Chapter 5: Bayesian Inference for Binomial Distribution

강의 목표

- 이항분포를 중심으로 베이지안 추론의 이해
- Parameter Estimation (모수 추정)
  - Point Estimation (점추정)
  - Confidence Interval (구간추정)
- Prediction (예측)

Beta Posterior Distribution

- Consider 40 flips of a coin having $\Pr(\mathit{Heads}) = \theta$.
- Suppose we observe 15 "heads".
- We model the count of heads as binomial:

$$p(X = x \mid \theta) = \binom{40}{x} \theta^x (1-\theta)^{40-x}, \quad x = 0, 1, ..., 40.$$

- Let's use a uniform prior for θ:

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

Beta Posterior Distribution

- Then the posterior is:

$$\begin{aligned} \pi(\theta \mid x) &\propto p(\theta) L(\theta \mid x) \\ &= \binom{40}{x} \theta^x (1-\theta)^{40-x} \\ &\propto \theta^x (1-\theta)^{40-x}, \quad 0 \leq \theta \leq 1. \end{aligned}$$

- This is a beta distribution for θ with parameters $x+1$ and $10-x+1$.
- Since $x = 15$ here, $\pi(\theta \mid x = 15)$ is beta $(16, 26)$.
- Then the point estimation for θ is:

$$\begin{aligned} \mathrm{Mode}(\theta \mid X_1, ..., X_n) &= 15/(15+25) = 0.375 \\ \mathbb{E}(\theta \mid X_1, ..., X_n) &= 16/(16+26) = 0.381 \end{aligned}$$

Beta Posterior Distribution

- Posterior distribution is a combination of prior information of θ and data.
- In this example,
  - Prior: 특정한 θ에 차별을 두지 않는다.
  - Data: θ가 0.375에 가까울 수록 확률이 높다.

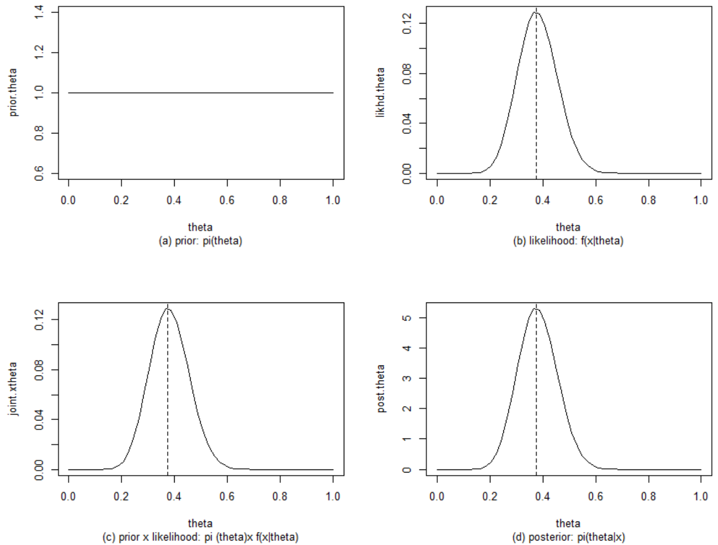
Prior vs Likelihood vs Posterior

```
> # theta ~ Beta(a, b)
> a=1 ; b=1
> # x|theta - B(n, theta)
> n=40 ; x=15
> # a discretization of the possible theta values
> theta = seq(0, 1, length=50)
> prior.theta = dbeta(theta, a, b)
> # prob of data\theta(likelihood)
> likhd.theta = dbinom ( x, n, theta)
> # joint prob of data & theta
> joint.xtheta = prior.theta*likhd.theta
> # posterior of theta
> post.theta = dbeta(theta, a+x, b+n-x)
```

Prior vs Likelihood vs Posterior

```
par (mfrow=c(2, 2)) # set up a 2x2 plotting window plot
plot (theta, prior.theta, type="l",
sub="(a) prior: pi(theta)")
plot(theta, likhd.theta, type="l",
sub="(b) likelihood: f(x|theta)")
abline(v=x/n, lty=2)
plot(theta, joint.xtheta, type="l",
sub="(c) prior x likelihood: pi (theta)x f(x|theta)")
abline(v=(a+x-1)/(a+b+n-2), lty=2)
plot (theta, post.theta, type="l",
sub="(d) posterior: pi(theta|x)")
abline(v=(a+x-1)/(a+b+n-2), lty=2)
```

