi(\theta \mid x) &\propto p(\theta)L(\theta \mid x) \\
&= \binom{10}{x}\theta^x(1-\theta)^{10-x} \\
&\propto \theta^x(1-\theta)^{10-x}, \quad 0 \le \theta \le 1.
\end{aligned}$$

- This is a beta distribution for θ with parameters $x+1$ and $10-x+1$.
- Since $x=2$ here, $\pi(\theta \mid x=2)$ is beta $(3,9)$.
- The 0.025 and 0.975 quantiles of a beta $(3,9)$ are (.0602, .5178), which is a 95% credible interval for θ.

Example 3: Quantile-Based Interval

$N(\theta, 2^2)$분포로부터 16개의 표본을 추출한 결과 $\bar{X} = 0.3$이었다. θ에 대한 무정보 사전분포 (non-informative)로 $\pi(\theta) = 1$을 가정하고, 이를 $\bar{\mathbf{X}} \mid \theta \sim N(\theta, 2^2/16)$과 합성하여 사후분포를 유도하여라. 그리고 대응되는 95% 베이지안 신뢰구간을 구해 보자.

Example 3: Quantile-Based Interval

$$\pi(\theta \mid \bar{X}) \quad \propto \quad f(\bar{X} \mid \theta)\pi(\theta)$$

Example 3: Quantile-Based Interval

$$\begin{aligned}\pi(\theta \mid \bar{X}) &\propto f(\bar{X} \mid \theta)\pi(\theta) \\ &\propto \exp\left(-\frac{1}{2 \times 0.25}(0.3 - \theta)^2\right).\end{aligned}$$

Hence the posterior distribution is Normal$(0.3, 0.5^2)$

Example 3: Quantile-Based Interval

- $\mathcal{C}$은 유일하지 않을 수 있다.

Example 3: Quantile-Based Interval

- $\mathcal{C}$은 유일하지 않을 수 있다.
- 가장 좋은 신뢰구간은 어떤 것일까?

Highest Posterior Density (HPD) Intervals

- Note that values of θ around 0.3 have much higher posterior probability than values around 7.5.
- A better approach here is to create our interval of θ-values having the Highest Posterior Density.

Highest Posterior Density (HPD) Intervals

- Definition: A $100(1-\alpha)\%$ HPD interval for θ is a subset $\mathcal{C} \in \Theta$ defined by

$$\mathcal{C} = \{\theta : \pi(\theta \mid x) \geq k\}$$

where k is the largest number such that

$$\int_{\theta:\pi(\theta|x)\geq k} \pi(\theta \mid x) d\theta = 1-\alpha.$$

Highest Posterior Density (HPD) Intervals

- Definition: A $100(1-\alpha)\%$ HPD interval for θ is a subset $\mathcal{C} \in \Theta$ defined by

$$\mathcal{C} = \{\theta : \pi(\theta \mid x) \geq k\}$$

  where k is the largest number such that

$$\int_{\theta:\pi(\theta|x)\geq k} \pi(\theta \mid x) d\theta = 1 - \alpha.$$

- The value k can be thought of as a horizontal line placed over the posterior density whose intersection(s) with the posterior define regions with probability $1-\alpha$.

Highest Posterior Density (HPD) Intervals

- Definition: A $100(1-\alpha)\%$ HPD interval for θ is a subset (θ_1, θ_2) defined by
  1. $P(\theta_1 < \theta < \theta_2 \mid X) = 1 - \alpha$.
  2. 만약 $\theta_a \in (\theta_1, \theta_2)$이고 $\theta_b \notin (\theta_1, \theta_2)$ 이면 $\Pr(\theta_a \mid x) > \Pr(\theta_b \mid x)$.

Highest Posterior Density (HPD) Intervals

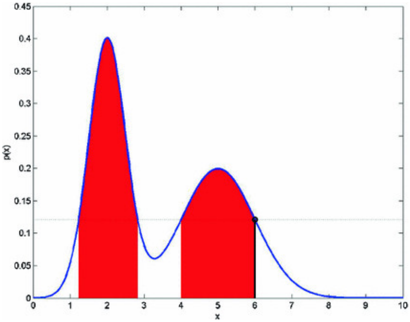
- Definition: A $100(1-\alpha)\%$ HPD interval for θ is a subset (θ_1, θ_2) defined by
  1. $P(\theta_1 < \theta < \theta_2 \mid X) = 1 - \alpha$.
  2. 만약 $\theta_a \in (\theta_1, \theta_2)$이고 $\theta_b \notin (\theta_1, \theta_2)$ 이면 $\Pr(\theta_a \mid x) > \Pr(\theta_b \mid x)$.
- 최대사후구간 (HPD Interval)은 주어진 신뢰도를 만족하는 베이지안 구간 중 최대한 Posterior density 값이 높은 θ들의 합집합이다.

