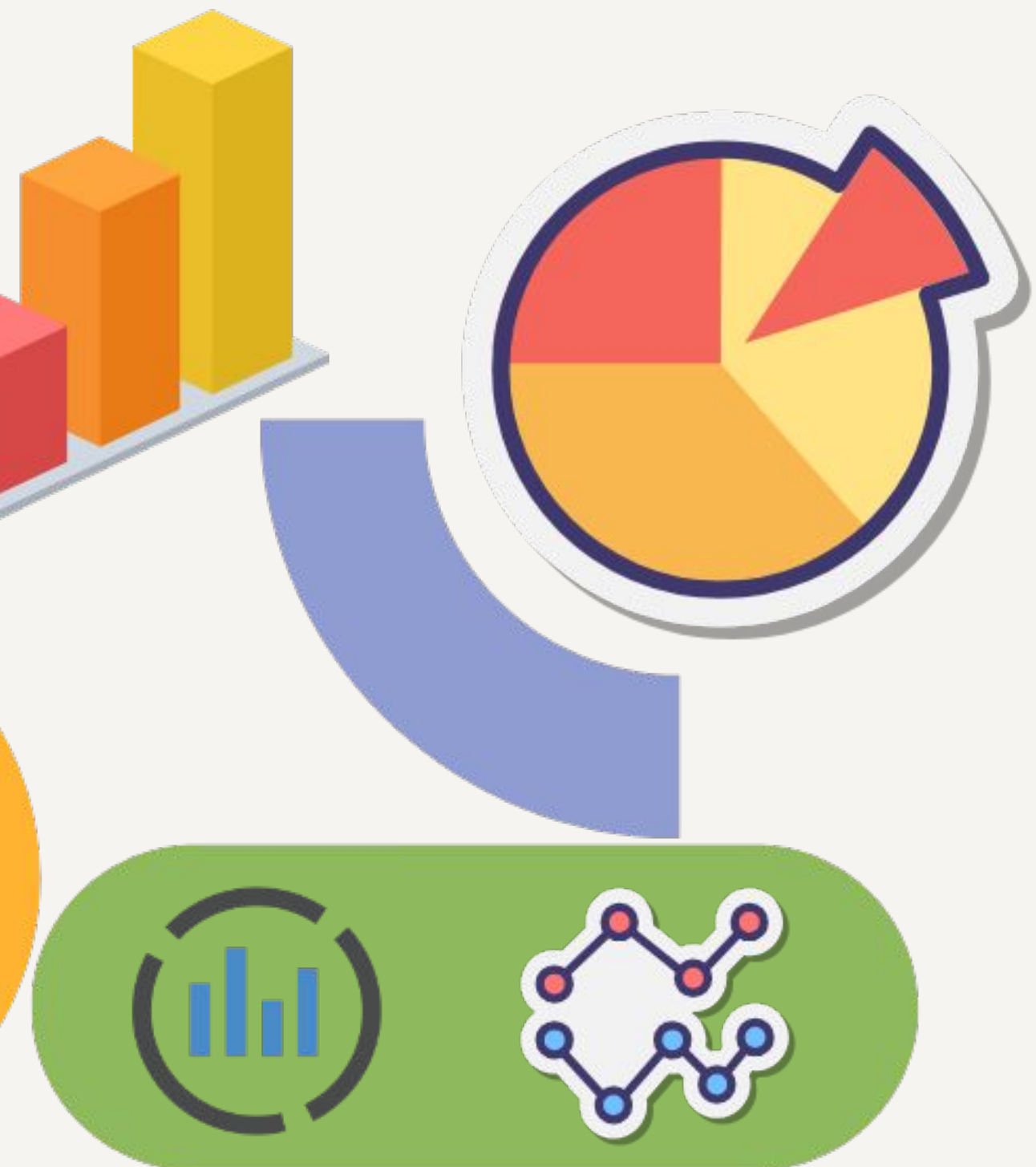


Bayesian Statistics

하지운

Department of Statistics



INDEX

Review

01

Big Question

02

Statistical Inference &
Nonparametric Statistics

03

Bayesian Statistics

04

Markov Chain Monte Carlo

05

Reference

06

Review

- **Monte Carlo Method is a class of computational algorithm that rely on repeated random sampling to compute results.**
- **Monte Carlo methods, or Monte Carlo experiments, are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.**
- **Developed in the 1940's by Neumann, Ulam, and Metropolis, while they were working on Manhattan Project in the Los Alamos National Laboratory.**
- **Named from the Monte Carlo casino.**
- **Also called parametric bootstrap.**

Review

- **Algorithm: Estimation of Standard Error of $\hat{\theta}$.**

- **1. For $j = 1, \dots, m$,**

- (1) Generate a sample $\underline{x}^{(j)} = (x_1^{(j)}, \dots, x_n^{(j)}) \sim f(x)$.**

- (2) Compute $\hat{\theta}^{(j)}$**

- **2. $\widehat{SE}(\hat{\theta}) = \sqrt{\frac{\sum_{j=1}^m (\hat{\theta}^{(j)} - \bar{\hat{\theta}})^2}{m-1}}$, where $\bar{\hat{\theta}} = \frac{1}{m} \sum_{j=1}^m \hat{\theta}^{(j)}$.**

$\Rightarrow$ Standard Deviation of $(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(m)})$

Review

- **Suppose that X_1, X_2 have independent and identically distributed from $N(0,1)$.**
- **We want to estimate $\theta = E(X_1 - X_2)$.**
- **Algorithm**
- **1. For $j = 1, \dots, m$,**
- **Generate a sample $(x_1^{(j)}, x_2^{(j)}) \sim N(0,1)$**
- **Compute $\hat{\theta}^{(j)} = |x_1^{(j)} - x_2^{(j)}|$.**
- **2. Compute $\bar{\hat{\theta}}$ and $S.E.(\hat{\theta})$.**

Review

```
MCtheta.R = function(m){  
  theta = numeric(m) # same as theta = [] with m elements  
  for (i in 1:m){  
    x = rnorm(2,0,1) #random 2 numbers  
    theta[i] = abs(x[1]-x[2])  
  }  
  mean_theta = mean(theta)  
  s.e.theta = sqrt(sum(theta-mean_theta)^2/(m-1))  
  return(s.e.theta)  
}  
MCtheta.R(1000)  
#Theoretical value = 0
```

```
> MCtheta.R = function(m){  
+   theta = numeric(m) # same as theta = [] with m elements  
+   for (i in 1:m){  
+     x = rnorm(2,0,1) #random 2 numbers  
+     theta[i] = abs(x[1]-x[2])  
+   }  
+   mean_theta = mean(theta)  
+   s.e.theta = sqrt(sum(theta-mean_theta)^2/(m-1))  
+   return(s.e.theta)  
+ }  
> MCtheta.R(1000)  
[1] 1.531489e-15
```

```
> MCtheta.R(1000)  
[1] 5.549893e-16  
> MCtheta.R(1000)  
[1] 2.058378e-15  
> MCtheta.R(1000)  
[1] 3.020828e-16  
> MCtheta.R(1000)  
[1] 2.409637e-15  
> MCtheta.R(1000)  
[1] 7.20081e-16
```

```
> MCtheta.R(5000)  
[1] 1.38339e-15  
> MCtheta.R(5000)  
[1] 5.682733e-15  
> MCtheta.R(5000)  
[1] 5.004385e-15  
> MCtheta.R(5000)  
[1] 4.509757e-15  
> MCtheta.R(5000)  
[1] 1.713142e-15
```

```
> MCtheta.R(10000)  
[1] 5.988842e-15  
> MCtheta.R(10000)  
[1] 9.054321e-15  
> MCtheta.R(10000)  
[1] 6.724957e-15  
> MCtheta.R(10000)  
[1] 9.892582e-15  
> MCtheta.R(10000)  
[1] 8.017321e-15  
> #Theoretical value = 0
```




Big Question

- **How can we believe that result?**
- **How much do we set the credibility for that result?**
- **How can we compute the confidence interval?**



Statistical Inference

- **What is Hypothesis Test?**
- **For the parameter space $\Theta = \Theta_0 \cup \Theta_1$, determine if**
$$H_0: \theta \in \Theta_0 \text{ versus } H_1: \theta \in \Theta_1$$
- **We define rejection region (or critical region) R is constructed as that we reject H_0 whenever $X \in R$, i.e.,**
 - 1. Reject H_0 if $X \in R$**
 - 2. Retain H_0 if $X \notin R$**
- **Typically, a rejection region is specified using a test statistic**
$$T = T(X)$$



Statistical Inference

- Let R be a rejection region.

- Then, $\phi(x) = I(x \in R) = \begin{cases} 1, & \text{reject } H_0 \\ 0, & \text{retain } H_0 \end{cases}$ is the test function.

- For a given test $\phi(x) = I(x \in R)$,

$$\beta(\theta) = \mathbb{E}_\theta \phi(X) = \mathbb{P}_\theta(X \in R), \quad \theta \in \Theta$$

is called the power function of the test.

- For a given test $\phi(x) = I(x \in R)$ and $\alpha \in (0,1)$, if

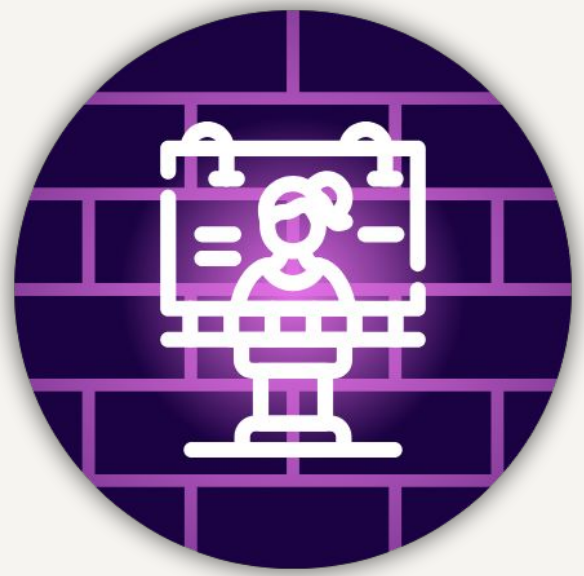
$$\max_{\theta_0 \in \Theta_0} \beta(\theta_0) = \max_{\theta_0 \in \Theta_0} \mathbb{E}_{\theta_0} \phi(X) = \max_{\theta_0 \in \Theta_0} \mathbb{P}_{\theta_0}(X \in R) = \alpha$$



What is p-value?

- The probability of observing a test statistic at least as extreme as the one actually observed under H_0 .
- In other words, we reject H_0 if p-value is less than significant level α .





Example for Confidence Interval

- $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$.
- **Suppose σ^2 is known (μ is unknown).**
- **Then we can construct confidence interval (CI)**
with significant level α .





Example for Confidence Interval

- $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.
- **Suppose μ is known (σ^2 is unknown).**
- **Then we can construct confidence interval (CI) with significant level α .**

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$



★ Example (Nonparametric Statistics)

max: 85, mean = 70, my score = 55, otherwise.

- We have 7 observed data from population as follows:

70,45,35,60,85,55,55

- We want to test $H_0 : \theta = 100$ vs $H_1 : \theta = 50$, where $\theta = \text{Var}(X)$

- Algorithm

- 1. Compute $\hat{\theta}_{obs} = \frac{1}{7} \sum_{i=1}^7 (x_i)^2 - \left(\frac{1}{7} \sum_{i=1}^7 x_i\right)^2$.

- 2. Create new dataset $x_i^* = \bar{x} + \frac{(x_i - \bar{x})\sqrt{57}}{\sqrt{\hat{\theta}_{obs}}}$ where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.



★ Example (Nonparametric Statistics)

- **Algorithm**

- **3. For $j = 1, \dots, m$,**

(1) Generate a sample $x^{(j)} = (x_1^{(j)}, \dots, x_7^{(j)})$ by resampling from x_i^* .

(2) Compute $\hat{\theta}^{(j)} = \frac{1}{7} \sum_{i=1}^7 (x_i^{(j)})^2 - (\frac{1}{7} \sum_{i=1}^7 x_i^{(j)})^2$.

- **4. Compute $p = \frac{1}{m} \sum_{i=1}^m I(\hat{\theta}^{(j)} \geq \hat{\theta}_{obs})$.**



★ Example (Nonparametric Statistics)

```
var_stat <- function(data) {  
  return(mean(data^2) - mean(data)^2) #calculate sample var  
}  
Bootstrap = function(m, H0_value, alpha){  
  x = c(70,45,35,60,85,55,55) #our sample  
  n = length(x) #sample size  
  theta_obs = var_stat(x) #standard value  
  x_bar = mean(x)  
  s = sqrt(theta_obs)  
  x_star = x_bar + (x - x_bar)/s * sqrt(H0_value) #new dataset  
  theta = numeric(m)  
  for (j in 1:m){  
    x_j = sample(x_star, n, replace = T) #resampling  
    theta[j] = var_stat(x_j)  
  }  
  
  p_value = sum(theta >= theta_obs) / m  
  test_result = ifelse(p_value >= alpha, T, F)  
  return(c(p_value, test_result))  
}  
  
result = Bootstrap(10000, 100, 0.05)  
  
print(paste("Bootstrap p-value:", result[1]))  
ifelse(result[2], "Result: p-value >= alpha. Fail to reject H0.",  
       "Result: p-value < alpha. Reject H0.")  
  
> print(paste("Bootstrap p-value:", result[1]))  
[1] "Bootstrap p-value: 0.001"  
> ifelse(result[2], "Result: p-value >= alpha. Fail to reject H0.",  
+         "Result: p-value < alpha. Reject H0.")  
[1] "Result: p-value < alpha. Reject H0."
```



Result from Example

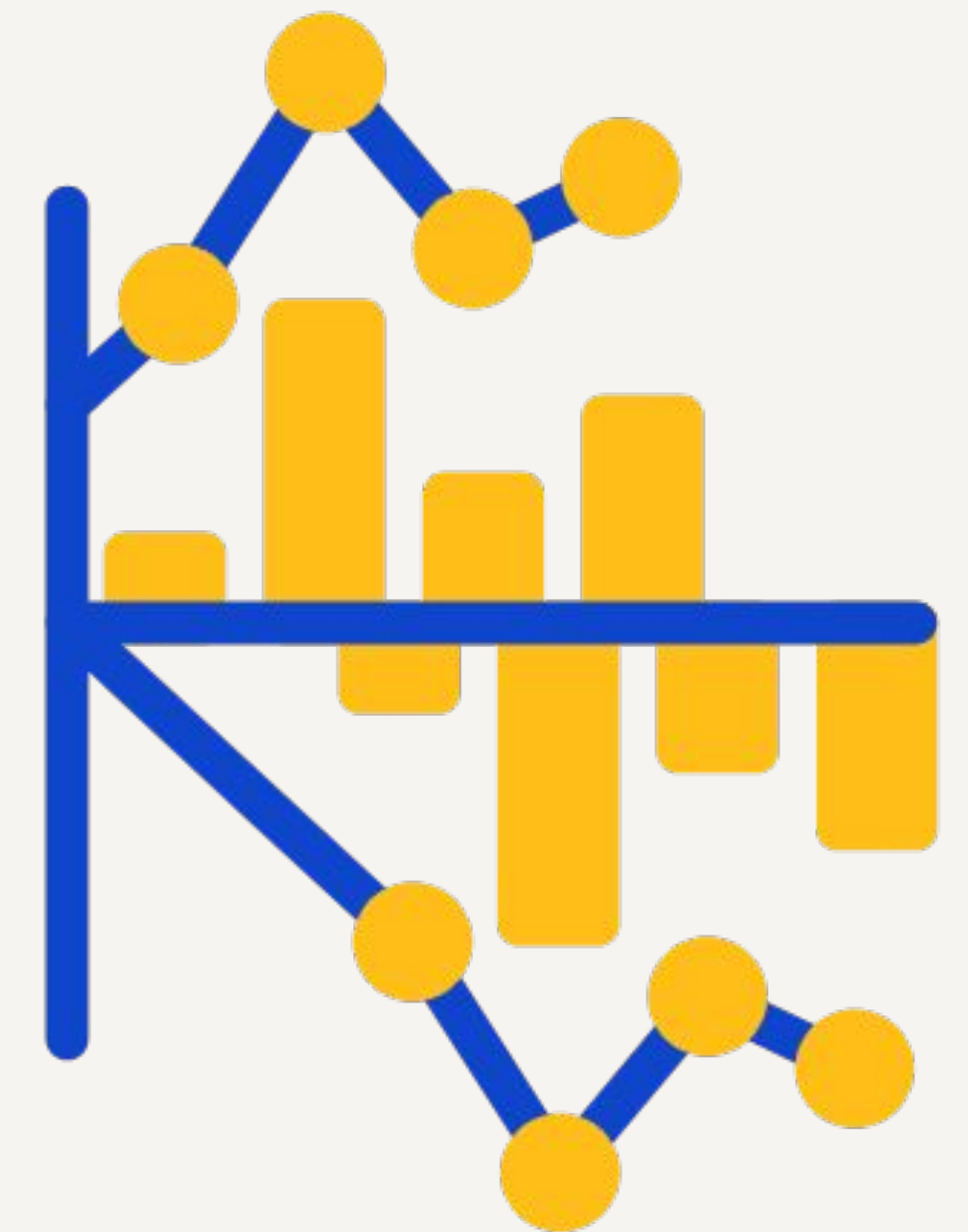
- **We reject $H_0 : \theta = 100$. Awful!**