Prior vs Likelihood vs Posterior



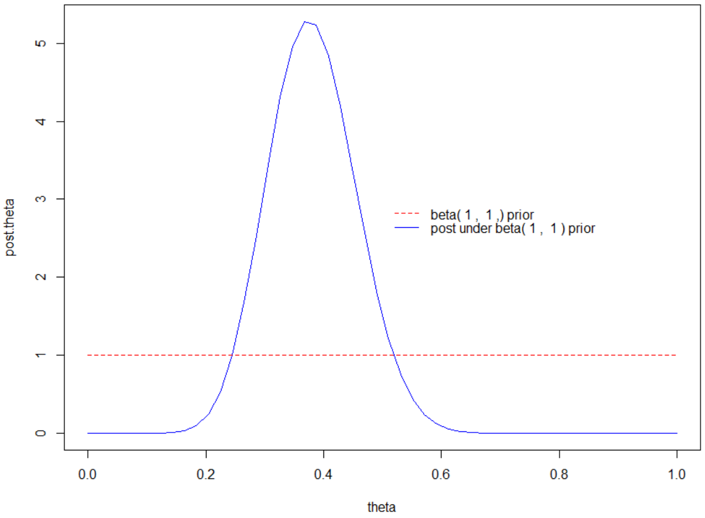
Prior vs Likelihood vs Posterior

- Likelihood와 Uniform $\times$ Likelihood는 동일
- Posterior와 Uniform $\times$ Likelihood은 세로축만 다름
- Posterior와 Likelihood의 평균은 다름.

Prior vs Likelihood vs Posterior

```
> par(mfrow=c(1, 1))
> plot(theta, post.theta, type="l", col="blue")
> lines(theta, prior.theta, col="red", lty=2)

> legend(.5, 3, legend=c(paste("beta(",a,", ",b,",) prior"),
+         paste("post under beta(",a, ", ",b,") prior")),
+         lty=c(2, 1), col=c("red", "blue"), bty="n")
```

Prior vs Likelihood vs Posterior



Prior vs Likelihood vs Posterior

- 사전정보: 어떤 특정한 θ 에 대하여 차별을 두지 않음
- 사후정보: θ가 0.375에 가까운 값일 확률이 매우 높다.

Monte Carlo Method

- It is often very difficult to find the posterior distribution.
- Solution: Monte Carlo Method.
- Through the **simulation**, find information of the posterior distribution.

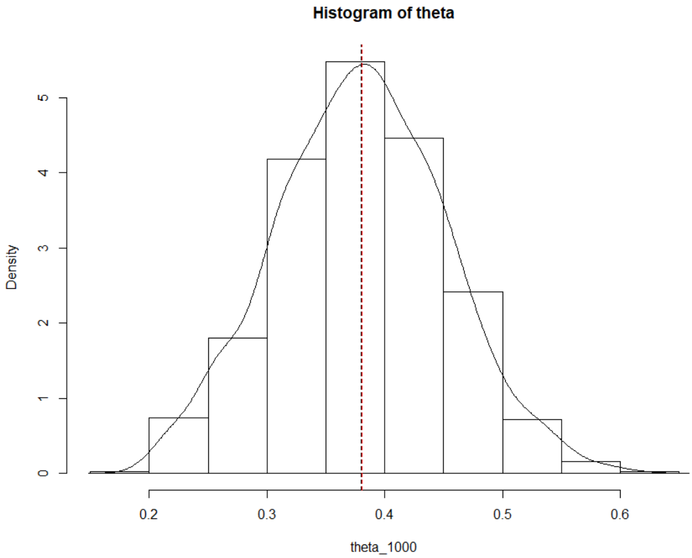
Monte Carlo Method

```
> theta_1000 = rbeta(1000, a+x, b+n-x) # generate posterior samples
> quantile(theta_1000, c(.025, .975)) # simulation-based quantiles
2.5%     97.5%
0.2412677 0.5268378
> qbeta(c(.025, .975), a+x, b+n-x) # theoretical quantiles
[1] 0.2420110 0.5306375
> mean(theta); var(theta)  # simulation-based estimates
[1] 0.379344
[1] 0.005324879
> # theoretical estimates
> (a+x)/(a+b+n); (a+x)*(b+n-x)/((a+b+n+1)*(a+b+n)^2)
[1] 0.3809524
[1] 0.005484364
```

