7.2.3 Bayesian Estimation.

$$X_1, \cdots, X_n \overset{iid}{\sim} f_x(x|\theta)$$

1. Prior Distribution. $\theta \sim \pi(\theta)$

2. Conditional Distribution (Likelihood) $\underset{\sim}{x} \,|\, \theta \sim f_{\underset{\sim}{x}}(\underset{\sim}{x}|\theta)$

Goal : to find the posterior distribution

$$\boxed{\pi(\theta|\underset{\sim}{x})} \;!!!$$

How? $$f_{\underset{\sim}{x},\theta}(\underset{\sim}{x},\theta) = \pi(\theta) \times f_{\underset{\sim}{x}}(\underset{\sim}{x}|\theta)$$

$$\pi(\theta|\underset{\sim}{x}) = \frac{f_{\underset{\sim}{x},\theta}(\underset{\sim}{x},\theta)}{m_{\underset{\sim}{x}}(\underset{\sim}{x})}$$

where $m_{\underset{\sim}{x}}(\underset{\sim}{x}) = \int f_{\underset{\sim}{x},\theta}(\underset{\sim}{x},\theta)\, d\theta$

$$\pi(\theta|\underset{\sim}{x}) = C \times e^{-\frac{(\theta-\bar{x}_n)^2}{2}} \qquad \theta \sim N(\bar{x}_n, 1)$$

Applications of invariance: In Example 7.9,

- $\overline{X}^2$ is the MLE of β^2
- $1/\overline{X}$ is the MLE of $1/\beta$
- For t fixed, $e^{-t/\overline{X}}$ is the MLE of $S_X(t|\beta) = e^{-t/\beta}$, the **survivor function** of X at t.

7.2.3 Bayesian estimation

Remark: Non-Bayesians think of inference in the following way:

$$\text{Observe } \mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}|\theta) \quad \longrightarrow \quad \text{Use } \mathbf{x} \text{ to make statement about } \theta.$$

In this paradigm, the model parameter θ is fixed (and unknown). I have taken θ to be a scalar here for ease of exposition.

Bayesians do not consider the parameter θ to be fixed. They regard θ as random, having its own probability distribution. Therefore, Bayesians think of inference in this way:

$$\text{Model } \theta \sim \pi(\theta) \quad \longrightarrow \quad \text{Observe } \mathbf{X}|\theta \sim f_{\mathbf{X}}(\mathbf{x}|\theta) \quad \longrightarrow \quad \text{Update with } \pi(\theta|\mathbf{x}).$$

The model for θ on the front end is called the **prior distribution**. The model on the back end is called the **posterior distribution**. The posterior distribution **combines** prior information (supplied through the prior model) and the observed data $\mathbf{x}$. For a Bayesian, all inference flows from the posterior distribution.

Important: Here are the relevant probability distributions that arise in a Bayesian context. These are given "in order" as to how the Bayesian uses them. Continue to assume that θ is a scalar.

1. **Prior distribution:** $\theta \sim \pi(\theta)$. This distribution incorporates the information available about θ before any data are observed.
2. **Conditional distribution:** $\mathbf{X}|\theta \sim f_{\mathbf{X}}(\mathbf{x}|\theta)$. This is the distribution of $\mathbf{X}$, but now viewed conditionally on θ:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= L(\theta|\mathbf{x}) \\ &\overset{\text{iid}}{=} \prod_{i=1}^{n} f_{X|\theta}(x_i|\theta). \end{aligned}$$

   Mathematically, the conditional distribution is the same as the likelihood function.

3. **Joint distribution:** This distribution describes how $\mathbf{X}$ and θ vary jointly. From the definition of a conditional distribution,

$$f_{\mathbf{X},\theta}(\mathbf{x},\theta) = \underbrace{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)}_{\text{likelihood}} \ \underbrace{\pi(\theta)}_{\text{prior}}.$$

4. **Marginal distribution.** This describes how $\mathbf{X}$ is distributed marginally. From the definition of a marginal distribution,

$$\begin{aligned} m_{\mathbf{X}}(\mathbf{x}) &= \int_{\Theta} f_{\mathbf{X},\theta}(\mathbf{x},\theta)d\theta \\ &= \int_{\Theta} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)d\theta, \end{aligned}$$

where Θ is the "support" of θ (remember, we are now treating θ as a random variable).