- **Another questions:**

A. Do we have to iterate those methods to test our expected value?

B. What if we switch H_0 and H_1 ?

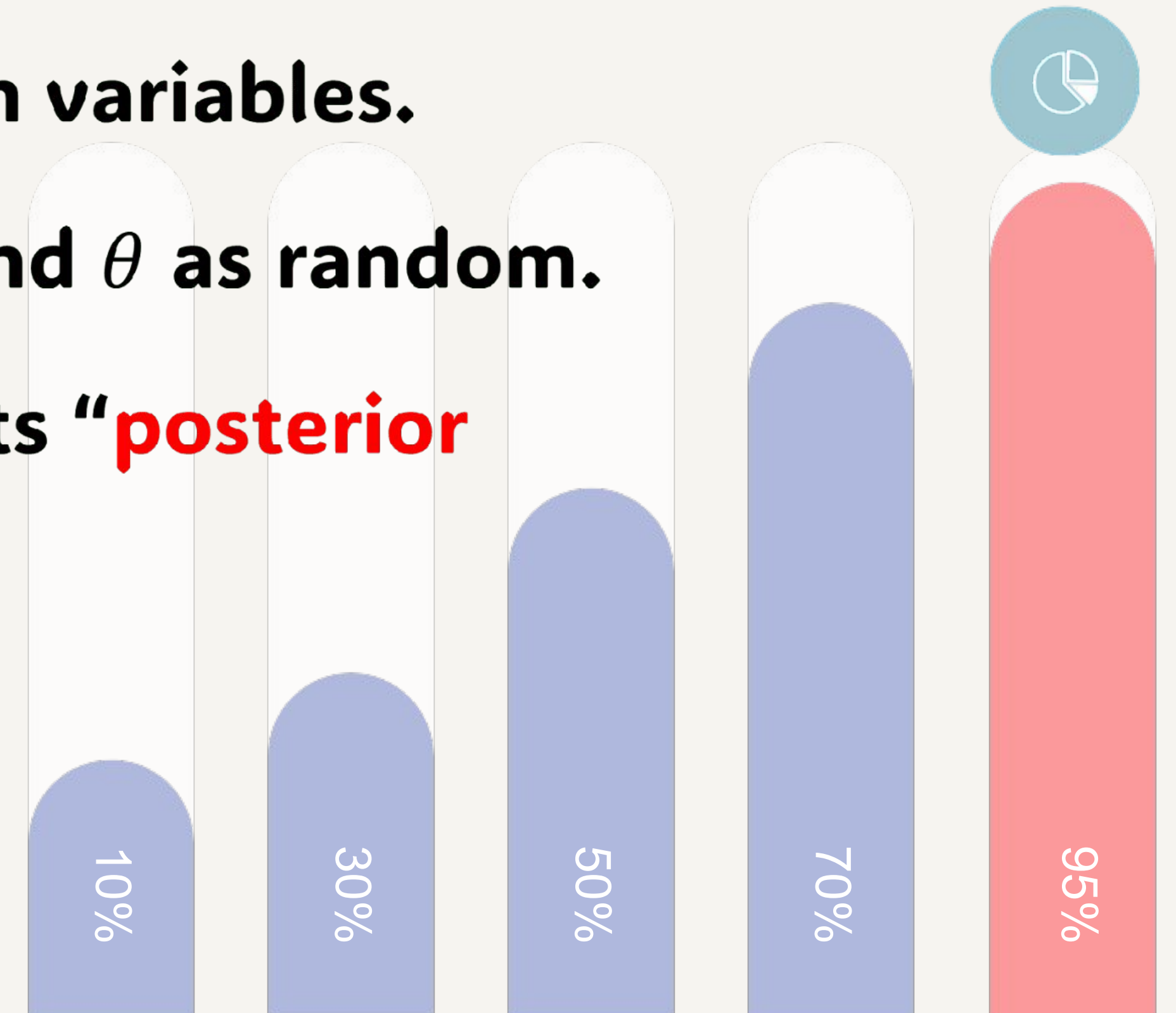
```
> result = Bootstrap(10000, 50, 0.05)
> print(paste("Bootstrap p-value:", result[1]))
[1] "Bootstrap p-value: 0"
> ifelse(result[2], "Result: p-value >= alpha. Fail to reject H0.",
+         "Result: p-value < alpha. Reject H0.")
[1] "Result: p-value < alpha. Reject H0."
```





Bayesian Statistics

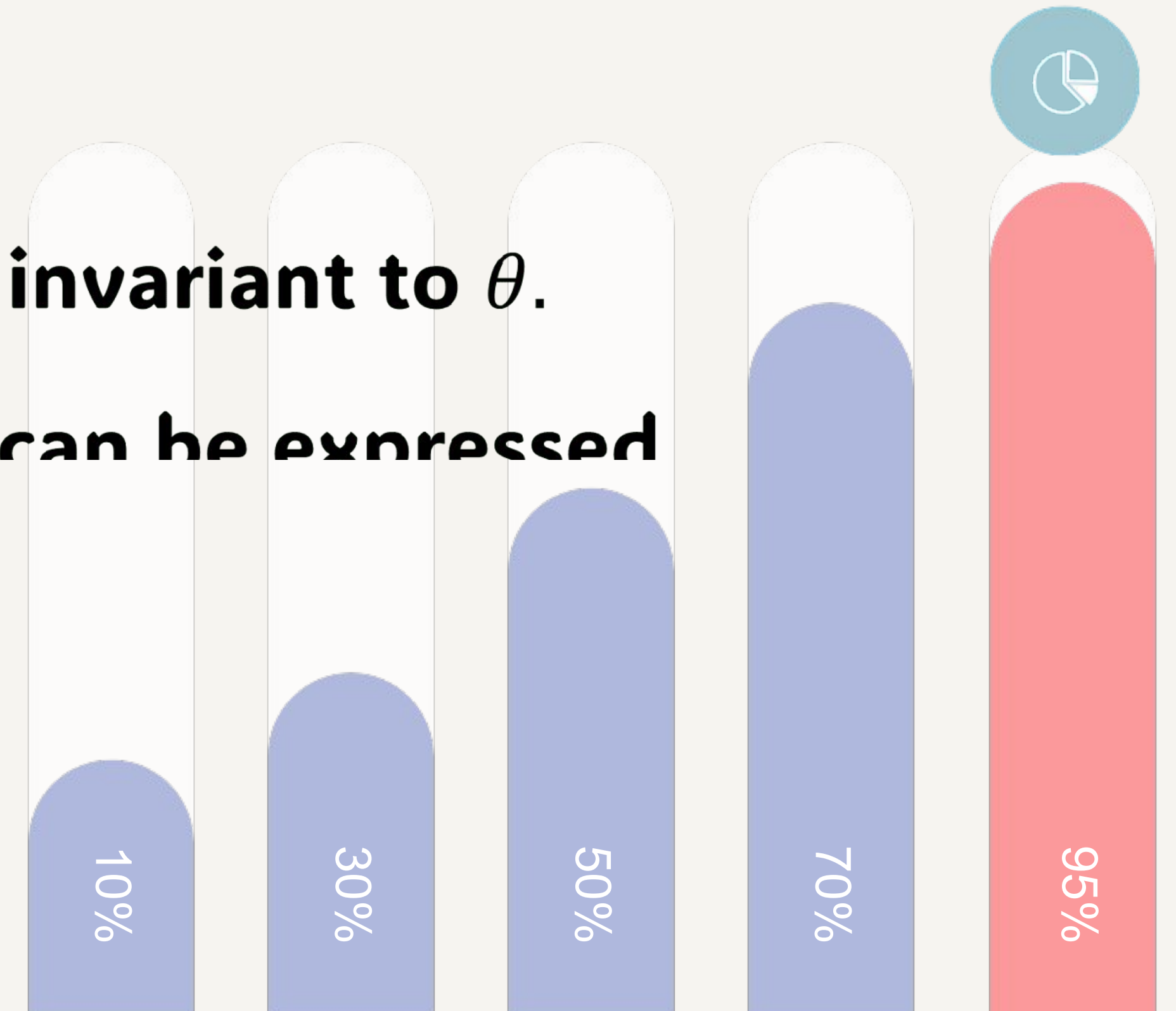
- Bayesians quantify uncertainty about fixed but unknown parameters, θ , by treating them as random variables.
- Thus, Bayesians view the data x as fixed and θ as random.
- Inference about θ is performed based on its “**posterior distribution**”, not an estimator.





Bayesian Statistics

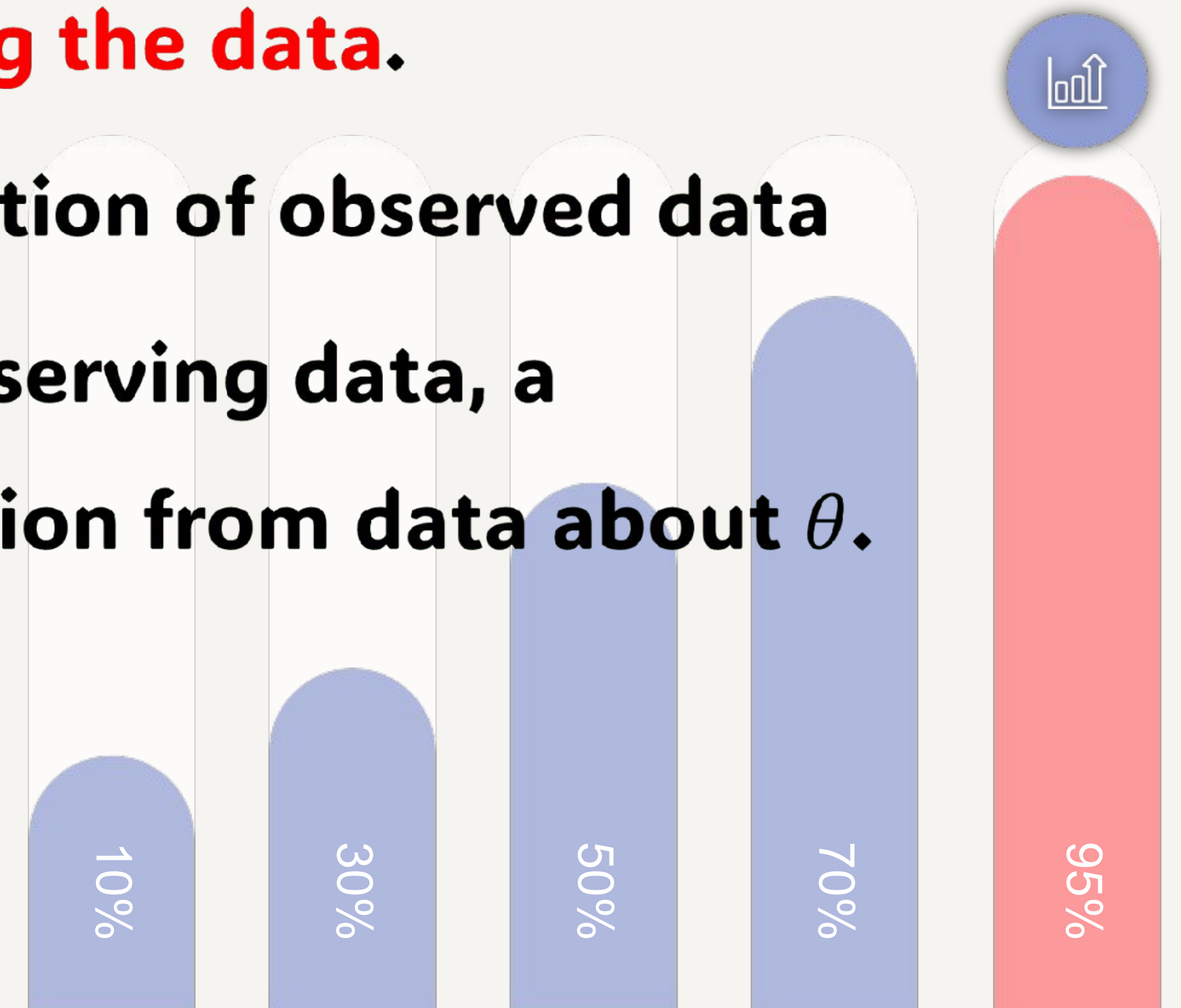
- In other words, we infer unknown parameter θ based on observed data x .
- And the distribution of x [$f(x|\theta)$] must be invariant to θ .
- Bayesians believe that every information can be expressed





Components of Bayesian Inference

- **Prior distribution:** $\theta \sim \pi(\theta)$, a distribution that represents the analyst's beliefs about θ **before observing the data.**
- **Model or Likelihood:** $x \sim f(x|\theta)$, a distribution of observed data
- **Posterior distribution:** $\theta \sim \pi(\theta|x)$, after observing data, a distribution that represents the information from data about θ .





Bayes' Theorem

- **We know this theorem: If $B_1, \dots, B_n$ are mutually exclusive and the partition of probability space S , and $P(B_i) > 0$ for all i , then for the event A such that $P(A) > 0$,**

$$P(B_k|A) = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^n P(B_i)P(A|B_i)}$$



Bayes' Theorem

- **We can use that for calculating posterior distribution.**

- $\pi(\theta|x) = \frac{\pi(\theta) \cdot f(x|\theta)}{m(x)}$, **where** $m(x) = \int \pi(\theta) \cdot f(x|\theta) d\theta$.

- **However, we do not need exact form of pdf when computing.**

$$\pi(\theta|x) \propto \pi(\theta) \cdot f(x|\theta)$$

- **From posterior, we can compute:**

$$\mathbb{E}(\theta|x) = \int \theta \pi(\theta|x) d\theta, \text{Var}(\theta|x) = \int \theta^2 \pi(\theta|x) d\theta - \{\mathbb{E}(\theta|x)\}^2$$



Example

Suppose we repeat Bernoulli trial, $X_i|\theta \sim iid Ber(\theta)$, and observe $(x_1, x_2, \dots, x_{12}) = (1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1)$

Suppose the prior distribution is uniform distribution.

Find the posterior distribution $\pi(\theta|x)$ and posterior mean.

Sol) $\pi(\theta) = 1, \theta \in [0, 1]$,

$$L(\theta|x) = \prod_{i=1}^n f(x_i|\theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

$$\pi(\theta|x) \propto \pi(\theta) \cdot f(x|\theta) = 1 \cdot \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$

$$= \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$$



Example (Cont'd)

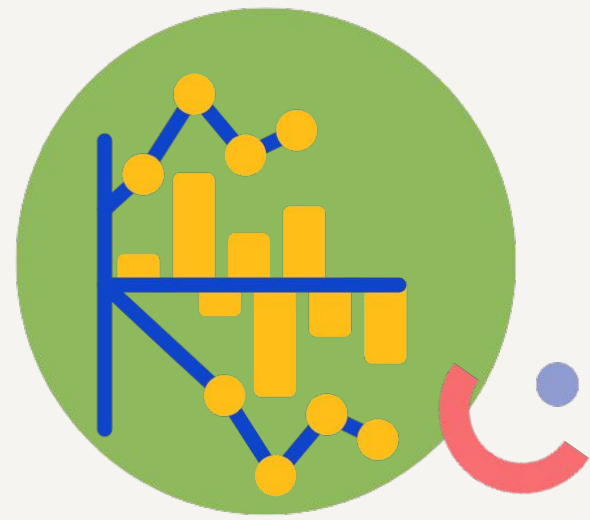
Suppose we repeat Bernoulli trial, $X_i|\theta \sim iid Ber(\theta)$, and observe $(x_1, x_2, \dots, x_{12}) = (1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1)$

Suppose the prior distribution is uniform distribution.

Find the posterior distribution $\pi(\theta|x)$ and posterior mean.

Sol) $\pi(\theta|x) \propto \pi(\theta) \cdot f(x|\theta) = 1 \cdot \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}$
 $= \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \leftarrow$ **Beta distribution**

$$(\otimes) \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \mathbb{E}(x) = \frac{\alpha}{\alpha+\beta}$$
$$\therefore \pi(\theta|x) = \text{Beta}\left(\sum_{i=1}^n x_i + 1, n - \sum_{i=1}^n x_i + 1\right) = \text{Beta}(10, 4), \mathbb{E}(\theta|x) = \frac{5}{7}$$



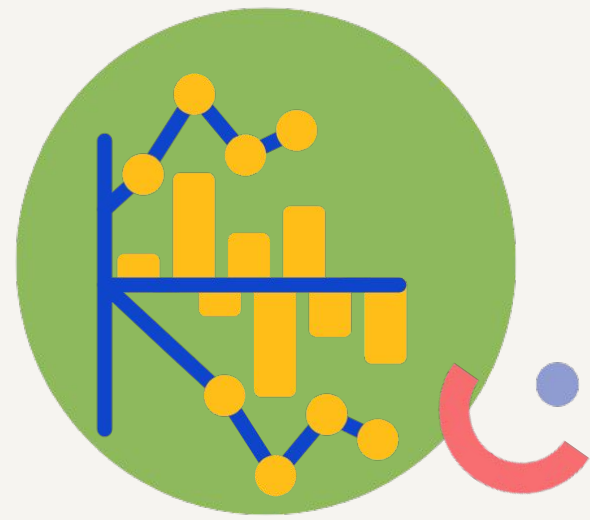
Likelihood principle



- The **likelihood principle** (LP) asserts that for inference on θ , all of the evidence about the parameter θ is contained in the likelihood function for the given data.
- If two experiments that yield equal (or proportional) likelihoods, i.e., $\exists c > 0$ such that, given the observed data,

$$L_1(\theta) = c \cdot L_2(\theta), \quad \forall \theta,$$

• . . . • . . . • . . . • . . .



Sufficiency principle

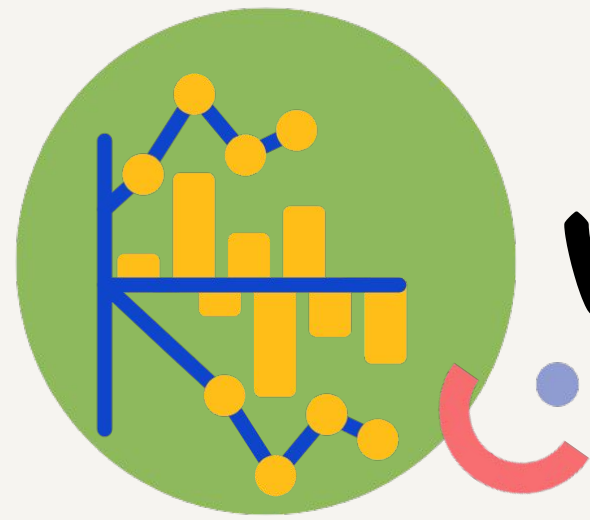


- **Factorization Theorem (FT): A statistic $T(X)$ is an SS iff**

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

for some non-negative functions g and h .

