4. 베이지안 추론

이경재

인하대학교 통계학과

March 20, 2019

Statistical Inference

- θ: parameter of interest (unknown)
- X: random variable
- x: a realization value of X. The observation (data).
- A statistical model is a distribution for X given the parameter θ, i.e., $f(x \mid \theta)$.
- Goal: inference about the parameter θ, based on data x.

Statistical Inference

- θ: parameter of interest (unknown)
- X: random variable
- x: a realization value of X. The observation (data).
- A statistical model is a distribution for X given the parameter θ, i.e., $f(x \mid \theta)$.
- Goal: inference about the parameter θ, based on data x.

Likelihood Function

- The likelihood function: $L(\theta \mid x) = f(x \mid \theta)$.
- $L(\theta \mid x)$ is a function of θ showing that how "likely" is the parameter value θ to have produced the *observed* data x.
  - (e.g.) We have two possible parameter values θ_1 and θ_2. If $f(x \mid \theta_1) > f(x \mid \theta_2)$, which one is more likely to have produced the data?
- It is not a probability density function for θ.

Likelihood Function

- The likelihood function: $L(\theta \mid x) = f(x \mid \theta)$.
- $L(\theta \mid x)$ is a function of θ showing that how "likely" is the parameter value θ to have produced the *observed* data x.
  - (e.g.) We have two possible parameter values θ_1 and θ_2. If $f(x \mid \theta_1) > f(x \mid \theta_2)$, which one is more likely to have produced the data?
- It is not a probability density function for θ.

Likelihood Function

- The likelihood function: $L(\theta \mid x) = f(x \mid \theta)$.
- $L(\theta \mid x)$ is a function of θ showing that how "likely" is the parameter value θ to have produced the *observed* data x.
  - (e.g.) We have two possible parameter values θ_1 and θ_2. If $f(x \mid \theta_1) > f(x \mid \theta_2)$, which one is more likely to have produced the data?
- It is not a probability density function for θ.

Maximum Likelihood Estimator (MLE)

In classical statistics,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta \mid X)$$

is the maximum likelihood estimator (MLE) of θ.

- (e.g.) 동전을 100회 던졌을때 앞면이 100회 연속 나왔다고 한다. 이 자료에 기반하면, 다음 중 어느 것이 동전을 던졌을때 앞면이 나올 확률 p로 더 가능성이 있을까?
  1. $p = 0$
  2. $p = 0.5$
  3. $p = 1$

Likelihood Principle

(Birnbaum, 1962) In statistical experiments, *all* of the evidence about the parameter θ is contained in the likelihood function.

- 통계적 실험에서 자료 x가 가지고 있는 θ에 관한 정보는 가능도함수에 *모두* 포함되어 있다.
- Two experiments that yield equal (or proportional) likelihoods, i.e., $\exists c > 0$ such that

$$L_1(\theta) = cL_2(\theta), \quad \forall\theta,$$

should produce equivalent inference about θ.

Example: Likelihood Principle

- Let $X_1, \ldots, X_{10} \overset{iid}{\sim} Ber(\theta)$ and $X = \sum_{i=1}^{10} X_i$.
- Then we have $X \sim B(10, \theta)$.
- If we observe $(x_1, \ldots, x_{10}) = (1, 1, 0, 0, 0, 0, 0, 0, 0, 1)$, i.e., $x = 3$,

$$f(x = 3 \mid \theta) \quad = \quad \binom{10}{3}\theta^3(1-\theta)^7.$$

Example: Likelihood Principle

- We will calculate the MLE.
- Note that the log likelihood function is concave.
- Then the MLE is given by the solution of the following:

$$\frac{d}{d\theta}\ell(\theta \mid x = 3) = 3\frac{1}{\theta} - 7\frac{1}{1-\theta} = 0.$$

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

Example: Likelihood Principle

- We will calculate the MLE.
- Note that the log likelihood function is concave.
- Then the MLE is given by the solution of the following:

$$\frac{d}{d\theta}\ell(\theta \mid x = 3) = 3\frac{1}{\theta} - 7\frac{1}{1-\theta} = 0.$$

