## Lecture Notes

# BAYESIAN ANALYSIS

## DSA 8505

## Strathmore University

Lecturer: Prof. Jacob Ong'ala

# Contents

# 4 Bayesian Estimation and Loss Functions

Suppose we want to estimate the proportion $p$ of voters who support a particular candidate. We begin with a prior belief about $p$ and collect data from a random sample of voters.

- **Prior Distribution**: Assume that $p$ follows a Beta distribution:

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

  where $B(\alpha, \beta)$ is the Beta function, and $\alpha, \beta > 0$ are shape parameters reflecting prior beliefs about $p$.

- **Likelihood Function**: Suppose we observe $X$ successes (votes for the candidate) in $n$ trials. The likelihood function is:

  $$f_X(x|p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

- **Posterior Distribution**: Using Bayes' theorem, the posterior distribution is:

  $$\pi(p|x) \propto f_X(x|p)\pi(p),$$

  which simplifies to:

  $$\pi(p|x) = \frac{p^{x+\alpha-1}(1-p)^{n-x+\beta-1}}{B(x+\alpha, n-x+\beta)},$$

  showing that $p|x \sim Beta(x+\alpha, n-x+\beta)$.

**Illustration**: If we initially believe that $p$ is likely close to 0.5, we could use a prior $Beta(2,2)$. After observing $x = 6$ successes in $n = 10$ trials, the posterior becomes $Beta(8,6)$, shifting our belief to reflect the observed data.

$$\textbf{Prior: } p \sim \text{Beta}(2,2)$$
$$\textbf{Observed data: } x = 6 \text{ successes out of } n = 10 \text{ trials}$$

### Posterior Distribution

For the Beta–Binomial conjugate update:

$$p \mid D \sim \text{Beta}(\alpha + x,\ \beta + n - x)$$

Substituting $\alpha = 2,\ \beta = 2,\ x = 6,\ n = 10$:

$$p \mid D \sim \text{Beta}(2+6,\ 2+10-6) = \text{Beta}(8,6)$$

### New Updated Value of $p$ (Posterior Mean)

The posterior mean is:

$$E(p \mid D) = \frac{\alpha}{\alpha + \beta}$$

Thus:

$$E(p \mid D) = \frac{8}{8+6} = \frac{8}{14} = 0.5714$$

$$\boxed{p_{\text{new}} = 0.5714}$$

Also,suppose we want to estimate the mean $\mu$ of a population based on sample data $x_1, x_2, \ldots, x_n$, assuming the population variance $\sigma^2$ is known.

- **Prior Distribution**: Assume $\mu \sim N(\mu_0, \sigma_0^2)$, reflecting prior knowledge about the mean.

$$P(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

- **Likelihood Function**: If the data are normally distributed $X_i \sim N(\mu, \sigma^2)$, the likelihood is:

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

$$f_X(x|\mu) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

- **Posterior Distribution**: Using Bayes' theorem, the posterior is:

$$P(\mu|x) \propto \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right)\left(\frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right)\right)$$

  Thus, the posterior distribution is:

$$\mu|x \sim \mathcal{N}\left(\mu_n, \sigma_n^2\right)$$

$$\mu|x \sim N\left(\frac{\frac{\mu_0}{\sigma_0^2} + \frac{\bar{x}}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}}, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right),$$

  where $\bar{x}$ is the sample mean.

**Illustration**: If $\mu_0 = 50$, $\sigma_0^2 = 25$, and we observe $n = 10$ samples with $\bar{x} = 55$ and $\sigma^2 = 16$, the posterior mean is:

$$\mu|x = \frac{\frac{50}{25} + \frac{10 \cdot 55}{16}}{\frac{1}{25} + \frac{10}{16}} = 53.57,$$

indicating an updated belief closer to the observed data.

## 4.1 Loss Functions

In statistics, a *loss function*, denoted $L(\theta, a)$, quantifies the cost or penalty associated with estimating a parameter $\theta$ as $a$. It reflects how "bad" the estimate $a$ is if the true value of the parameter is $\theta$. Different choices of loss functions allow us to tailor the estimation process to specific objectives.

In the Bayesian framework, the goal is to minimize the expected posterior loss, $h(a)$, which is defined as:

$$h(a) = \int L(\theta, a)\pi(\theta|x)\, d\theta,$$

where:

- $\pi(\theta|x)$ is the posterior distribution of $\theta$, given the observed data $x$.
- The integral represents averaging the loss over all possible values of $\theta$, weighted by the posterior probability of each $\theta$.

The value of $a$ that minimizes $h(a)$ is called the *Bayes estimator*.

### 4.1.1 Common Loss Functions

Different loss functions lead to different Bayes estimators. Below are three widely used loss functions and their implications.

1. **Squared Error Loss**:

$$L(\theta, a) = (\theta - a)^2.$$

This loss function penalizes large errors more severely than small ones, making it ideal when we want to minimize the average size of squared deviations.

*Bayes Estimator*: It turns out that the posterior mean minimizes the expected squared error loss:

$$\hat{\theta} = \mathbb{E}[\theta|x] = \int \theta\pi(\theta|x)\, d\theta.$$

**Why Does the Posterior Mean Minimize the Expected Squared Error Loss?**

**1. Start with the Expected Posterior Loss** The expected posterior loss under squared error loss is defined as:

$$h(a) = \int L(\theta, a)\pi(\theta|x)\, d\theta,$$

where:

- $L(\theta, a) = (\theta - a)^2$ is the squared error loss function.
- $\pi(\theta|x)$ is the posterior distribution of $\theta$, given the data $x$.
- $h(a)$ represents the "average penalty" incurred when estimating $\theta$ as $a$.