- **(Sufficiency Principle)** The information contained in X and $T(X)$



Weak conditionality principle

- Suppose we can perform one of two experiments, E_1 or E_2 , to collect data for estimating θ . Each experiment is selected with a probability of 0.5.
- **(Weak conditionality principle)** The actual inference about θ should depend only on the experiment E_j that is actually performed.
- Many statisticians consider this a desirable principle.
- Bayesian inference always satisfies this principle!

Birnbaum's proof(1962)



- Birnbaum showed that
Weak Conditionality Principle + Sufficiency Principle
 $\Leftrightarrow$ **Likelihood Principle**
- Note that Bayesian inference is based on the likelihood principle.
- This is a huge result because weak conditionality principle and sufficiency principle would seem appealing to most statisticians.



Bayesian Inference

- How can we do hypothesis test on Bayesian Inference?
- We will consider two hypothesis at the same time, not just assuming H_0 is true.

Step 1: Priors

$$\pi_0 = \pi(H_0) = \pi(\theta \in \Theta_0)$$
$$\pi_1 = \pi(H_1) = \pi(\theta \in \Theta_1)$$

- $\pi(H_0)$: the prior probability that H_0 is true.
- $\pi(H_1)$: the prior probability that H_1 is true.
- $\pi(H_0) + \pi(H_1) = 1$

- This process is depended on the analyzer's belief of the information of data before observing them.

Step 2: Priors

- Under each hypothesis, we use the following priors:

$$\pi(\theta|H_0) \text{ such that } \pi(\Theta_0|H_0) = 1,$$

$$\pi(\theta|H_1) \text{ such that } \pi(\Theta_1|H_1) = 1,$$

- By considering Step 1 and Step 2 in turn, it is equivalent to consider the following mixture prior:

$$\pi(\theta) \equiv \pi(H_0)\pi(\theta|H_0) + \pi(H_1)\pi(\theta|H_1)$$

Posterior

$$p_0 = \alpha_0 = \pi(H_0|x) = \pi(\theta \in \Theta_0 | x)$$

$$p_1 = \alpha_1 = \pi(H_1|x) = \pi(\theta \in \Theta_1 | x)$$

- $\pi(H_0|x)$: the posterior probability that H_0 is true.
- $\pi(H_1|x)$: the posterior probability that H_1 is true.
- Bayesian hypothesis testing: **support the hypothesis having larger posterior probability.**



Posterior

$$p_i = \alpha_i = \pi(H_i|x) = \pi(\theta \in \Theta_i | x), \quad i = 0,1$$

$g_i(\theta)d\theta$: a **density function under Θ_i**

$$\alpha_i \propto \pi_i \cdot f(x|H_i) \propto \pi_i \cdot m_i(x) = \pi_i \cdot \int_{\theta \in \Theta_i} f(x|\theta) g_i(\theta) d\theta$$

$$\alpha_1 = \frac{\pi_1 \cdot m_1(x)}{\pi_0 \cdot m_0(x) + \pi_1 \cdot m_1(x)}, \alpha_0 = 1 - \alpha_1$$



Prior odds, Posterior odds

- **Prior odds:** $\frac{\pi_1}{\pi_0}$
- **Posterior odds:** $\frac{\alpha_1}{\alpha_0} = \frac{\pi_1}{\pi_0} \times \frac{m_1(x)}{m_0(x)}$
- **Bayes Factor:** $B_{10} = \frac{m_1(x)}{m_0(x)}, B_{01} = \frac{1}{B_{10}}$
$$\alpha_1 = \frac{\pi_1 \cdot B_{10}}{\pi_0 + \pi_1 \cdot B_{10}}, \alpha_0 = 1 - \alpha_1$$



Jeffrey's criteria

		Not worth more than a bare mention
		Positive/Substantial
		Strong
		Very Strong/Decisive



Several Point Estimators

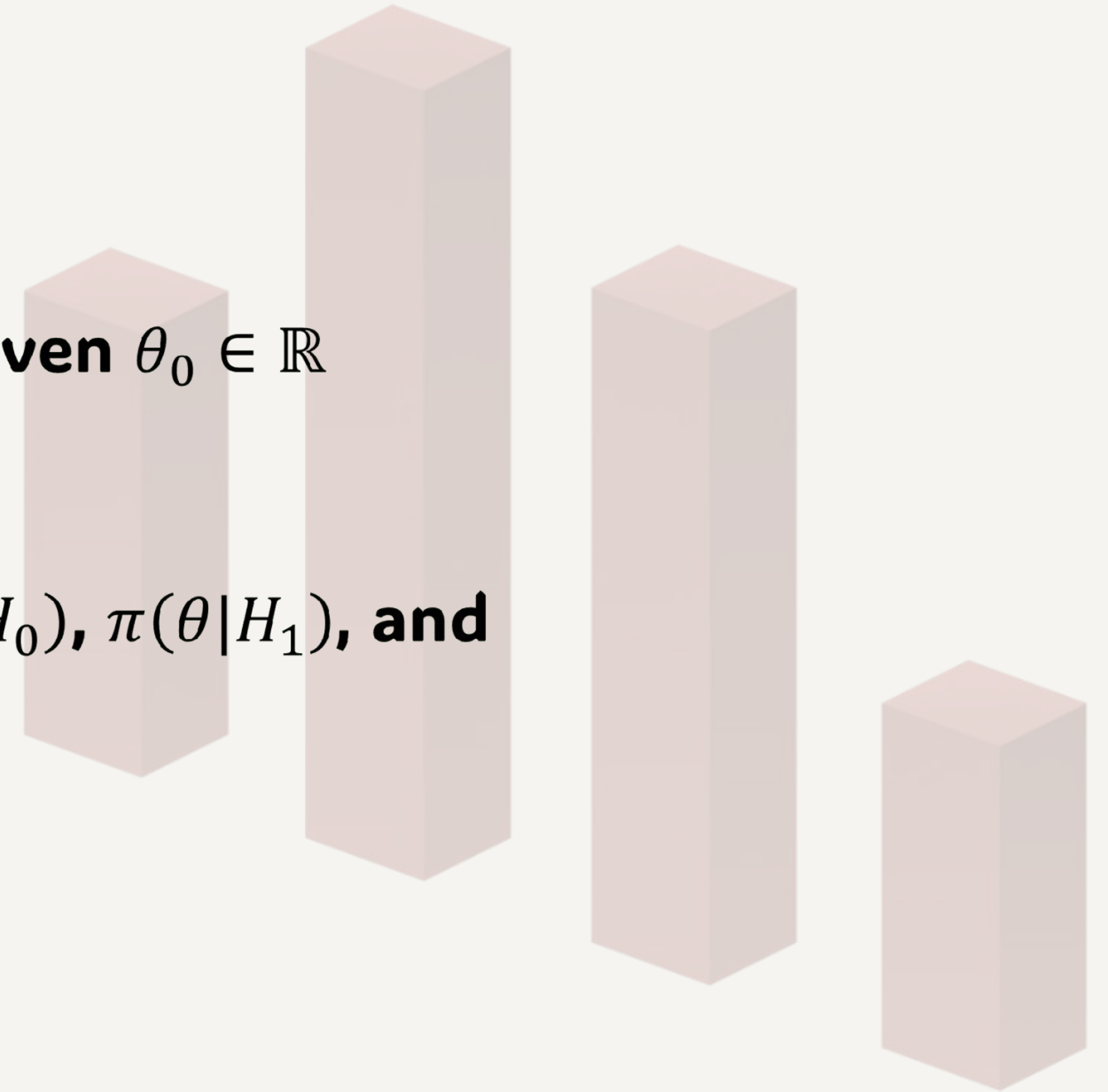
- $\hat{\theta}^{MLE} = \arg \max_{\theta} L(\theta|x)$, **where** $L(\theta|x) = \prod_{i=1}^n f(x_i|\theta)$ **or**
 $\hat{\theta}^{MLE} = \arg \max_{\theta} \mathcal{L}(\theta|x)$, **where** $\mathcal{L}(\theta|x) = \sum_{i=1}^n \log f(x_i|\theta)$
- $\hat{\theta}^B = \mathbb{E}(\theta|x)$ **(Bayesian Posterior Mean)**
- $\hat{\theta}^M = \pi_{0.5}(\theta|x)$
- <Maximum A Posteriori(MAP)>**
- $\hat{\theta}^{MAP} = \arg \max_{\theta} \pi(\theta|x)$ **[generalized MLE]**
- $\hat{\theta}^B = \arg \min_a \mathbb{E}(L(\theta, a)|x)$, **where** $L(\theta, a)$: **Loss function**

Two-sided Test

- **Suppose that** $X_1, \dots, X_n | \theta \sim iid N(\mu, \sigma^2)$.
- **We are interested in the hypothesis testing**

$$H_0 : \theta = \theta_0 \textbf{ vs } H_1 : \theta \neq \theta_0, \textbf{ for a given } \theta_0 \in \mathbb{R}$$

- **Then, we have** $\Theta_0 = \{\theta_0\}$ **and** $\Theta_1 = \mathbb{R} \setminus \{\theta_0\}$.
- **We need to determine** π_0, π_1 **and priors** $\pi(\theta|H_0), \pi(\theta|H_1)$, **and** $g_1(\theta)d\theta$, **a prior density function under** Θ_1 .



Two-sided Test

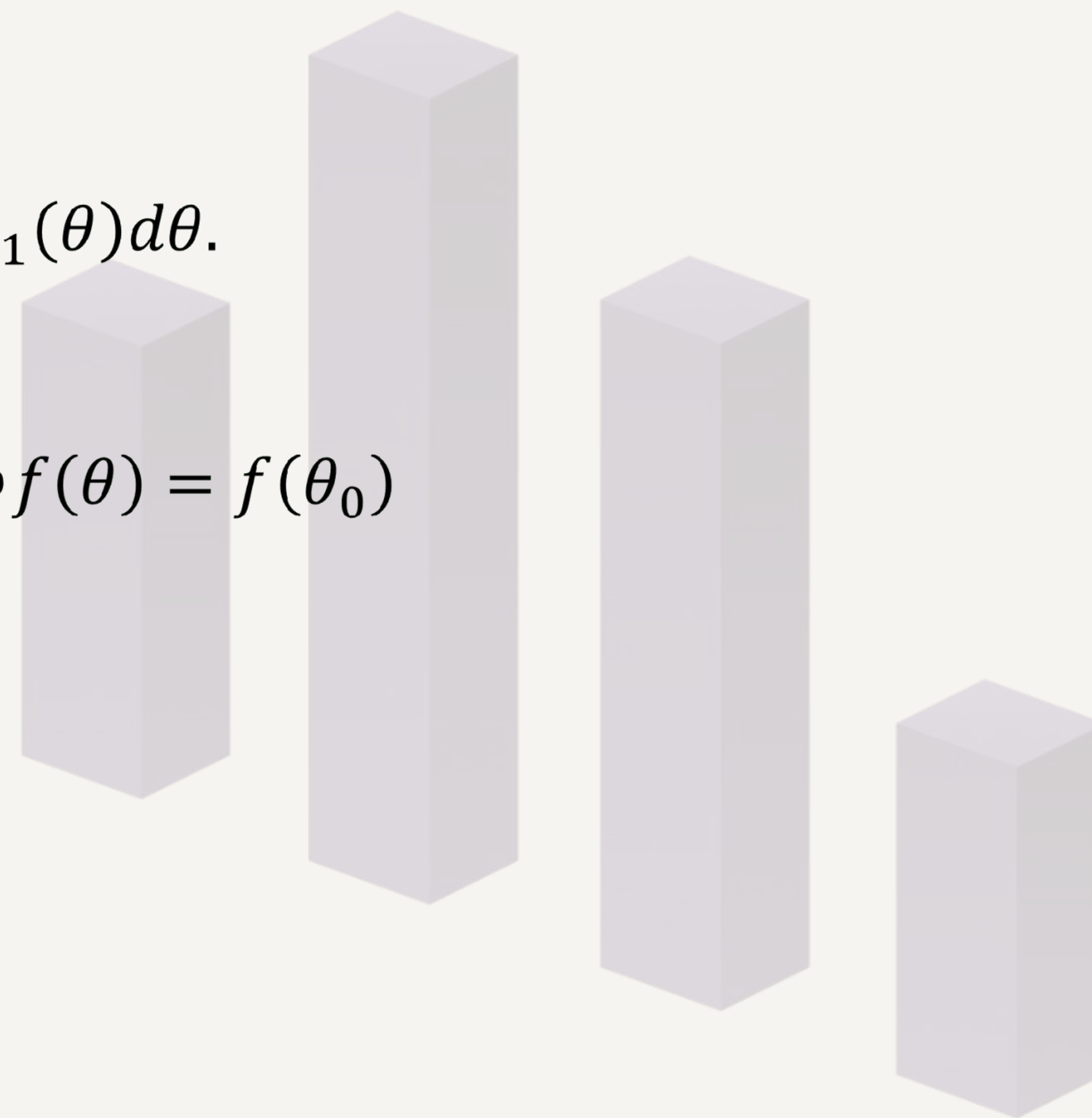
- Since $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \mathbb{R} \setminus \{\theta_0\}$,

$$\pi(\theta|H_0) = \delta_{\theta_0}(d\theta), \pi(\theta|H_1) = g_1(\theta)d\theta.$$

$$\ast \int_A \delta_{\theta_0}(d\theta) = \begin{cases} 1, & \theta_0 \in A \\ 0, & \text{o.w.} \end{cases}, \int f(\theta)\delta_{\theta_0}(d\theta) = \mathbb{E}^{\delta_{\theta_0}} f(\theta) = f(\theta_0)$$

- $\delta_{\theta_0}(d\theta)$: **Dirac measure**

- $\pi(d\theta) = \pi_0\delta_{\theta_0}(d\theta) + \pi_1g_1(\theta)d\theta$



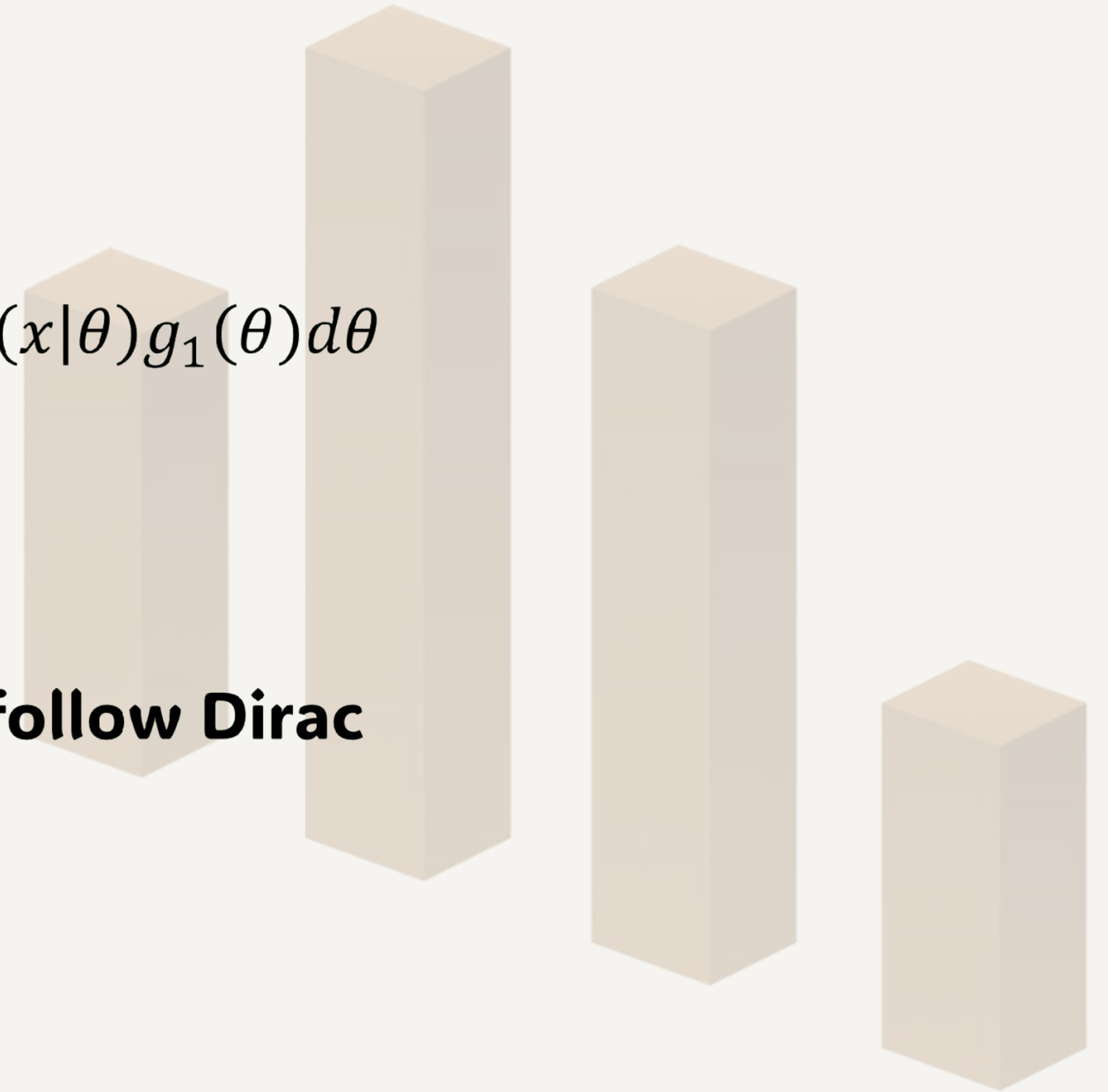
Two-sided Test

- **Bayes factor** $B_{10} = \frac{m_1(x)}{f(x|\theta_0)}$,

$$m_1(x) = \int_{\mathbb{R} \setminus \{\theta_0\}} f(x|\theta) g_1(\theta) d\theta = \int_{\mathbb{R}} f(x|\theta) g_1(\theta) d\theta$$

- $\alpha_1(x) = \frac{\pi_1 \cdot B_{10}}{\pi_0 + \pi_1 \cdot B_{10}}, \alpha_0(x) = 1 - \alpha_1(x)$

- **For one-sided test, it is much more simple, follow Dirac distribution.**



Credible Interval

- **Def)** For $\pi(\theta|x)$, θ 's $100(1 - \alpha)\%$ **credible set** is $C(x) \subseteq \Theta$ **iff**

$$\exists C(x) \text{ such that } P(\theta \in C(x)|x) \geq 1 - \alpha$$

- **How to find?**

1. Equal tail credible set

$$[A, B] \text{ such that } \pi(\theta \leq A|x) = \frac{\alpha}{2} = \pi(\theta \geq B|x)$$

2. Highest Posterior Density (HPD) credible set

For $\alpha \in [0,1]$, **if a set** $C_k = \{\theta \in \Theta: \pi(\theta|x) \geq k\}, k \geq 0$ **satisfies**



Example (Bayesian Statistics)

max: 85, mean = 70, my score = 55, otherwise.