5. **Posterior distribution.** This is the Bayesian's "updated" distribution of θ, given that the data $\mathbf{X} = \mathbf{x}$ have been observed. From the definition of a conditional distribution,

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f_{\mathbf{X},\theta}(\mathbf{x},\theta)}{m_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)}{\int_{\Theta} f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)d\theta}. \end{aligned}$$

Remark: The process of starting with $\pi(\theta)$ and performing the necessary calculations to end up with $\pi(\theta|\mathbf{x})$ is informally known as "turning the Bayesian crank." The distributions above can be viewed as steps in a "recipe" for posterior construction (i.e., start with the prior and the conditional, calculate the joint, calculate the marginal, calculate the posterior). We will see momentarily that not all steps are needed. In fact, in practice, computational techniques are used to essentially bypass Step 4 altogether. You can see that this might be desirable, especially if θ is a vector (and perhaps high-dimensional).

Example 7.10. Suppose that, conditional on θ, $X_1, X_2, ..., X_n$ are iid Poisson(θ), where the prior distribution for $\theta \sim$ gamma(a, b), a, b known. We now turn the Bayesian crank.

1. **Prior distribution.**

$$\pi(\theta) = \frac{1}{\Gamma(a)b^a}\,\theta^{a-1}e^{-\theta/b}I(\theta > 0).$$

2. **Conditional distribution.** For $x_i = 0, 1, 2, ...,$

$$f_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i}e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^n x_i}e^{-n\theta}}{\prod_{i=1}^n x_i!}.$$

Recall that this is the same function as the likelihood function.

3. **Joint distribution.** For $x_i = 0, 1, 2, ...,$ and $\theta > 0$,

$$\begin{aligned} f_{\mathbf{X},\theta}(\mathbf{x},\theta) &= f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\ &= \frac{\theta^{\sum_{i=1}^n x_i}e^{-n\theta}}{\cancel{\prod_{i=1}^n x_i!}}\,\frac{1}{\cancel{\Gamma(a)b^a}}\,\theta^{a-1}e^{-\theta/b} \\ &= \underbrace{\frac{1}{\cancel{\prod_{i=1}^n x_i!\,\Gamma(a)b^a}}}_{\text{does not depend on }\theta}\,\underbrace{\theta^{\sum_{i=1}^n x_i+a-1}e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}_{\hookrightarrow \text{ Gamma Kernel}}. \end{aligned}$$

• MLE maximizes likelihood

$f_{\underset{\sim}{X}|\theta}(\underset{\sim}{x}|\theta)$

• Posterior mode maximaizes the density of the posterior distribution

$\pi(\theta|\underset{\sim}{X})$

$= \frac{f_{\underset{\sim}{X}|\theta}(\underset{\sim}{X}|\theta)\pi(\theta)}{m_X(\underset{\sim}{X})}$

i.e. $f_{\underset{\sim}{X}|\theta}(\underset{\sim}{X}|\theta)\pi(\theta)$

4. Conclude: $\pi(\theta|\underset{\sim}{X}) = - - -.$ $\theta|\underset{\sim}{X} \sim \text{Gamma}\left(\Sigma X_i + a, \frac{1}{n+\frac{1}{b}}\right)$

4. **Marginal distribution.** For $x_i = 0, 1, 2, \ldots,$

$$\begin{aligned} m_{\mathbf{X}}(\mathbf{x}) &= \int_{\Theta} f_{\mathbf{X},\theta}(\mathbf{x},\theta)d\theta \\ &= \frac{1}{\prod_{i=1}^n x_i!\,\Gamma(a)b^a}\int_0^{\infty} \underbrace{\theta^{\sum_{i=1}^n x_i+a-1}e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}_{\text{gamma}(a^*,\,b^*)\text{ kernel}} d\theta, \end{aligned}$$

where

$$a^* = \sum_{i=1}^n x_i + a \quad \text{and} \quad b^* = \frac{1}{n+\frac{1}{b}}.$$

Therefore,

$$m_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\prod_{i=1}^n x_i!\,\Gamma(a)b^a}\,\Gamma\left(\sum_{i=1}^n x_i + a\right)\left(\frac{1}{n+\frac{1}{b}}\right)^{\sum_{i=1}^n x_i+a}.$$

5. **Posterior distribution.** For $\theta > 0$,

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f_{\mathbf{X},\theta}(\mathbf{x},\theta)}{m_{\mathbf{X}}(\mathbf{x})} \\ &= \frac{\dfrac{1}{\prod_{i=1}^n x_i!\,\Gamma(a)b^a}\,\theta^{\sum_{i=1}^n x_i+a-1}e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}{\dfrac{1}{\prod_{i=1}^n x_i!\,\Gamma(a)b^a}\,\Gamma\left(\sum_{i=1}^n x_i+a\right)\left(\dfrac{1}{n+\frac{1}{b}}\right)^{\sum_{i=1}^n x_i+a}} \\ &= \frac{1}{\Gamma\left(\sum_{i=1}^n x_i + a\right)\left(\frac{1}{n+\frac{1}{b}}\right)^{\sum_{i=1}^n x_i+a}}\,\theta^{\sum_{i=1}^n x_i+a-1}e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}, \end{aligned}$$

which we recognize as the gamma pdf with parameters

$$\begin{aligned} a^* &= \sum_{i=1}^n x_i + a \\ b^* &= \frac{1}{n+\frac{1}{b}}. \end{aligned}$$