**2. Substitute the Loss Function** Substituting $L(\theta, a) = (\theta - a)^2$ into $h(a)$, we get:

$$h(a) = \int (\theta - a)^2 \pi(\theta|x)\, d\theta.$$

**3. Expand the Squared Term** Expanding $(\theta - a)^2$ using the formula for squares gives:

$$h(a) = \int \left(\theta^2 - 2\theta a + a^2\right) \pi(\theta|x)\, d\theta.$$

Using the linearity of integration, split the integral into three terms:

$$h(a) = \int \theta^2 \pi(\theta|x)\, d\theta - 2a \int \theta \pi(\theta|x)\, d\theta + \int a^2 \pi(\theta|x)\, d\theta.$$

**4. Simplify the Terms** - The first term, $\int \theta^2 \pi(\theta|x)\, d\theta$, is the expected value of $\theta^2$, denoted $\mathbb{E}[\theta^2|x]$. - The second term, $\int \theta \pi(\theta|x)\, d\theta$, is the expected value of $\theta$, denoted $\mathbb{E}[\theta|x]$. - The third term, $\int a^2 \pi(\theta|x)\, d\theta$, simplifies to $a^2$, as $a$ is independent of $\theta$.

Thus, $h(a)$ becomes:

$$h(a) = \mathbb{E}[\theta^2|x] - 2a\mathbb{E}[\theta|x] + a^2.$$

**5. Minimize the Expected Posterior Loss** To find the value of $a$ that minimizes $h(a)$, take the derivative of $h(a)$ with respect to $a$ and set it equal to zero:

$$\frac{d}{da} h(a) = \frac{d}{da} \left(\mathbb{E}[\theta^2|x] - 2a\mathbb{E}[\theta|x] + a^2\right).$$

The derivative is:

$$\frac{d}{da} h(a) = -2\mathbb{E}[\theta|x] + 2a.$$

Setting this to zero:

$$-2\mathbb{E}[\theta|x] + 2a = 0.$$

Solve for $a$:

$$a = \mathbb{E}[\theta|x].$$

**6. Conclusion** The posterior mean, $\hat{\theta} = \mathbb{E}[\theta|x]$, minimizes the expected posterior loss under squared error loss.

### Intuitive Explanation

- Squared error loss penalizes larger errors more heavily than smaller ones. For example:
  - An error of 1 gives a loss of $1^2 = 1$,
  - An error of 2 gives a loss of $2^2 = 4$,
  - Larger errors grow much faster in their penalty.

- The posterior mean balances these penalties by finding the ”center of gravity” of the posterior distribution.
- If you choose a point to the left or right of the mean, the penalties grow asymmetrically, increasing the overall loss.

**Example 1: Posterior Mean for a Normal Distribution**

Suppose the posterior distribution is

$$\pi(\theta \mid x) \sim \mathcal{N}(10, 4),$$

where the posterior mean is 10 and the posterior variance is 4.

- The posterior mean is:
$$\mathbb{E}[\theta \mid x] = 10.$$
- Under squared error loss,
$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2,$$
the Bayes estimator that minimizes the posterior expected loss is the posterior mean. Hence,
$$\hat{\theta}_B = \mathbb{E}[\theta \mid x] = 10.$$
- If we estimate $\theta$ as 9 or 11, then the squared error increases:
$$L(\theta, 9) = (\theta - 9)^2, \qquad L(\theta, 11) = (\theta - 11)^2.$$
Since these are not equal to the posterior mean, their posterior expected losses are higher than that of $\hat{\theta} = 10$, and hence they lead to a larger average loss.

**Example 2**

Suppose the posterior distribution of $\theta$ is $\pi(\theta|x) \sim N(5, 2^2)$ (normal distribution with mean 5 and variance 4).

1. The posterior mean is:

$$\mathbb{E}[\theta|x] = 5.$$

2. To verify, consider another estimate $a = 6$. The expected squared error loss becomes:

$$h(6) = \int (\theta - 6)^2 \pi(\theta|x)\, d\theta.$$

but

$$\mathbb{E}[(\theta - 6)^2|x] = (\mathbb{E}[\theta|x] - 6)^2 + \mathrm{Var}(\theta|x).$$

Using properties of the normal distribution:

$$h(6) = (6 - 5)^2 + \text{Variance} = 1 + 4 = 5.$$

3. If $a = 5$, the loss is:

$$h(5) = (5 - 5)^2 + \text{Variance} = 0 + 4 = 4.$$

Thus, the posterior mean minimizes the loss compared to any other estimate.

2. **Absolute Error Loss**:

   $$L(\theta, a) = |\theta - a|.$$

   This loss function treats all errors equally, regardless of their size. It is often used when outliers might skew results under squared error loss.

   *Bayes Estimator*: The posterior median minimizes the expected absolute error loss:

   $$\hat{\theta} = \text{Median}(\pi(\theta|x)).$$

   The **posterior median** minimizes the *expected absolute error loss.* Let us demonstrate this statement step by step.

### Absolute Error Loss Function

The *absolute error loss function* is defined as:

$$L(\theta, a) = |\theta - a|,$$

where:

- $\theta$: True value of the parameter.
- $a$: Estimated value (in this case, the Bayes estimator $\hat{\theta}$).

The Bayesian approach minimizes the expected loss:

$$h(a) = \int |\theta - a| \pi(\theta|x) \, d\theta,$$

where $\pi(\theta|x)$ is the posterior distribution of $\theta$, given the observed data $x$.