- We have 7 observed data from population as follows:

70,45,35,60,85,55,55

- We want to test

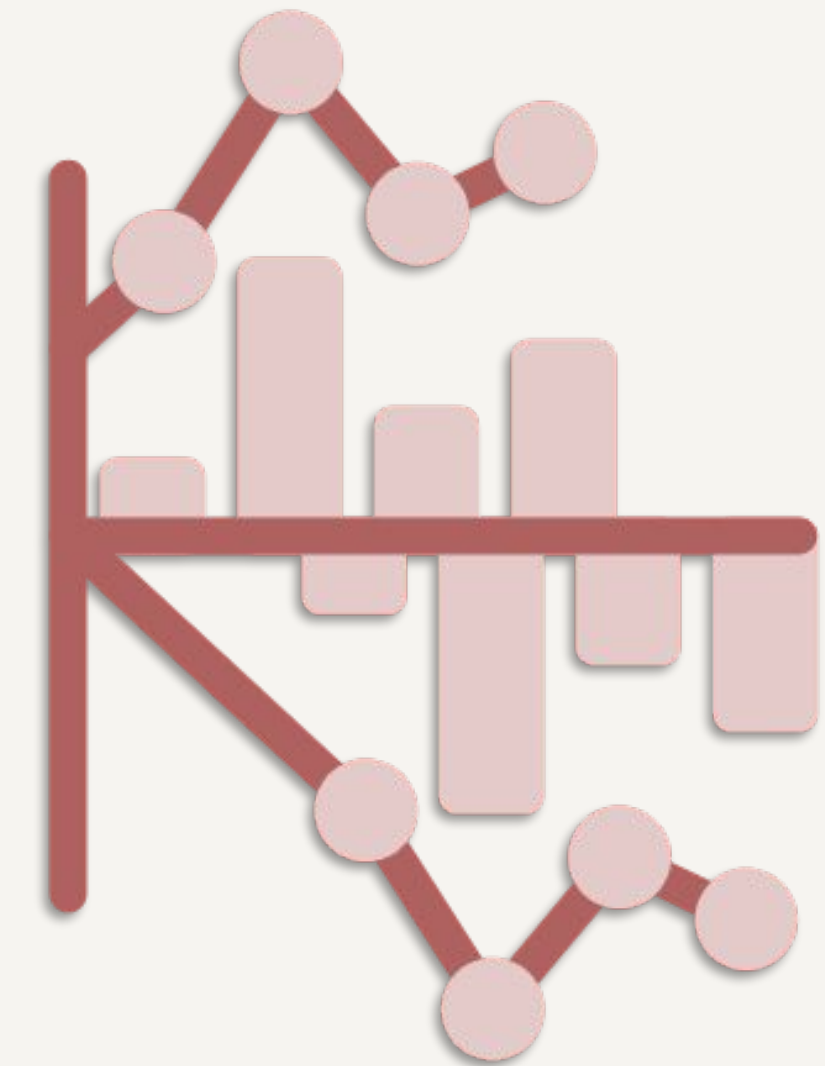
$H_0 : x_1, \dots, x_7 \sim iid N(70, 100)$ **vs** $H_1 : x_1, \dots, x_7 \sim iid N(70, 200)$.

- Suppose that we have $\pi_0 = \pi_1 = 0.5$.

Sol) From observed data, we have $s^2 = 2625, n = 7$.

Calculate likelihood function under each hypothesis!

$$L(\sigma^2 | x) = \prod_{i=1}^n f(x_i | \sigma^2)$$



Example (Bayesian Statistics)

max: 85, mean = 70, my score = 55, otherwise.

- We have 7 observed data from population as follows:

70, 45, 35, 60, 85, 55, 55

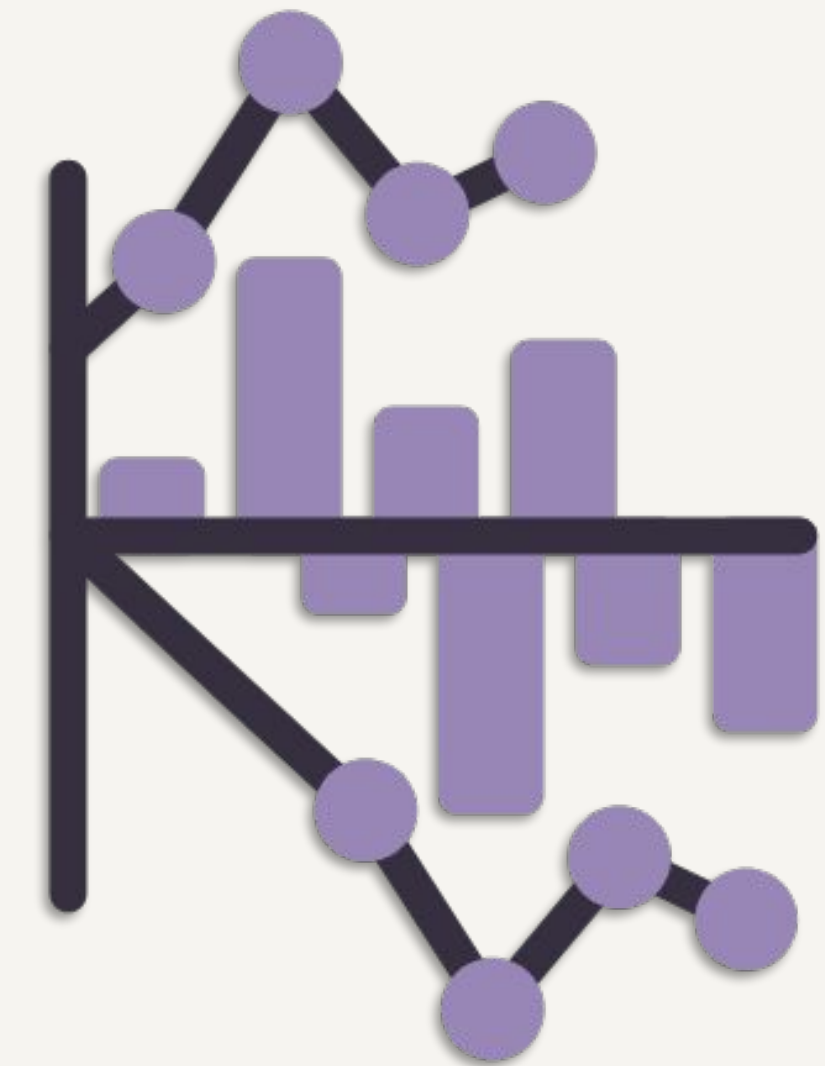
- We want to test

$H_0 : x_1, \dots, x_7 \sim iid N(70, 100)$ **vs** $H_1 : x_1, \dots, x_7 \sim iid N(70, 200)$.

- Suppose that we have $\pi_0 = \pi_1 = 0.5$.

Sol) Under H_0 , $n = 7, \mu = 70, \sigma^2 = 100$

$$\begin{aligned} L(\sigma_0^2 | x) &= \prod_{i=1}^n f(x_i | \sigma_0^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_0^2}\right) \\ &= (200\pi)^{-3.5} \exp(-13.125) \end{aligned}$$



Example (Bayesian Statistics)

max: 85, mean = 70, my score = 55, otherwise.

- We have 7 observed data from population as follows:

70, 45, 35, 60, 85, 55, 55

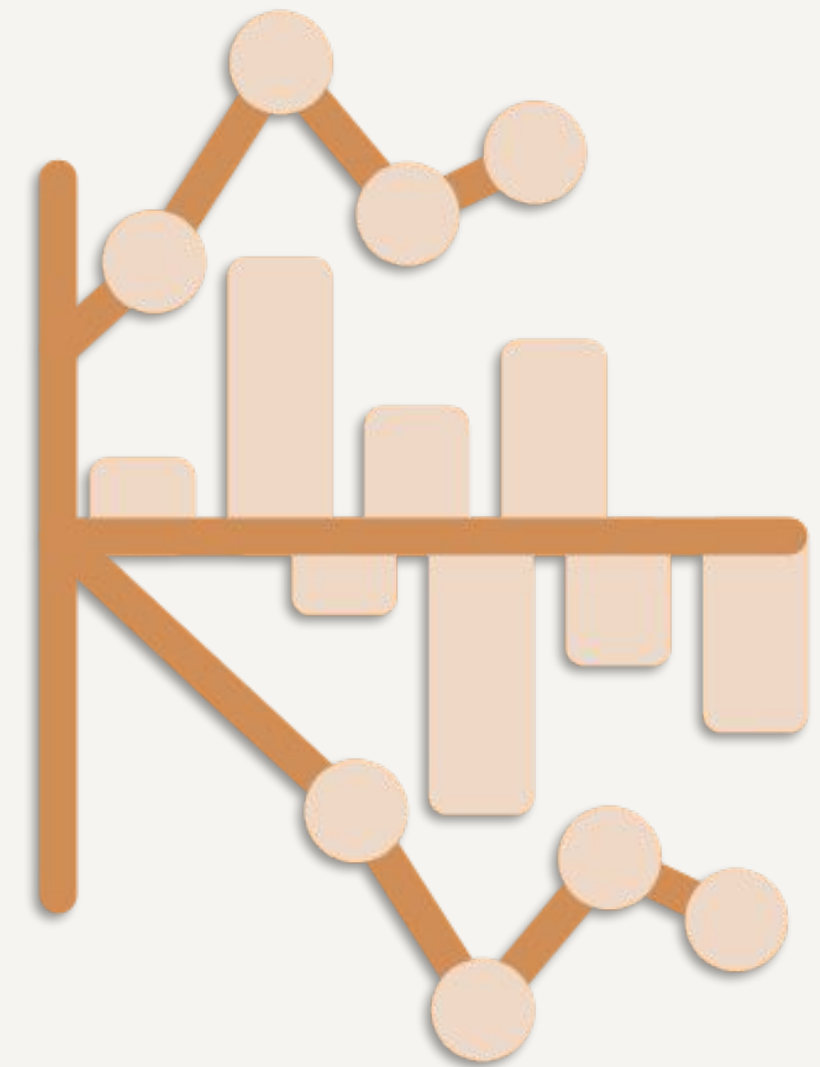
- We want to test

$H_0 : x_1, \dots, x_7 \sim iid N(70, 100)$ **vs** $H_1 : x_1, \dots, x_7 \sim iid N(70, 200)$.

- Suppose that we have $\pi_0 = \pi_1 = 0.5$.

Sol) Under H_1 , $n = 7, \mu = 70, \sigma^2 = 200$

$$\begin{aligned} L(\sigma_1^2 | x) &= \prod_{i=1}^n f(x_i | \sigma_1^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_1^2}\right) \\ &= (400\pi)^{-3.5} \exp(-6.5625) \end{aligned}$$



Example (Bayesian Statistics)

max: 85, mean = 70, my score = 55, otherwise.

- We have 7 observed data from population as follows:

70, 45, 35, 60, 85, 55, 55

- We want to test

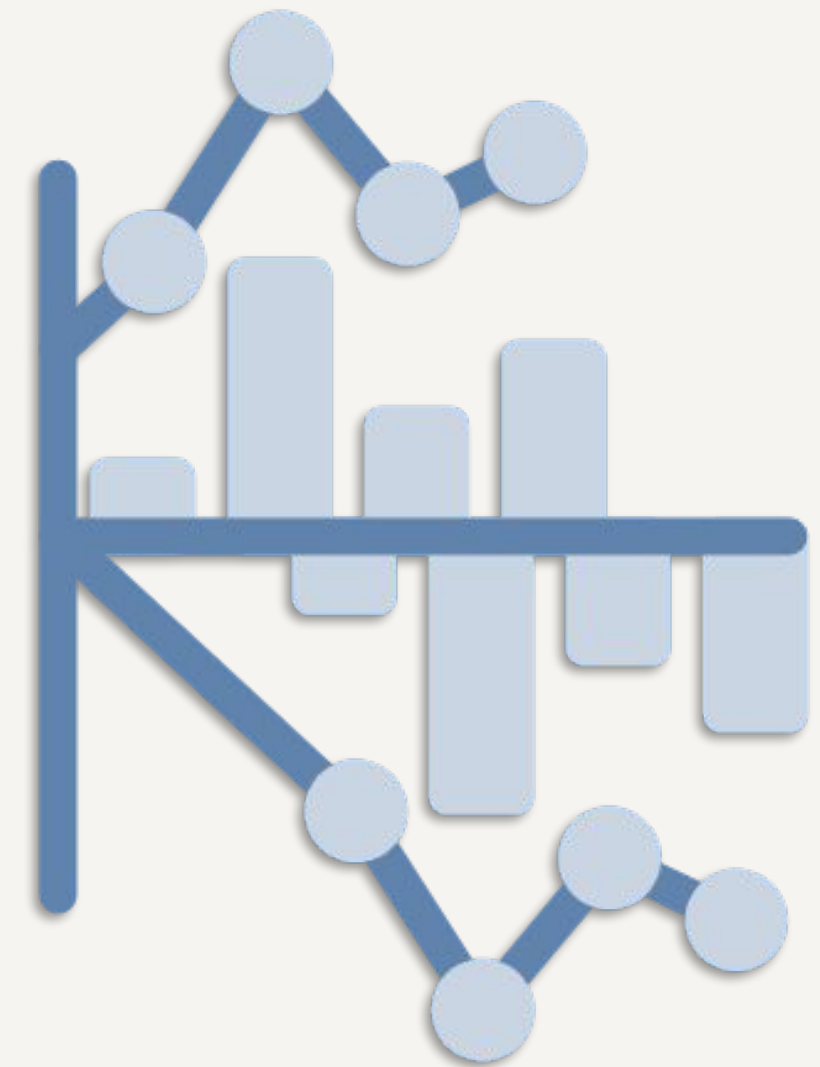
$H_0 : x_1, \dots, x_7 \sim iid N(70, 100)$ vs $H_1 : x_1, \dots, x_7 \sim iid N(70, 200)$.

- Suppose that we have $\pi_0 = \pi_1 = 0.5$.

Sol) Calculate Bayes Factor $B_{10} = \frac{L(\sigma_1^2 | x)}{L(\sigma_0^2 | x)} \approx 62.58$.

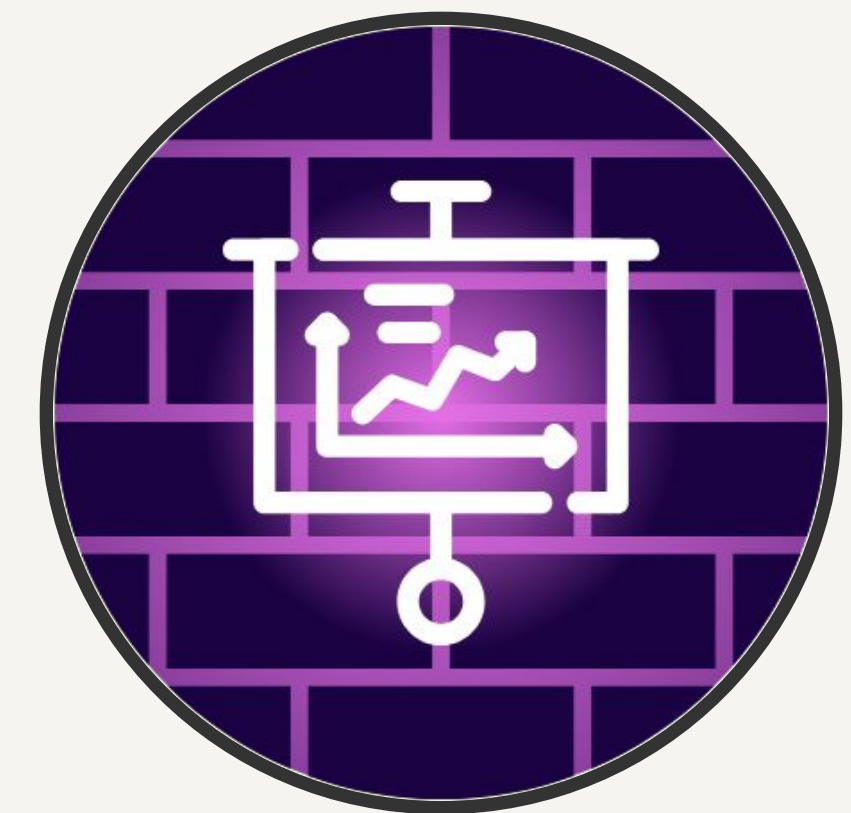
Note that $\frac{\alpha_1}{\alpha_0} = \frac{\pi_1}{\pi_0} \times \frac{m_1(x)}{m_0(x)}$. Then posterior odds is about 62.58.

By Jeffrey's criteria, H_1 is **strong**. ($\alpha_1 = \frac{62.58}{1+62.58} \approx 0.9843$, $\alpha_0 =$



Several priors

- $\pi(\mathbb{R}) = 1$ [Improper prior]
- Jeffrey's prior
- $\pi(\sigma)d\sigma \propto \frac{1}{\sigma} d\sigma$ **[noninformative prior]**
- $\pi^J(\theta) \propto |\det I(\theta)|^{0.5}$, where $I(\theta) = \text{Var}_\theta(\frac{d}{d\theta} \log f(x|\theta))$
- We will not go deep about priors.