Example 3: HPD Interval

- From the previous example, the posterior dist is Normal$(0.3, (0.5)^2)$
- HPD Interval is $0.3 \pm 1.96 \times 0.5 = (-0.68, 1.28)$
- For Normal distribution, the HPD interval has the minimum distance than other credible sets

Highest Posterior Density Intervals

- The HPD region will be an interval when the posterior is unimodal.
- If the posterior is multimodal, the HPD region might be a **discontiguous** set.



How to Find HPD Interval

- 예를 통해 보면 구간의 경계값들에서 사후밀도함수값이 동일함을 알 수 있다.
- 즉 주어진 신뢰도를 만족하는 구간 중 최대한 사후밀도함수값이 높은 θ값을 모으기 위하여 가상의 수평 막대를 사후밀도함수의 최댓값에서투 점차 아래로 내리면서 만나는 점들 사이의 면적을 계산하여 면적이 최초로 $(1-\alpha)$와 동일 하는 구간이 HPD interval이 된다.

How to Find HPD Interval

- 수리적으로 찾는 방법은 매우 어렵다.
- 그래서 근사적으로 찾는 방법이 권장 된다.
- 앞으로 근사적으로 HPD interval을 찾는 세가지 방법을 고려하겠다.

Case 1: How to Find HPD Interval

- Suppose that the posterior is symmetric and unimodal.
- Then consider the $\alpha/2$ and $1 - \alpha/2$ percentile.
- If the posterior distributions are well-known, the existing packages can be exploited.
- Otherwise some sampling methods can be used.

Case 1: How to Find HPD Interval

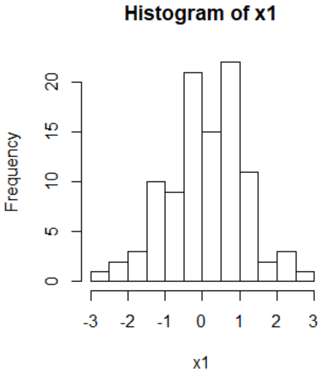
- Suppose that the posterior is symmetric and unimodal.
- Then consider the $\alpha/2$ and $1 - \alpha/2$ percentile.
- If the posterior distributions are well-known, the existing packages can be exploited.
- Otherwise some sampling methods can be used.

Case 1: Example

```
> n = 100
> x1 <-rnorm(n, 0, 1)
> quantile(x1, c(.025, .975))
2.5%      97.5%
-1.959474  2.269712
>
> n = 1000
> x2 <-rnorm(n, 0, 1)
> quantile(x2, c(.025, .975))
2.5%      97.5%
-1.928400  1.894172
```

Case 1: Example

```
> par(mfrow = c(1,2))
> hist(x1);hist(x2)
```



Case 2: Grid Search Method (격자점 방법)

- Main idea: Consider θ as N distinct values $(\theta_1, \theta_2, ...\theta_N)$.

Case 2: Grid Search Method (격자점 방법)

- Main idea: Consider θ as N distinct values $(\theta_1, \theta_2, ...\theta_N)$.
- Calculate

$$\widehat{\pi}(\theta_i \mid x) = \frac{\pi(\theta_i) f(x \mid \theta_i)}{\sum_i \pi(\theta_i) f(x \mid \theta_i)}.$$

Case 2: Grid Search Method (격자점 방법)

- Main idea: Consider θ as N distinct values $(\theta_1, \theta_2, ...\theta_N)$.
- Calculate

$$\widehat{\pi}(\theta_i \mid x) = \frac{\pi(\theta_i) f(x \mid \theta_i)}{\sum_i \pi(\theta_i) f(x \mid \theta_i)}.$$

- Find M such that

$$M := \min \left\{ m \;\middle|\; \sum_{j=1}^{m} \widehat{\pi}(\theta_i \mid x)^{\text{ordered}} \geq 1 - \alpha \right\}$$

