

# BAYESIAN STATISTICS A

## Foundations of Bayesian Statistics

---

Teruo Nakatsuma

Fall 2025

Faculty of Economics, Keio University

# Aims Of This Course

1. Learn basic principles of Bayesian learning.
2. Learn how to conduct statistical inference (point estimation, interval estimation, hypothesis testing, prediction) in the Bayesian way.
3. Learn basic principles of Markov chain Monte Carlo (MCMC) methods.
4. Hands-on practice of Python and PyMC.

# Reading List i

## 1. Introduction to Bayesian statistics

- Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2013). *Bayesian Data Analysis*, 3rd ed., Chapman & Hall/CRC.

Download the book [here](#).

- Hoff, P.D. (2010). *A First Course in Bayesian Statistical Methods*, Springer.

Download the book [here](#) (Keio account required).

- Martin, O. (2024). *Bayesian Analysis with Python: A practical guide to probabilistic modeling*, 3rd ed., Packt Publishing.

## 2. Advanced topics in Bayesian statistics

- Chan, J., Koop, G., Poirier, D.J. and Tobias, J.L. (2020). *Bayesian Econometric Methods*, 2nd ed., Cambridge University Press.
- Durbin, J. and Koopman, S.J. (2012). *Time Series Analysis by State Space Methods*, 2nd ed., Oxford University Press.
- Prado, R. and West, M. (2021). *Time Series: Modeling, Computation, and Inference*, 2nd ed., Chapman & Hall/CRC.
- Rossi, P.E., Allenby, G.E. and McCulloch, R. (2005). *Bayesian Statistics and Marketing*, Wiley.

# Definition of Probability

## Axiomatic Definition of Probability

Suppose  $\Omega$  is a sample space (a set of all possible outcomes) and  $\mathcal{F}$  is a ( $\sigma$ -)field (a collection of all relevant events) on  $\Omega$ .

**Axiom 1.** For any event  $A \in \mathcal{F}$ ,  $P(A) \geq 0$ .

**Axiom 2.**  $P(\Omega) = 1$ .

**Axiom 3.** For any pairwise disjoint events  
 $A_1, \dots, A_n \in \mathcal{F}$ ,

$$P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n).$$

When  $\mathcal{F}$  is a  $\sigma$ -field,  $n$  must be infinite.

# Interpretations Of Probability

- **Classical Interpretation**

Suppose every outcome is equally likely. The probability of an event **A** is defined as

$$P(A) = \frac{\text{\# of all outcomes in } A}{\text{\# of all outcomes in } \Omega}.$$

- **Frequentist Interpretation**

The probability of **A** is the limit of the ratio of occurrence in infinitely repeated trials, i.e.,

$$P(A) = \lim_{n \rightarrow \infty} \frac{\text{\# of occurrence of } A \text{ up to the } n\text{-th trial}}{n}.$$

- **Bayesian Interpretation**

**P(A)** is a researcher's degree of belief in **A**.

# Subjective Probability

- Bayesian statistics relies on the concept called **subjective probability**.
- The subjective probability still satisfies all mathematical properties of probability.
- Bruno de Finetti proved that probability could be derived as a subjective measure of uncertainty in relation to bookmaking.
- His founding is called the **Dutch Book Theorem**.

# Dutch Book Theorem

Consider the following bet on event  $A$ .

- The amount of money the bettor bets is  $P(A)$ .
- If  $A$  occurs, the bookmaker will pay 1 to the bettor.
- If  $A$  does not occur, the bettor will receive nothing.

If either bettor or bookmaker gains for sure, the bet is called a **Dutch book**. de Finetti's Dutch Book Theorem claims that a subjective measure of uncertainty must be a probability measure; otherwise it leads to a Dutch book.



# Payoff Function

The payoff function of the bet is defined as follows.

$$G_A = \begin{cases} 1 - P(A), & \text{if } A \text{ occurs;} \\ -P(A), & \text{otherwise.} \end{cases}$$

$G_A$  is rearranged as

$$\begin{aligned} G_A(\omega) &= (1 - P(A))\mathbb{1}_A(\omega) - P(A)(1 - \mathbb{1}_A(\omega)) \\ &= \mathbb{1}_A(\omega) - P(A), \end{aligned} \tag{1}$$

where

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & (\omega \in A), \\ 0, & (\omega \notin A). \end{cases}$$

# Proof of Axiom 1

Suppose  $\mathbf{P}(A) < 0$ . Then

$$G_A(\omega) = \begin{cases} 1 - \mathbf{P}(A) > 0, & (\omega \in A); \\ -\mathbf{P}(A) > 0, & (\omega \notin A). \end{cases}$$

Therefore this bet is a Dutch book. Hence  $\mathbf{P}(A) \geq 0$ . ■

## Proof of Axiom 2

Suppose we can bet on  $\Omega$ . From the definition of the payoff function (1), we have

$$G_{\Omega}(\omega) = \mathbb{1}_{\Omega}(\omega) - P(\Omega),$$

Since any  $\omega$  belongs to  $\Omega$ ,  $\mathbb{1}_{\Omega}(\omega) = 1$ . Thus

$$G_{\Omega}(\omega) = 1 - P(\Omega).$$

- If  $P(\Omega) < 1$ ,  $G_{\Omega}(\omega) = 1 - P(\Omega) > 0$ . So the bettor will always gain.
- If  $P(\Omega) > 1$ ,  $G_{\Omega}(\omega) = 1 - P(\Omega) < 0$ . So the bettor will always lose.

Therefore  $P(\Omega) = 1$ . ■

## Proof of Axiom 3 i

It is sufficient to show  $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$  if  $n$  is finite in Axiom 3. Define  $\mathbf{C} = A \cup B$  and consider the following payoff function:

$$\begin{aligned} G(\omega) &= G_A(\omega) + G_B(\omega) - G_C(\omega) \\ &= \mathbf{1}_A(\omega) - \mathbf{P}(A) + \mathbf{1}_B(\omega) - \mathbf{P}(B) - (\mathbf{1}_C(\omega) - \mathbf{P}(C)). \end{aligned}$$

Define further

- $\omega_1 \in A$ :  $A$  occurs
- $\omega_2 \in B$ :  $B$  occurs
- $\omega_3 \in C^c$ : either  $A$  or  $B$  does not occur

## Proof of Axiom 3 ii

Suppose  $\mathbf{P}(A \cup B) > \mathbf{P}(A) + \mathbf{P}(B)$ . Then

$$G(\omega_1) = \mathbf{P}(C) - \mathbf{P}(A) - \mathbf{P}(B) > 0,$$

$$G(\omega_2) = \mathbf{P}(C) - \mathbf{P}(A) - \mathbf{P}(B) > 0,$$

$$G(\omega_3) = \mathbf{P}(C) - \mathbf{P}(A) - \mathbf{P}(B) > 0.$$

Therefore the bettor will gain for sure.

If  $\mathbf{P}(A \cup B) < \mathbf{P}(A) + \mathbf{P}(B)$ , on the other hand, the bettor will lose for sure.

Thus  $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$ . ■

# Population, Sample, Parameter

1. In statistics, the **population** is any subject (not necessarily a group) which a researcher tries to analyze.
2. The **sample** is a collection of data related to the population. In a typical situation, the sample is assumed to be randomly and independently extracted from the population.
3. The **parameter** represents a property of the population to be analyzed. The parameter is unknown to the researcher.
4. The goal of statistics is to obtain useful insights about the parameter of the population with the sample extracted from it.

# Population Distribution

We regard the population as a probability distribution and call it the **population distribution**. Then we can interpret the sample as a set of random variables following the population distribution, and the parameters as variables which determine the “shape” of the population distribution. Let  $D = (x_1, \dots, x_n)$  denote the sample, and  $\theta$  denote the parameter of the population distribution. To indicate that the shape of the population distribution depends on  $\theta$ , the population p.m.f. or p.d.f. is denoted by  $p(x_i|\theta)$  where each  $x_i$  ( $i = 1, \dots, n$ ) is called an **observation** and supposed to be a realized value of the random variable following the population distribution.  $n$  is often called the **sample size**.

## Example: Bernoulli Distribution

Consider a random variable  $X_i$  ( $i = 1, \dots, n$ ) such that

$$X_i = \begin{cases} 1, & \text{Head is obtained;} \\ 0, & \text{Tail is obtained,} \end{cases}$$

and suppose  $\mathbf{P}(X_i = 1) = \theta$  and  $\mathbf{P}(X_i = 0) = 1 - \theta$ . Then  $X_i$  follows the **Bernoulli distribution** with the p.m.f.:

$$p(x_i|\theta) = \theta^{x_i} (1 - \theta)^{1-x_i}, \quad x_i = 0, 1.$$



# Likelihood

Suppose the sample  $D = (x_1, \dots, x_n)$  are taken from a population distribution where the parameter  $\theta$  is a set of unknown parameters. The joint p.m.f. or the joint p.d.f. of  $D$  is denoted by

$$p(D|\theta) = p(x_1, \dots, x_n|\theta).$$

In particular, if observations are independent of each other,

$$p(D|\theta) = p(x_1|\theta) \times \dots \times p(x_n|\theta) = \prod_{i=1}^n p(x_i|\theta).$$

When we regard  $p(D|\theta)$  as a function of  $\theta$ , it is called the **likelihood** or **likelihood function**.

## Example: Bernoulli Distribution

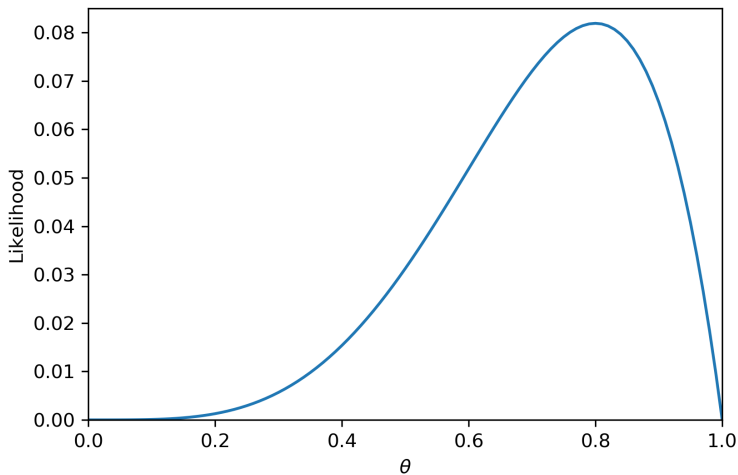
Suppose  $D = (x_1, \dots, x_n)$  is independently generated from the same Bernoulli distribution. Then

$$p(D|\theta) = \prod_{i=1}^n p(x_i|\theta) = \prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^y (1 - \theta)^{n-y},$$

where  $y = \sum_{i=1}^n x_i$ .

Suppose we have  $(x_1, x_2, x_3, x_4, x_5) = (1, 0, 1, 1, 1)$ . The value of  $p(D|\theta)$  depends on the value of  $\theta$ .

$\theta$	0.1000	0.2000	0.3000	0.4000	0.5000
$p(D \theta)$	0.0001	0.0013	0.0057	0.0154	0.0312
$\theta$	0.6000	0.7000	0.8000	0.9000	
$p(D \theta)$	0.0518	0.0720	0.0819	0.0656	



**Figure 1:** The likelihood of  $\theta$  in the Bernoulli distribution

# Interpretation Of The Likelihood

Given the sample  $D$ , the likelihood  $p(D|\theta)$  is regarded as a kind of “plausibility” of a specific value of  $\theta$ .

For example, the likelihood of  $\theta = 0.9$  is **0.0656** while that of  $\theta = 0.4$  in the previous example is **0.0154**. We may say that **0.9** is about 4 times more plausible than **0.4** as the true value of  $\theta$ .