### Why the Posterior Median Minimizes Absolute Error Loss

To find the $a$ that minimizes $h(a)$, we split the integral into two parts:

$$h(a) = \int_{-\infty}^{a} (a - \theta)\pi(\theta|x) \, d\theta + \int_{a}^{\infty} (\theta - a)\pi(\theta|x) \, d\theta.$$

- For $\theta < a$, the loss is $a - \theta$, so this is represented by the first integral.
- For $\theta > a$, the loss is $\theta - a$, so this is represented by the second integral.

To minimize $h(a)$, differentiate with respect to $a$:

$$\frac{d}{da} h(a) = \int_{-\infty}^{a} \pi(\theta|x) \, d\theta - \int_{a}^{\infty} \pi(\theta|x) \, d\theta.$$

At the minimum, $\frac{d}{da} h(a) = 0$, which implies:

$$\int_{-\infty}^{a} \pi(\theta|x) \, d\theta = \int_{a}^{\infty} \pi(\theta|x) \, d\theta = 0.5.$$

This means $a$ must split the posterior distribution into two equal parts, with 50% of the probability on each side. Therefore:

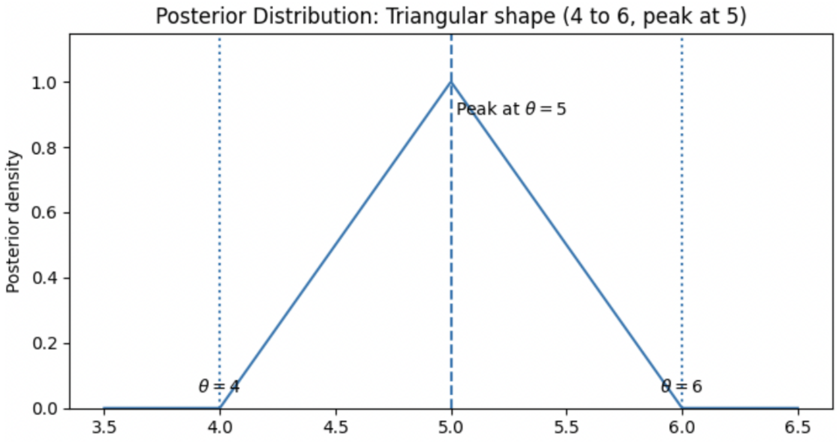
$$\hat{\theta} = \text{Median}(\pi(\theta|x)).$$

### Example: Posterior Median for a Triangular Distribution

Suppose the posterior distribution of $\theta$ is symmetric and triangular, defined as:

$$\pi(\theta|x) = \begin{cases} 2(1 - |\theta - 5|), & 4 \le \theta \le 6, \\ 0, & \text{otherwise.} \end{cases}$$

**Step 1: Visualize the Distribution** The posterior peaks at $\theta = 5$ and tapers linearly to 0 at $\theta = 4$ and $\theta = 6$.



## Step 2: Find the Posterior Median

The posterior density is triangular on the interval $[4, 6]$ with peak at $\theta = 5$, given by:

$$\pi(\theta \mid x) = 2\,(1 - |\theta - 5|)\,, \qquad 4 \le \theta \le 6.$$

The posterior median $\hat{\theta}$ satisfies:

$$P(\theta \le \hat{\theta} \mid x) = 0.5 \quad \iff \quad \int_4^{\hat{\theta}} \pi(\theta \mid x)\, d\theta = 0.5.$$

Substituting the posterior density,

$$\int_4^{\hat{\theta}} 2(1 - |\theta - 5|)\, d\theta = 0.5.$$

### Step 2.1: Simplify the absolute value

For $\theta \le 5$,

$$|\theta - 5| = 5 - \theta.$$

Hence,

$$1 - |\theta - 5| = 1 - (5 - \theta) = \theta - 4.$$

Therefore, for $4 \le \theta \le 5$,

$$\pi(\theta \mid x) = 2(\theta - 4).$$

**Step 2.2: Integrate**

$$\int_4^{\hat{\theta}} 2(\theta - 4)\, d\theta = 2 \int_4^{\hat{\theta}} (\theta - 4)\, d\theta.$$

Let $u = \theta - 4$. Then $du = d\theta$. When $\theta = 4$, $u = 0$; when $\theta = \hat{\theta}$, $u = \hat{\theta} - 4$.

$$2 \int_0^{\hat{\theta}-4} u\, du = 2 \left[ \frac{u^2}{2} \right]_0^{\hat{\theta}-4} = \left[ u^2 \right]_0^{\hat{\theta}-4} = (\hat{\theta} - 4)^2.$$

Thus,

$$\int_4^{\hat{\theta}} 2(\theta - 4)\, d\theta = (\hat{\theta} - 4)^2.$$

**Step 2.3: Set equal to 0.5 and solve**

$$(\hat{\theta} - 4)^2 = 0.5.$$

Taking square roots,

$$\hat{\theta} - 4 = \pm\sqrt{0.5}.$$

Since $\hat{\theta} \geq 4$, we take the positive root:

$$\hat{\theta} = 4 + \sqrt{0.5}.$$

Therefore, the posterior median is

$$\boxed{\hat{\theta} = 4 + \sqrt{0.5} \approx 4.7071.}$$

**Note:** The value $\hat{\theta} = 4.5$ would occur if the integral were set equal to 0.25, not 0.5.

**Step 3: Verify**

The posterior median is

$$\hat{\theta} = 4 + \sqrt{0.5} \approx 4.7071,$$

not 4.5. To verify that it is the median, we compute the posterior probability to the left of $\hat{\theta}$:

For $\theta \leq 5$, the posterior density simplifies to

$$\pi(\theta \mid x) = 2(\theta - 4), \qquad 4 \leq \theta \leq 5.$$

Hence,

$$P(\theta \leq \hat{\theta} \mid x) = \int_4^{\hat{\theta}} 2(\theta - 4)\, d\theta = (\hat{\theta} - 4)^2.$$

Substitute $\hat{\theta} = 4 + \sqrt{0.5}$:

$$P(\theta \leq \hat{\theta} \mid x) = \left(\sqrt{0.5}\right)^2 = 0.5.$$

Therefore,

$$P(\theta \leq \hat{\theta} \mid x) = 0.5 \quad \text{and} \quad P(\theta \geq \hat{\theta} \mid x) = 0.5,$$

which confirms that $\hat{\theta} = 4 + \sqrt{0.5}$ splits the posterior distribution into two equal probability halves.