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

Example: Likelihood Principle

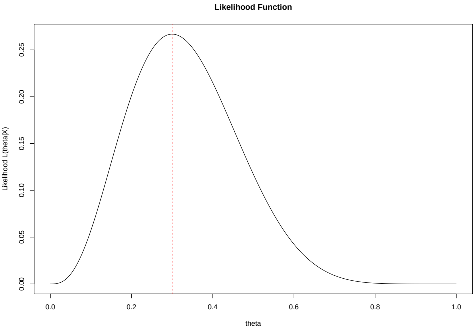
- We will calculate the MLE.
- Note that the log likelihood function is concave.
- Then the MLE is given by the solution of the following:

$$\frac{d}{d\theta}\ell(\theta \mid x = 3) = 3\frac{1}{\theta} - 7\frac{1}{1-\theta} = 0.$$

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

Example: Likelihood Principle

```
> 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 )
```



Example: Likelihood Principle

- 성공확률이 θ인 베르누이 시행을 3번째 성공이 나올 때까지 실험을 계속하기로 한다면 실패횟수 x는 음이항 분포 $NB(3, \theta)$를 따르게 된다.
- If we observe $(x_1, \dots, x_{10}) = (1, 1, 0, 0, 0, 0, 0, 0, 0, 1)$ from $NB(3, \theta)$, then

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

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

Example: Likelihood Principle

- MLE depends only on the part proportional to θ.
  - Binomial distribution: $\theta^3(1-\theta)^7$
  - Negative binomial distribution: $\theta^3(1-\theta)^7$
- MLE follows the likelihood principle.

Example: Likelihood Principle (Cont'd)

- Suppose we wish to test

$$H_0 : \theta = 0.5 \quad \text{v.s.} \quad H_1 : \theta > 0.5$$

  with $\alpha = 0.05$.

- For $X_1, X_2, \ldots \overset{iid}{\sim} Ber(\theta)$, we observe 9 heads and 3 tails.
- Under the binomial model $B(12, \theta)$, ($L_1(\theta) = \binom{12}{9}\theta^9(1-\theta)^3$)

$$\text{p-value} \quad = \quad P_{H_0}(X \geq 9) \; = \; 0.075. \qquad \text{(We accept } H_0\text{!)}$$

- Under the negative binomial model $NB(3, \theta)$, ($L_2(\theta) = \binom{12+9-1}{9}\theta^9(1-\theta)^3$)

$$\text{p-value} \quad = \quad P_{H_0}(X \geq 9) \; = \; 0.0325. \qquad \text{(We reject } H_0\text{!)}$$

- Thus, the classical significance test violates the likelihood principle.

Example: Likelihood Principle (Cont'd)

- Suppose we wish to test

$$H_0 : \theta = 0.5 \quad \text{v.s.} \quad H_1 : \theta > 0.5$$

  with $\alpha = 0.05$.

- For $X_1, X_2, \ldots \overset{iid}{\sim} Ber(\theta)$, we observe 9 heads and 3 tails.
- Under the binomial model $B(12, \theta)$, $(L_1(\theta) = \binom{12}{9}\theta^9(1-\theta)^3)$

$$\text{p-value} = P_{H_0}(X \geq 9) = 0.075. \qquad (\text{We accept } H_0!)$$

- Under the negative binomial model $NB(3, \theta)$, $(L_2(\theta) = \binom{12+9-1}{9}\theta^9(1-\theta)^3)$

$$\text{p-value} = P_{H_0}(X \geq 9) = 0.0325. \qquad (\text{We reject } H_0!)$$

- Thus, the classical significance test violates the likelihood principle.

Example: Likelihood Principle (Cont'd)

- Suppose we wish to test

$$H_0 : \theta = 0.5 \quad \text{v.s.} \quad H_1 : \theta > 0.5$$

with $\alpha = 0.05$.