Monte Carlo Method

```
> hist(theta_1000, prob=T, main="Histogram of theta")

> lines(density(theta_1000))

> mean.theta = mean(theta_1000)

> abline(v=mean.theta, lty=2)

> abline(v=(a+x)/(a+b+n), lty=2, col = "red")
```

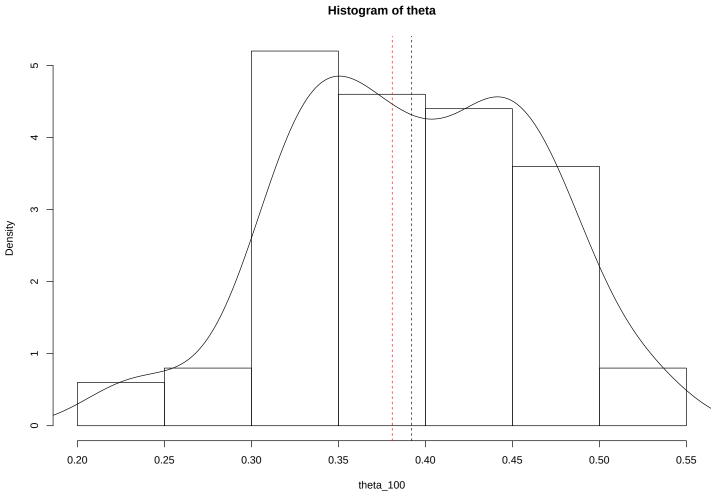
Monte Carlo Method



Monte Carlo Method

```
theta_100 = rbeta(100, a+x, b+n-x) # generate posterior samples
hist(theta_100, prob=T, main="Histogram of theta")
lines(density(theta_100))
mean.theta = mean(theta_100)
abline(v=mean.theta, lty=2)
abline(v=(a+x)/(a+b+n), lty=2, col = "red")
```

Monte Carlo Method



Monte Carlo Method

```
> ## log odds ratio
> a=b=1
> X=15;n=40
> theta=rbeta(10000,a+x,b+n-x)
> eta=log(theta/(1-theta))
> hist(eta, prob=T, main="Histogram of eta")
> lines(density(eta), lty=2)
> mean(eta); var(eta)
[1] -0.4947897
[1] 0.1035466
```

Beta/Binomial Bayesian Model

- Suppose we observe n independent Bernoulli(p) r.v.'s $X_1, ..., X_n$. We wish to estimate the "success probability" p via the Bayesian approach.
- We will use a $beta(a, b)$ prior for p and show this is a conjugate prior.
- Consider the r.v. $Y = \sum_{i=1}^{n} X_i$. This is $Binomial(n, p)$ distribution.
- We first write the joint density of Y and p.

Complete Derivation of Beta/Binomial Model

$$
\begin{aligned}
f(y, p) &= f(y \mid p) f(p) \\
&= \left[\binom{n}{y} p^y (1-p)^{n-y}\right] \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}\right] \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{y+a-1} (1-p)^{n-y+b-1}
\end{aligned}
$$

Derivation of Beta/Binomial Model

The marginal density of Y.

$$
\begin{aligned}
f(y) &= \int_0^1 f(y,p)dp \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 p^{y+a-1}(1-p)^{n-y+b-1}dp \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)} \\
&\times \int_0^1 \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}p^{y+a-1}(1-p)^{n-y+b-1}dp \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}.
\end{aligned}
$$

Derivation of Beta/Binomial Model

- Then the posterior $\pi(p \mid y) = f(p \mid y)$ is

$$\frac{f(y,p)}{f(y)} = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)} p^{y+a-1}(1-p)^{n-y+b-1}$$

  for some $0 \leq p \leq 1$.