To make comparison between two competing values of  $\theta$ , say  $\theta_0$  and  $\theta_1$ , we introduce the **likelihood ratio**:

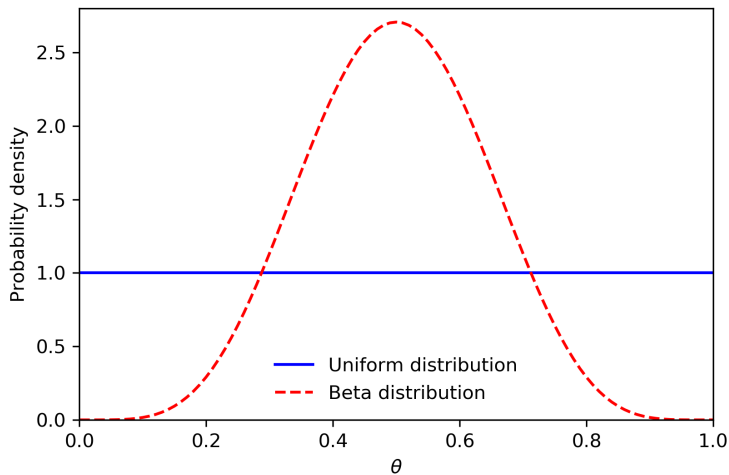
$$\text{likelihood ratio} = \frac{p(D|\theta_0)}{p(D|\theta_1)}.$$

# Prior Knowledge On Parameters

In practice, researchers often have information on unknown parameters before they start analysis. For example,

- $\theta$  must take a value between 0 and 1 because it is probability;
- in case of tossing a coin,  $\theta$  is supposed to be 50% if the coin is fair.

In Bayesian statistics, we construct a distribution of unknown parameters that reflect our prior knowledge on their true values. This is called the **prior distribution**. Let  $p(\theta)$  denote the prior distribution.



**Figure 2:** Prior distributions of  $\theta$  in the Bernoulli distribution

The **uniform distribution**  $\text{Uniform}(a, b)$  is

$$p(x|a, b) = \begin{cases} \frac{1}{b-a}, & (a \leq x \leq b); \\ 0, & (\text{otherwise}). \end{cases}$$

In the above figure, we set  $a = 0$  and  $b = 1$ .

The **beta distribution**  $\text{Beta}(\alpha, \beta)$  is

$$p(x|\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 \leq x \leq 1.$$

where  $B(\alpha, \beta)$  is the beta function:

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.$$

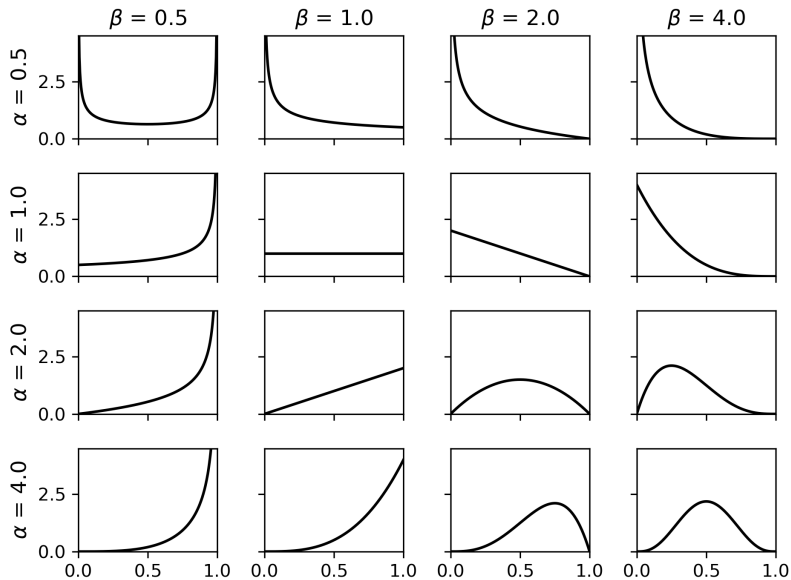


Figure 3: Beta distributions with various  $(\alpha, \beta)$



# Bayes' Theorem And Posterior Distribution

Suppose  $p(\theta, D)$  is the joint distribution of  $\theta$  and  $D$ . Using the definition of the conditional distribution, we have

$$p(\theta, D) = p(\theta|D)p(D) = p(D|\theta)p(\theta).$$

Arranging the middle and the right-hand side of the above equation, we have

## Bayes' Theorem

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

- The above formula is called **Bayes' theorem**.
- $p(\theta|D)$  is called the **posterior distribution**.
- $p(D)$  is called the **normalizing constant**.

# Marginal Likelihood

By marginalizing the joint distribution  $p(\theta, D)$ , we have

$$p(D) = \int_{\Theta} p(\theta, D) d\theta = \int_{\Theta} p(D|\theta)p(\theta) d\theta.$$

Thus  $p(D)$  is interpreted as “averaged likelihood” in terms of the prior  $p(\theta)$ . In this sense,  $p(D)$  is called the **marginal likelihood**.

Then Bayes’ theorem is rewritten as

## Bayes’ Theorem (Alternative Form)

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta) d\theta} \propto p(D|\theta)p(\theta).$$

We can ignore  $p(D)$  since it does not depend on  $\theta$ .

## Example: Bernoulli Distribution

Suppose the prior distribution is  $\text{Beta}(\alpha_0, \beta_0)$ .

The posterior distribution of  $\theta$  is given by

$$\begin{aligned} p(\theta|D) &\propto p(D|\theta)p(\theta) \\ &\propto \theta^y (1-\theta)^{n-y} \times \theta^{\alpha_0-1} (1-\theta)^{\beta_0-1} \\ &\propto \theta^{y+\alpha_0-1} (1-\theta)^{n-y+\beta_0-1} \\ &\propto \theta^{\alpha_\star-1} (1-\theta)^{\beta_\star-1}, \\ \alpha_\star &= y + \alpha_0, \quad \beta_\star = n - y + \beta_0. \end{aligned}$$

This is the beta distribution  $\text{Beta}(\alpha_\star, \beta_\star)$ .

# Bayesian Learning

Bayes' theorem is rearranged as

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

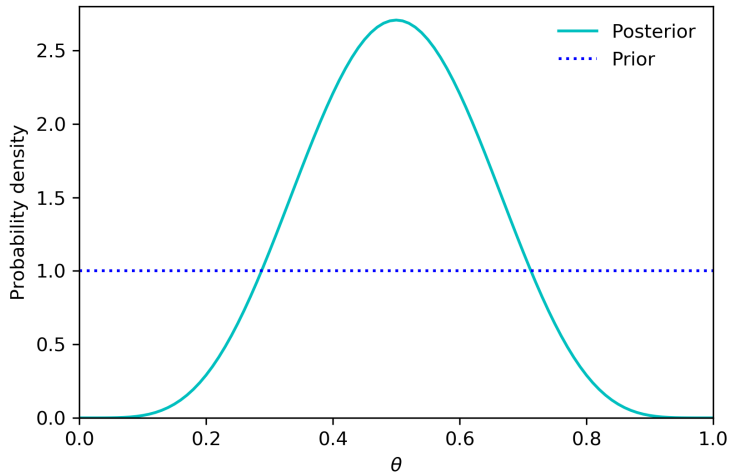
Therefore

$$\frac{p(\theta|D)}{p(\theta)} \gtrless 1 \quad \text{if and only if} \quad \frac{p(D|\theta)}{p(D)} \gtrless 1.$$

Since

$$\begin{cases} \frac{p(\theta|D)}{p(\theta)} > 1, & \text{plausibility of } \theta \text{ is increased;} \\ \frac{p(\theta|D)}{p(\theta)} < 1, & \text{plausibility of } \theta \text{ is decreased,} \end{cases}$$

The plausibility of  $\theta$  depends on whether  $p(D|\theta)$  is higher than  $p(D)$ .



**Figure 4:** Posterior distributions of the probability of success

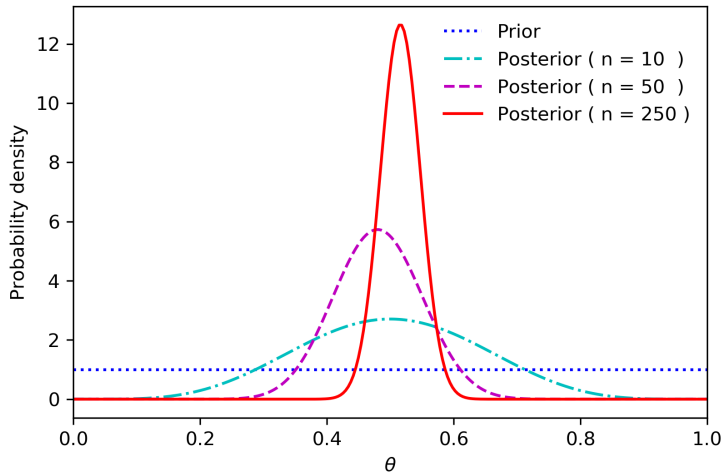


Figure 5: Sequential updating of the posterior distribution

# Bayesian Inference With The Posterior Distribution

The posterior distribution  $p(\theta|D)$  embodies all available information about unknown parameters,  $\theta$ . When the number of parameters to be analyzed is relatively small, displaying graphs of all (marginal) posterior distributions may be sufficient to convey useful insights on the parameters to readers.

However, when we need to analyze many parameters, it is impractical and pointless to show all graphs on the parameters in an article or report. In practice, we calculate and report several “summary statistics” that show us key characteristics of the posterior distribution. We call them the **posterior statistics**.

# Point Estimation

On many occasions, we need to report one particular value of the parameter we regard as the most plausible guess. This type of value is called an **estimate** and a procedure to obtain an estimate is called **point estimation**.

In Bayesian statistics, an estimate of the parameter is defined as a value that minimizes the **expected loss**.

$$\delta_{\star} = \arg \min_{\delta} \mathbb{E}_{\theta}[L(\theta, \delta)|D] = \arg \min_{\delta} \int_{\Theta} L(\theta, \delta) p(\theta|D) d\theta,$$

where  $L$  is the **loss function** and  $\Theta$  is a set of all possible values of  $\theta$  (**parameter space**). In case of the Bernoulli probability,  $\Theta = \{\theta : 0 \leq \theta \leq 1\}$ .



# Examples Of Loss Functions

loss function	$L(\theta, \delta)$	point estimate
quadratic loss	$(\theta - \delta)^2$	posterior mean
absolute loss	$ \theta - \delta $	posterior median
0-1 loss	$1 - \mathbb{1}_\theta(\delta)$	posterior mode

where  $\mathbb{1}_\theta(\delta)$  is the **indicator function** such that

$$\mathbb{1}_\theta(\delta) = \begin{cases} 1, & (\delta = \theta), \\ 0, & (\delta \neq \theta). \end{cases}$$

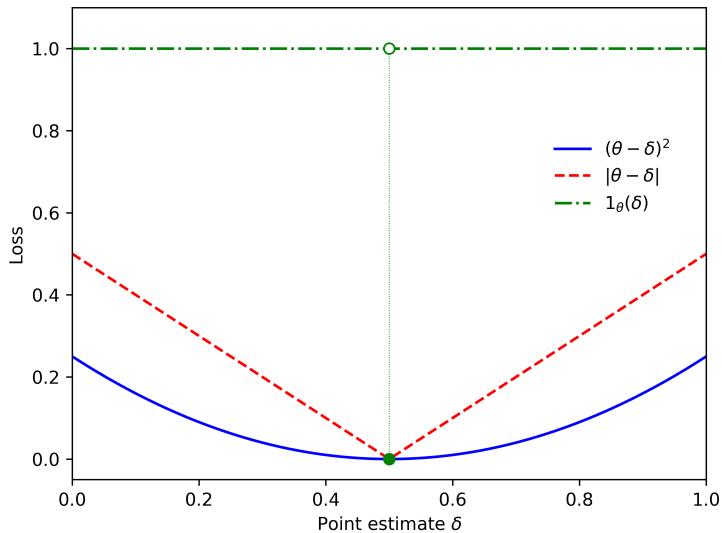


Figure 6: Examples of loss functions

# Mean, Median, Mode

The mean of the distribution is the weighted average of all possible values  $\theta$  may take, i.e.,

$$\mathbf{E}_{\theta}[\theta|D] = \int_{\Theta} \theta p(\theta|D) d\theta.$$

The median of the distribution is a point that divides the distribution in half, i.e.,


$$\mathbf{P}(\theta \leq \mathbf{Median}_{\theta}|D) = 50\%.$$

The mode of the distribution is the highest point of the density, i.e.,

$$\mathbf{Mode}_{\theta} = \mathbf{arg\,max}_{\theta} p(\theta|D).$$

# Proof: Point Estimation with the Quadratic Loss

$$\begin{aligned}\mathbb{E}_\theta[L(\theta, \delta)|D] &= \mathbb{E}_\theta[(\theta - \delta)^2|D] \\&= \mathbb{E}_\theta[(\theta - \mathbb{E}_\theta[\theta|D] + \mathbb{E}_\theta[\theta|D] - \delta)^2|D] \\&= \mathbb{E}_\theta[(\theta - \mathbb{E}_\theta[\theta|D])^2 + 2(\theta - \mathbb{E}_\theta[\theta|D])(\mathbb{E}_\theta[\theta|D] - \delta) \\&\quad + (\mathbb{E}_\theta[\theta|D] - \delta)^2|D] \\&= \mathbb{E}_\theta[(\theta - \mathbb{E}_\theta[\theta|D])^2|D] + (\mathbb{E}_\theta[\theta|D] - \delta)^2.\end{aligned}$$

The first term in the last equation is the posterior variance of  $\theta$ , which must be positive. Thus the expected loss is minimized if the second term in the last equation is zero. 