- For $X_1, X_2, \ldots \stackrel{iid}{\sim} Ber(\theta)$, we observe 9 heads and 3 tails.
- Under the binomial model $B(12, \theta)$, $(L_1(\theta) = \binom{12}{9}\theta^9(1-\theta)^3)$

$$\text{p-value} \quad = \quad P_{H_0}(X \geq 9) \;=\; 0.075. \qquad (\text{We accept } H_0!)$$

- Under the negative binomial model $NB(3, \theta)$, $(L_2(\theta) = \binom{12+9-1}{9}\theta^9(1-\theta)^3)$

$$\text{p-value} \quad = \quad P_{H_0}(X \geq 9) \;=\; 0.0325. \qquad (\text{We reject } H_0!)$$

- Thus, the classical significance test violates the likelihood principle.

Likelihood Ratio

- Suppose we have two candidates θ_a and θ_b.
- What if $L(\theta \mid x)$ is not differentiable?
- How to compare two values for θ?
- Likelihood ratio:

$$\frac{f(x \mid \theta_a)}{f(x \mid \theta_b)} = \frac{L(\theta_a \mid x)}{f(\theta_b \mid x)}.$$

Example: Likelihood Ratio

Suppose $X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1)$ and let $X = (X_1, ..., X_n)$. We have two candidates θ_a and θ_b.

- Likelihood ratio (LR)

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

Example: Likelihood Ratio

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

- The LR is given by

$$\begin{aligned}\ell(\theta_a \mid x) - \ell(\theta_b \mid x) &= n(\theta_a - \theta_b)\bar{x} - \frac{1}{2}n(\theta_a^2 - \theta_b^2) \\ &= 10(0 - 1)0.1 - \frac{1}{2}10(0 - 1) \\ &= -1 + 5 = 4.\end{aligned}$$

- Thus, we can conclude that θ_a is more likely to be a true value.

Sufficient Statistic (SS: 충분통계량)

Suppose $X \sim f(x \mid \theta)$. A statistic $T(X)$ is called sufficient statistic (SS) if

$$f(x \mid \theta, T(x)) = f(x \mid T(x)).$$

- It means that $T(X)$ has all information about θ.
- A SS is not unique, so it is important to find the SS having minimal dimensionality (minimal sufficient statistic: MSS).

Sufficient Statistic: factorization theorem

(Fisher-Neyman factorization theorem) A statistics $T(X)$ is a SS if and only if (iff)

$$f(x \mid \theta) = g(T(x) \mid \theta) h(x)$$

for some nonnegative functions g and h.

- The above theorem implies

$$\begin{aligned} \frac{f(x \mid \theta_a)}{f(x \mid \theta_b)} &= \frac{g(T(x) \mid \theta_a) h(x)}{g(T(x) \mid \theta_b) h(x)} \\ &= \frac{g(T(x) \mid \theta_a)}{g(T(x) \mid \theta_b)}. \end{aligned}$$

- Thus, the LR depends only on $g(T(X) \mid \theta)$, the conditional distribution of SS $T(X)$ given θ.

Bayesian Inference

- Bayesian inference uses prior distribution $\pi(\theta)$ as well as the likelihood function $f(x \mid \theta)$.
- What is the difference between frequentist and Bayesian inference?
  - (F): θ is an unknown constant.
  - (B): θ is unknown & a random variable.

Bayesian Inference

- Bayesian inference uses prior distribution $\pi(\theta)$ as well as the likelihood function $f(x \mid \theta)$.
- What is the difference between frequentist and Bayesian inference?
  - (F): θ is an unknown constant.
  - (B): θ is unknown & a random variable.