Normal model

- Consider we do not know $\mu \in \mathbb{R}, \sigma^2 > 0$ of $N(\mu, \sigma^2)$.
- Priors are given:

$$\mu | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right), \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2} \sigma_0^2\right)$$

- Compute posterior distribution. (Just enjoy the process)

$$\pi(\mu, \sigma^2 | \mathbf{x}) \propto \pi(\mu | \sigma^2, \mathbf{x}) \cdot \pi(\sigma^2 | \mathbf{x}) \cdot \prod_{i=1}^n f(x_i | \mu, \sigma^2)$$

$$= \left(\frac{\kappa_0}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{1}{2} \cdot \frac{\kappa_0}{\sigma^2} (\mu - \mu_0)^2\right) \times \frac{\left(\frac{\nu_0}{2} \sigma_0^2\right)^{\left(\frac{\nu_0}{2}\right)}}{\Gamma\left(\frac{\nu_0}{2}\right)} (\sigma^2)^{\left(-\frac{\nu_0}{2}-1\right)} \exp\left(-\frac{\nu_0}{2} \frac{\sigma_0^2}{\sigma^2}\right)$$

Normal model (Cont'd)

$$= \left(\frac{\kappa_0}{2\pi\sigma^2} \right)^{1/2} \exp \left(-\frac{1}{2} \cdot \frac{\kappa_0}{\sigma^2} (\mu - \mu_0)^2 \right) \times \frac{\left(\frac{\nu_0}{2} \sigma_0^2 \right)^{\left(\frac{\nu_0}{2} \right)}}{\Gamma \left(\frac{\nu_0}{2} \right)} (\sigma^2)^{\left(-\frac{\nu_0}{2} - 1 \right)} \exp \left(-\frac{\nu_0}{2} \frac{\sigma_0^2}{\sigma^2} \right) \\ \times \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp \left(-\frac{1}{2\sigma^2} ((n-1)s^2 + n(\bar{x} - \mu)^2) \right)$$

$$\propto (\sigma^2)^{-\frac{n+\nu_0+2}{2}} \exp \left(-\frac{1}{2} \cdot \frac{\kappa_0}{\sigma^2} (\mu - \mu_0)^2 - \frac{\nu_0}{2} \frac{\sigma_0^2}{\sigma^2} - \frac{1}{2\sigma^2} ((n-1)s^2 + n(\bar{x} - \mu)^2) \right)$$

$$= (\sigma^2)^{-\frac{n+\nu_0+2}{2}} \exp \left(-\frac{1}{2\sigma^2} (\kappa_0(\mu - \mu_0)^2 + \nu_0\sigma_0^2 + (n-1)s^2 + n(\bar{x} - \mu)^2) \right)$$

$$= (\sigma^2)^{-\frac{n+\nu_0+2}{2}} \exp \left(-\frac{1}{2\sigma^2} ((\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n} \right)^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{x} - \mu_0)^2 + (n-1)s^2 \right)$$

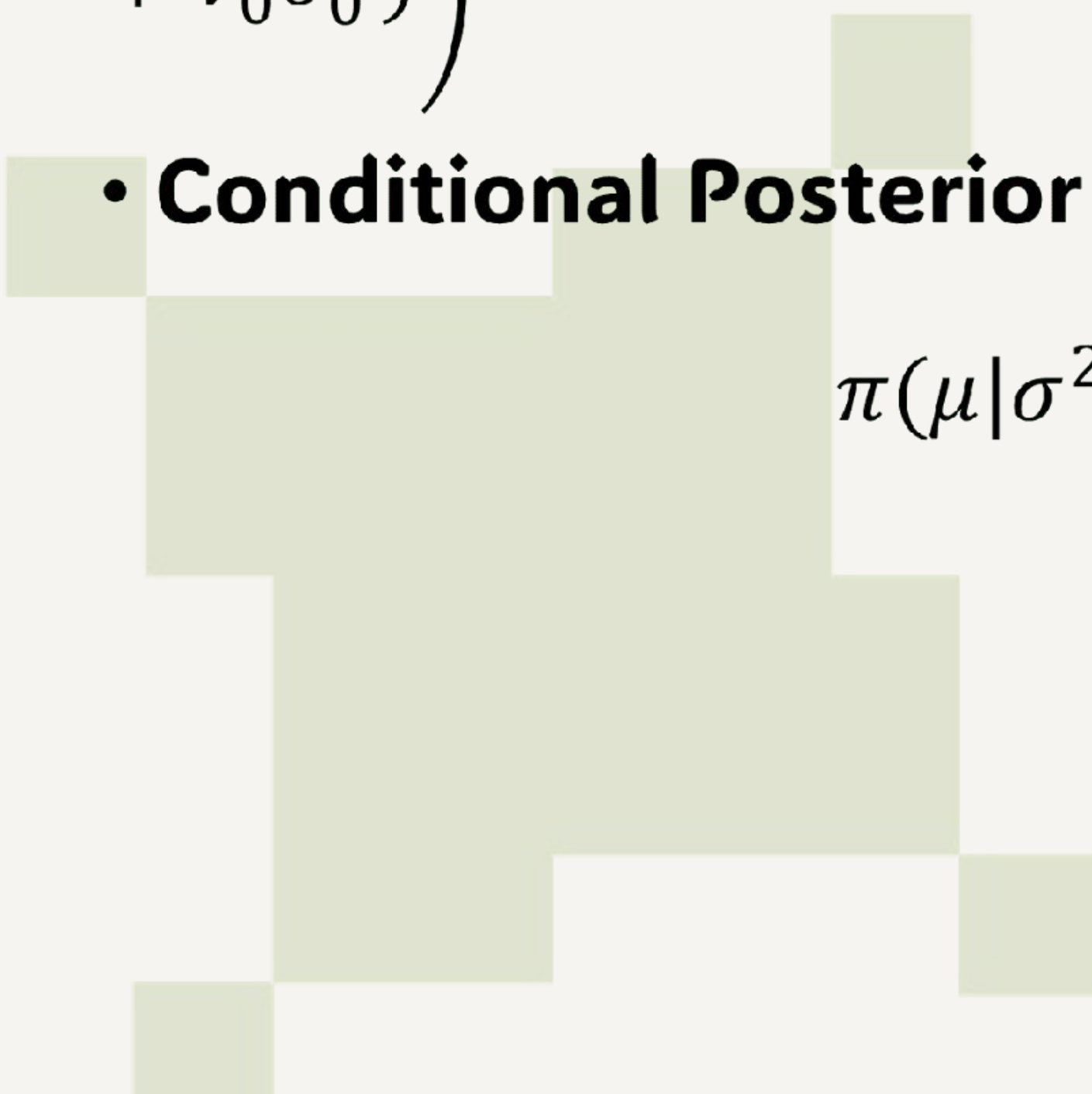
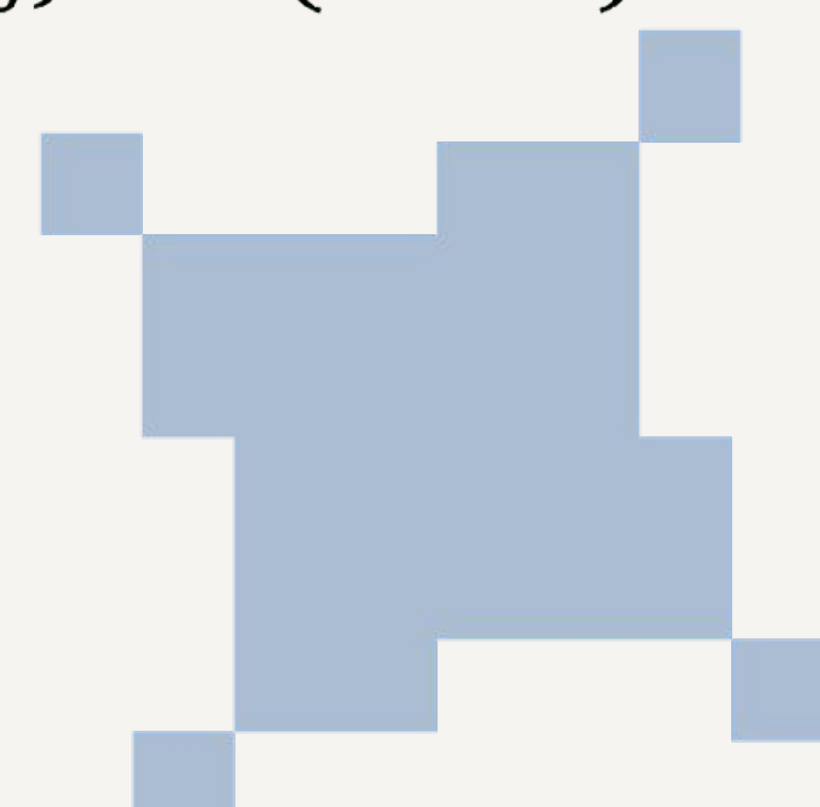
Normal model (Cont'd)

$$\pi(\mu, \sigma^2 | \mathbf{x})$$

$$\propto (\sigma^2)^{-\frac{n+\nu_0+2}{2}} \exp \left(-\frac{1}{2\sigma^2} \left((\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0+n} \right)^2 + \frac{\kappa_0 n}{\kappa_0+n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2 \right) \right)$$

- **Conditional Posterior distributions**

$$\pi(\mu | \sigma^2, \mathbf{x}) \propto \exp \left(-\frac{1}{2\sigma^2} (\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0+n} \right)^2 \right)$$



Normal model (Cont'd)

$$\pi(\mu, \sigma^2 | \mathbf{x})$$

$$\propto (\sigma^2)^{-\frac{n+\nu_0+2}{2}} \exp \left(-\frac{1}{2\sigma^2} \left((\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n} \right)^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2 \right) \right)$$

- **Conditional Posterior distributions**

$$\pi(\sigma^2 | \mu, \mathbf{x})$$

$$= Inv$$

$$= Gamma \left(\frac{n + \nu_0 + 1}{2}, \left(-\frac{1}{2} \left((\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n} \right)^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2 \right) \right) \right)$$

Normal model (Cont'd)

- **Marginal Posterior distributions**

$$\pi(\sigma^2 | \mathbf{x}) = \int \pi(\mu, \sigma^2 | \mathbf{x}) d\mu \propto \int (\sigma^2)^{-\frac{n+\nu_0+2}{2}} \exp \left(-\frac{1}{2\sigma^2} \left((\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0+n\bar{x}}{\kappa_0+n} \right)^2 + \frac{\kappa_0 n}{\kappa_0+n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2 \right) \right) d\mu$$

$$= (\sigma^2)^{-\frac{n+\nu_0+1}{2}} \exp \left(-\frac{1}{2\sigma^2} \left(\frac{\kappa_0 n}{\kappa_0+n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2 \right) \right) \times$$

$$\int (\sigma^2)^{-\frac{1}{2}} \exp \left(-\frac{1}{2\sigma^2} (\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0+n\bar{x}}{\kappa_0+n} \right)^2 \right) d\mu$$

$$\propto (\sigma^2)^{-\frac{n+\nu_0}{2}-1} \exp \left(-\frac{1}{2\sigma^2} \left(\frac{\kappa_0 n}{\kappa_0+n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2 \right) \right)$$

$$\pi(\sigma^2 | \mathbf{x}) = \text{Inv-Gamma} \left(\frac{n+\nu_0}{2}, \frac{1}{2} \left(\frac{\kappa_0 n}{\kappa_0+n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2 \right) \right)$$

Normal model (Cont'd)

- **Marginal Posterior distributions**

$$\pi(\mu|\mathbf{x}) = \int \pi(\mu, \sigma^2|\mathbf{x}) d\sigma^2 \propto \int (\sigma^2)^{-\frac{n+\nu_0+2}{2}} \exp\left(-\frac{1}{2\sigma^2} \left((\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0+n\bar{x}}{\kappa_0+n}\right)^2 + \frac{\kappa_0 n}{\kappa_0+n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2\right)\right) d\sigma^2$$

$$= \frac{\Gamma\left(\frac{\nu_0+n+1}{2}\right)}{\left(\frac{1}{2} \left((\kappa_0+n) \left(\mu - \frac{\kappa_0\mu_0+n\bar{x}}{\kappa_0+n}\right)^2 + \frac{\kappa_0 n}{\kappa_0+n} (\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2\right)\right)^{\frac{\nu_0+n+1}{2}}}$$

Example (Bayesian Statistics)

max: 85, mean = 70, my score = 55, otherwise.

- We have 7 observed data from population as follows:

70, 45, 35, 60, 85, 55, 55

- We want to find the 95% credible set for variance σ^2 .
- Suppose that $\pi(\sigma^2)d\sigma^2 \propto \frac{1}{\sigma^2} d\sigma^2$ [noninformative prior]
- Suppose that data follow $N(70, \sigma^2)$.
- Sol) We need to compute posterior distribution.
- $\pi(\sigma^2|x) \propto \pi(\sigma^2) \cdot L(\sigma^2|x)$

Example (Bayesian Statistics)

- $\pi(\sigma^2|x) \propto \pi(\sigma^2) \cdot L(\sigma^2|x)$

$$L(\sigma^2|x) = \prod_{i=1}^7 f(x_i|\sigma^2) = \prod_{i=1}^7 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - 70)^2}{2\sigma^2}\right)$$

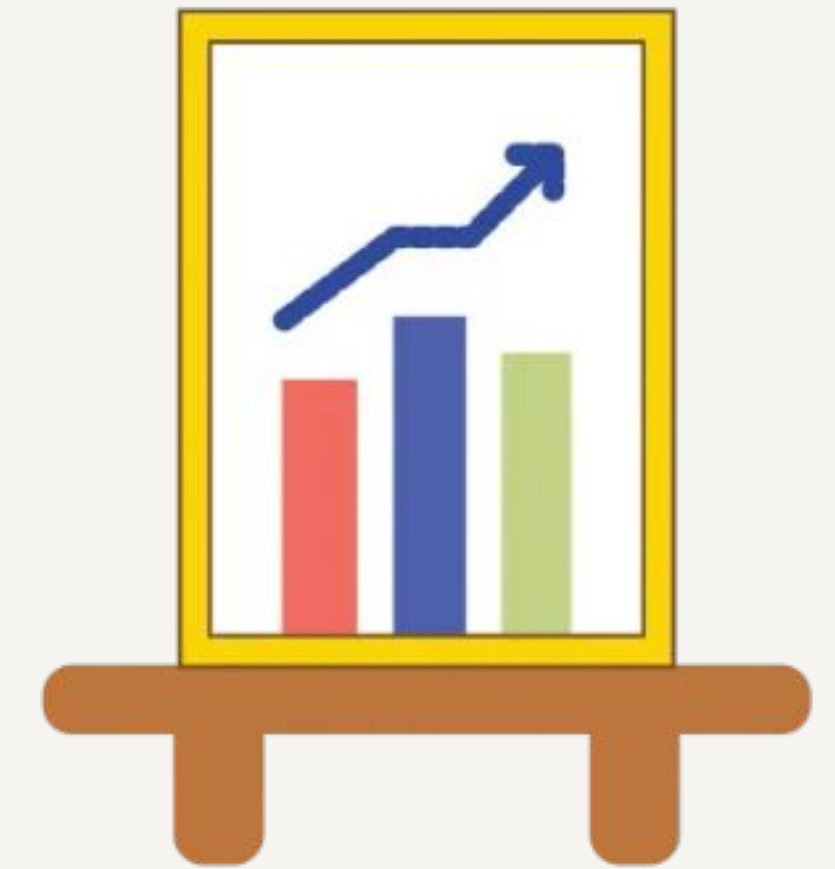
$$\propto (\sigma^2)^{-\frac{7}{2}} \exp\left(-\frac{2625}{2\sigma^2}\right)$$

- $\pi(\sigma^2)d\sigma^2 \propto \frac{1}{\sigma^2} d\sigma^2$

- **Then,**

$$\pi(\sigma^2|x) \propto (\sigma^2)^{-\frac{9}{2}} \exp\left(-\frac{2625}{2\sigma^2}\right)$$

- **Thus** $\pi(\sigma^2|x) = \text{Inv-Gamma}\left(\frac{7}{2}, \frac{2625}{2}\right)$



Example (Bayesian Statistics)

- Then, $\sigma^2 | X \sim \text{Inv-Gamma} \left(\frac{7}{2}, \frac{2625}{2} \right) \Leftrightarrow \frac{1}{\sigma^2} | X \sim \text{Gamma} \left(\frac{7}{2}, \frac{2625}{2} \right)$

- Note that for $\alpha, \beta > 0$,

$$X \sim \text{Gamma}(\alpha, \beta) \Rightarrow 2\beta X \sim \chi_{2\alpha}^2$$

- Then,

$$\frac{2625}{\sigma^2} | X \sim \chi_7^2$$

- Then, $P \left(\chi_{7,0.025}^2 \leq \frac{2625}{\sigma^2} \leq \chi_{7,0.975}^2 \right) = 0.05$



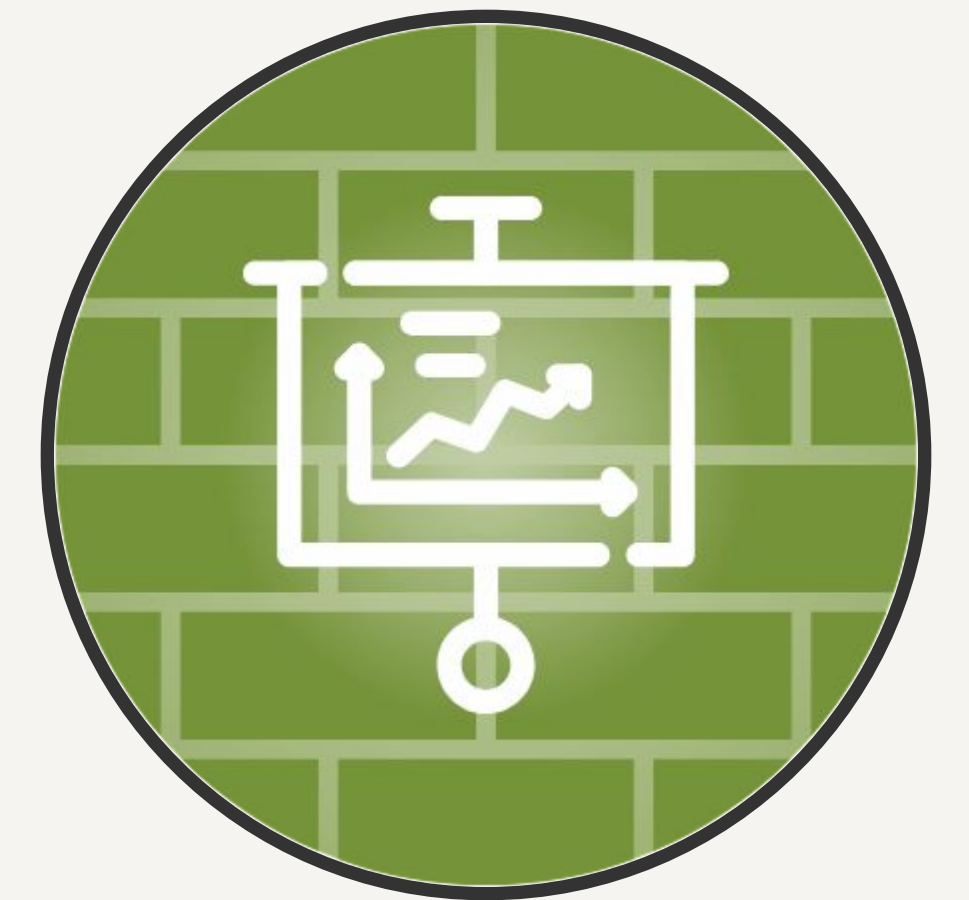
Problem

- Q) What if we do not use Normal model?
- A) We cannot acknowledge what the posterior distribution is. Also, we might not be able to compute the posterior distribution.
- Remind that even using Normal model, we got *Normal-Inv-gamma distribution*.
- To deal with this computation problem, we need new method!