Thus, under absolute error loss

$$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|,$$

the estimator

$$\boxed{\hat{\theta} = 4 + \sqrt{0.5} \approx 4.7071}$$

minimizes the posterior expected loss, since the Bayes estimator under absolute loss is the posterior median.

3. **0-1 Loss**:

This loss function assigns no penalty if the estimate is exactly correct but imposes a fixed penalty otherwise. It is best suited when we want the most likely value of $\theta$.

The **0-1 loss function** is defined as:

$$L(\hat{\theta}, \theta) = \begin{cases} 0, & \text{if } \hat{\theta} = \theta, \\ 1, & \text{if } \hat{\theta} \neq \theta. \end{cases}$$

This loss function assigns:

- Zero loss if the estimate $\hat{\theta}$ is equal to the true parameter $\theta$,
- A loss of 1 if $\hat{\theta} \neq \theta$.

In Bayesian decision theory, the goal is to minimize the **expected loss**, which is given by:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = \int L(\hat{\theta}, \theta)\pi(\theta|x)\, d\theta.$$

Using the definition of $L(\hat{\theta}, \theta)$, this integral splits into two cases:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = \int_{\theta \neq \hat{\theta}} 1 \cdot \pi(\theta|x)\, d\theta + \int_{\theta = \hat{\theta}} 0 \cdot \pi(\theta|x)\, d\theta.$$

The second term vanishes because it is multiplied by 0, leaving:

$$\mathbb{E}[L(\hat{\theta}, \theta)|x] = \int_{\theta \neq \hat{\theta}} \pi(\theta|x)\, d\theta.$$

The posterior distribution $\pi(\theta|x)$ integrates to 1 over all possible values of $\theta$, i.e.,

$$\int_{\theta \in \mathbb{R}} \pi(\theta|x)\, d\theta = 1.$$

This integral can be split into two disjoint regions:

$$\int_{\theta\in\mathbb{R}} \pi(\theta|x)\,d\theta = \int_{\theta=\hat{\theta}} \pi(\theta|x)\,d\theta + \int_{\theta\neq\hat{\theta}} \pi(\theta|x)\,d\theta.$$

Rearranging, we find:

$$\int_{\theta\neq\hat{\theta}} \pi(\theta|x)\,d\theta = 1 - \pi(\hat{\theta}|x).$$

Thus, the expected loss under the 0-1 loss function simplifies to:

$$\mathbb{E}[L(\hat{\theta},\theta)|x] = 1 - \pi(\hat{\theta}|x).$$

For the 0-1 loss, this simplifies to:

$$\mathbb{E}[L(\hat{\theta},\theta)|x] = 1 - \pi(\hat{\theta}|x),$$

where $\pi(\hat{\theta}|x)$ is the posterior probability density at $\hat{\theta}$.

## MAP as the Minimizer of Expected 0-1 Loss

To minimize the risk, we want to minimize:

$$1 - \pi(\hat{\theta}|x).$$

Since 1 is constant, minimizing the loss is equivalent to maximizing the posterior probability $\pi(\theta|x)$. That is:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta|x).$$

Thus, the MAP estimate corresponds to the **posterior mode**—the value of $\theta$ that has the highest posterior probability.

MAP stands for **Maximum A Posteriori**. It refers to the estimate of a parameter $\theta$ that maximizes the posterior distribution $\pi(\theta|x)$, given observed data $x$.

This shows that the **posterior mode** minimizes the expected loss under the 0-1 loss function, making it the optimal choice when the loss is defined in this way.

## Key Points

- The MAP estimate provides the parameter value that is most probable under the posterior distribution $\pi(\theta|x)$.
- Under the 0-1 loss, the expected loss is minimized by choosing $\hat{\theta}$ to maximize $\pi(\theta|x)$, which corresponds to the posterior mode.
- MAP incorporates both prior information ($\pi(\theta)$) and data ($\pi(x|\theta)$) to provide a balanced estimate.

**Example:** For a bimodal posterior distribution (e.g., $\pi(\theta|x)$ peaks at $\theta = 2$ and $\theta = 5$), the MAP estimate would be the mode with the highest probability density.