Prior Distribution

- There are two types of prior distribution $\pi(\theta)$:
  1. Subjective prior,
  2. Objective prior.
- When we have prior information or personal belief in θ, subjective priors can be used.
- When we don't have any prior information about θ, objective prior may be more appropriate.
- (e.g.) Albert가 아빠일 확률

Prior Distribution

- There are two types of prior distribution $\pi(\theta)$:
  1. Subjective prior,
  2. Objective prior.
- When we have prior information or personal belief in θ, subjective priors can be used.
- When we don't have any prior information about θ, objective prior may be more appropriate.
- (e.g.) Albert가 아빠일 확률

Prior Distribution

- There are two types of prior distribution $\pi(\theta)$:
  1. Subjective prior,
  2. Objective prior.
- When we have prior information or personal belief in θ, subjective priors can be used.
- When we don't have any prior information about θ, objective prior may be more appropriate.
- (e.g.) Albert가 아빠일 확률

Prior Distribution

- There are two types of prior distribution $\pi(\theta)$:
  1. Subjective prior,
  2. Objective prior.
- When we have prior information or personal belief in θ, subjective priors can be used.
- When we don't have any prior information about θ, objective prior may be more appropriate.
- (e.g.) Albert가 아빠일 확률

Example 1: Choice of Prior Distribution

- θ: 어느 공장에서 생산되는 제품의 불량품
- 과거의 경험으로부터, 우리는 $\theta \approx 0.2$인 것을 알고 있다.
- $\theta \in (0, 1)$이므로 $\theta \sim Beta(\alpha, \beta)$로 선택할 수 있고,

$$E(\theta) = \frac{\alpha}{\alpha + \beta}, \quad \theta \sim Beta(\alpha, \beta)$$

이므로 $\alpha = 2, \beta = 7$로 선택하면 $E(\theta) = 0.2$를 만족시킬 수 있다.

Example 2: Choice of Prior Distribution

- θ: 어느 광물의 나이
- 과거의 추정 결과에 따르면, θ는 대략 370 ± 21백만 년이다.
- $\theta \in \mathbb{R}^+$이지만, θ가 정규분포를 따른다고 가정해보자(Why?).
- 기존 추정 결과를 이용하면,

$$\theta \sim N(370, 21^2)$$

  로 선택할 수 있다.
- 위 사전분포에서 θ가 음수가 될 확률은 매우 낮다($< 0.1^{70}$).

Bayesian Inference

- Goal: inference about θ given the observed data.
- Posterior:
$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{f(x)},$$
where $f(x) = \int \pi(\theta) f(x \mid \theta) d\theta$.

Bayesian Inference

- (Use of SS) By the factorization theorem,

$$\begin{aligned}
\pi(\theta \mid x) &= \frac{\pi(\theta) f(x \mid \theta)}{f(x)} \\
&= \frac{\pi(\theta) f(x \mid \theta)}{\int \pi(\theta) f(x \mid \theta) d\theta} \\
&= \frac{\pi(\theta) g(T(x) \mid \theta) h(x)}{\int \pi(\theta) g(T(x) \mid \theta) h(x) d\theta} \\
&= \frac{\pi(\theta) g(T(x) \mid \theta)}{\int \pi(\theta) g(T(x) \mid \theta) d\theta}.
\end{aligned}$$

- Thus, it suffices to know the conditional distribution of SS $T(X)$.

Posterior Distribution

- Bayesian inference is based on the posterior distribution

$$\pi(\theta \mid x) = \frac{\pi(\theta) f(x \mid \theta)}{f(x)}.$$

- We update the prior information about θ ($\pi(\theta)$) as we have more information by observing the data ($f(x \mid \theta)$).

Posterior Summaries

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

- (Posterior mean)

$$\mathrm{E}(\theta \mid x) = \int \theta \cdot \pi(\theta \mid x) d\theta.$$

- (Posterior variance)

$$\begin{aligned}
\mathrm{Var}(\theta \mid x) &= \mathrm{E}\left\{(\theta - \mathrm{E}(\theta \mid x))^2 \mid x\right\} \\
&= \int (\theta - \mathrm{E}(\theta \mid x))^2 \pi(\theta \mid x) d\theta \\
&= \mathrm{E}(\theta^2 \mid x) - \mathrm{E}(\theta \mid x)^2
\end{aligned}$$

- If θ is discrete, sums would replace the integrals.