Markov Chain Monte Carlo

- The **Markov Chain Monte Carlo (MCMC)** methods can be used to generate a draw from a distribution that approximates f .
- In fact, they are more properly viewed as methods for generating a sample from which expectations of functions of X can be reliably be estimated.
- MCMC methods are distinguished from Monte Carlo techniques in that it relies on dependent samples.



Markov Chain

- Let $(X_t, t \in T)$ be a stochastic process with state space $S = \mathbb{R}^d$, and T is time space.

- We say that (X_t) satisfies the **Markov property** if

$$\mathbb{P}(X_t \in A | X_0 = x_0, \dots, X_t = x_t) = \mathbb{P}(X_t \in A | X_t = x_t)$$

for all measurable set $A \in S, \forall t \geq 1$ and all $x_0, \dots, x_t \in S$.

- We say that $(X_t, t \in T)$ is **Markov Chain (MC)** if

$$[X_{t+1} | X_t, \dots, X_0] = [X_{t+1} | X_t], \forall t \in T$$

Note that for $(X_t, t \in T)$: Markov Chain,

$$\begin{aligned} [X_{t+1} | X_t, \dots, X_0] &= [X_0][X_1 | X_0][X_2 | X_1, X_0] \cdots [X_{t+1} | X_t, \dots, X_0] \\ &= [X_0][X_1 | X_0][X_2 | X_1] \cdots [X_{t+1} | X_t] \end{aligned}$$



Markov Chain

- **Definition (Transition Kernel)**
- $K: S \times S \rightarrow \mathbb{R}$ is called the **Transition Kernel** of Markov Chain (X_t) if

$$\mathbb{P}(X_{t+1} \in A | X_t = x_t) = \int_A K(x_t, x_{t+1}) dx_{t+1}, A \subseteq S$$

$$\text{or } K_t(x, dy) = \mathbb{P}(X_{t+1} \in dy | X_t = x_t)$$

$$\text{where } K(x_t, x_{t+1}) = [X_{t+1} | X_t]$$

- **Definition (n -step Transition Kernel)**
- **We call**

$$K^{(n)}(x, A) = \mathbb{P}(X_{t+m} \in A | X_t = x_t),$$



Markov Chain

- **Definition (Irreducibility)**

- **If state space S is discrete, there exists $m \in \mathbb{N}$ such that**

$$\mathbb{P}(X_m = y | X_0 = x) > 0, \forall x, y \in S$$

- **This means that, somehow there is a route that from any starting point to any final point in state space.**

- **Definition (Recurrence)**

- **For (X_t) , A is recurrence if**

$$\forall x \in A, \mathbb{E}(\sum_{t=0}^{\infty} I(X_t \in A) | X_0 = x) = \infty$$

- **This means that, start from starting point and infinitely many visits the starting point.**



Markov Chain

- **Definition (Invariant distribution)**
- μ is called the **invariant distribution** of kernel K if

$$\mu(A) = \int_A \int_S K(x, y) \mu(x) dx dy, \quad \forall A \subseteq S$$

- **Theorem**
- **Let** $X_0 \sim \mu$ **and** $(X_t)_{t \geq 0}$ **be a MC with the transition kernel** K .
- **If** μ **is the invariant distribution** K , **then**

$$X_i \sim \mu, \quad i = 0, 1, \dots$$

- **Ergodic Theorem**

- **If** $S = \mathbb{R}^d$, (X_t) 's **invariant distribution is** μ **and** (X_t) **is**
 μ **–irreducibility MC, then for function** g **such that** $\int |g| d\mu < \infty$,

$$\mathbb{P}_x \left(\frac{1}{N} \sum_{t=0}^{N-1} g(X_t) \rightarrow \int g d\mu, \text{ as } N \rightarrow \infty \right) = 1, \text{ almost all } x(\text{initial value})$$



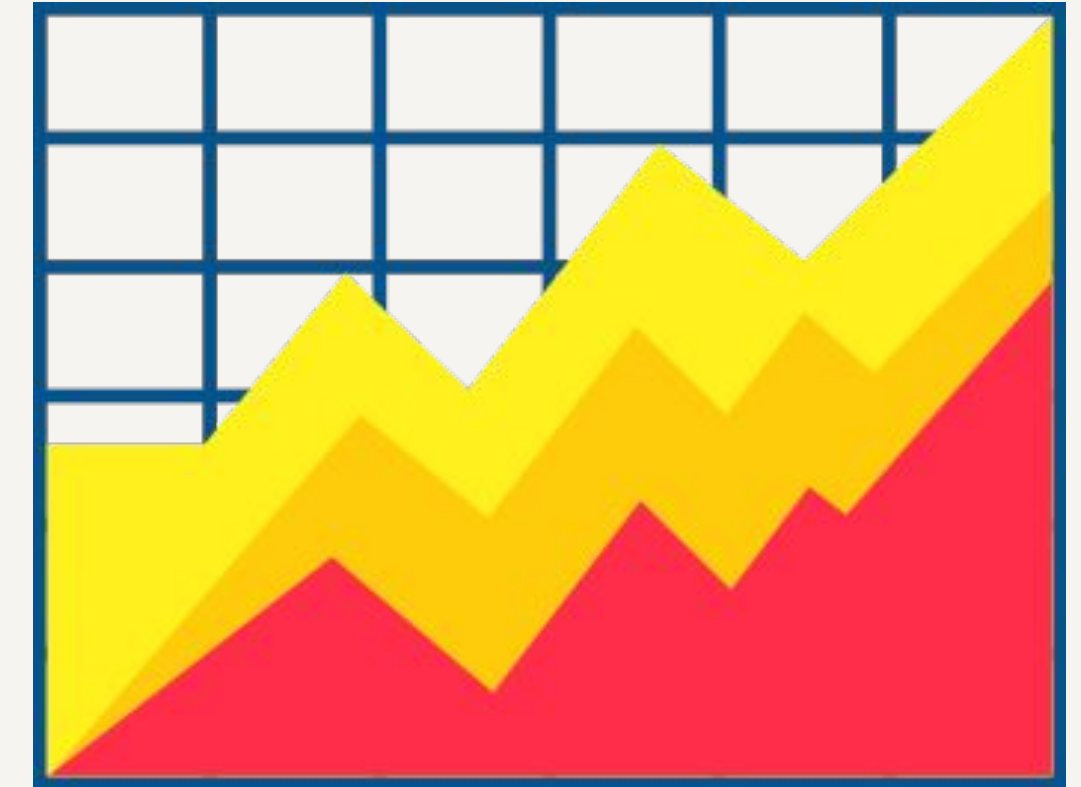


MCMC-Gibbs Sampler/Sampling

- The goal of MCMC is to generate Markov Chain that has posterior distribution as invariant distribution.
- It suffices to find the transition kernel that has posterior distribution as invariant distribution.
- We must consider two conditions from Ergodic Theorem:
 1. Irreducibility
 2. invariant distribution(μ)
- We do not actually focus on first one.
- Also, we should give up iid.

MCMC-Gibbs Sampler/Sampling

- **Gibbs Sampling's idea**
- **Sample θ from the full conditional posteriors and use it to approximate the joint posterior.**
- **Example)** $\pi(\theta_1, \theta_2) = \pi(\theta_1) \cdot \pi(\theta_2 | \theta_1)$
- $\theta^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}) \rightarrow \theta_2^{(2)} \sim \pi(\theta_2 | \theta_1 = \theta_1^{(1)}) \rightarrow (\theta_1^{(1)}, \theta_2^{(2)}) \sim \pi(\theta_1, \theta_2)$
- $\theta_1^{(2)} \sim \pi(\theta_1 | \theta_2 = \theta_2^{(2)}) \rightarrow (\theta_1^{(2)}, \theta_2^{(2)}) \sim \pi(\theta_1, \theta_2)$
- $\theta_2^{(3)} \sim \pi(\theta_2 | \theta_1 = \theta_1^{(2)}) \rightarrow (\theta_1^{(2)}, \theta_2^{(3)}) \sim \pi(\theta_1, \theta_2)$
- $\theta_1^{(3)} \sim \pi(\theta_1 | \theta_2 = \theta_2^{(3)}) \rightarrow (\theta_1^{(3)}, \theta_2^{(3)}) \sim \pi(\theta_1, \theta_2)$



MCMC-Gibbs Sampler/Sampling

- **Gibbs Sampling Algorithm**

1. **Set initial value** $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_k^{(0)})$

2. **For** $t = 1, \dots, m$,

$$\theta_1^{(t)} \sim \pi \left(\theta_1 | \theta_2 = \theta_2^{(t-1)}, \dots, \theta_k = \theta_k^{(t-1)} \right)$$

$$\theta_2^{(t)} \sim \pi \left(\theta_2 | \theta_1 = \theta_1^{(t)}, \dots, \theta_k = \theta_k^{(t-1)} \right)$$

$\vdots$

$$\theta_k^{(t)} \sim \pi \left(\theta_k | \theta_1 = \theta_1^{(t)}, \dots, \theta_{k-1} = \theta_{k-1}^{(t)} \right)$$

3. **Approximate** $\pi(\theta)$ **by** $(\theta^{(0)}, \dots, \theta^{(m)})$.

This algorithm performs ***systematic sweep***.



Example (MCMC-Gibbs Sampling)

max: 85, mean = 70, my score = 55, otherwise.

- We have 7 observed data from population as follows:

70, 45, 35, 60, 85, 55, 55

- We want to estimate μ, σ^2 .
- Suppose that we have noninformative priors:

$$\mu | \sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right), \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0}{2} \sigma_0^2\right),$$
$$\mu_0 = 0, \kappa_0 = \nu_0 = \sigma_0^2 = 1$$

Sol) From observed data, we have $\bar{x} = \frac{405}{7}, s^2 = 2625, n = 7$.

Priors: $\mu | \sigma^2 \sim N(0, \sigma^2), \sigma^2 \sim \text{Inv-Gamma}(0.5, 0.5)$



Example (MCMC-Gibbs Sampling)

We know conditional distributions:

$$\pi(\mu|\sigma^2, \mathbf{x}) = N\left(\frac{\kappa_0\mu_0 + n\bar{x}}{\kappa_0 + n}, \frac{\sigma^2}{\kappa_0 + n}\right) = N\left(\frac{405}{8}, \frac{\sigma^2}{8}\right)$$

$$\pi(\sigma^2|\mu, \mathbf{x}) = \text{Inv-Gamma}\left(\frac{n+\nu_0+1}{2}, \left(-\frac{1}{2}((\kappa_0+n)\left(\mu - \frac{\kappa_0\mu_0+n\bar{x}}{\kappa_0+n}\right)^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{x} - \mu_0)^2 + (n-1)s^2 + \nu_0\sigma_0^2)\right)\right)$$

Algorithm

1. **Set initial value** $\theta^{(0)} = (\mu^{(0)}, (\sigma^2)^{(0)})$
2. **For** $t = 1, \dots, m$,
$$\mu^{(t)} \sim \pi(\mu|\sigma^2 = (\sigma^2)^{(t-1)}), (\sigma^2)^{(t)} \sim \pi(\sigma^2|\mu = \mu^{(t)})$$
3. **Approximate** $\pi(\theta)$ **by** $(\theta^{(0)}, \dots, \theta^{(m)})$.



Example (MCMC-Gibbs Sampling)

```
# Gibbs-Sampler
set.seed(123)

# 1. data
X <- c(70, 45, 35, 60, 85, 55, 55)
n <- length(X)
x_bar <- mean(X)

# prior parameters (kappa_0 = 1, v_0 = 1, mu_0 = 0, sigma_0^2 = 1)
k0 <- 1
mu0 <- 0
v0 <- 1
alpha0 <- v0 / 2 # 0.5
beta0 <- (v0 * 1) / 2 # 0.5

# 2. Gibbs-Sampler parameter
kn <- k0 + n
mu_n <- (k0 * mu0 + n * x_bar) / kn
alpha_n <- (v0 + n + 1) / 2

# 3. Gibbs-Sampler setting
N_iter <- 100
burn_in <- 10
S <- matrix(nrow = N_iter, ncol = 2)
colnames(S) <- c("mu", "sigma2")
mu_t <- mean(X)

# 4. Loop
for (t in 1:N_iter) {
  sum_sq_dev <- sum((X - mu_t)^2)
  beta_n <- beta0 + 0.5 * sum_sq_dev + 0.5 * k0 * (mu_t - mu0)^2
  sigma2_t <- 1 / rgamma(1, shape = alpha_n, rate = beta_n)
  mu_var_n <- sigma2_t / kn
  mu_t <- rnorm(1, mean = mu_n, sd = sqrt(mu_var_n))
  S[t, ] <- c(mu_t, sigma2_t)
}
```

```
# 5. Burn-in 샘플 제거
S_post_burn <- S[(burn_in + 1):N_iter, ]

# 6. 결과 요약
summary(S_post_burn)

# 7. 95% 신용 구간
apply(S_post_burn, 2, quantile, probs = c(0.025, 0.975))
```

```
> summary(S_post_burn)
```

	mu	sigma2
Min.	:27.42	Min. : 202.2
1st Qu.	:46.16	1st Qu.: 460.5
Median	:50.54	Median : 673.1
Mean	:51.80	Mean : 771.2
3rd Qu.	:57.30	3rd Qu.: 976.8
Max.	:77.16	Max. : 2404.0

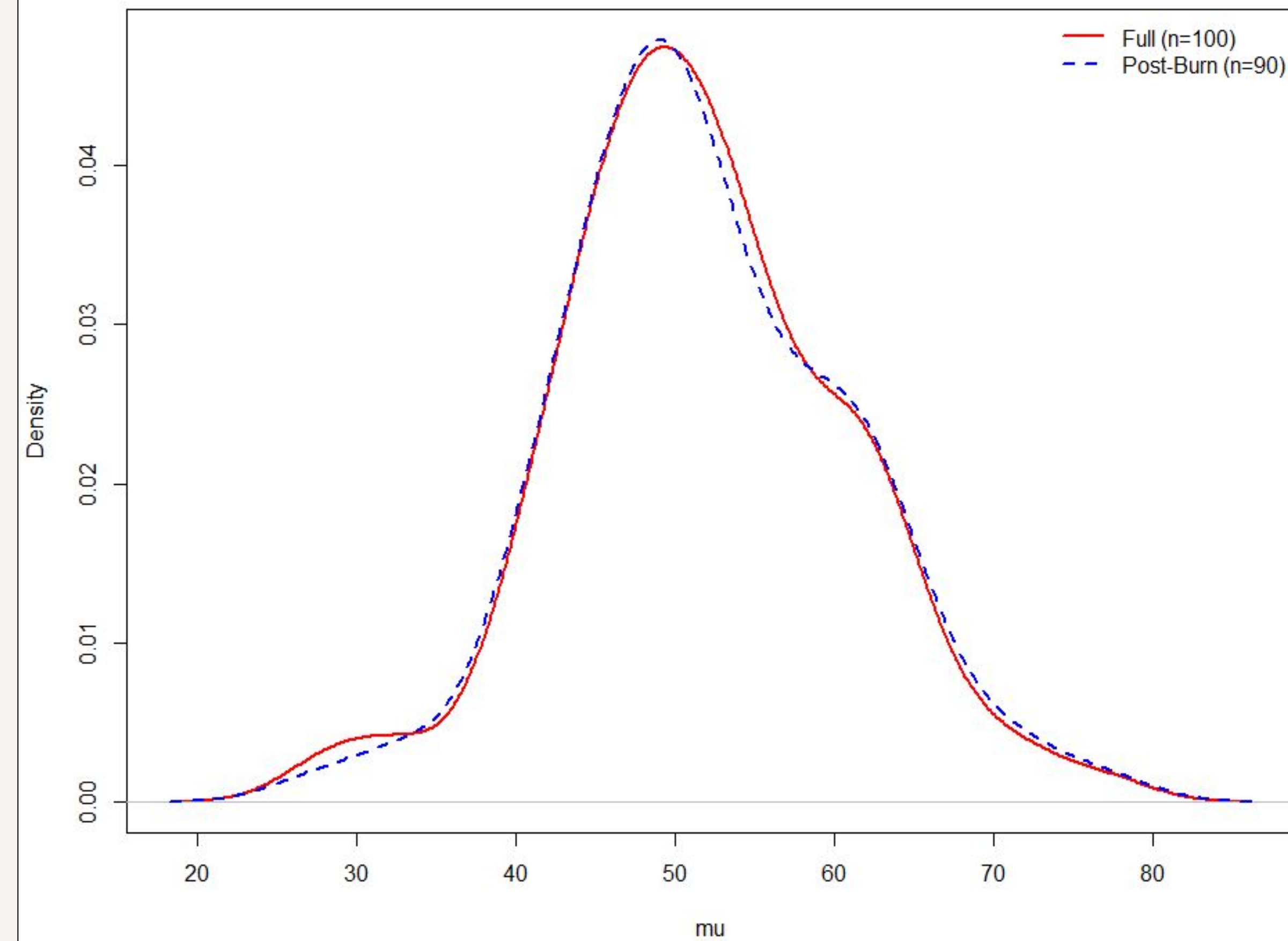
```
> apply(S_post_burn, 2, quantile, probs = c(0.025, 0.975))
```