## Worked Example: 0–1 Loss and MAP Estimation

### Problem Setup (Discrete Parameter Case)

Suppose a diagnostic system classifies the severity level of a condition into three possible categories:

$$\theta \in \{\theta_1, \theta_2, \theta_3\} = \{\text{Low}, \text{Moderate}, \text{High}\}.$$

After observing data $x$, the posterior probabilities are given as:

$$\pi(\theta = \text{Low} \mid x) = 0.20, \qquad \pi(\theta = \text{Moderate} \mid x) = 0.55, \qquad \pi(\theta = \text{High} \mid x) = 0.25.$$

### 0–1 Loss Function

The 0–1 loss assigns no loss for a correct estimate and a loss of 1 for an incorrect estimate:

$$L(\hat{\theta}, \theta) = \begin{cases} 0, & \hat{\theta} = \theta, \\ 1, & \hat{\theta} \neq \theta. \end{cases}$$

### Step 1: Compute the Posterior Risk

The posterior expected loss (posterior risk) for a decision $\hat{\theta}$ is:

$$R(\hat{\theta} \mid x) = \mathbb{E}[L(\hat{\theta}, \theta) \mid x] = \sum_{\theta} L(\hat{\theta}, \theta)\, \pi(\theta \mid x).$$

### Step 2: Evaluate Risk for Each Possible Estimate

**Case 1: Choose $\hat{\theta} = \text{Low}$**

$$R(\text{Low} \mid x) = 0 \cdot \pi(\text{Low} \mid x) + 1 \cdot \pi(\text{Moderate} \mid x) + 1 \cdot \pi(\text{High} \mid x)$$

$$R(\text{Low} \mid x) = 0 + 0.55 + 0.25 = 0.80.$$

**Case 2: Choose $\hat{\theta} = \text{Moderate}$**

$$R(\text{Moderate} \mid x) = 1 \cdot \pi(\text{Low} \mid x) + 0 \cdot \pi(\text{Moderate} \mid x) + 1 \cdot \pi(\text{High} \mid x)$$

$$R(\text{Moderate} \mid x) = 0.20 + 0 + 0.25 = 0.45.$$

**Case 3: Choose $\hat{\theta} = \text{High}$**

$$R(\text{High} \mid x) = 1 \cdot \pi(\text{Low} \mid x) + 1 \cdot \pi(\text{Moderate} \mid x) + 0 \cdot \pi(\text{High} \mid x)$$

$$R(\text{High} \mid x) = 0.20 + 0.55 + 0 = 0.75.$$

### Step 3: Select the Bayes Estimator

We choose the estimate that minimizes the posterior risk:

$$R(\text{Low} \mid x) = 0.80, \quad R(\text{Moderate} \mid x) = 0.45, \quad R(\text{High} \mid x) = 0.75.$$

The minimum risk occurs at $\hat{\theta} = \text{Moderate}$. Therefore,

$$\boxed{\hat{\theta}_{\text{0-1}} = \text{Moderate}.}$$

### Step 4: Show that the Bayes Estimator is the MAP

The MAP estimator is defined by:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta \mid x).$$

Since

$$\max\{0.20, 0.55, 0.25\} = 0.55,$$

we obtain

$$\boxed{\hat{\theta}_{\text{MAP}} = \text{Moderate.}}$$

### Conclusion

Thus, under 0–1 loss, the Bayes estimator coincides with the MAP estimator:

$$\boxed{\hat{\theta}_{\text{0-1}} = \hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta \mid x).}$$

### Important Note (Continuous Parameter Case)

For continuous parameters, $\pi(\theta \mid x)$ is a **density** not a probability, hence $\pi(\hat{\theta} \mid x)$ is not a posterior probability at a single point. In such cases, the MAP estimator is obtained by using a small-neighborhood 0–1 loss:

$$L(\theta, \hat{\theta}) = \begin{cases} 0, & |\theta - \hat{\theta}| < \varepsilon, \\ 1, & \text{otherwise.} \end{cases}$$

which yields:

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta \mid x).$$

## Note:

- Different loss functions serve different purposes. The choice of a loss function should align with the problem's objectives and the consequences of estimation errors.
- The posterior mean, median, and mode correspond to different optimality criteria:
  - **Mean**: Minimizes squared error loss, sensitive to outliers.
  - **Median**: Minimizes absolute error loss, robust to outliers.
  - **Mode**: Maximizes posterior density, ideal for identifying the most probable value.
- In practice, understanding the nature of the problem and the properties of the posterior distribution is crucial for choosing the appropriate loss function.

## 4.2 Bayesian Point and Interval Estimation

In Bayesian inference, estimation is based on the posterior distribution $\pi(\theta|x)$, which combines prior beliefs with observed data through Bayes' theorem. Bayesian estimation can be categorized into *point estimation* and *interval estimation.*

### 4.2.1 Point Estimation

A Bayesian point estimate is a single value that best represents the unknown parameter $\theta$ based on the posterior distribution. The choice of the optimal point estimate depends on the loss function, which quantifies the penalty for estimation errors. Different loss functions lead to different Bayesian estimators:

- **Posterior Mean**: Given by

$$\hat{\theta}_{\text{mean}} = \mathbb{E}[\theta|x] = \int \theta\pi(\theta|x)d\theta.$$

  This estimator minimizes the **squared error loss**, which penalizes large deviations quadratically. It provides a measure of central tendency and is useful when the posterior distribution is symmetric.

- **Posterior Median**: Defined as

$$P(\theta \leq \hat{\theta}_{\text{median}}|x) = P(\theta \geq \hat{\theta}_{\text{median}}|x) = 0.5.$$

  This estimator minimizes the **absolute error loss**, making it more robust to skewed distributions compared to the posterior mean.

- **Posterior Mode (MAP Estimate)**: Given by

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \pi(\theta|x).$$

  This estimator minimizes the **0-1 loss function**, which considers only whether the estimate is correct or incorrect. The MAP estimate corresponds to the mode of the posterior distribution and is useful when seeking the most probable parameter value.

### 4.2.2 Interval Estimation

While point estimates provide a single best estimate of $\theta$, they do not account for uncertainty. *Credible intervals* offer a Bayesian alternative to frequentist confidence intervals. Given a confidence level $1 - \alpha$, a credible interval satisfies:

$$P(\theta \in [a, b]|x) = 1 - \alpha.$$

Unlike confidence intervals, which rely on long-run frequency properties, credible intervals have a direct probabilistic interpretation: the true parameter $\theta$ lies within the interval with probability $1 - \alpha$, given the observed data.