Example

$$\begin{aligned} X \mid \theta &\sim B(10, \theta) \quad \text{(likelihood)} \\ \theta &\sim Unif(0, 1) \quad \text{(prior)} \end{aligned}$$

- We have observed $x = 3$.
- Then the posterior density function is

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

Example

- The resulting posterior density is the density function of $Beta(4,8)$, i.e.,

$$\theta \mid x \quad \sim \quad Beta(4,8).$$

- The posterior mean is $\frac{4}{4+8} = \frac{1}{3}$.
- The posterior standard deviation is $\sqrt{\frac{4\times 8}{(4+8)^2(4+8+1)}} = 0.13$.

Review: Confidence Interval (신뢰구간)

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

$$P\big(L(X) < \theta < U(X) \mid \theta\big) \quad = \quad 1-\alpha.$$

- It means that if we observe $X^{(1)}, \ldots, X^{(N)} \mid \theta \overset{iid}{\sim} f(x \mid \theta)$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} I(L(X^{(i)} < \theta < U(X^{(i)}))) \quad = \quad 1-\alpha.$$

- Note that for a given observation $X = x$,

$$\text{the probability of } \theta \in (L(x), U(x)) = \begin{cases} 0 & \text{if } \theta \notin (L(x), U(x)) \\ 1 & \text{if } \theta \in (L(x), U(x)). \end{cases}$$

Credible Interval (신용구간)

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

$$\pi(L(x) < \theta < U(x) \mid x) \quad = \quad 1-\alpha.$$

- (Interpretation) The probability that θ lies in $(L(x), U(x))$.
- It does not consider the future data that have not been observed, but focuses on the current data that have been observed.
- The frequentist interpretation is less desirable if we are performing inference about θ based on a single interval.

Credible Set

(General definition) For some positive α, a $(1-\alpha)100\%$ credible set for θ is

$$\pi(\theta \in C_\alpha \mid x) = \int_{C_\alpha} \pi(\theta \mid x) d\theta = 1 - \alpha.$$

- If θ is discrete, C_α is

$$C_\alpha = \text{argmin}_{C'_\alpha} \left\{ \pi(\theta \in C'_\alpha \mid x) : \pi(\theta \in C'_\alpha \mid x) \geq 1 - \alpha \right\}.$$

- One could find multiple C_α, i.e., $(1-\alpha)100\%$ credible set may not be unique.

Quantile-based Interval

- θ_L^*: the $\alpha/2$ posterior quantile for θ, i.e., $P(\theta < \theta_L^* \mid x) = \alpha/2$.
- θ_U^*: the $1 - \alpha/2$ posterior quantile for θ, i.e., $P(\theta > \theta_U^* \mid x) = \alpha/2$.
- Then (θ_L^*, θ_U^*) is a $100(1-\alpha)\%$ credible interval for θ since

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

Example: Quantile-based Interval

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

$$X \mid \theta \quad \sim \quad B(10, \theta).$$

- Let’s use a uniform prior for θ:

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

Example: Quantile-based Interval

- Then the posterior is

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

- This is a beta distribution with parameters $x + 1$ and $10 - x + 1$.
- Since $x = 2$ here, $\pi(\theta \mid x)$ 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 2: Quantile-based Interval

Consider the normal model

$$X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} N(\theta, 2^2).$$

- Suppose $n = 16$ and we observe $\bar{x} = 16^{-1} \sum_{i=1}^{16} x_i = 0.3$.
- Assume the non-informative prior $\pi(\theta) \propto 1$.
- Calculate a 95% credible interval for θ.

