Last time: ① Examples of Finding MLE:

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta; x_1, \cdots, x_N) \text{ or } l(\theta; x_1, \cdots, x_N)$$

where $\underline{X}_1, \cdots, \underline{X}_N \overset{iid}{\sim} f_{\underline{X}}(x; \theta_0)$ for some $\theta_0 \in \Theta$

② $$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmax}}\, E\left[l(\theta; \underline{X}) \right] \quad \text{where } \underline{X} \sim f_{\underline{X}}(x; \theta_0)$$

where $l(\theta; x) = \ln\left(f_{\underline{X}}(x; \theta) \right)$

By the WLLN:

$$\frac{1}{N} \sum_{i=1}^{N} l(\theta; X_i) \xrightarrow{P} E\left[l(\theta; \underline{X}) \right] \text{ for all } \theta \in \Theta$$

③ Bias, Variance and MSE:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

where $Bias(\hat{\theta}) = E[\hat{\theta}] - \theta_0 = E[\hat{\theta} - \theta_0]$

$\hat{\theta}$ unbiased $\Longleftrightarrow Bias(\hat{\theta}) = 0$ for all $\theta_0 \in \Theta$

$\Rightarrow MSE(\hat{\theta}) = Var(\hat{\theta})$

There is a trade-off between bias and variance.

Today: ① Efficiency + Cramer-Rao Lower Bound

② Consistency

③ Bayesian Estimation

① Efficiency: See previous lecture notes

"Lecture-Nov09-Sheng.pdf"

② Consistency:

An estimator $\hat{\theta}_N$ based on $X_1, \cdots, X_N$ is consistent if $\hat{\theta}_N \xrightarrow{P} \theta_0$:

For every $\varepsilon > 0$, $\lim_{N \to \infty} P(|\hat{\theta}_N - \theta_0| > \varepsilon) = 0$

Example 1: $X_1, X_2, \ldots \overset{iid}{\sim} Ber(\theta_0)$

let $\hat{\theta}_N = \frac{1}{N} \sum_{i=1}^{N} X_i$

$$P(|\hat{\theta}_N - \theta_0| > \varepsilon) = P(|\hat{\theta}_N - \theta_0|^2 > \varepsilon^2)$$

Markov's Inequality $\longrightarrow$ $\leq \frac{E[|\hat{\theta}_N - \theta_0|^2]}{\varepsilon^2}$

$E[\hat{\theta}_N] = \theta_0 \longrightarrow$ $= \frac{Var(\hat{\theta}_N)}{\varepsilon^2}$

$$= \frac{\theta_0(1-\theta_0)}{N\varepsilon^2} \longrightarrow 0 \quad \text{as } N \to \infty$$

So $\hat{\theta}_N$ is consistent.

Example 2: $X_1, X_2, \ldots \overset{iid}{\sim} N(\mu, \sigma^2)$

Let $\hat{\mu}_N = \frac{1}{N} \sum_{i=1}^{N} X_i \sim N(\mu, \frac{\sigma^2}{N})$

$$P(|\hat{\mu}_N-\mu|>\varepsilon)$$

$$=P\left(\frac{|\hat{\mu}_N-\mu|}{\frac{\sigma}{\sqrt{N}}}>\frac{\varepsilon}{\frac{\sigma}{\sqrt{N}}}\right)$$

$$=P\left(|Z_N|>\frac{\sqrt{N}\varepsilon}{\sigma}\right) \qquad \left[Z_N=\frac{\hat{\mu}_N-\mu}{\frac{\sigma}{\sqrt{N}}}\sim N(0,1)\right]$$

$$=2\left(1-\Phi\left(\frac{\sqrt{N}\varepsilon}{\sigma}\right)\right)\to 0 \qquad \left[\begin{array}{l}\Phi(\cdot)=\text{standard normal CDF}\\ \lim_{z\to\infty}\Phi(z)=1\end{array}\right]$$

Let $\hat{\sigma}_N^2=\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2$

$$E[\hat{\sigma}_N^2]=\frac{N-1}{N}\sigma^2, \quad Var(\hat{\sigma}_N^2)=\frac{2(N-1)}{N^2}\sigma^4$$

$$\Rightarrow MSE(\hat{\sigma}_N^2)=E\left[|\hat{\sigma}_N^2-\sigma^2|^2\right]=\frac{2N-1}{N^2}\sigma^4$$

So $$P(|\hat{\sigma}_N^2-\sigma^2|>\varepsilon)\le\frac{E\left[|\hat{\sigma}_N^2-\sigma^2|^2\right]}{\varepsilon^2}$$

$$=\frac{2N-1}{N^2}\frac{\sigma^4}{\varepsilon^2}\to 0 \quad \text{as } N\to\infty$$

Biased but consistent example!