Case 2: Grid Search Method (격자점 방법)

```
HPDgrid = function(prob, level = 0.95){
  prob.sort = sort(prob, decreasing = T)
  M = min( which(cumsum(prob.sort)>=level) )
  height = prob.sort[M]
  HPD.index = which( prob >= height)
  HPD.level = sum(prob[HPD.index])
  res = list( index = HPD.index, level = HPD.level )
  return(res)
}
```

Case 2: Grid Search Method (격자점 방법)

Suppose that the posterior distributions satisfies

$$f(\theta \mid x) \propto \exp\left(-2(\theta - 0.3)^2\right).$$

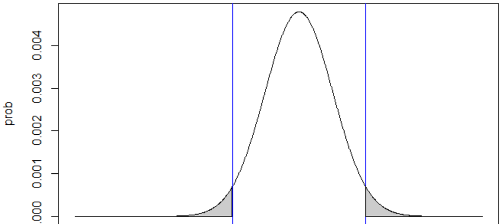
```
> N = 1001

> theta = seq(-3, 3, length = N)

> prob = exp(-0.5/0.25*(theta-0.3)^2)

> prob = prob/sum(prob)

> alpha = 0.05; level = 1-alpha
```



Case 2: Grid Search Method (격자점 방법)

```
HPD = HPDgrid(prob, level)

HPDgrid.hat = c( min(theta[HPD$index]),
                                max(theta[HPD$index])  )

HPDgrid.hat

-0.678  1.278
```

Case 2: Grid Search Method (격자점 방법)

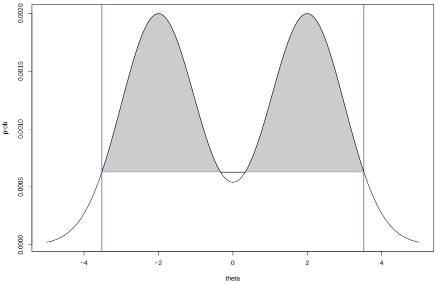
```
par(mfrow = c(1,1))
plot(theta, prob, type ="l", ylab = "prob", xlab ="theta",
xlim = c(-3,3))
abline(v = HPDgrid.hat, col = 'blue')
polygon(x = c(theta[which(theta < HPDgrid.hat[1])],
   rev(theta[which(theta < HPDgrid.hat[1])]) ),
   y = c(prob[which(theta < HPDgrid.hat[1])],
   rep( 0, sum(theta < HPDgrid.hat[1]) )), col = "grey80")
polygon(x = c(theta[which(theta > HPDgrid.hat[2])],
   rev(theta[which(theta > HPDgrid.hat[2])]) ),
   y = c(prob[which(theta > HPDgrid.hat[2])],
   rep( 0, sum(theta > HPDgrid.hat[2]) )),
   col = "grey80")
}
```

Case 2: Grid Search Method (격자점 방법)

Suppose that the posterior distributions satisfies

$$f(\theta \mid x) \propto \exp\left(-(\theta-2)^2/2\right) + \exp\left(-(\theta+2)^2/2\right).$$

```
N = 1001
theta = seq(-5, 5, length = N)
#prob = exp(-0.5/0.25*(theta-0.3)^2)
prob = exp(-1/2*(theta+2)^2) + exp(-1/2*(theta-2)^2)
prob = prob/sum(prob)
alpha = 0.10; level = 1-alpha
```



Case 2: Grid Search Method (격자점 방법)

```
> HPD = HPDgrid(prob, level)
> HPDgrid.hat = c( min(theta[HPD$index]), max(theta[HPD$index])  )
> HPDgrid.hat
[1] -3.52  3.52
> theta[which(prob == min( prob[HPD$index] ) )]
[1] -0.33  0.33
```

Case 2: Grid Search Method (격자점 방법)

```
par(mfrow = c(1,1))
plot(theta, prob, type ="l", ylab = "prob", xlab ="theta", xlim = c(-5,5))
abline(v = HPDgrid.hat, col = 'blue')
polygon(x = c(theta[HPD$index]) , y = c(prob[HPD$index] ), col = "grey80")
```

Case 2: Grid Search Method (격자점 방법)

- It is very useful for the multivariate or multimodal θ.
- It is difficult for find the optimal HPD interval when the posterior density is wiggly.
- It is hard to calculate all possible values for θ if $\theta \in \mathbb{R}$.