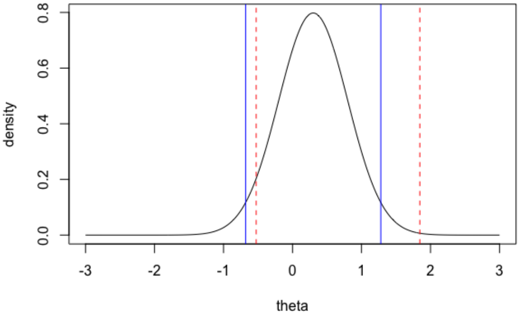
Example 2: Quantile-based Interval

The posterior is

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

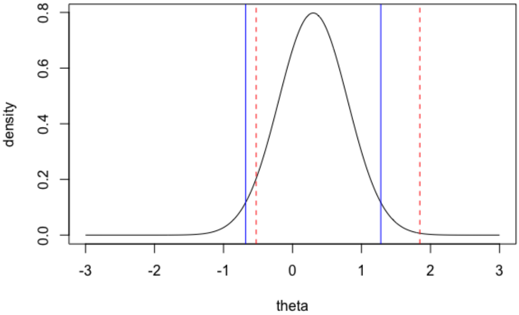
Hence the posterior distribution is $N(0.3, (0.5)^2)$.

Example 2: Quantile-based Interval




```
> theta = seq(-3,3, length=500)
> plot(theta, dnorm(theta, 0.3,0.5), type="l", ylab="density")
> abline(v=qnorm(c(0.049, 0.999), 0.3,0.5), lty=2, col=2)
> abline(v=qnorm(c(0.025, 0.975), 0.3,0.5), lty=1, col=4)
```

Example 2: Quantile-based Interval



- Intuitively, the blue credible set is better than the red one.
- How can we choose a “good” credible interval?

Highest Posterior Density (HPD) Set

A 100$(1-\alpha)$% HPD set for θ is a subset $C_\alpha \in \Theta$ defined by

$$C_\alpha = \{\theta : \pi(\theta \mid x) \geq k\},$$

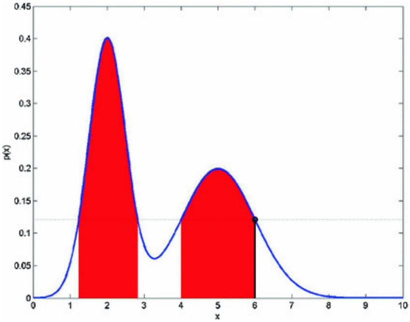
where k is the largest number such that

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

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

Highest Posterior Density (HPD) Set

- 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) Interval

A 100$(1-\alpha)$% HPD interval for θ is an interval (θ_a, θ_b) satisfying

1. $\pi(\theta_a < \theta < \theta_b \mid x) = 1 - \alpha$.
2. If $\theta_1 \in (\theta_a, \theta_b)$ and $\theta_2 \notin (\theta_a, \theta_b)$, then

$$\pi(\theta_1 \mid x) \quad > \quad \pi(\theta_2 \mid x).$$

- 즉, 최대사후구간(HPD interval)은 주어진 신뢰도를 만족하는 베이지안 구간 중 최대한 사후 밀도함수값이 높은 θ들의 집합이다.
- 사후 밀도함수값이 높으므로, 우량의 θ를 많이 포함하고 있다고 해석 가능하다.
- 100$(1-\alpha)$% credible interval 중 가장 짧은 구간을 제공한다. (Why?)

Highest Posterior Density (HPD) Interval

A $100(1-\alpha)\%$ HPD interval for θ is an interval (θ_a, θ_b) satisfying

1. $\pi(\theta_a < \theta < \theta_b \mid x) = 1 - \alpha$.
2. If $\theta_1 \in (\theta_a, \theta_b)$ and $\theta_2 \notin (\theta_a, \theta_b)$, then

$$\pi(\theta_1 \mid x) \quad > \quad \pi(\theta_2 \mid x).$$

- 즉, 최대사후구간(HPD interval)은 주어진 신뢰도를 만족하는 베이지안 구간 중 최대한 사후 밀도함수값이 높은 θ들의 집합이다.
- 사후 밀도함수값이 높으므로, 우량의 θ를 많이 포함하고 있다고 해석 가능하다.
- $100(1-\alpha)\%$ credible interval 중 가장 짧은 구간을 제공한다. (Why?)