- Clearly this posterior is a $\mathit{beta}(y+a, n-y+b)$ distribution.

Inference with Beta/Binomial Model

- Consider letting $\hat{p}$ be the posterior mean.
- The mean of the (posterior) beta distribution is:

$$\hat{p} = \frac{y+a}{y+a+n-y+b} = \frac{y+a}{a+b+n}.$$

- Note that it can be decomposed

$$\hat{p} = \underbrace{\frac{n}{a+b+n}\left(\frac{y}{n}\right)}_{\text{sample mean}} + \underbrace{\frac{a+b}{a+b+n}\left(\frac{a}{a+b}\right)}_{\text{prior mean}}.$$

Inference with Beta/Binomial Model

- So the Bayes estimator $\hat{p}$ is a weighted average of the usual frequentist estimator (sample mean) and the prior mean.
- As n increases, the sample data are weighted more heavily and the prior information less heavily.
- In general, with Bayesian estimation, as the sample size increases, the likelihood dominates the prior.

Characteristics of Beta/Binomial Model

- Easy to derive the posterior distribution
- Easy to apply Monte Carlo method
- Easy to add new data
- Restricted form of the prior distribution

Prediction for Beta/Binomial Model

- 데이터 $X_1, X_2, ..., X_n$이 주어졌을 때, 다음 관측치 $\mathbf{X}_{n+1}$에 대한 예측 확률.

$$
\begin{aligned}
& P(\mathbf{X}_{n+1} = 1 \mid X_1, X_2, ..., X_n) \\
= & \int P(\mathbf{X}_{n+1} = 1 \mid \theta, X_1, X_2, ..., X_n)\pi(\theta \mid X_1, X_2, ..., X_n) \\
= & \int P(\mathbf{X}_{n+1} = 1 \mid \theta)\pi(\theta \mid X_1, X_2, ..., X_n) \\
= & \int \theta\pi(\theta \mid X_1, X_2, ..., X_n) \\
= & \mathbb{E}(\theta \mid X_1, X_2, ..., X_n)
\end{aligned}
$$

Distribution of Prediction

- 데이터 $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$이 주어졌을 때, 다음 관측치 $\mathbf{X}_{n+1}$에 대한 예측 확률.

$$
\begin{aligned}
& P(\mathbf{X}_{n+1} = 1 \mid X_1, X_2, ..., X_n) \\
= & \int P(\mathbf{X}_{n+1} = 1 \mid \theta, X_1, X_2, ..., X_n)\pi(\theta \mid X_1, X_2, ..., X_n) \\
= & \int P(\mathbf{X}_{n+1} = 1 \mid \theta)\pi(\theta \mid X_1, X_2, ..., X_n) \\
= & \int \theta\pi(\theta \mid X_1, X_2, ..., X_n) \\
= & \mathbb{E}(\theta \mid X_1, X_2, ..., X_n)
\end{aligned}
$$

Distribution of Prediction: Beta-Binomial distribution

- 데이터 $x_1, x_2, ..., x_n$이 주어졌을 때, 다음 관측치 $Z = \mathbf{X}_{n+1} + \mathbf{X}_{n+2} + ...\mathbf{X}_{n+m}$에 대한 예측 확률.

$$P(Z \mid X_1, X_2, ..., X_n) = \binom{m}{z} \frac{\Gamma(a+b+n)}{\Gamma(a+\sum x_i)\Gamma(b+n-\sum x_i)} \times \frac{\Gamma(a+\sum x_i+Z)\Gamma(b+n-\sum x_i+m-Z)}{\Gamma(a+b+n+m)}. \quad \text{(숙제)}$$

- 위의 예측 분포를 베타-이항분포 (Beta-Binomial distribution) 이라고 한다.