Case 3: Bayesian Posterior Sampling

- Posterior sampling histogram 이 density function 과 유사하다는 성질을 이용
- 예를 들어 1000개의 사후표본이 주어졌을때, 95% CI는 950개의 표본을 포함할 것이다.
- 1000개의 θ 오른차순으로 정렬하여 $(\theta_1, ..., \theta_{1000})$이라고 하자.
- 이 때 가능한 신뢰구간은 $(\theta_1, \theta_{950}), (\theta_2, \theta_{951}), (\theta_3, \theta_{953}), ...$이 된다.
- 이 중에 가장 짧은 구간을 근사적 HPD interval로 취할 수 있다.

Case 3: Bayesian Posterior Sampling

```
HPDsample = function(theta, level = 0.95){
N = length(theta)
theta.sort = sort(theta)
M = ceiling(N*level)
nCI=N-M
CI.width = rep(0, nCI)
for(i in 1:nCI) CI.width[i] = theta.sort[i+M] - theta.sort[i]
index = which.min(CI.width)
HPD = c(theta.sort[index], theta.sort[index+M])
return(HPD)
}
```

Case 3: Bayesian Posterior Sampling

- Suppose that the posterior distribution is $\theta \sim N(0,1)$.

```
> N = 1000
> theta = rnorm(N, 0, 1)
> alpha = 0.05
> level = 1-alpha
> HPDsample(theta)
[1] -1.632139  2.141612
```

Case 3: Bayesian Posterior Sampling

- Suppose that the posterior distribution is $\theta \sim N(0, 1)$.

```
> N = 10000
> theta = rnorm(N, 0, 1)
> alpha = 0.05
> level = 1-alpha
> HPDsample(theta)
[1] -1.909751  1.967354
```

Case 3: Bayesian Posterior Sampling

Pros.

- 많은 경우 θ의 posterior distribution이 매우 복잡하여 percentile을 직접 찾을 수 없다.
- Grid search method의 경우, 도메인이 무한인 경우 사용하기 어렵다.

Cons.

- Unimodal에서만 사용 가능하다.
- 다변량 모수에 대한 다차원 사후구간을 찾는데에 적용할 수 없다.

Weakness of Frequentist

분산이 $\sigma^2 = 1$인 정규분포의 평균 θ를 추정하고자 한다. 표본을 얻기 전에 먼저 동전을 던져 앞면이 나오면 표본을 2개만 취하고, 뒷면이 나오면 표본을 1000개 취하기로 하였다. 즉 표본크기 n은 각각 확률 $\frac{1}{2}$로 2, 아니면 1000이 될 것이다. 이 실험에서 θ에 대한 추정치는 표본의 평균 $\bar{X}$가 적절하며 $\bar{X}$의 정확도를 측정하는 통계량으로는 $\bar{X}$의 분산이 적절 할 것이다. $\bar{X}$의 분산은

$$
\begin{aligned}
\mathrm{Var}(\bar{X}) &= \frac{1}{2}\mathrm{Var}(\bar{X} \mid n = 2) + \frac{1}{2}\mathrm{Var}(\bar{X} \mid n = 1000) \\
&= \frac{1}{2}(\sigma^2/2 + \sigma^2/1000) \approx 1/4.
\end{aligned}
$$

Weakness of Frequentist

만약 동전의 결과가 뒷면이고 따라서 1000개의 표본을 취한 결과가 $\bar{X} = 0.1$이었다고 하자. 고전적 통계추론에 의하면 θ에 대한 추정치는 0.1이고 추정오차는 $\sqrt{\frac{1}{4}} = 0.5$로 결론 짓는다. 이미 1000개의 표본을 취했다는 것을 안 상태에서, 추정오차를 $\sqrt{\frac{1}{1000}} = 0.03$아닌 0.5를 합리적인 추정오차라고 할 수 있겠는가?

Weakness of Frequentist

두 변수 X_1, X_2는 $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$를 따른다. 고전적 통계추론에서 θ대한 95% 신뢰구간을 구하면, 적절한 양의 상수 C에 대하여 $\bar{X} \pm C$의 형태를 가진다. 만약 두변수의 관측값이 각각, $X_1 = 1, X_2 = 2$라면, θ가 1.5임이 확실하다. 이때 우리가 신뢰계수를 100%가 아닌 95%로 보아야 하는가?