Highest Posterior Density (HPD) Interval

A $100(1-\alpha)\%$ HPD interval for θ is an interval (θ_a, θ_b) satisfying

1. $\pi(\theta_a < \theta < \theta_b \mid x) = 1 - \alpha$.
2. If $\theta_1 \in (\theta_a, \theta_b)$ and $\theta_2 \notin (\theta_a, \theta_b)$, then

$$\pi(\theta_1 \mid x) \quad > \quad \pi(\theta_2 \mid x).$$

- 즉, 최대사후구간(HPD interval)은 주어진 신뢰도를 만족하는 베이지안 구간 중 최대한 사후 밀도함수값이 높은 θ들의 집합이다.
- 사후 밀도함수값이 높으므로, 우량의 θ를 많이 포함하고 있다고 해석 가능하다.
- $100(1-\alpha)\%$ credible interval 중 가장 짧은 구간을 제공한다. (Why?)

Highest Posterior Density (HPD) Interval

A 100$(1-\alpha)$% HPD interval for θ is an interval (θ_a, θ_b) satisfying

1. $\pi(\theta_a < \theta < \theta_b \mid x) = 1 - \alpha$.
2. If $\theta_1 \in (\theta_a, \theta_b)$ and $\theta_2 \notin (\theta_a, \theta_b)$, then

$$\pi(\theta_1 \mid x) \quad > \quad \pi(\theta_2 \mid x).$$

- 즉, 최대사후구간(HPD interval)은 주어진 신뢰도를 만족하는 베이지안 구간 중 최대한 사후 밀도함수값이 높은 θ들의 집합이다.
- 사후 밀도함수값이 높으므로, 우량의 θ를 많이 포함하고 있다고 해석 가능하다.
- $100(1-\alpha)$% credible interval 중 가장 짧은 구간을 제공한다. (Why?)

How to Find HPD Interval

- When θ is continuous, the boundaries of HPD interval have the same posterior density values.
- We consider an imaginary horizontal bar and moving it downward until the posterior probability between the points becomes $1 - \alpha$.
- It may be hard to find the HPD interval, so one can calculate approximate HPD interval in this case.

Method 1: Quantile-based Method

- Suppose that the posterior is symmetric and unimodal.
- Consider the $\alpha/2$ and $1 - \alpha/2$ percentiles.
- If the posterior distribution is well-known, the existing packages can be exploited.
- Otherwise some sampling methods can be used.

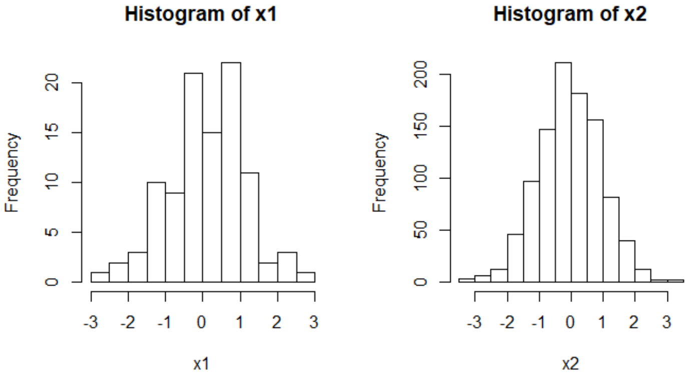
Method 1: Quantile-based Method

```
> 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
```

Method 1: Quantile-based Method

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



Method 2: Grid Search Method (격자점 방법)

- (Main idea)
  1. Consider θ as N distinct values $\{\theta_1, ...\theta_N\}$
  2. Approximate the posterior density function $\pi(\theta \mid x)$ with the normalized posterior probabilities on $\{\theta_1, ...\theta_N\}$
- Calculate

$$\widehat{\pi}(\theta_i \mid x) = \frac{\pi(\theta_i) f(x \mid \theta_i)}{\sum_{i=1}^{N} \pi(\theta_i) f(x \mid \theta_i)}.$$

- Find M such that

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

Method 2: Grid Search Method (격자점 방법)