The common credible intervals is Equal-Tailed Interval

- **Equal-Tailed Interval**: An equal-tailed credible interval ensures that the probability of $\theta$ falling below the lower bound $a$ is the same as the probability of it exceeding the upper bound $b$. Mathematically, it satisfies:

$$P(\theta < a|x) = P(\theta > b|x) = \frac{\alpha}{2}.$$

This means that a total probability mass of $\alpha$ is excluded from the interval, with $\frac{\alpha}{2}$ in each tail. The remaining probability $1-\alpha$ is contained within the interval $[a, b]$, capturing the central $(1-\alpha)$ fraction of the posterior distribution.

**Key Features of the Equal-Tailed Interval:**

- **Symmetric Tail Probabilities:** The interval cuts off an equal probability $\frac{\alpha}{2}$ from both tails of the posterior distribution. This ensures that extreme values are treated symmetrically.

- **Application in Bayesian Inference:** It is useful when the posterior distribution is symmetric (e.g., a normal posterior), in which case the equal-tailed interval often coincides with the highest posterior density (HPD) interval.
- **Limitations:** If the posterior distribution is skewed, the equal-tailed interval may exclude more probable values while including less likely ones. Additionally, in the case of a multi-modal posterior (having multiple peaks), it may not capture the most significant mode.

## Worked Example: Constructing a 95% Credible Interval for a Normal Posterior

**Example:** Suppose that after observing data $x$, the posterior distribution of the parameter $\theta$ is Normal:

$$\theta \mid x \sim \mathcal{N}(\mu, \sigma^2),$$

where:

- $\mu$ is the posterior mean,
- $\sigma^2$ is the posterior variance (so $\sigma$ is the posterior standard deviation).

We want to construct a **95% equal-tailed credible interval** for $\theta$.

### Step 1: Define the credible interval and confidence level

A $100(1-\alpha)\%$ credible interval $[a, b]$ satisfies:

$$P(a < \theta < b \mid x) = 1 - \alpha.$$

For a 95% credible interval:

$$1 - \alpha = 0.95 \quad \Rightarrow \quad \alpha = 0.05.$$

### Step 2: Equal-tailed condition

An **equal-tailed** credible interval places equal posterior probability in both tails:

$$P(\theta < a \mid x) = \frac{\alpha}{2}, \qquad P(\theta > b \mid x) = \frac{\alpha}{2}.$$

Since $\alpha = 0.05$,

$$P(\theta < a \mid x) = P(\theta > b \mid x) = \frac{0.05}{2} = 0.025.$$

Thus, the lower bound $a$ is the 2.5% posterior quantile and the upper bound $b$ is the 97.5% posterior quantile.

### Step 3: Standardize the posterior distribution

Because $\theta \mid x \sim \mathcal{N}(\mu, \sigma^2)$, we standardize using:

$$Z = \frac{\theta - \mu}{\sigma}.$$

Then:

$$Z \sim \mathcal{N}(0, 1).$$

### Step 4: Translate the probability statement

We want:

$$P(a < \theta < b \mid x) = 0.95.$$

Subtract $\mu$ and divide by $\sigma$:

$$P\left(\frac{a - \mu}{\sigma} < \frac{\theta - \mu}{\sigma} < \frac{b - \mu}{\sigma} \mid x\right) = 0.95.$$

But $\frac{\theta - \mu}{\sigma} = Z$, hence:

$$P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) = 0.95.$$

### Step 5: Use standard normal quantiles

Since the credible interval is equal-tailed,

$$P(Z < z_{0.025}) = 0.025, \qquad P(Z < z_{0.975}) = 0.975.$$

From the standard normal table:

$$z_{0.975} \approx 1.96, \qquad z_{0.025} = -1.96.$$

Thus:

$$P(-1.96 < Z < 1.96) = 0.95.$$

### Step 6: Convert back to $\theta$

Replace $Z$ by $\frac{\theta-\mu}{\sigma}$:

$$P\left(-1.96 < \frac{\theta-\mu}{\sigma} < 1.96\right) = 0.95.$$

Multiply by $\sigma$:

$$P(-1.96\sigma < \theta - \mu < 1.96\sigma) = 0.95.$$

Add $\mu$:

$$P(\mu - 1.96\sigma < \theta < \mu + 1.96\sigma) = 0.95.$$

Therefore, the 95% equal-tailed credible interval is:

$$[a, b] = [\mu - 1.96\sigma,\ \mu + 1.96\sigma].$$

### Final Answer

$$\boxed{[a, b] = [\mu - 1.96\sigma,\ \mu + 1.96\sigma]}$$

### Numerical Illustration (for teaching)

Assume that the posterior distribution is:

$$\theta \mid x \sim \mathcal{N}(10, 4).$$

Then:

$$\mu = 10, \qquad \sigma^2 = 4 \Rightarrow \sigma = \sqrt{4} = 2.$$

Compute the bounds:

$$a = \mu - 1.96\sigma = 10 - 1.96(2) = 10 - 3.92 = 6.08,$$

$$b = \mu + 1.96\sigma = 10 + 1.96(2) = 10 + 3.92 = 13.92.$$

Hence the 95% credible interval is:

$$\boxed{[6.08,\ 13.92]}$$

**Interpretation:** There is a 95% posterior probability that the parameter $\theta$ lies between 6.08 and 13.92, given the observed data $x$.