## Proof: Point Estimation with the Absolute Loss i

$$\begin{aligned}\mathbb{E}_{\theta}[L(\theta, \delta)|D] &= \int_{-\infty}^{\infty} |\theta - \delta| p(\theta|D) d\theta \\&= \int_{-\infty}^{\delta} (\delta - \theta) p(\theta|D) d\theta + \int_{\delta}^{\infty} (\theta - \delta) p(\theta|D) d\theta \\&= \delta P(\delta|D) - \int_{-\infty}^{\delta} \theta p(\theta|D) d\theta \\&\quad + \int_{\delta}^{\infty} \theta p(\theta|D) d\theta - \delta[1 - P(\delta|D)] \\&= 2\delta P(\delta|D) - \delta - 2 \int_{-\infty}^{\delta} \theta p(\theta|D) d\theta + \int_{-\infty}^{\infty} \theta p(\theta|D) d\theta.\end{aligned}$$

## Proof: Point Estimation with the Absolute Loss ii

By applying the integration by part, we have

$$\begin{aligned}\mathbf{E}_\theta[L(\theta, \delta)|D] &= 2\delta P(\delta|D) - \delta \\ &\quad - 2 \left\{ \theta P(\theta|D) \Big|_{-\infty}^{\delta} - \int_{-\infty}^{\delta} P(\theta|D) d\theta \right\} + \mathbf{E}_\theta[\theta|D] \\ &= 2 \int_{-\infty}^{\delta} P(\theta|D) d\theta - \delta + \mathbf{E}_\theta[\theta|D].\end{aligned}$$

The first-order condition for minimization is

$$\nabla_\delta \mathbf{E}_\theta[L(\theta, \delta)|D] = 2P(\delta_\star|D) - 1 = 0.$$

Thus we have  $P(\delta_\star|D) = \frac{1}{2}$ , i.e.,  $\delta_\star$  must be **Median** $_\theta$ . ■

# Proof: Point Estimation with the 0-1 Loss

For any  $\epsilon > 0$ , define

$$L_{\epsilon}(\theta, \delta) = 1 - \mathbb{1}_{[\theta - \epsilon, \theta + \epsilon]}(\delta) = \begin{cases} 0, & (\theta - \epsilon \leq \delta \leq \theta + \epsilon), \\ 1, & (\delta < \theta - \epsilon, \theta + \epsilon < \delta). \end{cases}$$

Note that  $\lim_{\epsilon \rightarrow 0} L_{\epsilon}(\theta, \delta) = 1 - \mathbb{1}_{\theta}(\delta)$ . Then we have

$$\mathbb{E}_{\theta}[L_{\epsilon}(\theta, \delta)|D] = \int_{-\infty}^{\infty} L_{\epsilon}(\theta, \delta)p(\theta|D)d\theta = 1 - \int_{\delta - \epsilon}^{\delta + \epsilon} p(\theta|D)d\theta,$$

which is minimized for  $\delta$  that maximizes

$$\int_{\delta - \epsilon}^{\delta + \epsilon} p(\theta|D)d\theta = \mathbf{P}(\delta - \epsilon \leq \theta \leq \delta + \epsilon|D).$$

For infinitesimally small  $\epsilon$ , such  $\delta$  must be  $\mathbf{Mode}_{\theta}$ . ■

# Posterior Probability

The probability that the true value of  $\theta$  is within a region in the parameter space,  $S_0 \subset \Theta$ , is given by

$$\mathbf{P}(\theta \in S_0|D) = \int_{S_0} p(\theta|D)d\theta.$$

Such a probability is often called the **posterior probability**.

When the region is an interval,  $S_0 = \{\theta : a \leq \theta \leq b\}$ , we have

$$\mathbf{P}(a \leq \theta \leq b|D) = \int_a^b p(\theta|D)d\theta.$$



# Credible Interval (CI)

It is tempting to state that the true value of the parameter must be within an interval with very high posterior probability (say 95%). However, there exist infinitely many intervals with 95% probability because the posterior distribution of the parameter is continuous. Thus we need extra conditions to pin down a unique interval with high posterior probability.

The **credible interval** of  $\theta$  is an interval  $[a_c, b_c]$  such that

1.  $\mathbf{P}(a_c \leq \theta \leq b_c | D) = 1 - c,$
2.  $\mathbf{P}(\theta < a_c | D) = \frac{c}{2}$  and  $\mathbf{P}(\theta > b_c | D) = \frac{c}{2}.$

Set  $c = 0.05$  for the 95% CI.

# Highest Posterior Density Interval (HPDI)

The **highest posterior density interval** of  $\theta$  is an interval  $[a_c, b_c]$  such that

1.  $P(a_c \leq \theta \leq b_c | D) = 1 - c$ ,
2. for any pair  $(\theta, \theta')$  such that  $\theta \in [a_c, b_c]$  and  $\theta' \notin [a_c, b_c]$ ,  $p(\theta | D) > p(\theta' | D)$  must hold.

In particular, if the distribution is unimodal (it has the unique mode), the HPDI must satisfy

$$\begin{aligned} P(a_c \leq \theta \leq b_c | D) &= 1 - c, \\ p(a_c | D) &= p(b_c | D). \end{aligned}$$

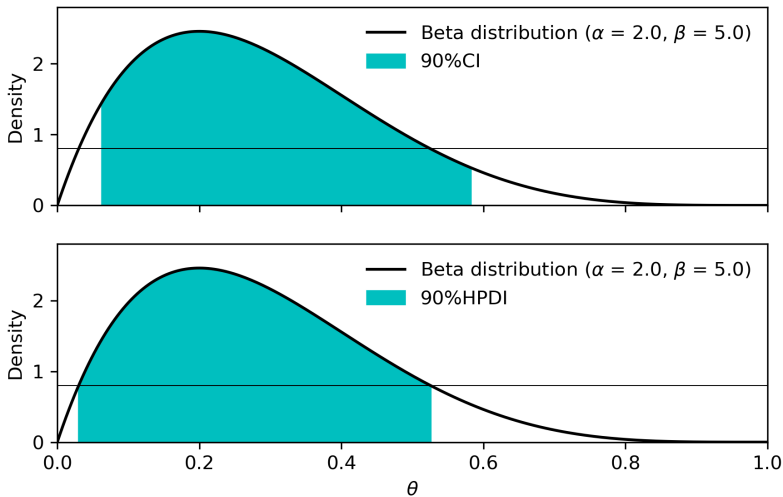


Figure 7: Comparison between CI and HPDI

# Hypotheses On Parameters

In statistics, either Bayesian or frequentist, a hypothesis on the parameters is a region or interval where the true value of the parameter is supposed to be located. For example,

- $\{\theta : 0.5 \leq \theta \leq 1\},$
  - $\theta = 0.5,$
  - $\theta \neq 0.5$
- $$\Leftrightarrow \{\theta : 0 \leq \theta < 0.5\} \cup \{\theta : 0.5 < \theta \leq 1\}.$$

In general, a hypothesis  $H_i$  under which the true value of  $\theta$  is located in a region  $S_i \subset \Theta$  is expressed as

$$H_i : \theta \in S_i, \quad i = 0, 1, 2, \dots$$

# Hypothesis Testing

In Bayesian statistics, plausibility of a hypothesis is measured by the posterior probability that the true value of  $\theta$  is located in  $S_i$ , that is,

$$\mathbf{P}(H_i|D) = \mathbf{P}(\theta \in S_i|D) = \int_{S_i} p(\theta|D)d\theta.$$

Competing hypotheses can be compared by using the **posterior odds ratio**:

$$\text{Posterior odds ratio} = \frac{\mathbf{P}(H_i|D)}{\mathbf{P}(H_j|D)}, \quad i \neq j.$$

Unlike the frequentist approach, Bayesian hypothesis testing does not involve the level of significance, the power and that dreaded P-value!

# Bayes Factor

One catch of the posterior odds ratio is that it is affected by the prior information. If the prior information is biased in favor of one hypothesis, the posterior odds ratio is also biased for that hypothesis.

To control the impact of the prior information, the **Bayes factor** is often used. It is defined as

$$\text{Bayes factor} = \mathbf{B}_{ij} = \frac{\mathbf{P}(H_i|D)}{\mathbf{P}(H_j|D)} \div \frac{\mathbf{P}(H_i)}{\mathbf{P}(H_j)},$$

where  $\mathbf{P}(H_i) = \int_{S_i} p(\theta) d\theta$  and  $\mathbf{P}(H_i)/\mathbf{P}(H_j)$  is called the **prior odds ratio**. Note that the Bayes factor is equivalent to the posterior odds ratio if the prior odds ratio is one.

# Scale Of Bayes Factor By Jeffreys (1961)

We compare  $H_i$  against  $H_j$  ( $i \neq j$ ). We suppose  $H_i$  is the hypothesis we keep unless we have no strong evidence against it.  $H_j$ , on the other hand, is the hypothesis we want to check whether it is supported by the evidence.

Rank	Bayes factor $B_{ij}$	Support for $H_j$
0	$0 < \log_{10}(B_{ij})$	Rejected
1	$-\frac{1}{2} < \log_{10}(B_{ij}) < 0$	Barely worth mentioning
2	$-1 < \log_{10}(B_{ij}) < -\frac{1}{2}$	Substantial
3	$-\frac{3}{2} < \log_{10}(B_{ij}) < -1$	Strong
4	$-2 < \log_{10}(B_{ij}) < -\frac{3}{2}$	Very strong
5	$\log_{10}(B_{ij}) < -2$	Decisive

# Two-Sided Test

On some occasions, we need to check whether the true value of  $\theta$  is exactly equal to a particular value, say **0.5** ( $\theta$  must be **0.5** if the coin is fair). For this purpose, we need to compare  $H_0 : \theta = 0.5$  against  $H_1 : \theta \neq 0.5$ .

As a general setup, we consider

$$\begin{cases} H_0 : & \theta = \theta_0, \\ H_1 : & \theta \neq \theta_0. \end{cases}$$

For these hypothesis, however, it is meaningless to construct the Bayes factor because

$$\mathbf{P}(\theta = \theta_0) = \mathbf{P}(\theta = \theta_0|D) = 0,$$

and prior and posterior odds ratio are identical to zero.



# Spike-And-Slab Prior

To avoid this problem, we introduce a **spike-and-slab prior**:

$$p(\theta) = p_0 \delta(\theta - \theta_0) + (1 - p_0) f(\theta), \quad 0 < p_0 < 1,$$

where  $f(\cdot)$  is a continuous distribution of  $\theta$  and  $\delta(\cdot)$  is the Dirac delta function such that

- for any continuous function  $g(x)$ ,  
$$\int_{-\infty}^{\infty} g(x) \delta(x) dx = g(0);$$
- $$\int_{-\infty}^{\infty} \delta(x) dx = 1;$$
- $\delta(x) = 0$  only if  $x \neq 0$ .

# Savage-Dickey Density Ratio

With the spike-and-slab prior, the prior odds ratio is  $\frac{p_0}{1-p_0}$  and the posterior odds ratio is

$$\text{Posterior odds ratio} = \frac{p_0 f(\theta_0|D)}{(1-p_0)f(\theta_0)},$$

where  $f(\theta|D)$  is the posterior distribution when  $\theta \neq \theta_0$ , i.e.,

$$f(\theta|D) = \frac{p(D|\theta)f(\theta)}{\int_{\Theta} p(D|\theta)f(\theta)d\theta}.$$

Then the Bayes factor is given by

$$B_{01} = \frac{f(\theta_0|D)}{f(\theta_0)},$$

which is called the **Savage-Dickey density ratio** (SDDR).