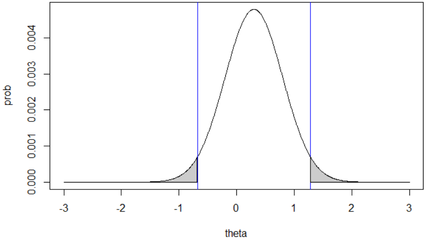
```
> HPDgrid = function(prob, level = 0.95){
>  prob.sort = sort(prob, decreasing = T)
>  M = min( which(cumsum(prob.sort)>=level) )
>  height = prob.sort[M]
>  HPD.index = which( prob >= height)
>  HPD.level = sum(prob[HPD.index])
>  res = list( index = HPD.index, level = HPD.level )
>  return(res)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

Method 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
```



Method 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
```

Method 2: Grid Search Method (격자점 방법)

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

Method 3: Posterior Sampling

- 특정 분포로부터 샘플들로 이루어진 히스토그램이 밀도함수와 유사하다는 성질을 이용
- 1000개의 사후표본이 주어졌을때, 95% CI는 950개의 표본을 포함
- 1000개의 θ 오름차순으로 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.

Method 3: Posterior Sampling

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

Weakness of Frequentist

- 편이, 분산, 신뢰구간, 가설검정의 오차확률 등은 모든 가능한 X값에 대하여 적분이나 합의 형식을 취한 값들
- 즉, 현재 주어진 관측치가 아니라 실험이나 표본조사를 무한히 반복했을 때 발생할 수 있는 가능한 모든 관측치들을 고려하여 얻어지는 것들
- 고전적 통계추론은 가상적인 반복실험을 가정하기 때문에 때로 납득하기 어려운 결과를 제공하기도 한다.

Example1 : Weakness of Frequentist

- 분산이 $\sigma^2 = 1$인 정규분포의 평균 θ를 추정하고자 한다.
- 동전을 던져 앞면이 나오면 표본을 2개만 취하고, 뒷면이 나오면 표본을 1000개 취하기로 하였다.
- θ에 대한 추정치는 표본의 평균 $\bar{X}$가 적절하며 $\bar{X}$의 정확도를 측정하는 통계량으로는 $\bar{X}$의 분산이 적절 할 것이다.
- 이 실험에서 $\bar{X}$의 분산은

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

Example 1: 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?

Example 2: Weakness of Frequentist

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

Weak Conditionality Principle

- Suppose one can perform either of two experiments E_1 and E_2, both pertaining to θ and the actual experiment is conducted is the mixed experiment of first choosing $J = 1, 2$ with probability 0.5 each independent of θ.
- Then, perform E_J.
- The actual inference about θ obtained overall mixed experiment should depend only on the experiment E_J that is actually performed.

Weak Conditionality Principle

- Suppose one can perform either of two experiments E_1 and E_2, both pertaining to θ and the actual experiment is conducted is the mixed experiment of first choosing $J = 1, 2$ with probability 0.5 each independent of θ.
- Then, perform E_J.
- The actual inference about θ obtained overall mixed experiment should depend only on the experiment E_J that is actually performed.

Birnbaum's Proof (1962)

- (Sufficiency Principle) The information contained in X and $T(X)$ are the same.
- In discrete models,

$$\begin{aligned} & \text{Weak Conditionality Principle} + \text{Sufficiency Principle} \\ \Longleftrightarrow \quad & \text{Likelihood Principle.} \end{aligned}$$

- Bayesian inference is based on the likelihood principle.

Birnbaum's Proof (1962)

- (Sufficiency Principle) The information contained in X and $T(X)$ are the same.
- In discrete models,

$$\begin{aligned} & \text{Weak Conditionality Principle} + \text{Sufficiency Principle} \\ \iff & \text{Likelihood Principle.} \end{aligned}$$

- Bayesian inference is based on the likelihood principle.

Birnbaum's Proof (1962)

- (Sufficiency Principle) The information contained in X and $T(X)$ are the same.
- In discrete models,

$$\begin{aligned} & \text{Weak Conditionality Principle} + \text{Sufficiency Principle} \\ \Longleftrightarrow \quad & \text{Likelihood Principle.} \end{aligned}$$

- Bayesian inference is based on the likelihood principle.