That is, the posterior distribution

$$\theta|\mathbf{X}=\mathbf{x} \sim \text{gamma}\left(\sum_{i=1}^n x_i + a,\ \frac{1}{n+\frac{1}{b}}\right).$$

Remark: Note that the shape and scale parameters of the posterior distribution $\pi(\theta|\mathbf{x})$ depend on

- a and b, the prior distribution parameters (i.e., the "hyperparameters")
- the data $\mathbf{x}$ through the sufficient statistic $t(\mathbf{x}) = \sum_{i=1}^n x_i$.

In this sense, the posterior distribution combines information from the prior and the data.

Q: In general, which functional of $\pi(\theta|\mathbf{x})$ should we use as a point estimator?
A: Answering this question technically would require us to discuss **loss functions** (see Section 7.3.4, CB). In practice, it is common to use one of

$$\begin{aligned}
\widehat{\theta}_B &= E(\theta|\mathbf{X}=\mathbf{x}) \longrightarrow \text{posterior mean} \\
\widetilde{\theta}_B &= \text{med}(\theta|\mathbf{X}=\mathbf{x}) \longrightarrow \text{posterior median} \\
\widehat{\theta}^*_B &= \text{mode}(\theta|\mathbf{X}=\mathbf{x}) \longrightarrow \underline{\text{posterior mode.}}
\end{aligned}$$

Note that in Example 7.10 (the Poisson-gamma example), the posterior mean equals

$$\begin{aligned}
\widehat{\theta}_{B\,\text{Mean}} = E(\theta|\mathbf{X}=\mathbf{x}) &= \frac{\sum_{i=1}^n x_i + a}{n + \frac{1}{b}} \\
&= \underline{\left(\frac{nb}{nb+1}\right)\overline{x} + \left(\frac{1}{nb+1}\right)ab.} \quad \approx \bar{x}
\end{aligned}$$

$$\hat{\theta}_{\text{mode}} = \frac{\sum_{i=1}^{n} X_i + a - 1}{n + \frac{1}{b}} \approx \bar{x}$$

That is, the posterior mean is a **weighted average** of the sample mean $\overline{x}$ and the prior mean ab. Note also that as the sample size n increases, more weight is given to the data (through $\overline{x}$) and less weight is given to the the prior (through the prior mean).

Remark: In Example 7.10, we wrote the joint distribution (in **Step 3**) as

$$\begin{aligned}
f_{\mathbf{X},\theta}(\mathbf{x},\theta) &= f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\
&= \frac{\theta^{\sum_{i=1}^n x_i} e^{-n\theta}}{\prod_{i=1}^n x_i!} \, \frac{1}{\Gamma(a)b^a} \, \theta^{a-1} e^{-\theta/b} \\
&= \underbrace{\frac{1}{\prod_{i=1}^n x_i! \, \Gamma(a)b^a}}_{\text{does not depend on } \theta} \; \underbrace{\theta^{\sum_{i=1}^n x_i + a - 1} e^{-\theta/\left(n+\frac{1}{b}\right)^{-1}}}_{\text{gamma}(a^*,b^*) \text{ kernel}}.
\end{aligned}$$

At this step, we can clearly identify the kernel of the posterior distribution. We can therefore skip calculating the marginal distribution $m_{\mathbf{X}}(\mathbf{x})$ in Step 4, because we know $m_{\mathbf{X}}(\mathbf{x})$ does not depend on θ. Because of this, it is common to write, in general,

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta) \\
&= L(\theta|\mathbf{x})\pi(\theta).
\end{aligned}$$

The posterior distribution is proportional to the likelihood function times the prior distribution. A (classical) Bayesian analysis requires these two functions $L(\theta|\mathbf{x})$ and $\pi(\theta)$ only.

Remark: Suppose $\mathbf{X}|\theta \sim f_{\mathbf{X}|\theta}(x|\theta)$. If $T = T(\mathbf{X})$ is sufficient, we can write

$$f_{\mathbf{X}|\theta}(x|\theta) = g(t|\theta)h(\mathbf{x}),$$

by the Factorization Theorem. Therefore, the posterior distribution

$$\begin{aligned}
\boxed{\pi(\theta|\mathbf{x}) \propto f_{\mathbf{X}|\theta}(\mathbf{x}|\theta)\pi(\theta)} \\
\propto g(t|\theta)\pi(\theta).
\end{aligned}$$

This shows that the posterior distribution will depend on the data $\mathbf{x}$ through the value of the sufficient statistic $t = T(\mathbf{x})$. We can therefore write the posterior distribution as depending on t only; i.e.,

$$\pi(\theta|t) \propto f_{T|\theta}(t|\theta)\pi(\theta),$$

and restrict attention to the (sampling) distribution of $T = T(\mathbf{X})$ from the beginning.