## Proof: SDDR i

The posterior distribution with the spike-and-slab prior is derived as follows.

$$\begin{aligned} p(\theta|D) &= \frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)d\theta} \\ &= \frac{p(D|\theta) \{p_0\delta(\theta - \theta_0) + (1 - p_0)f(\theta)\}}{\int_{\Theta} p(D|\theta) \{p_0\delta(\theta - \theta_0) + (1 - p_0)f(\theta)\} d\theta} \\ &= \mathcal{K}^{-1} \{p_0p(D|\theta)\delta(\theta - \theta_0) + (1 - p_0)p(D|\theta)f(\theta)\}, \end{aligned}$$

where  $\mathcal{K} = p_0p(D|\theta_0) + (1 - p_0) \int_{\Theta} p(D|\theta)f(\theta)d\theta$ .

## Proof: SDDR ii

The posterior distribution is rewritten as

$$\begin{aligned} p(\theta|D) &= \frac{p_0 p(D|\theta_0)}{\mathcal{K}} \delta(\theta - \theta_0) \\ &\quad + \frac{(1 - p_0) \int_{\Theta} p(D|\theta) f(\theta) d\theta}{\mathcal{K}} \frac{p(D|\theta) f(\theta)}{\int_{\Theta} p(D|\theta) f(\theta) d\theta} \\ &= P(\theta = \theta_0|D) \delta(\theta - \theta_0) + P(\theta \neq \theta_0|D) f(\theta|D), \end{aligned}$$

where

$$P(\theta = \theta_0|D) = \mathcal{K}^{-1} p_0 p(D|\theta_0),$$

$$P(\theta \neq \theta_0|D) = \mathcal{K}^{-1} (1 - p_0) \int_{\Theta} p(D|\theta) f(\theta) d\theta.$$

## Proof: SDDR iii

Thus the posterior odds ratio is

$$\frac{P(\theta = \theta_0|D)}{P(\theta \neq \theta_0|D)} = \frac{p_0}{1-p_0} \times \frac{p(D|\theta_0)}{\int_{\Theta} p(D|\theta)f(\theta)d\theta} = \frac{p_0}{1-p_0} \times \frac{f(\theta_0|D)}{f(\theta_0)}.$$

Therefore the SDDR is obtained as

$$B_{01} = \frac{P(\theta = \theta_0|D)}{P(\theta \neq \theta_0|D)} \div \frac{P(\theta = \theta_0)}{P(\theta \neq \theta_0)} = \frac{f(\theta_0|D)}{f(\theta_0)}.$$



# Predictive Distribution i

Let  $\tilde{x}$  denote an unrealized/future value of the population distribution  $p(x|\theta)$ . Since it is a random variable, we can consider the joint distribution of  $\tilde{x}$  and the previous data  $D = (x_1, \dots, x_n)$ :

$$p(\tilde{x}, x_1, \dots, x_n) = p(\tilde{x}, D),$$

Then, from the definition of the conditional probability, we have

$$p(\tilde{x}, D) = p(\tilde{x}|D)p(D) \quad \Rightarrow \quad p(\tilde{x}|D) = \frac{p(\tilde{x}, D)}{p(D)}.$$

## Predictive Distribution ii

Furthermore, both  $p(D)$  and  $p(\tilde{x}, D)$  are regarded as the marginal likelihood given  $D$  and  $(\tilde{x}, D)$  respectively, that is,

$$p(D) = \int_{\Theta} p(D|\theta)p(\theta)d\theta,$$
$$p(\tilde{x}, D) = \int_{\Theta} p(\tilde{x}, D|\theta)p(\theta)d\theta.$$

In sum, we have

$$p(\tilde{x}|D) = \frac{\int_{\Theta} p(\tilde{x}, D|\theta)p(\theta)d\theta}{\int_{\Theta} p(D|\theta)p(\theta)d\theta}.$$

## Predictive Distribution iii

This is called the **predictive distribution** of  $\tilde{\mathbf{x}}$ . In particular, if  $\tilde{\mathbf{x}}$  and  $\mathbf{D}$  are independent, we have

$$p(\tilde{\mathbf{x}}, \mathbf{D} | \theta) = p(\tilde{\mathbf{x}} | \theta) p(\mathbf{D} | \theta).$$

Thus the predictive distribution of  $\tilde{\mathbf{x}}$  is rearranged as

$$\begin{aligned} p(\tilde{\mathbf{x}} | \mathbf{D}) &= \frac{\int_{\Theta} p(\tilde{\mathbf{x}} | \theta) p(\mathbf{D} | \theta) p(\theta) d\theta}{\int_{\Theta} p(\mathbf{D} | \theta) p(\theta) d\theta} \\ &= \int_{\Theta} p(\tilde{\mathbf{x}} | \theta) \frac{p(\mathbf{D} | \theta) p(\theta)}{\int_{\Theta} p(\mathbf{D} | \theta) p(\theta) d\theta} d\theta \\ &= \int_{\Theta} p(\tilde{\mathbf{x}} | \theta) p(\theta | \mathbf{D}) d\theta. \end{aligned}$$



# Predictive Distribution (Bernoulli Model) i

Let us derive the predictive distribution for the Bernoulli distribution.

$$\begin{aligned} p(\tilde{x}|D) &= \int_{\Theta} p(\tilde{x}|\theta)p(\theta|D)d\theta \\ &= \int_0^1 \theta^{\tilde{x}}(1-\theta)^{1-\tilde{x}} \frac{\theta^{\alpha_{\star}-1}(1-\theta)^{\beta_{\star}-1}}{B(\alpha_{\star},\beta_{\star})} d\theta \\ &= \frac{\int_0^1 \theta^{\tilde{x}+\alpha_{\star}-1}(1-\theta)^{\beta_{\star}-\tilde{x}} d\theta}{B(\alpha_{\star},\beta_{\star})} \\ &= \frac{B(\alpha_{\star} + \tilde{x}, \beta_{\star} - \tilde{x} + 1)}{B(\alpha_{\star}, \beta_{\star})}. \end{aligned}$$

## Predictive Distribution (Bernoulli Model) ii

Using

$$B(\alpha + 1, \beta) = \frac{\alpha}{\alpha + \beta} B(\alpha, \beta),$$

$$B(\alpha, \beta + 1) = \frac{\beta}{\alpha + \beta} B(\alpha, \beta),$$

we have

$$P(\tilde{X} = 1|D) = \frac{B(\alpha_{\star} + 1, \beta_{\star})}{B(\alpha_{\star}, \beta_{\star})} = \frac{\alpha_{\star}}{\alpha_{\star} + \beta_{\star}},$$

$$P(\tilde{X} = 0|D) = \frac{B(\alpha_{\star}, \beta_{\star} + 1)}{B(\alpha_{\star}, \beta_{\star})} = \frac{\beta_{\star}}{\alpha_{\star} + \beta_{\star}}.$$

Finally

$$p(\tilde{x}|D) = \left( \frac{\alpha_{\star}}{\alpha_{\star} + \beta_{\star}} \right)^{\tilde{x}} \left( \frac{\beta_{\star}}{\alpha_{\star} + \beta_{\star}} \right)^{1-\tilde{x}}.$$

This is the Bernoulli distribution with  $\theta = \frac{\alpha_{\star}}{\alpha_{\star} + \beta_{\star}}$ .

# Poisson Distribution

The p.m.f. of a Poisson distribution is

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots, \lambda > 0,$$

$$\mathbf{E}[X] = \mathbf{Var}[X] = \lambda.$$

The Poisson distribution is used to model the number of occurrences in a fixed time interval of rare events such as

- traffic accidents
- crimes
- arrival of customers

# Gamma Distribution

We use a **gamma distribution**

$$\lambda \sim \mathbf{Gamma}(\alpha_0, \beta_0),$$

as the prior of  $\lambda$  in the Poisson distribution.

The p.d.f. of a gamma distribution  $\mathbf{Gamma}(\alpha, \beta)$  is given by

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad x > 0, \alpha > 0, \beta > 0,$$

where  $\Gamma(\cdot)$  is the gamma function:

$$\Gamma(x) = \int_0^\infty z^{x-1} e^{-z} dz.$$

# Likelihood

The likelihood of  $\lambda$  given  $D = (x_1, \dots, x_n)$  is

$$\begin{aligned} p(D|\lambda) &= \prod_{i=1}^n p(x_i|\lambda) \\ &= \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \frac{\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}}{\prod_{i=1}^n x_i!}, \\ &= \frac{\lambda^y e^{-n\lambda}}{\prod_{i=1}^n x_i!}, \quad y = \sum_{i=1}^n x_i. \end{aligned}$$

# Posterior Distribution

Applying Bayes' theorem, we have

$$\begin{aligned} p(\lambda|D) &\propto p(D|\lambda)p(\lambda) \\ &\propto \lambda^y e^{-n\lambda} \times \lambda^{\alpha_0-1} e^{-\beta_0\lambda} \\ &\propto \lambda^{y+\alpha_0-1} e^{-(n+\beta_0)\lambda} \\ &\propto \lambda^{\alpha_\star-1} e^{-\beta_\star\lambda} \\ \alpha_\star &= y + \alpha_0, \quad \beta_\star = n + \beta_0. \end{aligned}$$

This is the gamma distribution **Gamma**( $\alpha_\star, \beta_\star$ ).

# Natural Conjugate Prior

Some of you already noticed the following observations.

- In case of the Bernoulli distribution, if we use a beta distribution as the prior for the probability of success, the posterior is also a beta distribution.
- In case of the Poisson distribution, if we use a gamma distribution as the prior for the mean, the posterior is also a gamma distribution.

A class of prior distributions that make the posterior distribution belong to the same family of the prior distribution is called the **natural conjugate prior** or **conjugate prior**.



# Predictive Distribution

The predictive distribution for the Poisson distribution is

$$p(\tilde{x}|D) = \binom{\tilde{x} + \alpha_{\star} - 1}{\alpha_{\star} - 1} \left( \frac{\beta_{\star}}{\beta_{\star} + 1} \right)^{\alpha_{\star}} \left( \frac{1}{\beta_{\star} + 1} \right)^{\tilde{x}},$$
$$\tilde{x} = 0, 1, 2, \dots$$

This is a **negative-binomial distribution**.

## Proof i

Following the definition of the predictive distribution, we have

$$\begin{aligned} p(\tilde{x}|D) &= \int_0^\infty \frac{\lambda^{\tilde{x}} e^{-\lambda}}{\tilde{x}!} \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \lambda^{\alpha_\star-1} e^{-\beta_\star \lambda} d\lambda \\ &= \frac{\beta_\star^{\alpha_\star}}{\tilde{x}! \Gamma(\alpha_\star)} \int_0^\infty \lambda^{\tilde{x}+\alpha_\star-1} e^{-(\beta_\star+1)\lambda} d\lambda \\ &= \frac{\Gamma(\tilde{x} + \alpha_\star)}{\tilde{x}! \Gamma(\alpha_\star)} \frac{\beta_\star^{\alpha_\star}}{(\beta_\star + 1)^{\tilde{x}+\alpha_\star}}. \end{aligned}$$

## Proof ii

Since  $\Gamma(n) = (n-1)!$ , if  $\alpha_\star$  is a natural number, we obtain

$$\frac{\Gamma(\tilde{x} + \alpha_\star)}{\tilde{x}! \Gamma(\alpha_\star)} = \frac{(\tilde{x} + \alpha_\star - 1)!}{\tilde{x}! (\alpha_\star - 1)!} = \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1}.$$

Therefore

$$\begin{aligned} p(\tilde{x}|D) &= \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1} \frac{\beta_\star^{\alpha_\star}}{(\beta_\star + 1)^{\tilde{x} + \alpha_\star}} \\ &= \binom{\tilde{x} + \alpha_\star - 1}{\alpha_\star - 1} \left( \frac{\beta_\star}{\beta_\star + 1} \right)^{\alpha_\star} \left( \frac{1}{\beta_\star + 1} \right)^{\tilde{x}}. \end{aligned}$$



# Exponential Distribution

The p.d.f. of an exponential distribution  $\text{Exp}(\lambda)$  is

$$p(x|\lambda) = \lambda e^{-\lambda x},$$

$$0 < x < \infty, \quad \lambda > 0,$$

$$\mathbf{E}[X] = \frac{1}{\lambda}, \quad \text{Var}[X] = \frac{1}{\lambda^2}.$$

The exponential distribution is used to model the length of duration.