Example: Beta-Binomial distribution

- 앞선 동전 던지기 실험에서, 앞으로 10번 던졌을때 성공횟수 Z에 대한 예측 분포.
- Frequentist:

$$P(Z = z \mid \hat{\theta} = 0.375) = \binom{10}{z} 0.375^z (1-0.375)^{10-z}, \quad z = 0, ..., 10.$$

- Bayesian:

$$\begin{aligned} & P(Z = z \mid X_1, X_2, ..., X_n) \\ & = \binom{10}{z} \frac{\Gamma(1+1+40)}{\Gamma(1+15)\Gamma(1+40-15)} \\ & \times \frac{\Gamma(1+15+z)\Gamma(1+40-15+10-z)}{\Gamma(1+1+40+10)}. \end{aligned}$$

Example: Beta-Binomial distribution

- Bayesian:

$$P(Z = z \mid X_1, X_2, ..., X_n) = \binom{10}{z} \frac{\Gamma(42)}{\Gamma(16)\Gamma(26)} \frac{\Gamma(16+z)\Gamma(36-z)}{\Gamma(52)}.$$

- Prediction for Z is $\mathbb{E}(Z \mid X_1, ..., X_n) = 3.8095$.
- 베이지안 예측평균이 빈도론자 예측평균에 비해서 약간 크다.

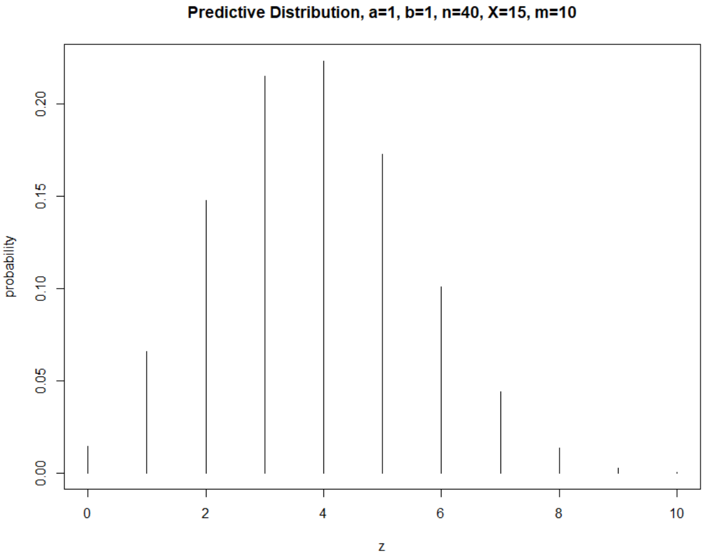
Example: Beta-Binomial distribution

- 빈도론자 예측분산: 2.3438.
- 베이지안 예측분산: 2.8558.
- 베이지안 예측분산이 빈도론자 예측분산에 비해서 약간 크다.
- 베이지안은 θ의 변동성을 고려했기 때문.
- 고전적 예측에서는 예측분산을 작게 추정하는 (underestimate) 문제가 발생.

Example: Beta-Binomial distribution

```
> ## beta binomial distribution ####
> a=b=1
> n=40;x=15
> m=10;z=c(0:10)
> pred.z = gamma(m+1)/gamma(z+1)/gamma(m-z+1)*beta(a+z+x,
+          b+n-x+m-z)/beta(a+x, b+n-x)
> plot(z, pred.z, xlab="z", ylab="probability", type="h")
> title("Predictive Distribution, a=1, b=1, n=40, X=15, m=19")
```

Example: Beta-Binomial distribution



Monte Carlo Method Example

$$f(Z = z \mid X_1, ..., X_n) = \mathbb{E}(f(z \mid \theta)).$$

We find an approximate $\mathbb{E}(f(z \mid \theta))$ using Monte Carlo method.

- First Method: Suppose that we sample $\theta_1, ..., \theta_N$ from the posterior distribution.

$$\widehat{f}(z \mid X_1, ..., X_n) = \frac{1}{N}\sum_{i=1}^{N} f(z \mid \theta_i) = \frac{1}{N}\sum_{i=1}^{N} \binom{m}{z} \theta_i^z (1-\theta_i)^{m-z}.$$

Monte Carlo Method Example

$$f(Z = z \mid X_1, ..., X_n) = \mathbb{E}(f(z \mid \theta)).$$

We find an approximate $\mathbb{E}(f(z \mid \theta))$ using Monte Carlo method.