Example 7.11. Suppose that $X_1, X_2, ..., X_n$ are iid Bernoulli(θ), where the prior distribution for $\theta \sim \text{beta}(a, b)$, a, b known. We know that

$$T = T(\mathbf{X}) = \sum_{i=1}^{n} X_i$$

is a sufficient statistic for the Bernoulli family and that $T \sim b(n, \theta)$. Therefore, for $t = 0, 1, 2, ..., n$ and $0 < \theta < 1$, the posterior distribution

$$\begin{aligned}
\pi(\theta|t) &\propto f_{T|\theta}(t|\theta)\pi(\theta) \\
&= \binom{n}{t}\theta^t(1-\theta)^{n-t}\,\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\theta^{a-1}(1-\theta)^{b-1} \\
&= \underbrace{\binom{n}{t}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{does not depend on } \theta}\;\underbrace{\theta^{t+a-1}(1-\theta)^{n-t+b-1}}_{\text{beta}(a^*, b^*)\text{ kernel}},
\end{aligned}$$

where $a^* = t + a$ and $b^* = n - t + b$. From here, we can immediately conclude that the posterior distribution

$$\theta|T = t \sim \text{beta}(t + a, n - t + b),$$

where $t = T(\mathbf{x}) = \sum_{i=1}^{n} x_i$.

$$\hat{\theta}_{mean} = \frac{t+a}{t+a+n-t+b} = \frac{t+a}{n+a+b}$$

$$= \frac{\sum X_i + a}{n+a+b}$$

Discussion: In Examples 7.10 and 7.11, we observed the following occurrence:

- Example 7.10. $\theta \sim$ gamma (prior) $\longrightarrow$ $\theta|\mathbf{X} = \mathbf{x} \sim$ gamma (posterior).
- Example 7.11. $\theta \sim$ beta (prior) $\longrightarrow$ $\theta|T = t \sim$ beta (posterior).

Definition: Let $\mathcal{F} = \{f_X(x|\theta) : \theta \in \Theta\}$ denote a class of pdfs or pmfs. A class Π of prior distributions is said to be a **conjugate prior family** for $\mathcal{F}$ if the posterior distribution also belongs to Π.

As we have already seen in Examples 7.10 and 7.11,

Gama Prior

↓

- The gamma family is conjugate for the Poisson family. → Gama Posterior
- The beta family is conjugate for the binomial family. Beta Prior → Beta Posterior

Example 7.12. Suppose $X_1, X_2, ..., X_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, where $-\infty < \mu < \infty$ and $\sigma^2 > 0$.

- If σ^2 is known, a conjugate prior for μ is

$$\mu \sim \mathcal{N}(\xi, \tau^2), \qquad \xi, \tau^2 \text{ known.}$$

- If μ is known, a conjugate prior for σ^2 is

$$\sigma^2 \sim \text{IG}(a, b) \qquad a, b \text{ known.}$$

7.3 Methods of Evaluating Estimators

7.3.1 Bias, variance, and MSE

Definition: Suppose $W = W(\mathbf{X})$ is a point estimator. We call W an **unbiased estimator** of θ if

$$E_\theta(W) = \theta \quad \text{for all } \theta \in \Theta.$$

More generally, we call W an unbiased estimator of $\tau(\theta)$ if

$$E_\theta(W) = \tau(\theta) \quad \text{for all } \theta \in \Theta.$$

Definition: The **mean-squared error** (**MSE**) of a point estimator $W = W(\mathbf{X})$ is

$$\begin{aligned} \text{MSE}_\theta(W) &= E_\theta[(W - \theta)^2] \\ &= \text{var}_\theta(W) + [E_\theta(W) - \theta]^2 \\ &= \text{var}_\theta(W) + \text{Bias}^2_\theta(W), \end{aligned}$$

where $\text{Bias}_\theta(W) = E_\theta(W) - \theta$ is the **bias** of W as an estimator of θ. Note that if W is an unbiased estimator of θ, then for all $\theta \in \Theta$,

$$E_\theta(W) = \theta \implies \text{Bias}_\theta(W) = E_\theta(W) - \theta = 0.$$

In this case,

$$\text{MSE}_\theta(W) = \text{var}_\theta(W).$$

Remark: In general, the MSE incorporates two components:

- $\text{var}_\theta(W)$; this measures **precision**
- $\text{Bias}_\theta(W)$; this measures **accuracy**.

Obviously, we prefer estimators with small MSE because these estimators have small bias (i.e., high accuracy) and small variance (i.e., high precision).