# Posterior Distribution i

Suppose we have the i.i.d. sample  $D = (x_1, \dots, x_n)$  from  $\mathbf{Exp}(\lambda)$ .

The likelihood is

$$p(D|\lambda) = \prod_{i=1}^n p(x_i|\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}.$$

We use the gamma distribution  $\mathbf{Gamma}(\alpha_0, \beta_0)$  as the prior of  $\lambda$ .

## Posterior Distribution ii

The posterior distribution of  $\lambda$  is derived as

$$\begin{aligned}p(\lambda|D) &\propto p(D|\lambda)p(\lambda) \\&\propto \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \times \lambda^{\alpha_0-1} e^{-\beta_0 \lambda} \\&\propto \lambda^{n+\alpha_0-1} e^{-(\sum_{i=1}^n x_i + \beta_0) \lambda} \\&\propto \lambda^{\alpha_\star-1} e^{-\beta_\star \lambda},\end{aligned}$$

$$\alpha_\star = n + \alpha_0, \quad \beta_\star = \sum_{i=1}^n x_i + \beta_0,$$

which is the gamma distribution **Gamma**( $\alpha_\star, \beta_\star$ ).

# Predictive Distribution

$$\begin{aligned}p(\tilde{x}|D) &= \int_0^\infty p(\tilde{x}|\lambda)p(\lambda|D)d\lambda \\&= \int_0^\infty \lambda e^{-\lambda\tilde{x}} \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \lambda^{\alpha_\star-1} e^{-\beta_\star\lambda} d\lambda \\&= \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \int_0^\infty \lambda^{\alpha_\star} e^{-(\tilde{x}+\beta_\star)\lambda} d\lambda = \frac{\beta_\star^{\alpha_\star}}{\Gamma(\alpha_\star)} \frac{\Gamma(\alpha_\star+1)}{(\tilde{x}+\beta_\star)^{\alpha_\star+1}} \\&= \frac{\alpha_\star \beta_\star^{\alpha_\star}}{(\tilde{x}+\beta_\star)^{\alpha_\star+1}},\end{aligned}$$

which is a Pareto distribution.

# Alternative Parameterization i

In many textbooks, the exponential distribution is defined As

$$p(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}},$$

$$0 < x < \infty, \quad \theta > 0,$$

$$\mathbb{E}[x] = \theta, \quad \text{Var}[X] = \theta^2.$$

Obviously, reparameterization  $\lambda = s(\theta) = 1/\theta$  leads to the aforementioned expression of the exponential distribution.



## Alternative Parameterization ii

We can derive the posterior distribution of  $\theta$  by applying the change-of-variable formula. The Jacobian is

$$|\nabla s(\theta)| = \frac{1}{\theta^2}.$$

Thus the posterior distribution of  $\theta$  is

$$\begin{aligned} p(\theta|D) &= \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} s(\theta)^{\alpha^*-1} \exp[-\beta^* s(\theta)] |\nabla s(\lambda)| \\ &= \frac{\beta^{*\alpha^*}}{\Gamma(\alpha^*)} \theta^{-(\alpha^*+1)} \exp\left(-\frac{\beta^*}{\theta}\right), \end{aligned}$$

which is called the **inverse Gamma distribution**.

## Alternative Parameterization iii

In general, the p.d.f. of the inverse Gamma distribution **Inv.Gamma**( $\alpha, \beta$ ) is given by

$$p(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right), \quad x > 0, \quad \alpha > 0, \quad \beta > 0.$$

Note that

$$X \sim \text{Gamma}(\alpha, \beta) \quad \Rightarrow \quad \frac{1}{X} \sim \text{Inv.Gamma}(\alpha, \beta).$$

# Normal Distribution

The p.d.f. of a normal distribution  $\text{Normal}(\mu, \sigma^2)$  is

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right],$$

$$-\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma^2 > 0,$$

$$\mathbf{E}[X] = \mu, \quad \mathbf{Var}[X] = \sigma^2.$$

- The normal distribution is the mainstay of statistics.
- Many economic data are supposed to follow a normal distribution, though it is not the case for financial data.
- Many sophisticated statistical models are built upon the normal distribution.
- The normal distribution is often a limit of the other distribution — the central limit theorem.

# Normal-Inverse-Gamma Distribution i

The natural conjugate prior for  $(\mu, \sigma^2)$  is

$$\mu|\sigma^2 \sim \text{Normal}\left(\mu_0, \frac{\sigma^2}{n_0}\right), \sigma^2 \sim \text{Inv.Gamma}\left(\frac{\nu_0}{2}, \frac{\lambda_0}{2}\right),$$

which is called the **normal-inverse-gamma distribution**:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2),$$

$$p(\mu|\sigma^2) = \sqrt{\frac{n_0}{2\pi\sigma^2}} \exp\left[-\frac{n_0(\mu - \mu_0)^2}{2\sigma^2}\right],$$

$$p(\sigma^2) = \frac{\left(\frac{\lambda_0}{2}\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} (\sigma^2)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\lambda_0}{2\sigma^2}\right).$$

# Likelihood i

The likelihood of  $(\mu, \sigma^2)$  is

$$\begin{aligned} p(D|\mu, \sigma^2) &= \prod_{i=1}^n p(x_i|\mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right]. \end{aligned}$$

## Likelihood ii

Since

$$\begin{aligned}\sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 \\&= \sum_{i=1}^n \left\{ (x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2 \right\} \\&= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2,\end{aligned}$$

the likelihood is rewritten as

$$p(D|\mu, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right].$$

# Posterior Distribution Of $(\mu, \sigma^2)$

Applying Bayes' theorem, we have

$$\begin{aligned} & p(\mu, \sigma^2 | D) \\ & \propto p(D | \mu, \sigma^2) p(\mu | \sigma^2) p(\sigma^2) \\ & \propto (\sigma^2)^{-\frac{n}{2}} \exp \left[ -\frac{\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2} \right] \\ & \quad \times (\sigma^2)^{-\frac{1}{2}} \exp \left[ -\frac{n_0(\mu - \mu_0)^2}{2\sigma^2} \right] \times (\sigma^2)^{-\left(\frac{\nu_0}{2} + 1\right)} \exp \left[ -\frac{\lambda_0}{2\sigma^2} \right] \\ & \propto (\sigma^2)^{-\frac{n+\nu_0+3}{2}} \exp \left[ -\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right. \right. \\ & \quad \left. \left. + n_0(\mu - \mu_0)^2 + \lambda_0 \right\} \right]. \end{aligned}$$

# Completing The Square

“Completing the square” is referred to as a transformation of a quadratic functions:

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}.$$

By completing the square, we have

$$\begin{aligned} & n(\bar{x} - \mu)^2 + n_0(\mu - \mu_0)^2 \\ &= (n + n_0)\mu^2 - 2(n\bar{x} + n_0\mu_0)\mu + n\bar{x}^2 + n_0\mu_0^2 \\ &= (n + n_0)\left(\mu - \frac{n\bar{x} + n_0\mu_0}{n + n_0}\right)^2 + \frac{nn_0}{n + n_0}(\mu_0 - \bar{x})^2. \end{aligned}$$



# Joint Posterior Distribution i

Therefore the joint posterior distribution of  $(\mu, \sigma^2)$  is

$$p(\mu, \sigma^2 | D) \propto (\sigma^2)^{-\frac{1}{2}} \exp \left[ -\frac{n_{\star}(\mu - \mu_{\star})^2}{2\sigma^2} \right] \\ \times (\sigma^2)^{-\left(\frac{\nu_{\star}}{2} + 1\right)} \exp \left( -\frac{\lambda_{\star}}{2\sigma^2} \right),$$

where

$$\mu_{\star} = \frac{n\bar{x} + n_0\mu_0}{n + n_0}, \quad n_{\star} = n + n_0, \quad \nu_{\star} = n + \nu_0, \\ \lambda_{\star} = \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{nn_0}{n + n_0} (\mu_0 - \bar{x})^2 + \lambda_0.$$

## Joint Posterior Distribution ii

This is also a normal-inverse-gamma distribution.

$$\begin{aligned}\mu|\sigma^2, D &\sim \text{Normal}\left(\mu_\star, \frac{\sigma^2}{n_\star}\right), \\ \sigma^2|D &\sim \text{Inv.Gamma}\left(\frac{\nu_\star}{2}, \frac{\lambda_\star}{2}\right).\end{aligned}$$

# Integrating Out Nuisance Parameters

To proceed to posterior inference on the mean  $\mu$ , we need the marginal posterior distribution  $p(\mu|D)$ . This is obtained by integrating out the other parameter  $\sigma^2$  as

$$\begin{aligned} p(\mu|D) &= \int_0^\infty p(\mu, \sigma^2|D) d\sigma^2 \\ &= \int_0^\infty p(\mu|\sigma^2, D) p(\sigma^2|D) d\sigma^2. \end{aligned}$$

A parameter that is not the primary subject of our analysis is called the nuisance parameter.

# Marginal Posterior Distribution Of $\mu$

The marginal posterior distribution of  $\mu$  is derived as

$$\begin{aligned} p(\mu|D) &= \int_0^\infty p(\mu|\sigma^2, D)p(\sigma^2|D)d\sigma^2 \\ &= \sqrt{\frac{n_\star}{2\pi}} \frac{\left(\frac{\lambda_\star}{2}\right)^{\frac{\nu_\star}{2}}}{\Gamma\left(\frac{\nu_\star}{2}\right)} \int_0^\infty (\sigma^2)^{-\left(\frac{\nu_\star+1}{2}+1\right)} \exp\left[-\frac{n_\star(\mu - \mu_\star)^2 + \lambda_\star}{2\sigma^2}\right] d\sigma^2 \\ &= \frac{\sqrt{n_\star}\lambda_\star^{\frac{\nu_\star}{2}} 2^{-\frac{\nu_\star+1}{2}} \Gamma\left(\frac{\nu_\star+1}{2}\right)}{\Gamma\left(\frac{\nu_\star}{2}\right) \sqrt{\pi}} \left[\frac{n_\star(\mu - \mu_\star)^2 + \lambda_\star}{2}\right]^{-\frac{\nu_\star+1}{2}} \\ &= \frac{\Gamma\left(\frac{\nu_\star+1}{2}\right)}{\Gamma\left(\frac{\nu_\star}{2}\right)} \sqrt{\frac{n_\star}{\pi\lambda_\star}} \left[1 + \frac{n_\star(\mu - \mu_\star)^2}{\lambda_\star}\right]^{-\frac{\nu_\star+1}{2}}. \end{aligned}$$

## Marginal Posterior Distribution Of $\mu$ ii

The integral is evaluated with the following formula:

$$\int_0^{\infty} x^{-(\alpha+1)} e^{-\frac{\beta}{x}} dx = \beta^{-\alpha} \Gamma(\alpha).$$

The marginal posterior distribution  $p(\mu|D)$  is a (Student's)  $t$ -distribution:

$$\mu|D \sim t(\nu_{\star}, \mu_{\star}, \tau_{\star}^2), \quad \tau_{\star}^2 = \frac{\lambda_{\star}}{\nu_{\star} n_{\star}},$$

In general, the p.d.f. of the  $t$ -distribution  $t(\nu, \mu, \sigma^2)$  is

$$p(x|\nu, \mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2}) \sqrt{\pi\nu\sigma^2}} \left[ 1 + \frac{(x-\mu)^2}{\nu\sigma^2} \right]^{-\frac{\nu+1}{2}}.$$

# Student's $t$ -Distribution

The  $t$ -distribution in introductory statistics is defined as

$$T = \frac{Z}{\sqrt{V/\nu}}, \quad Z \sim \text{Normal}(0, 1), \quad V \sim \chi^2(\nu), \quad Z \perp V,$$

where  $\chi^2(\nu)$  represents the  $\chi^2$ -distribution with degree of freedom  $\nu$  and “ $\perp$ ” implies independence. In our notation,  $T \sim t(\nu, 0, 1)$ .