**Interpretation:** Since the posterior is normal (symmetric), the equal-tailed interval places 2.5% probability in each tail, ensuring that extreme values are treated symmetrically. This means:

- There is a 2.5% probability that $\theta$ is less than $\mu - 1.96\sigma$. - There is a 2.5% probability that $\theta$ is greater than $\mu + 1.96\sigma$. - The middle 95% of the probability mass is contained within $[a, b]$.

## Highest Posterior Density (HPD/HDI) Credible Intervals

In Bayesian inference, uncertainty about an unknown parameter $\theta$ after observing data $x$ is represented by the posterior distribution:

$$\pi(\theta \mid x).$$

A key inferential objective is to summarize this posterior uncertainty using a **credible interval** (or more generally a credible set), for example at the 95% level:

$$P(\theta \in C \mid x) = 0.95.$$

The most commonly used credible interval is the **equal-tailed credible interval**. However, this interval is not always optimal, particularly when the posterior distribution is **skewed**, **bounded**, or **multimodal**. In such cases, the preferred Bayesian alternative is the:

**Highest Posterior Density (HPD) interval**
(or Highest Density Interval, HDI).

### Formal Definition

Let $\pi(\theta \mid x)$ be the posterior density of $\theta$. A set $C_\gamma$ is called a $\gamma$-level HPD credible set if:

1. **Credibility condition:**

$$P(\theta \in C_\gamma \mid x) = \int_{\theta \in C_\gamma} \pi(\theta \mid x)\, d\theta = \gamma.$$

2. **Highest density condition:** For any $\theta_1 \in C_\gamma$ and $\theta_2 \notin C_\gamma$,

$$\pi(\theta_1 \mid x) \geq \pi(\theta_2 \mid x).$$

Thus, the HPD interval is a region containing $\gamma$ posterior probability mass such that all included points have posterior density at least as high as any excluded point.

## Equivalent Definition (Threshold Form)

The HPD set can be written as:

$$C_\gamma = \left\{\theta : \pi(\theta \mid x) \geq k_\gamma\right\},$$

where $k_\gamma$ is a constant chosen such that:

$$\int_{\pi(\theta|x) \geq k_\gamma} \pi(\theta \mid x)\, d\theta = \gamma.$$

**Interpretation:** The HPD set consists of all parameter values whose posterior density is above a certain cutoff $k_\gamma$, and the cutoff is chosen so that exactly 95% (or $\gamma$) of posterior mass is included.

## Key Properties of HPD Intervals

### Shortest Credible Interval Property

One important theoretical result is:

> Among all credible intervals with posterior probability $\gamma$, the HPD interval has the **smallest length**.

That is, if $C_\gamma$ is the HPD interval and $I_\gamma$ is any other $\gamma$-credible interval, then:

$$\text{Length}(C_\gamma) \leq \text{Length}(I_\gamma).$$

### Relationship to Posterior Mode

Because the HPD interval contains the highest density points, it always contains the **posterior mode** (MAP estimator):

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \pi(\theta \mid x).$$

## HPD for Common Posterior Distributions

### 4.2.3 Normal Posterior Case

If:

$$\theta \mid x \sim \mathcal{N}(\mu, \sigma^2),$$

then the posterior is symmetric and unimodal.

In this case, the HPD interval coincides with the equal-tailed credible interval:

$$C_{0.95} = [\mu - 1.96\sigma,\ \mu + 1.96\sigma]\,.$$

Thus, **for symmetric unimodal posteriors**, equal-tailed and HPD intervals are identical.

### 4.2.4 Skewed Posterior Case (Typical)

If the posterior is skewed (e.g., Gamma, Beta, Lognormal), then:

- the equal-tailed interval is not shortest,
- the HPD interval shifts toward the mode,
- the HPD interval will have unequal tail probabilities.

## 4.3 Computation of HPD Intervals

### 4.3.1 Method 1: Analytical Solution (when possible)

For certain distributions, HPD intervals can be derived analytically by solving:

$$\pi(a \mid x) = \pi(b \mid x),$$

and:

$$\int_a^b \pi(\theta \mid x)\, d\theta = \gamma.$$

These two equations ensure:

1. same density at the boundaries (HPD cutoff),
2. probability mass inside equals $\gamma$.

### 4.3.2 Method 2: Simulation / MCMC-based HPD (most common)

In modern Bayesian analysis, posterior samples

$$\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(M)}$$

are often available from MCMC.

A common HPD approximation method:

1. Sort samples:
$$\theta_{(1)} \le \theta_{(2)} \le \cdots \le \theta_{(M)}.$$
2. Let:
$$m = \lfloor \gamma M \rfloor.$$
3. Consider all intervals of length $m$:
$$[\theta_{(1)}, \theta_{(1+m)}],\ [\theta_{(2)}, \theta_{(2+m)}],\ \dots$$
4. Choose the shortest interval:
$$\text{HPD} = \arg\min_i \left(\theta_{(i+m)} - \theta_{(i)}\right).$$

This provides an accurate approximation of the HPD interval even for complex posteriors.

## Example (Skewed Posterior)

Suppose the posterior distribution is:

$$\theta \mid x \sim \text{Gamma}(\alpha = 4, \beta = 1),$$

which is right-skewed.

## Equal-tailed 95% credible interval

$$[a, b] = [Q_{0.025}, Q_{0.975}]\,,$$

where $Q_p$ is the posterior quantile at probability $p$.