	mu	sigma2
2.5%	34.12486	276.9116
97.5%	70.50514	1879.5174

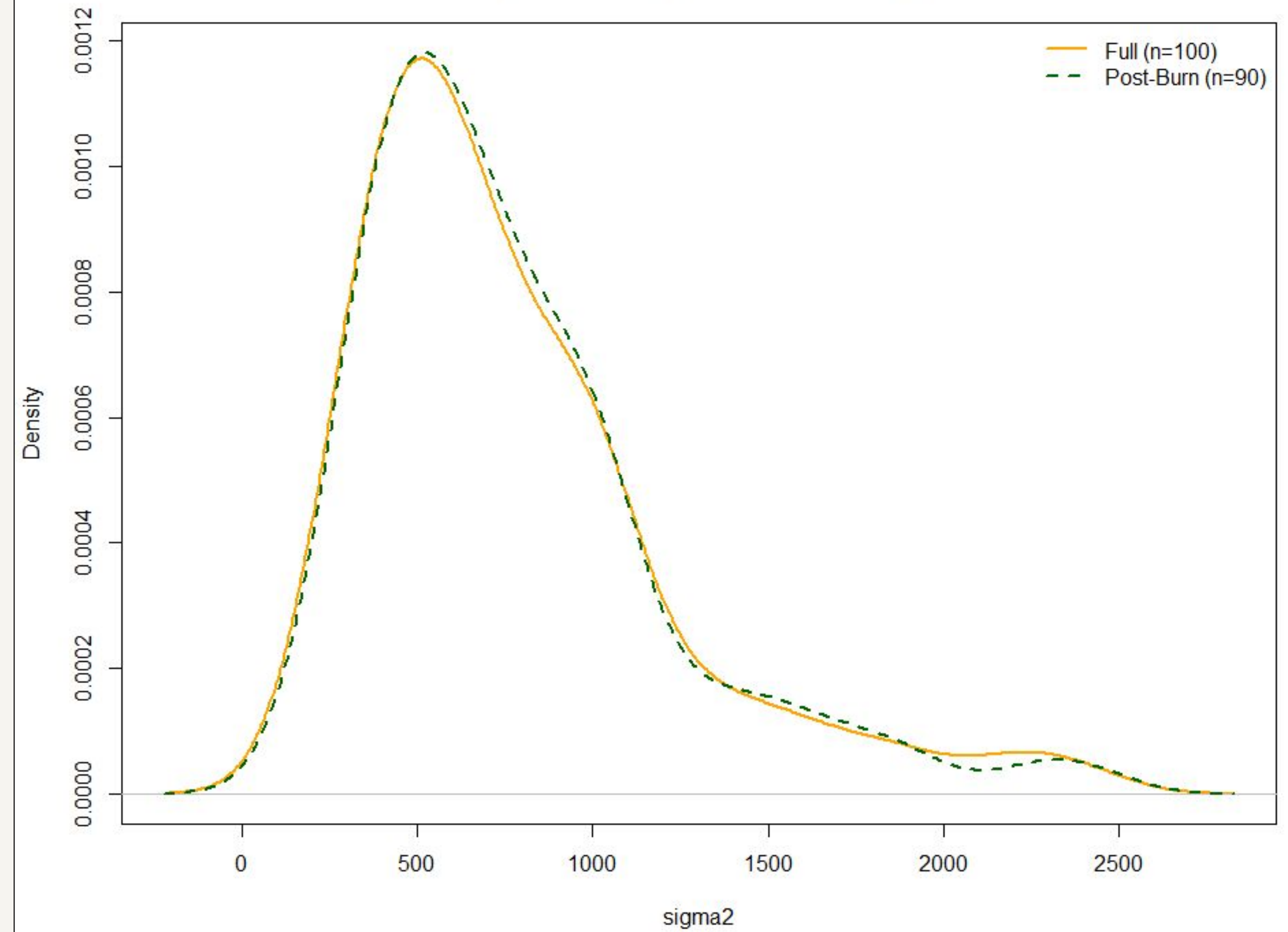
Example (MCMC-Gibbs Sampling)



Mu Density (Full vs Post-Burn)



Sigma2 Density (Full vs Post-Burn)



Example (MCMC-Gibbs Sampling)



```
# Gibbs-Sampler
set.seed(123)

# 1. data
X <- c(70, 45, 35, 60, 85, 55, 55)
n <- length(X)
x_bar <- mean(X)

# prior parameters (kappa_0 = 1, v_0 = 1, mu_0 = 0, sigma_0^2 = 1)
k0 <- 1
mu0 <- 0
v0 <- 1
alpha0 <- v0 / 2 # 0.5
beta0 <- (v0 * 1) / 2 # 0.5

# 2. Gibbs-Sampler parameter
kn <- k0 + n
mu_n <- (k0 * mu0 + n * x_bar) / kn
alpha_n <- (v0 + n + 1) / 2

# 3. Gibbs-Sampler setting
N_iter <- 1000
burn_in <- 100
S <- matrix(nrow = N_iter, ncol = 2)
colnames(S) <- c("mu", "sigma2")
mu_t <- mean(X)

# 4. Loop
for (t in 1:N_iter) {
  sum_sq_dev <- sum((X - mu_t)^2)
  beta_n <- beta0 + 0.5 * sum_sq_dev + 0.5 * k0 * (mu_t - mu0)^2
  sigma2_t <- 1 / rgamma(1, shape = alpha_n, rate = beta_n)
  mu_var_n <- sigma2_t / kn
  mu_t <- rnorm(1, mean = mu_n, sd = sqrt(mu_var_n))
  S[t, ] <- c(mu_t, sigma2_t)
}
```

```
# 5. Burn-in 샘플 제거
S_post_burn <- S[(burn_in + 1):N_iter, ]

# 6. 결과 요약
summary(S_post_burn)

# 7. 95% 신용 구간
apply(S_post_burn, 2, quantile, probs = c(0.025, 0.975))
```

```
> summary(S_post_burn)
```

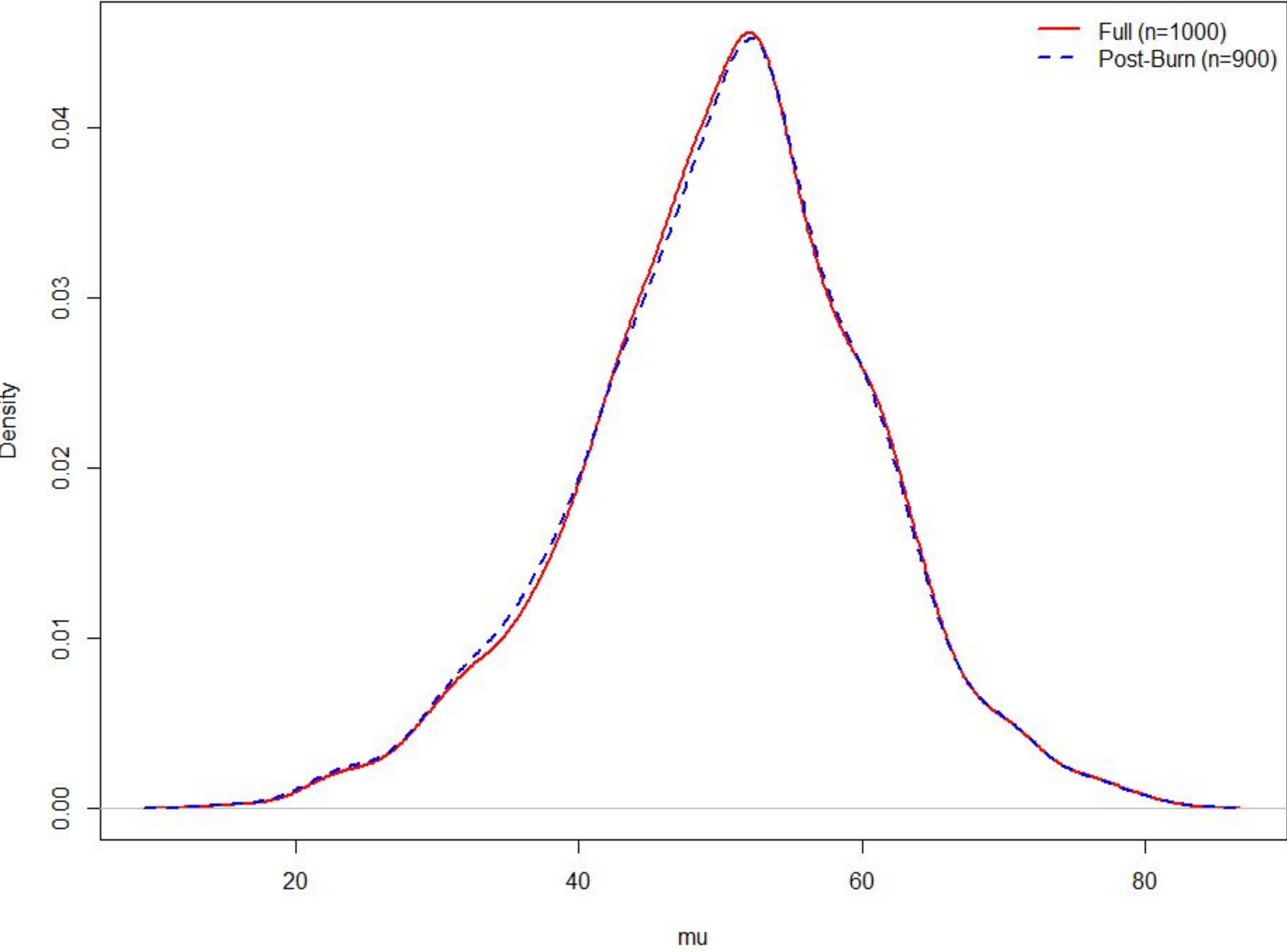
	mu	sigma2
Min.	:15.98	Min. : 146.8
1st Qu.	:43.96	1st Qu.: 455.3
Median	:51.09	Median : 654.1
Mean	:50.33	Mean : 783.7
3rd Qu.	:56.94	3rd Qu.: 940.2
Max.	:80.28	Max. : 6299.6

```
> apply(S_post_burn, 2, quantile, probs = c(0.025, 0.975))
```

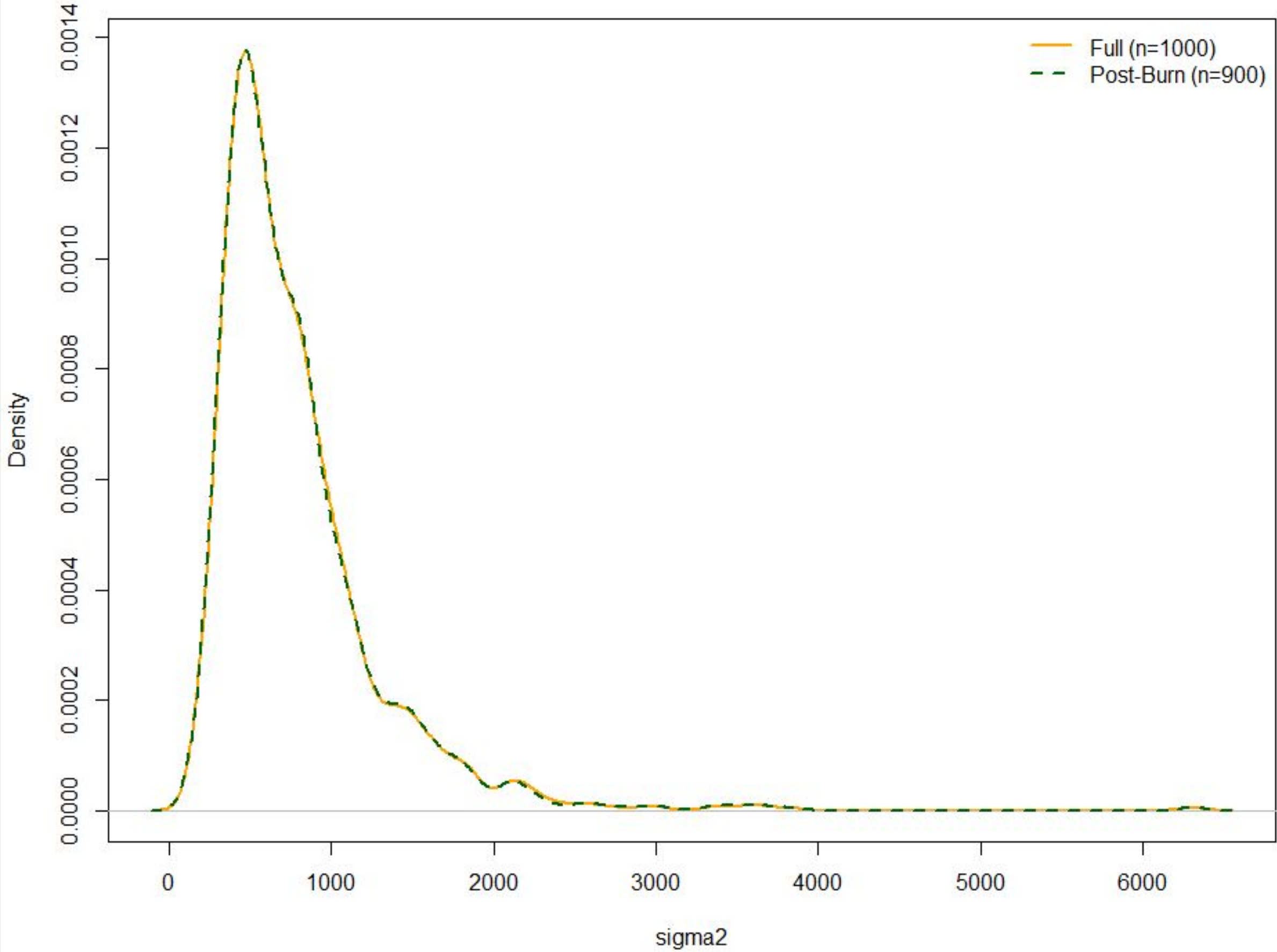
	mu	sigma2
2.5%	29.28998	255.1852
97.5%	70.35599	2078.7091

Example (MCMC-Gibbs Sampling)

Mu Density (Full vs Post-Burn)



Sigma2 Density (Full vs Post-Burn)



Example (MCMC-Gibbs Sampling)

```
# Gibbs-sampler
set.seed(123)

# 1. data
X <- c(70, 45, 35, 60, 85, 55, 55)
n <- length(X)
x_bar <- mean(X)

# prior parameters (kappa_0 = 1, v_0 = 1, mu_0 = 0, sigma_0^2 = 1)
k0 <- 1
mu0 <- 0
v0 <- 1
alpha0 <- v0 / 2 # 0.5
beta0 <- (v0 * 1) / 2 # 0.5

# 2. Gibbs-sampler parameter
kn <- k0 + n
mu_n <- (k0 * mu0 + n * x_bar) / kn
alpha_n <- (v0 + n + 1) / 2

# 3. Gibbs-sampler setting
N_iter <- 10000
burn_in <- 1000
S <- matrix(nrow = N_iter, ncol = 2)
colnames(S) <- c("mu", "sigma2")
mu_t <- mean(X)

# 4. Loop
for (t in 1:N_iter) {
  sum_sq_dev <- sum((X - mu_t)^2)
  beta_n <- beta0 + 0.5 * sum_sq_dev + 0.5 * k0 * (mu_t - mu0)^2
  sigma2_t <- 1 / rgamma(1, shape = alpha_n, rate = beta_n)
  mu_var_n <- sigma2_t / kn
  mu_t <- rnorm(1, mean = mu_n, sd = sqrt(mu_var_n))
  S[t, ] <- c(mu_t, sigma2_t)
}
```

```
# 5. Burn-in 샘플 제거
S_post_burn <- S[(burn_in + 1):N_iter, ]

# 6. 결과 요약
summary(S_post_burn)

# 7. 95% 신용 구간
apply(S_post_burn, 2, quantile, probs = c(0.025, 0.975))
```

```
> summary(S_post_burn)
```