In the same manner,  $T \sim t(\nu, \mu, \sigma^2)$  is defined as

$$T = \mu + \frac{\sigma Z}{\sqrt{V/\nu}}, \quad Z \sim \text{Normal}(0, 1), \quad V \sim \chi^2(\nu), \quad Z \perp V.$$

# Bayesian Inference On $\mu$

1. The posterior mean, median and mode are equal to  $\mu_\star$ .
2. The  $100 \times (1 - c)\%$  credible interval is expressed as

$$\mu_\star - t_{\frac{c}{2}} \tau_\star \leq \mu \leq \mu_\star + t_{\frac{c}{2}} \tau_\star,$$

where  $t_{\frac{c}{2}}$  is the critical value that satisfies

$$\mathbf{P} \left( T > t_{\frac{c}{2}} \right) = \frac{c}{2}, \quad T \sim \mathbf{t}(\nu_\star),$$

3. The credible interval is identical to the HPDI since the  $\mathbf{t}$  distribution is symmetric around  $\mu_\star$ .

# Predictive Distribution i

The predictive distribution for the normal distribution is

$$\tilde{x}|D \sim t(\nu_\star, \mu_\star, \tau_\star^2(1 + n_\star)).$$

*Proof:* We must evaluate

$$p(\tilde{x}|D) = \int_0^\infty \underbrace{\int_{-\infty}^\infty p(\tilde{x}|\mu, \sigma^2) p(\mu|\sigma^2, D) d\mu}_{g(\sigma^2)} p(\sigma^2|D) d\sigma^2.$$



## Predictive Distribution ii

First we derive the closed form of the integral:

$$\begin{aligned}g(\sigma^2) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\tilde{x} - \mu)^2}{2\sigma^2}\right] \sqrt{\frac{n_{\star}}{2\pi\sigma^2}} \exp\left[-\frac{n_{\star}(\mu - \mu_{\star})^2}{2\sigma^2}\right] d\mu \\&= \frac{\sqrt{n_{\star}}}{2\pi\sigma^2} \int_{-\infty}^{\infty} \exp\left[-\frac{(\tilde{x} - \mu)^2 + n_{\star}(\mu - \mu_{\star})^2}{2\sigma^2}\right] d\mu.\end{aligned}$$

Note that, by completing the square, we have

$$\begin{aligned}&(\tilde{x} - \mu)^2 + n_{\star}(\mu - \mu_{\star})^2 \\&= (1 + n_{\star})\mu^2 - 2(\tilde{x} + n_{\star}\mu_{\star})\mu + \tilde{x}^2 + n_{\star}\mu_{\star}^2 \\&= (1 + n_{\star})\left(\mu - \frac{\tilde{x} + n_{\star}\mu_{\star}}{1 + n_{\star}}\right)^2 + \frac{n_{\star}}{1 + n_{\star}}(\tilde{x} - \mu_{\star})^2.\end{aligned}$$

## Predictive Distribution iii

Thus

$$\begin{aligned} g(\sigma^2) &= \sqrt{\frac{n_{\star}}{2\pi\sigma^2(1+n_{\star})}} \exp\left[-\frac{n_{\star}(\tilde{x} - \mu_{\star})^2}{2\sigma^2(1+n_{\star})}\right] \\ &\quad \times \underbrace{\int_{-\infty}^{\infty} \sqrt{\frac{1+n_{\star}}{2\pi\sigma^2}} \exp\left[-\frac{1+n_{\star}}{2\sigma^2} \left(\mu - \frac{\tilde{x} + n_{\star}\mu_{\star}}{1+n_{\star}}\right)^2\right] d\mu}_1 \\ &= \sqrt{\frac{n_{\star}}{2\pi\sigma^2(1+n_{\star})}} \exp\left[-\frac{n_{\star}(\tilde{x} - \mu_{\star})^2}{2\sigma^2(1+n_{\star})}\right]. \end{aligned}$$

Note:  $g(\sigma^2)$  is the p.d.f. of **Normal**  $\left(\mu_{\star}, \sigma^2 \frac{1+n_{\star}}{n_{\star}}\right)$ .

# Predictive Distribution iv

Substituting this result for  $g(\sigma^2)$ , we have

$$\begin{aligned} p(\tilde{X}|D) &= \sqrt{\frac{n_{\star}}{2\pi(1+n_{\star})}} \frac{\left(\frac{\lambda_{\star}}{2}\right)^{\frac{\nu_{\star}}{2}}}{\Gamma\left(\frac{\nu_{\star}}{2}\right)} \\ &\quad \times \int_0^{\infty} (\sigma^2)^{-\left(\frac{\nu_{\star}+1}{2}+1\right)} \exp\left[-\frac{\frac{n_{\star}}{1+n_{\star}}(\tilde{X}-\mu_{\star})^2 + \lambda_{\star}}{2\sigma^2}\right] d\sigma^2 \\ &= \frac{\Gamma\left(\frac{\nu_{\star}+1}{2}\right)}{\Gamma\left(\frac{\nu_{\star}}{2}\right)} \sqrt{\frac{n_{\star}}{\pi\lambda_{\star}(1+n_{\star})}} \left[1 + \frac{n_{\star}(\tilde{X}-\mu_{\star})^2}{\lambda_{\star}(1+n_{\star})}\right]^{-\frac{\nu_{\star}+1}{2}} \\ &= \frac{\Gamma\left(\frac{\nu_{\star}+1}{2}\right)}{\Gamma\left(\frac{\nu_{\star}}{2}\right) \sqrt{\pi\nu_{\star}\tau_{\star}^2(1+n_{\star})}} \left[1 + \frac{(\tilde{X}-\mu_{\star})^2}{\nu_{\star}\tau_{\star}^2(1+n_{\star})}\right]^{-\frac{\nu_{\star}+1}{2}}, \quad \tau_{\star}^2 = \frac{\lambda_{\star}}{\nu_{\star}n_{\star}}. \blacksquare \end{aligned}$$

# Summary Of Bayesian Analysis

- $\theta$  —  $m \times 1$  vector of unknown parameters
- $D$  — data
- $p(x|\theta)$  — population distribution
- $p(\theta)$  — prior distribution
- $p(D|\theta)$  — likelihood
- $p(\theta|D)$  — posterior distribution

## Bayes' Theorem

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\mathbb{R}^m} p(D|\theta)p(\theta)d\theta}.$$

# Difficulties In Bayesian Analysis

1. In Bayesian analysis, we are required to evaluate multiple integrals to proceed statistical inference.
2. For example, we need  $\int_{\mathbb{R}^m} p(D|\theta)p(\theta)d\theta$  to derive the exact expression of the posterior distribution, but it is not available except for a limited number of examples (e.g., natural conjugate prior).
3. It is the case for the following quantities:
  - Marginal distribution:  $p(\theta_j|D) = \int_{\mathbb{R}^{m-1}} p(\theta|D)d\theta_{-j}$ .
  - Mean:  $\mathbf{E}_{\theta}[\theta_j|D] = \int_{-\infty}^{\infty} \theta_j p(\theta_j|D)d\theta_j$ .
  - Median:  $\int_{-\infty}^{\text{Median}_{\theta_j}} p(\theta_j|D)d\theta_j = \frac{1}{2}$ .
  - Probability:  $\mathbf{P}(a \leq \theta_j \leq b|D) = \int_a^b p(\theta_j|D)d\theta_j$ .
  - Predictive dist.:  $p(\tilde{x}|D) = \int_{\mathbb{R}^m} p(\tilde{x}|\theta)p(\theta|D)d\theta$ .

# MAP Estimation

One exception is the posterior mode:

$$\text{Mode}_\theta = \arg \max_{\theta \in \mathbb{R}^m} p(\theta|D).$$

Note that the posterior mode is equivalent to

$$\text{MAP}_\theta = \arg \max_{\theta \in \mathbb{R}^m} p(D|\theta)p(\theta),$$

which does not depend on the normalizing constant.

Since the exact expressions of the prior distribution and the likelihood are known in many applications, we can solve the above optimization problem without evaluating the normalizing constant.

The posterior mode is often called **MAP (maximum a posteriori) estimate**.

# Monte Carlo Methods i

Monte Carlo methods are widely applied numerical techniques to evaluate integrals. Suppose we need to evaluate the following expectation:

$$\mathbf{E}_{\theta}[h(\theta)] = \int_{\mathbb{R}^m} h(\theta) p(\theta|D) d\theta.$$

Examples:

- Mean  $\mathbf{E}_{\theta}[\theta_j|D]$ :  $h(\theta) = \theta_j$ .
- Probability  $\mathbf{P}(a \leq \theta_j \leq b|D)$ :  $h(\theta) = \mathbb{1}_{[a,b]}(\theta_j)$ .
- Predictive distribution  $p(\tilde{x}|D)$ :  $h(\theta) = p(\tilde{x}|\theta)$ .

## Monte Carlo Methods ii

Suppose we have a set of pseudo-random numbers generated from the posterior distribution  $p(\theta|D)$ ,  $\{\theta^{(1)}, \dots, \theta^{(T)}\}$ , which is called the Monte Carlo sample, and define

$$\hat{h} = \frac{1}{T} \sum_{t=1}^T h(\theta^{(t)}).$$

If the law of large numbers holds,

$$\hat{h} \rightsquigarrow \mathbf{E}_{\theta}[h(\theta)] \quad \text{as } T \rightarrow \infty,$$

where “ $\rightsquigarrow$ ” means the convergence in probability.



## Monte Carlo Methods iii

Therefore, when  $T$  is sufficiently large,  $\mathbf{E}_{\theta}[\mathbf{h}(\theta)]$  can be well approximated with  $\hat{\mathbf{h}}$ . This is the basic idea behind the Monte Carlo integration method.

Monte Carlo approximations of the posterior statistics are given as follows:

- Mean:  $\hat{\mathbf{E}}_{\theta}[\theta_j|D] = \frac{1}{T} \sum_{t=1}^T \theta_j^{(t)}$ .
- Probability:  $\hat{\mathbf{P}}(a \leq \theta_j \leq b|D) = \frac{1}{T} \sum_{t=1}^T \mathbb{1}_{[a,b]}(\theta_j^{(t)})$ .
- Predictive distribution:  $\hat{p}(\tilde{x}|D) = \frac{1}{T} \sum_{t=1}^T p(\tilde{x}|\theta^{(t)})$ .

Note that the Monte Carlo method merely gives us an approximated value of the true posterior statistic (the true one is obtained when  $T \rightarrow \infty$ ). So it is contaminated with numerical errors. These errors are measured by

## Numerical errors in Monte Carlo approximation

$$\text{SE}[\hat{h}] = \sqrt{\frac{1}{T(T-1)} \sum_{t=1}^T \left( h(\theta^{(t)}) - \hat{h} \right)^2}.$$

# Kernel Density Estimation

The marginal posterior p.d.f.  $p(\theta_j|D)$  can be evaluated with a kernel density estimation method:

$$\hat{p}(\theta_j) = \frac{1}{Tw} \sum_{t=1}^T K\left(\frac{\theta_j - \theta_j^{(t)}}{w}\right),$$

where  $K(\cdot)$  is a kernel function. In practice, the p.d.f. of the standard normal distribution:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}},$$

is often used as  $K(\cdot)$ , which is called the Gaussian kernel.  $w$  is the bandwidth and determines the smoothness of the evaluated density function.

# Monte Carlo Approximation Of Quantiles i

## Quantile

The  $100q\%$  quantile  $\theta_j^{[q]}$  is defined as

$$\mathbf{P} \left( \theta_j \leq \theta_j^{[q]} \middle| D \right) = q, \quad 0 < q < 1.$$

We can approximate the quantile  $\theta_j^{[q]}$  with

## Sample Quantile

$$\hat{\theta}_j^{[q]} = \max_{1 \leq t \leq T} \theta_j^{(t)} \quad \text{s.t.} \quad \frac{1}{T} \sum_{s=1}^T \mathbb{1}_{(-\infty, \theta_j^{(t)})} \left( \theta_j^{(s)} \right) \leq q,$$

# Monte Carlo Approximation Of Quantiles ii

For example, when  $T = 10,000$ ,  $\hat{\theta}_j^{[0.05]}$  is the 500th smallest value in the Monte Carlo sample.