## HPD 95% credible interval

The HPD interval is the **shortest** interval $[a, b]$ such that:

$$\int_a^b \pi(\theta \mid x)\, d\theta = 0.95,$$

and:

$$\pi(a \mid x) = \pi(b \mid x).$$

**Important point:** Unlike the equal-tailed interval, HPD will typically yield

$$P(\theta < a \mid x) \neq 0.025, \qquad P(\theta > b \mid x) \neq 0.025.$$

## Worked Example (Skewed Posterior): Gamma Posterior and HPD Interval

Suppose the posterior distribution of the parameter $\theta$ is:

$$\theta \mid x \sim \text{Gamma}(\alpha = 4, \beta = 1),$$

where $\alpha$ is the shape and $\beta$ is the rate parameter.

Since $\alpha = 4 > 1$, the Gamma posterior is **unimodal** and **right-skewed**.

### Step 1: Write the posterior density and CDF

The Gamma$(\alpha, \beta)$ posterior density is:

$$\pi(\theta \mid x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \qquad \theta > 0.$$

Substitute $\alpha = 4, \beta = 1$:

$$\pi(\theta \mid x) = \frac{1^4}{\Gamma(4)} \theta^3 e^{-\theta} = \frac{1}{6} \theta^3 e^{-\theta}, \qquad \theta > 0,$$

because $\Gamma(4) = 3! = 6$.

The CDF is:

$$F(\theta) = P(\Theta \le \theta \mid x) = \int_0^\theta \pi(t \mid x)\, dt.$$

### Step 2: Equal-tailed 95% credible interval

The 95% equal-tailed credible interval $[a_{ET}, b_{ET}]$ satisfies:

$$P(\theta < a_{ET} \mid x) = 0.025, \qquad P(\theta > b_{ET} \mid x) = 0.025.$$

Equivalently:

$$F(a_{ET}) = 0.025, \qquad F(b_{ET}) = 0.975.$$

Thus,

$$[a_{ET}, b_{ET}] = [Q_{0.025}, Q_{0.975}],$$

where $Q_p$ is the $p$-th posterior quantile.

**Numerical values (Gamma(4,1)):**

$$Q_{0.025} \approx 1.0899, \qquad Q_{0.975} \approx 8.7673.$$

Therefore the equal-tailed 95% credible interval is:

$$\boxed{[a_{ET}, b_{ET}] \approx [1.0899,\ 8.7673].}$$

**Remarks (important for teaching):**

- This interval has exactly 2.5% posterior mass in each tail.
- It does *not* guarantee that all points inside are more plausible than points outside.
- For skewed posteriors, it is usually **not the shortest** credible interval.

### Step 3: HPD 95% credible interval

The HPD interval $[a_{HPD}, b_{HPD}]$ is defined as the **shortest** interval such that:

$$\int_{a_{HPD}}^{b_{HPD}} \pi(\theta \mid x)\, d\theta = 0.95.$$

In addition, the HPD boundaries must satisfy the **equal-density condition**:

$$\pi(a_{HPD} \mid x) = \pi(b_{HPD} \mid x) = k,$$

for some cutoff $k$. This ensures that the interval contains only the **highest posterior density** values.

### Step 3.1: Express the HPD conditions using the Gamma density

Recall:

$$\pi(\theta \mid x) = \frac{1}{6}\theta^3 e^{-\theta}.$$

The HPD boundary condition becomes:

$$\frac{1}{6}a_{HPD}^3 e^{-a_{HPD}} = \frac{1}{6}b_{HPD}^3 e^{-b_{HPD}}.$$

Cancel $\frac{1}{6}$:

$$a_{HPD}^3 e^{-a_{HPD}} = b_{HPD}^3 e^{-b_{HPD}}.$$

Taking natural logarithms:

$$3\ln(a_{HPD}) - a_{HPD} = 3\ln(b_{HPD}) - b_{HPD}.$$

### Step 3.2: Probability content condition

The credibility condition is:

$$P(a_{HPD} < \theta < b_{HPD} \mid x) = 0.95,$$

i.e.,

$$F(b_{HPD}) - F(a_{HPD}) = 0.95.$$

Hence, the HPD interval is obtained by solving the system:

$$\begin{cases} F(b_{HPD}) - F(a_{HPD}) = 0.95, \\ a_{HPD}^3 e^{-a_{HPD}} = b_{HPD}^3 e^{-b_{HPD}}. \end{cases}$$

### Step 3.3: Numerical solution (HPD interval)

Solving the system numerically for $\theta \mid x \sim \text{Gamma}(4, 1)$ gives:

$$a_{HPD} \approx 0.7125, \qquad b_{HPD} \approx 7.9483.$$

Therefore, the 95% HPD credible interval is:

$$\boxed{[a_{HPD}, b_{HPD}] \approx [0.7125,\ 7.9483].}$$

We can verify the HPD boundary property:

$$\pi(a_{HPD} \mid x) \approx \pi(b_{HPD} \mid x) \approx 0.02956.$$

### Step 4: Compare Equal-tailed vs HPD

$$\text{Equal-tailed: } [1.0899,\ 8.7673]$$
$$\text{HPD: } [0.7125,\ 7.9483]$$

The HPD interval is shorter:

$$(8.7673 - 1.0899) \approx 7.6774, \qquad (7.9483 - 0.7125) \approx 7.2358.$$

Thus, the HPD interval has smaller width while still covering 95% posterior mass.

### Key take away (Very Important)

Unlike the equal-tailed interval, the HPD interval generally satisfies:

$$P(\theta < a_{HPD} \mid x) \neq 0.025, \qquad P(\theta > b_{HPD} \mid x) \neq 0.025.$$

This happens because the HPD interval is chosen to include the most plausible values (high density region), not to force equal probability in the tails.

### Conclusion

For skewed posteriors, the HPD interval is preferred since it is the shortest and avoids low-density regions.