How about unbiased but not consistent?

For example: $X_1, X_2, \ldots \overset{iid}{\sim}$ with mean μ, let $\hat{\mu}_N=X_N$

then $E[\hat{\mu}_N]=\mu$ but $\hat{\mu}_N\not\to\mu$ as $N\to\infty$.

③ Bayesian Estimation

When estimating a parameter $\theta_0 \in \Theta$, the MLE framework makes almost no assumptions.

It is often the case that some values of $\theta \in \Theta$ are a priori more likely than others.

Bayes rule allows us to incorporate information like this.

Let A, B be two events, since

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

then we have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

More generally, in Bayesian estimation, the unknown parameter θ is treated as a random variable, and prior information for θ is encoded in a distribution $f_\Theta(\theta)$:

$$\Theta \sim f_\Theta(\theta)$$

The observable random variable X is related to θ through the conditional distribution:

$$X \sim f_X(x \mid \Theta = \theta)$$

Given an observation $\Xi = x$, we update our model for Θ according to the Bayes rule:

likelihood

$$f_\Theta(\theta \mid \Xi = x) = \frac{f_\Xi(x \mid \Theta = \theta) f_\Theta(\theta)}{f_\Xi(x)}$$

prior

posterior

marginal likelihood

$$= \int_{\theta \in \Theta} f_\Xi(x \mid \Theta = \theta) f_\Theta(\theta)\, d\theta$$

How do we turn the posterior $f_\Theta(\theta \mid \Xi = x)$ into an estimate of θ? There are two popular approaches:

① Posterior mean or MMSE:

$$\hat{\theta}_{MMSE} = E[\Theta \mid \Xi = x] = \arg\min_{\theta} E[|\Theta - \theta|^2 \mid \Xi = x]$$

② Posterior mode or MAP:

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} f_\Theta(\theta \mid \Xi = x)$$

$$= \arg\max_{\theta \in \Theta} f_\Xi(x \mid \Theta = \theta) f_\Theta(\theta)$$

If $f_\Theta(\theta) = \frac{1}{|\Theta|}$ for all $\theta \in \Theta$, then

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} f_\Xi(x \mid \Theta = \theta) = \hat{\theta}_{MLE}$$

Example: $X_1, \dots, X_N \overset{iid}{\sim} Ber(\theta)$, $\theta \in Uniform(0,1)$

observations: $X_1 = x_1, \dots, X_N = x_N$

$$f_X(x_1, x_2, \dots, x_N \mid \Theta = \theta) f_\Theta(\theta) = \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i} \cdot 1$$

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

$$= \theta^{S_N} (1-\theta)^{N-S_N} \quad \text{where } S_N \triangleq \sum_{i=1}^N x_i$$

$$f_X(x_1, \dots, x_N) = \int_0^1 f_X(x_1, \dots, x_N \mid \Theta = \theta) f_\Theta(\theta) \, d\theta$$

$$= \int_0^1 \theta^{S_N} (1-\theta)^{N-S_N} d\theta$$

$$= B(S_N+1, N-S_N+1) \quad \text{where } B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \text{ is the Beta function}$$

therefore

$$f_\Theta(\theta \mid x_1, \dots, x_N) = \frac{f_X(x_1, \dots, x_N \mid \Theta = \theta) f_\Theta(\theta)}{f_X(x_1, \dots, x_N)}$$

$$= \frac{\theta^{S_N} (1-\theta)^{N-S_N}}{B(S_N+1, N-S_N+1)}$$

This is the PDF of the Beta$(S_N+1, N-S_N+1)$ distribution.

We compute explicitly the marginal $f_X(x_1, \dots, x_N)$ above, but this was not necessary to find the posterior. Indeed,

$$f_{\underline{X}}(x_1,\cdots,x_N \mid \Theta=\theta) f_{\Theta}(\theta) = \theta^{S_N}(1-\theta)^{N-S_N}$$

tells us the PDF of the posterior distribution of Θ.

Since $\theta \mid x_1,\cdots,x_N \sim \text{Beta}(S_N+1,\ N-S_N+1)$ has

mean $E[\theta \mid x_1,\cdots,x_N] = \dfrac{S_N+1}{(S_N+1)+(N-S_N+1)} = \dfrac{S_N+1}{N+2}$

mode $\dfrac{(S_N+1)-1}{(S_N+1)+(N-S_N+1)-2} = \dfrac{S_N}{N}$

then $\hat{\theta}_{MMSE} = \dfrac{S_N+1}{N+2}$ and $\hat{\theta}_{MAP} = \dfrac{S_N}{N}$