- Median:  $\hat{\theta}_j^{[0.5]}$ .
- $100(1 - q)\%$  credible interval:  $\hat{\theta}_j^{[\frac{q}{2}]} \leq \theta_j \leq \hat{\theta}_j^{[1 - \frac{q}{2}]}$ .
- $100(1 - q)\%$  HPDI [Chen and Shao (1999)]:

$$\hat{\theta}_j^{[\frac{t^*}{T}]} \leq \theta_j \leq \hat{\theta}_j^{[1 - q + \frac{t^*}{T}]}, \quad 1 \leq t \leq qT,$$

where

$$t^* = \arg \min_{1 \leq t \leq qT} \left| \hat{\theta}_j^{[\frac{t}{T}]} - \hat{\theta}_j^{[1 - q + \frac{t}{T}]} \right|.$$

# Markov Chain Monte Carlo (MCMC)

- So far we suppose we have the Monte Carlo sample generated from the posterior distribution.
- However it is not obvious how to generate random numbers from an unknown distribution.
- For this purpose, we apply **Markov chain sampling**.
- Markov chain sampling + Monte Carlo integration = **Markov chain Monte Carlo (MCMC)**
- MCMC is indispensable for Bayesian statistics.
- To understand MCMC, we need to know a Markov chain.

# Markov Chain Of Continuous Random Variables

Consider a sequence of continuous random variables  $\{X_t\}_{t=0}^{\infty}$ .  $X_t$  takes a real value in  $\mathcal{X} \subseteq \mathbb{R}$ .

$\{X_t\}_{t=0}^{\infty}$  is called a **Markov chain** if, given  $\{x_s\}_{s=0}^{t-1}$ , the conditional probability that  $X_t$  takes a real value in  $A \subseteq \mathcal{X}$  is expressed as

$$\begin{aligned} P(X_t \in A | X_0 = x_0, \dots, X_{t-1} = x_{t-1}) \\ = P(X_t \in A | X_{t-1} = x_{t-1}). \end{aligned}$$

# Properties i

- Time-homogeneity

For any  $A \subseteq \mathcal{X}$ ,  $x \in \mathcal{X}$ ,  $t \geq 0$ ,

$$\mathbf{P}(X_{t+1} \in A | X_t = x) = \mathbf{P}(X_t \in A | X_{t-1} = x).$$

- Regularity

Suppose  $\int_A f(x) dx > 0$  for any  $A \subseteq \mathcal{X}$ . A Markov chain  $\{X_t\}_{t=0}^\infty$  is **regular** with respect to  $f$  or  $f$ -regular if there exists a finite  $t \geq 1$  such that

$$\mathbf{P}(X_t \in A | X_0 = x) > 0 \text{ for any } x \in \mathcal{X}.$$



## Properties ii

- Aperiodicity

Consider any  $t \geq 1$  satisfies  $\mathbf{P}(X_t \in A | X_0 = x) > 0$  for any  $x \in A \subseteq \mathcal{X}$ . The greatest common divisor of such  $t$  is called the **period**. If the period is one for any  $A \subseteq \mathcal{X}$ , a Markov chain  $\{X_t\}_{t=0}^{\infty}$  is **aperiodic**.

- Recurrence

Define  $\tau_A = \inf\{t > 0 : X_t \in A\}$ .  $A$  is **recurrent** if  $\mathbf{P}(\tau_A < \infty | X_0 = x) = 1$  for any  $x \in A$ .  $\{X_t\}_{t=0}^{\infty}$  is **recurrent** with respect to  $f$  if  $A$  is recurrent for a  $f$ -regular Markov chain  $\{X_t\}_{t=0}^{\infty}$  and  $\int_A f(x) dx > 0$ .

# Transition Kernel i

The conditional p.d.f. of  $X_t$  given  $X_0, \dots, X_{t-1}$  is

$$f_t(x_t|x_0, \dots, x_{t-1}) = f(x_t|x_{t-1}),$$

which is called the transition kernel, and the right-hand side is often expressed as  $K(x_{t-1}, x_t)$ .

Suppose  $f_0(x_0)$  is the p.d.f. of the initial distribution.

Then the joint p.d.f. of  $\{X_s\}_{s=0}^t$  is

$$\begin{aligned} f(x_0, \dots, x_t) &= f_0(x_0)f_t(x_1|x_0)f_2(x_2|x_0, x_1) \times \dots \\ &\times f_t(x_t|x_0, \dots, x_{t-1}) = f_0(x_0) \prod_{s=1}^t K(x_{s-1}, x_s). \end{aligned}$$

## Transition Kernel ii

Note that

$$\begin{aligned}f_t(x_t) &= \int_{\mathcal{X}} f(x_{t-1}, x_t) dx_{t-1} \\&= \int_{\mathcal{X}} f(x_t | x_{t-1}) f_{t-1}(x_{t-1}) dx_{t-1} \\&= \int_{\mathcal{X}} f_{t-1}(x_{t-1}) K(x_{t-1}, x_t) dx_{t-1}.\end{aligned}$$

Define

$$f_t = f_{t-1} \circ K = \int_{\mathcal{X}} f_{t-1}(x_{t-1}) K(x_{t-1}, x_t) dx_{t-1}.$$

## Transition Kernel iii

Then

$$\begin{aligned}f_t &= f_{t-1} \circ K = \left\{ \int_{\mathcal{X}} f_{t-2}(x_{t-2}) K(x_{t-2}, x_{t-1}) dx_{t-2} \right\} \circ K \\&= \int_{\mathcal{X}} \left\{ \int_{\mathcal{X}} f_{t-2}(x_{t-2}) K(x_{t-2}, x_{t-1}) dx_{t-2} \right\} K(x_{t-1}, x_t) dx_{t-1} \\&= \int_{\mathcal{X}} f_{t-2}(x_{t-2}) \underbrace{\int_{\mathcal{X}} K(x_{t-2}, x_{t-1}) K(x_{t-1}, x_t) dx_{t-1}}_{K \circ K} dx_{t-2} \\&= f_{t-2} \circ (K \circ K) = f_{t-2} \circ K^2 \quad \dots \quad = f_0 \circ K^t, \\K^t &= \int_{\mathcal{X}} \dots \int_{\mathcal{X}} K(x_0, x_1) \dots K(x_{t-1}, x_t) dx_1 \dots dx_{t-1}.\end{aligned}$$

## Example: AR(1) Process i

AR(1) process

$$X_t = \phi X_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2), \quad |\phi| < 1, \quad \phi \neq 0.$$

This a Markov chain with the transition kernel:

$$f(x_t|x_{t-1}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_t - \phi x_{t-1})^2}{2\sigma^2} \right],$$

which is the p.d.f. of the normal distribution

**Normal**( $\phi x_{t-1}, \sigma^2$ ).

## Example: AR(1) Process ii

The joint p.d.f. of  $\{X_s\}_{s=0}^t$  is

$$\begin{aligned} f(x_0, \dots, x_t) &= f_0(x_0) \prod_{s=1}^t f(x_s | x_{s-1}) \\ &= f_0(x_0) \prod_{s=1}^t \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x_s - \phi x_{s-1})^2}{2\sigma^2} \right]. \end{aligned}$$

# Invariant Distribution

The **invariant distribution (density)**  $\bar{f}$  of a Markov chain with kernel  $K$  is

$$\bar{f}(\tilde{x}) = \int_{\mathcal{X}} \bar{f}(x) K(x, \tilde{x}) dx \quad \text{or} \quad \bar{f} = \bar{f} \circ K.$$

If a Markov chain is recurrent and aperiodic,

**Ergodicity:**  $\lim_{t \rightarrow \infty} \sup_{A \subseteq \mathcal{X}} \left| \int_A (f_t(x) - \bar{f}(x)) dx \right| = 0,$

**LLN:**  $\frac{1}{T} \sum_{t=1}^T h(X_t) \rightsquigarrow \int_{\mathcal{X}} h(x) \bar{f}(x) dx.$

**Detailed Balance Condition (DBC)**

$$\bar{f}(x) K(x, \tilde{x}) = \bar{f}(\tilde{x}) K(\tilde{x}, x), \quad \forall x, \tilde{x} \in \mathcal{X}.$$

# Proof Of DBC

By integrating both sides of DBC with respect to  $x$ , we have

$$\begin{aligned}\int_{\mathcal{X}} \bar{f}(x) K(x, \tilde{x}) dx &= \int_{\mathcal{X}} \bar{f}(\tilde{x}) K(\tilde{x}, x) dx \\ &= \bar{f}(\tilde{x}) \int_{\mathcal{X}} K(\tilde{x}, x) dx \\ &= \bar{f}(\tilde{x}),\end{aligned}$$

that is,

$$\bar{f}(\tilde{x}) = \int_{\mathcal{X}} \bar{f}(x) K(x, \tilde{x}) dx,$$

holds. This means that  $\bar{f}$  is the invariant density of  $K$ . ■



## Example: AR(1) Process i

The invariant distribution of an AR(1) process is

$$\text{Normal}\left(0, \frac{\sigma^2}{1 - \phi^2}\right),$$

with the p.d.f.:

$$\bar{f}(x) = \sqrt{\frac{1 - \phi^2}{2\pi\sigma^2}} \exp\left[-\frac{(1 - \phi^2)x^2}{2\sigma^2}\right].$$

It suffices to show that the above p.d.f. satisfies DBC.

## Example: AR(1) Process ii

$$\begin{aligned}\bar{f}(x)K(x, \tilde{x}) &= \frac{\sqrt{1-\phi^2}}{2\pi\sigma^2} \exp \left[ -\frac{(1-\phi^2)x^2 + (\tilde{x} - \phi x)^2}{2\sigma^2} \right] \\&= \frac{\sqrt{1-\phi^2}}{2\pi\sigma^2} \exp \left[ -\frac{x^2 - \phi^2 x^2 + \tilde{x}^2 - 2\phi x\tilde{x} + \phi^2 x^2}{2\sigma^2} \right] \\&= \frac{\sqrt{1-\phi^2}}{2\pi\sigma^2} \exp \left[ -\frac{\tilde{x}^2 - \phi^2 \tilde{x}^2 + x^2 - 2\phi x\tilde{x} + \phi^2 \tilde{x}^2}{2\sigma^2} \right] \\&= \frac{\sqrt{1-\phi^2}}{2\pi\sigma^2} \exp \left[ -\frac{(1-\phi^2)\tilde{x}^2 + (x - \phi\tilde{x})^2}{2\sigma^2} \right] \\&= \bar{f}(\tilde{x})K(\tilde{x}, x).\end{aligned}$$

## Example: AR(1) Process iii

As a numerical illustration, we simulate the following AR(1) process:

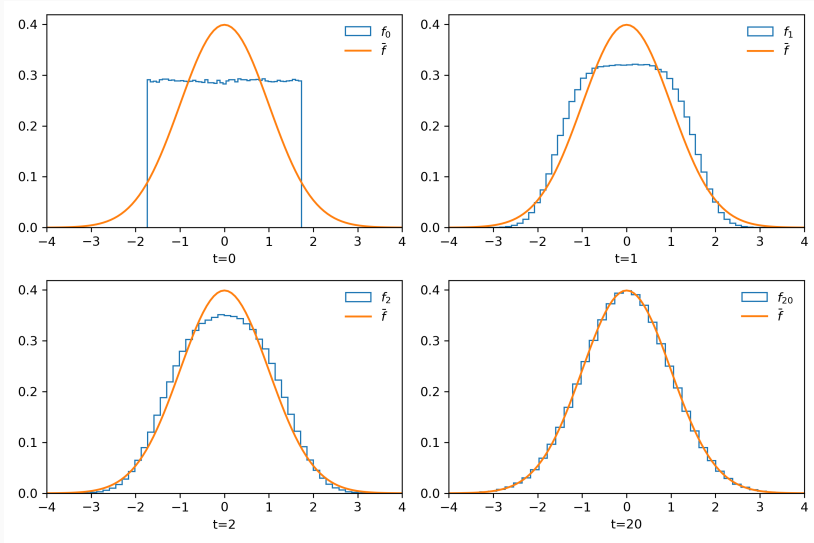
$$x_0 \sim \text{Uniform}(-\sqrt{3}, \sqrt{3}),$$

$$x_t = 0.9x_{t-1} + \epsilon_t, \epsilon_t \sim \text{Normal}(0, 0.19), t = 1, 2, \dots$$

The invariant distribution is the standard normal distribution **Normal**(0, 1) since

$$\frac{\sigma^2}{1 - \phi^2} = \frac{0.19}{1 - 0.9^2} = 1.$$

# Example: AR(1) Process iv



# Random Number Generation From A Markov Chain

## Markov chain sampling

Step 1. Set  $t = 1$  and  $\tilde{x}_0 \leftarrow f_0(x_0)$ .

Step 2.  $\tilde{x}_t \leftarrow K(\tilde{x}_{t-1}, x_t)$ .

Step 3. Increase  $t$  by 1 and go to Step 2.