- Second Method: Using the following property.

$$f(z, \theta \mid X_1, ..., X_n) = f(z \mid \theta)\pi(\theta \mid X_1, ..., X_n).$$

- We randomly choose N samples $(z_i, \theta_i)_{i=1}^N$ from

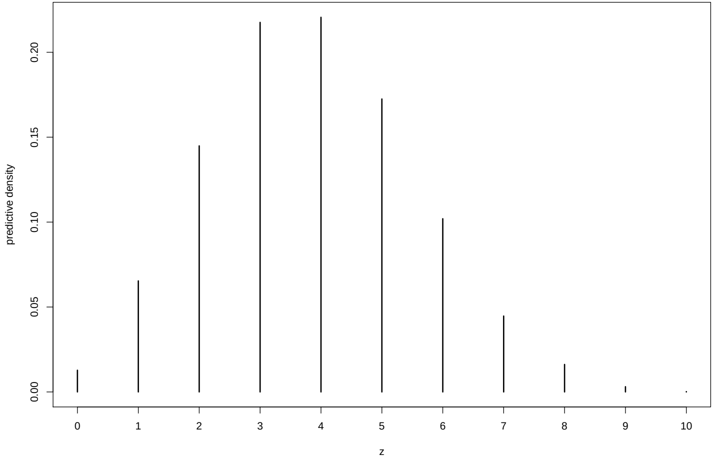
$$\begin{aligned} \theta_i &\sim Beta(a + x, b + n - x) \\ z_i \mid \theta_i &\sim Bin(10, \theta_i). \end{aligned}$$

Monte Carlo Method Example

- We choose only $\{z_i\}$.

```
> ### Monte Carlo Method ####
> a=b=1; X=15; n=40; m=10; N=10000
> theta = rbeta(N,a+x,b+n-x)
> pred.z=c(1: (m+1))*0
> for(z in c(0:m)) pred.z[z+1]=mean(dbinom(z,m, theta))
> zsample=rbinom(N, m, theta)
> plot(table(zsample)/N, type="h", xlab="z", ylab="predictive density",
main="")
> mean(zsample)
[1] 3.8373
> var(zsample)
[1] 2.891118
```

Monte Carlo Method Example



Bayesian Credible Interval

- Consider Beta posterior distribution.
- 시행횟수 $n = 10$.
- 관측성공횟수 $X = 2$.
- Non-informative prior $\theta \sim U(0, 1)$.

Bayesian C.I Example

Bayesian C.I using Grid Search Method

```
a=b=1

X=2; n=10;

theta = seq(0,1,length = 1001)

ftheta=dbeta(theta,a+X, n-X+b)

prob=ftheta/sum(ftheta)

HPD = HPDgrid(prob, 0.95)

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

HPD.grid

[1] 0.041 0.484
```

Classical C.I Example

Classical C.I using Quantile-based Method

```
install.packages("binom")

library(binom)

n=10; X=2

CI.exact=binom.confint(X, n, conf.level = 0.95, methods = c("exact"))

CI.exact=c(CI.exact$lower, CI.exact$upper)

CI.exact

[1] 0.02521073 0.55609546
```

Bayesian C.I vs Classical C.I

- Bayesian C.I:

$$P(\theta \in (0.041, 0.484) \mid X) = 0.95.$$

- Classical C.I:

$$P(\theta \in (0.025, 0.556) \mid X) \neq 0.95.$$

- The Bayesian C.I is shorter than the classical C.I because the Bayesian C.I exploits the prior.
- The Bayesian C.I is valid even if $X = 0$ or $X = n$.

Problem of Bayesian and Classical C.Is

- Bayesian C.I: it is very hard to find the HPD interval.
- Classical C.I: it sometimes provide meaningless interval.

Bayesian C.I vs Classical C.I

```
> HPD.approx=qbeta(c(0.025, 0.975),a+X, n-X+b)
> p=X/n
> CI.asympt=c(p-1.96*sqrt(p*(1-p)/n), p+1.96*sqrt(p*(1-p)/n))
> HPD.approx
[1] 0.06021773 0.51775585
> CI.asympt
[1] -0.04792257  0.44792257
```

Bayesian C.I vs Classical C.I