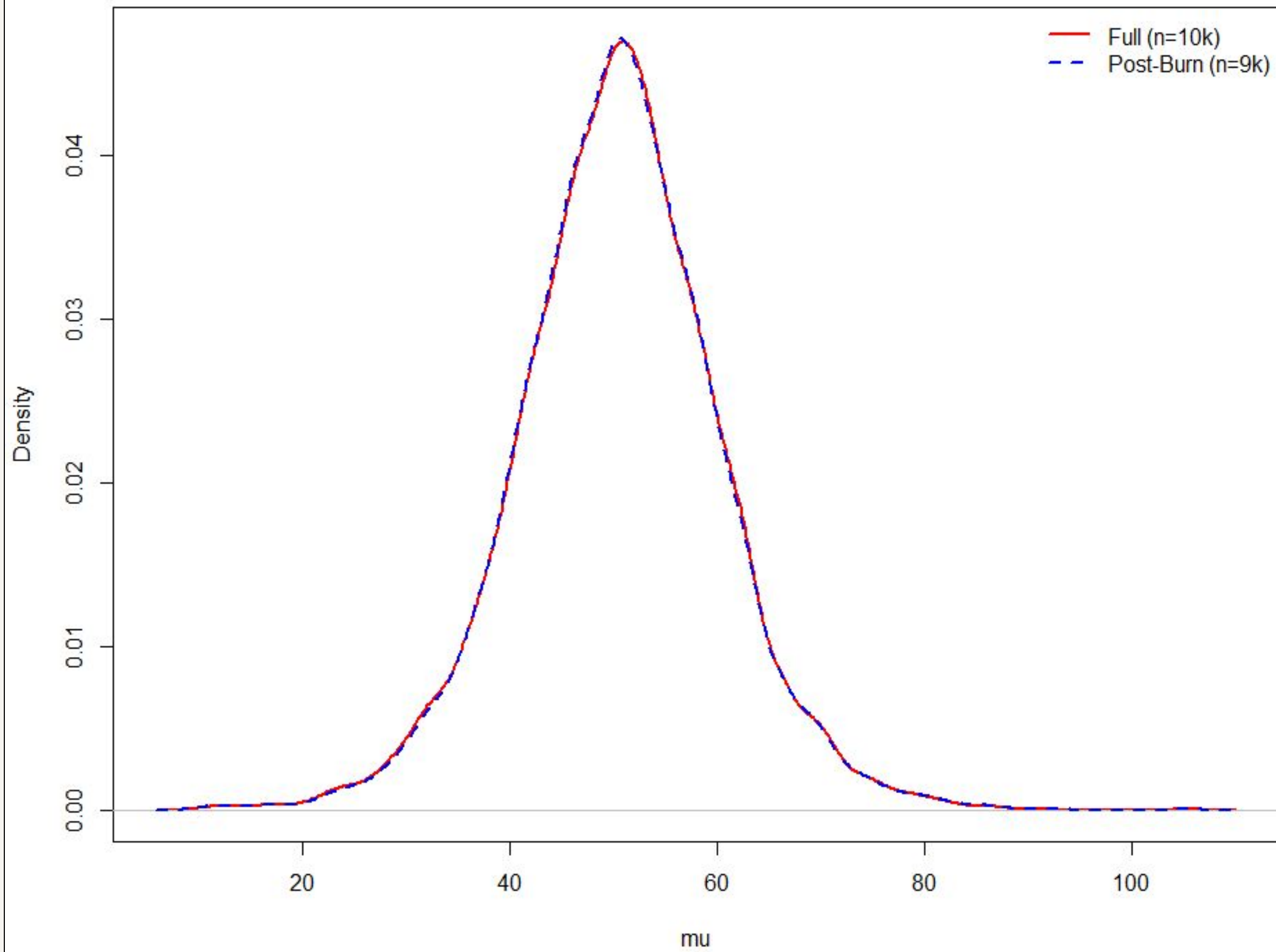
	mu	sigma2
Min.	:15.98	Min. : 146.8
1st Qu.	:43.96	1st Qu.: 455.3
Median	:51.09	Median : 654.1
Mean	:50.33	Mean : 783.7
3rd Qu.	:56.94	3rd Qu.: 940.2
Max.	:80.28	Max. : 6299.6

```
> apply(S_post_burn, 2, quantile, probs = c(0.025, 0.975))
```

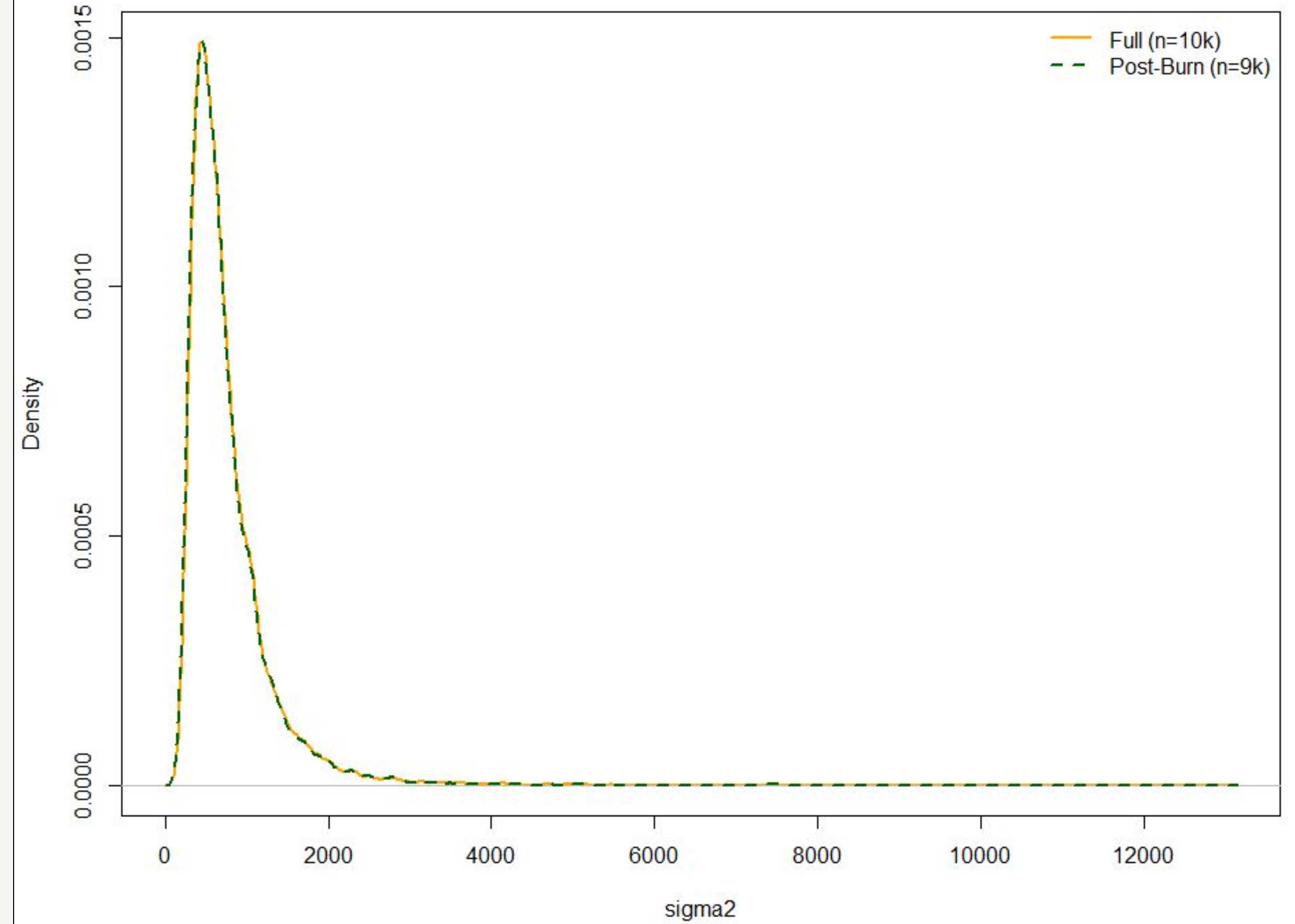
	mu	sigma2
2.5%	29.28998	255.1852
97.5%	70.35599	2078.7091

Example (MCMC-Gibbs Sampling)

Mu Density (Full vs Post-Burn)



Sigma2 Density (Full vs Post-Burn)



MCMC-Metropolis-Hastings

- To use the Gibbs sampler, full conditional posteriors,

$$\pi(\theta_i | \theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p, x) \text{ for } i = 1, \dots, p,$$

should be available.

- Q) Can we get an arbitrary kernel that has the invariant distribution $\pi(\theta|x)$?
- A) Metropolis-Hastings Algorithm is the solution!
- [Metropolis-Hastings(MH) Algorithm idea]
- We want to sample from the target distribution $\pi(x)$.
- Suppose we have a kernel $q(x, y)$ that is easy to generate samples and

$$\text{supp}(\pi) \subset \bigcup_{x \in \text{supp}(\pi)} \text{supp}\{q(x, \cdot)\}$$

- Can we use samples from $q(x, y)$ with any fixed initial value



MCMC-Metropolis-Hastings

- **Metropolis-Hastings(MH) Algorithm**

- **Step 1: Set $\theta^{(0)}$.**

- **Step 2: For $t = 1, 2, \dots, m$**

- 1. Sample a candidate value and a uniform r.v. as iid:**

$$\theta' \sim q(\theta^{(t-1)}, \theta'), u \sim U(0,1)$$

- 2. Compute the acceptance probability**

$$\alpha(\theta^{(t-1)}, \theta') = \min \left(1, \frac{\pi(\theta')q(\theta', \theta^{(t-1)})}{\pi(\theta^{(t-1)})q(\theta^{(t-1)}, \theta')} \right)$$

- 3. Determine the value of $\theta^{(t)}$**

$$\theta^{(t)} = \begin{cases} \theta', & \text{if } u \leq \alpha(\theta^{(t-1)}, \theta') \\ \theta^{(t-1)}, & \text{if } u > \alpha(\theta^{(t-1)}, \theta') \end{cases}$$

- **Step 3: Approximate $\pi(\theta)$ with $\theta^{(t)}, t = 1, 2, \dots, m$**



Extra) Betting

- This is a method that uses betting to determine probabilities, as a way to make the analyst think more seriously about the relative likelihoods they hold.
- Consider one wins $z \in \mathbb{R}$ dollars if event E occurs, and loses 1 dollar if event E^c occurs.
- Let $U: \mathbb{R} \rightarrow \mathbb{R}$ be the analyst's given utility for an amount of money.
- If this bet is fair, the expected utility must be 0:
$$0 = U(z)P(E) - U(1)(1 - P(E))$$
- Then,

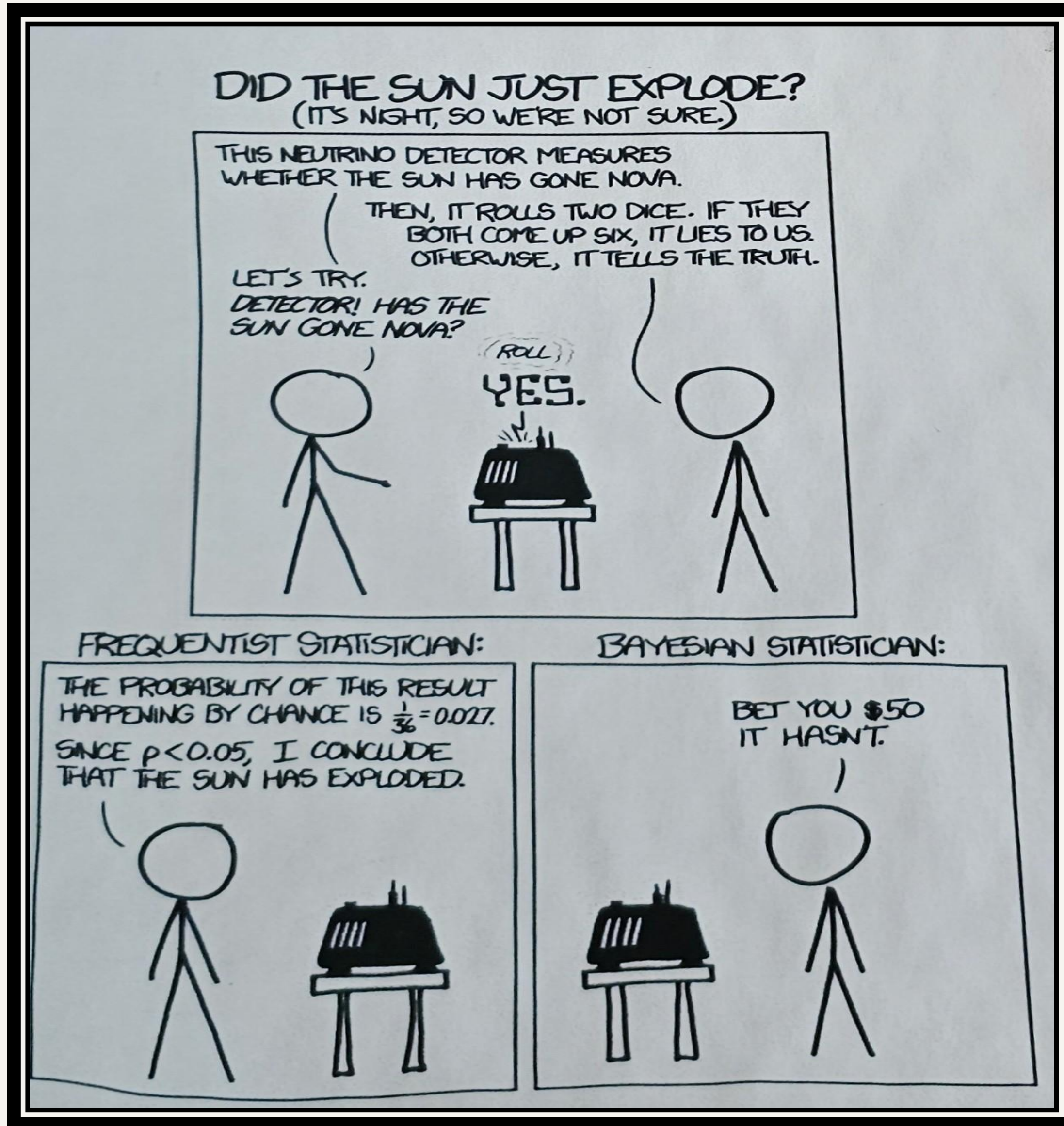
$$P(E) = \frac{U(1)}{U(1) + U(z)}$$

- If the utility is linear in z , that is, $U(z) = cz, c > 0$, then

$$P(E) = \frac{1}{1+z}$$



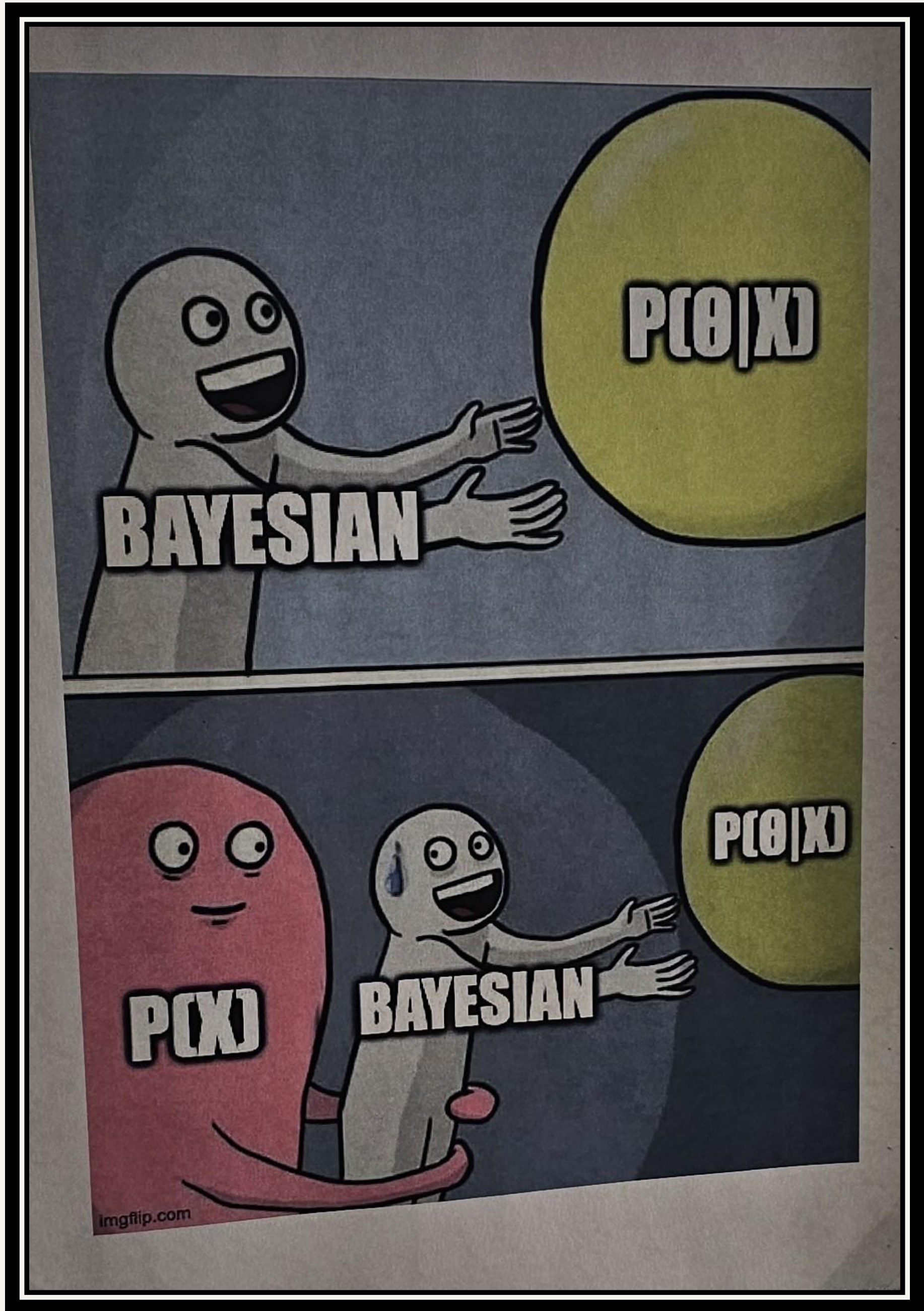
Extra) Betting- Example



Reference

- 이재용, & 이기재. (2022). *베이지 데이터 분석*. 방송통신대학교출판문화원.
- Lee, J. (2025). *Bayesian Statistics and Lab*. [Lecture notes]. Statistics. Seoul National University.
- Baek, C. (2025). Introduction to Stochastic Processes [Lecture notes]. Statistics. Sungkyunkwan University.
- Lee, K. (2025). Introduction to Bayesian Statistics [Lecture notes]. Statistics. Sungkyunkwan University.
- Lee, K. (2024). Introduction to Statistical Inference [Lecture notes]. Statistics. Sungkyunkwan University.





끝까지 들어 주셔서

감사합니다!