- Suppose we can generate  $\tilde{x}_t$  from a recurrent and aperiodic Markov chain with  $K$  and  $\bar{f}$ .
- After we repeat Step 1–3 sufficiently many times (say  $M$ ), the distribution of  $\tilde{x}_t$  will be very close to  $\bar{f}$ .
- $\mathbf{E}[h(X)] = \int_{\mathcal{X}} h(x) \bar{f}(x) dx$  can be approximated by  $\frac{1}{N} \sum_{t=M+1}^{M+N} h(\tilde{x}_t)$  due to ergodicity of the Markov chain.
- The first  $M$  runs of the Markov chain sampling is called the **burnin**.

# Convergence Diagnostics

Unfortunately Markov chain sampling does not always work, even though the distribution of generated random numbers must converge to the posterior distribution in theory.

To make it work properly, we need careful tuning and monitoring of the sampling algorithm. So we discuss how to check the convergence of generated chains.

- $n$  — the number of draws
- $m$  — the number of chains
- $\theta_{ij}$  — draw  $i$  in chain  $j$  ( $i = 1, \dots, n$ ,  $j = 1, \dots, m$ )
- $\{\theta_1, \dots, \theta_T\}$  — single-chain Monte Carlo sample

# Gelman-Rubin Convergence Diagnostic

Gelman-Rubin convergence diagnostic  $\hat{R}$

$$\hat{R} = \sqrt{\frac{\hat{V}}{\hat{W}}}, \quad \hat{V} = \frac{n-1}{n}\hat{W} + \frac{1}{n}\hat{B},$$

$$\hat{B} = \frac{n}{m-1} \sum_{j=1}^m (\bar{\theta}_{\cdot j} - \bar{\theta})^2, \quad \bar{\theta}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n \theta_{ij},$$

$$\hat{W} = \frac{1}{m} \sum_{j=1}^m s_j^2, \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (\theta_{ij} - \bar{\theta}_{\cdot j})^2.$$

If the mean of each chain is equal to each other,  $\frac{1}{n}\hat{B}$  will be close to zero when  $n$  is sufficiently large. Therefore  $\hat{R}$  must be close to one; otherwise, the convergence is doubtful.

# Numerical Errors For Correlated Samples

The numerical error of the sample mean is

$$\text{SE}[\bar{\theta}] = \sqrt{\frac{1}{T(T-1)} \sum_{t=1}^T (\theta_t - \bar{\theta})^2}, \quad \bar{\theta} = \frac{1}{T} \sum_{t=1}^T \theta_t.$$

In practice, we observe strong positive autocorrelations in the Monte Carlo sample, which makes the above  $\text{SE}[\bar{\theta}]$  underestimate the true numerical error. Instead we use

$$\text{SE}_S[\bar{\theta}] = \sqrt{\frac{1}{T} \left( \hat{\gamma}_0 + 2 \sum_{s=1}^S w\left(\frac{s}{S+1}\right) \hat{\gamma}_s \right)},$$

where  $\hat{\gamma}_s = \frac{1}{T} \sum_{t=s+1}^T (\theta_t - \bar{\theta})(\theta_{t-s} - \bar{\theta})$  and  $w(\cdot)$  is called a **lag window**.



# Lag Windows

- Rectangular lag window (used in PyMC by default)

$$w(x) = 1, \quad (|x| \leq 1).$$

- Triangular (Bartlett) lag window

$$w(x) = 1 - |x|, \quad (|x| \leq 1).$$

- Parzen lag window

$$w(x) = \begin{cases} 1 - 6x^2 + 6|x|^3, & (|x| < \frac{1}{2}); \\ 2(1 - |x|)^3, & (\frac{1}{2} \leq x \leq 1); \\ 0, & (\text{otherwise}). \end{cases}$$

# Numerical Inefficiency

Effective sample size

$$\hat{\tau}_e = \frac{T}{1 + 2 \sum_{s=1}^S w\left(\frac{s}{S+1}\right) \hat{\rho}_s}, \quad \hat{\rho}_s = \frac{\hat{\gamma}_s}{\hat{\gamma}_0}$$

If  $T$  is large enough,

$$\begin{aligned} \text{SE}_S[\bar{\theta}] &= \sqrt{\frac{\hat{\gamma}_0}{T} \left( 1 + 2 \sum_{s=1}^S w\left(\frac{s}{S+1}\right) \hat{\rho}_s \right)} \\ &\approx \text{SE}[\bar{\theta}] \sqrt{1 + 2 \sum_{s=1}^S w\left(\frac{s}{S+1}\right) \hat{\rho}_s} = \text{SE}[\bar{\theta}] \sqrt{\frac{T}{\hat{\tau}_e}}. \end{aligned}$$

Thus  $\text{SE}_S[\bar{\theta}]$  will be inflated if  $\hat{\tau}_e$  is small.

# Regression Model i

Suppose the conditional expectation of  $y_i$  is a linear function of  $k$  independent variables  $(x_{1i}, \dots, x_{ki})$ , i.e.,

$$E[y_i|x_i] = E[y_i|x_{1i}, \dots, x_{ki}] = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}.$$

In practice, we almost always put the constant term  $\beta_0$  in the model. By adding the error term, we have a **linear regression model**:

$$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2).$$

This is sometimes called the **multiple regression model**.

## Regression Model ii

Now define

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Then the regression model is summarized as

$$y = X\beta + \epsilon, \quad \epsilon \sim \text{Normal}(0, \sigma^2 \mathbf{I}).$$

The unknown parameters in the regression model are  $\theta = (\beta, \sigma^2)$ .

# Non-Standard Dependent Variables

In the regression model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i,$$

we implicitly suppose the dependent variable  $y_i$  in the regression model is continuous and can be either positive or negative. In other words,  $y_i$  is supposed to be a real-valued random variable.

As long as the normality assumption is valid and the conditional expectation of  $y_i$  is expressed as a linear function:

$$\mathbf{E}[y_i|x_i] = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki},$$

this assumption seems benign. In some cases, however, it is impractical to assume so.

# Binary Data

Suppose  $y_i$  takes either 1 or 0 with the constant probability, i.e.,

$$y_i = \begin{cases} 1, & \text{with probability } q_i; \\ 0, & \text{with probability } 1 - q_i, \end{cases}$$

which is called a **Bernoulli distribution**. This type of data appears in empirical analysis on decision making (e.g., consumer's choice) or events (e.g., bankruptcy).

In this case, the conditional expectation of  $y_i$  is  $q_i$ , the conditional probability  $\mathbf{P}(y_i = 1|x_i)$  itself. Since  $0 \leq q_i \leq 1$  by definition, it is unrealistic to assume that  $q_i$  is a linear function of  $x_i$ .

# Count Data

Suppose  $y_i$  takes non-negative integers,  $0, 1, 2, \dots$ . This type of data is used in analyzing occurrences of rare events such as traffic accidents, crimes, mechanical failures, and so forth.

It is often assume that  $y_i$  follows a **Poisson distribution**:

$$P(y_i = y | x_i) = \frac{e^{-\lambda_i} \lambda_i^y}{y!}, \quad y = 0, 1, 2, \dots, \quad \lambda_i > 0.$$

In this case, the conditional expectation of  $y_i$  is  $\lambda_i$ , which must be positive. Thus the linearity assumption on the conditional expectation of  $y_i$  seems inappropriate.

# Duration Data

Suppose  $y_i$  takes a positive real number. This type of data is used in analyzing the length of duration.

It is often assume that  $y_i$  follows an exponential distribution:

$$p(y_i|x_i) = \frac{1}{\theta_i} e^{-\frac{y_i}{\theta_i}}, \quad y_i > 0, \quad \theta_i > 0.$$

In this case, the conditional expectation of  $y_i$  is  $\theta_i$ , which must be positive. Thus the linearity assumption on the conditional expectation of  $y_i$  seems inappropriate.



# Generalized Linear Models

Let  $\mu_i$  denote the conditional expectation  $\mathbf{E}[y_i|x_i]$  and

$$\mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki},$$

for brevity. To relax the limitation of the linearity assumption  $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ , we introduce a transformation of the conditional expectation  $g(\cdot)$ :

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{or} \quad \mu_i = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

$g(\cdot)$  is called a **link function** and a regression-type model of the transformed conditional expectation is called a **generalized linear model (GLM)**.

## Link Functions

### 1. Logit link

$$\log \frac{\mu_i}{1 - \mu_i} = x_i^\top \beta \quad \Leftrightarrow \quad \mu_i = \frac{1}{1 + e^{-x_i^\top \beta}}.$$

### 2. Probit link

$$\Phi^{-1}(\mu_i) = x_i^\top \beta \quad \Leftrightarrow \quad \mu_i = \Phi(x_i^\top \beta),$$

where  $\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ .

### 3. Log link

$$\log \mu_i = x_i^\top \beta \quad \Leftrightarrow \quad \mu_i = e^{x_i^\top \beta}.$$

# Remarks

1. Since  $\mu_i = x_i^T \beta$  in the linear regression model,  $g(\mu_i) = \mu_i$ , which is called the **linear link function**.
2. Since both logit and probit link functions assure that  $\mu_i$  takes a value between 0 and 1, they are suitable for the binary data model.
3. When the probability  $p_i$  is transformed with the logit link function, such a GLM is called a **logit model**.
4. When the probability  $p_i$  is transformed with the probit link function, such a GLM is called a **probit model**.
5. Since the log link function assures that  $\mu_i$  is positive, it is used in the Poisson model of count data as well as the exponential model of duration data.

# Binary Choice Model

A GLM of Bernoulli binary data with either logit or probit link function is often called a **binary choice model**, though the binary data are not necessarily related to decision making.

## Examples

- Consumer's choice  
 $y_i = 1$ , if Consumer  $i$  owns an iPhone;  $0$ , otherwise.
- Labor force participation  
 $y_i = 1$ , if Person  $i$  works;  $0$ , otherwise.
- Bankruptcy  
 $y_i = 1$ , if Firm  $i$  goes bankrupt;  $0$ , otherwise.

# Likelihood in the Logit / Probit Model

Since the probability of  $y_i$  is expressed as

$$\mathbf{P}(y_i = y | x_i) = q_i^y (1 - q_i)^{1-y}, \quad q_i = g^{-1}(x_i^\top \boldsymbol{\beta}), \quad y = 0, 1,$$

where  $g$  is either logit or probit link function. The joint probability of  $\mathbf{D} = (y_1, \dots, y_n)$  is given by

$$p(\mathbf{D} | \boldsymbol{\beta}) = \prod_{i=1}^n q_i^{y_i} (1 - q_i)^{1-y_i}.$$

This is the likelihood in the logit / probit model. Since this likelihood is a complicated non-linear function of  $\boldsymbol{\beta}$ , it is impractical to derive the closed form of the posterior distribution.

# Poisson Regression Model

A GLM with Poisson count data with the log link function is called a **Poisson regression model**. Since the probability of  $y_i$  is expressed as

$$\mathbf{P}(y_i = y | x_i) = \frac{\lambda_i^y e^{-\lambda_i}}{y!}, \quad \lambda_i = e^{x_i^T \beta}, \quad y = 0, 1, 2, \dots,$$

the likelihood is given by

$$p(D|\beta) = \prod_{i=1}^n \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \propto \exp \left[ \sum_{i=1}^n \left\{ y_i x_i^T \beta - \exp(x_i^T \beta) \right\} \right].$$

Since the above likelihood is a complicated non-linear function of  $\beta$ , it is impractical to derive the closed form of the posterior distribution.

# Exponential Duration Model

Suppose an observed length of duration  $y_i$  follows the following exponential distribution:

$$p(y_i|x_i) = \frac{1}{\theta_i} e^{-\frac{y_i}{\theta_i}}, \quad \theta_i = e^{x_i^\top \beta}, \quad y_i > 0,$$

where the conditional mean is  $\mathbf{E}[y_i|x_i] = \theta_i = e^{x_i^\top \beta}$ . Then the likelihood is given by

$$p(D|\beta) = \prod_{i=1}^n \frac{1}{\theta_i} e^{-\frac{y_i}{\theta_i}} = \exp \left[ - \sum_{i=1}^n \left\{ x_i^\top \beta + y_i \exp(-x_i^\top \beta) \right\} \right].$$

Since the above likelihood is a complicated non-linear function of  $\beta$ , it is impractical to derive the closed form of the posterior distribution.