2.1 Big O little o Notation

Deterministic: Let $a_n = a_1, a_2, \ldots$ be a sequence

1. $a_n = o(1)$ means $a_n \to 0$ as $n \to \infty$
2. $a_n = o(b_n)$ means $\frac{a_n}{b_n} \to 0$ as $n \to \infty$. Or equivalently, $\frac{a_n}{b_n} = o(1)$.
   Examples:
   - If $a_n = \frac{1}{n}$, then $a_n = o(1)$
   - If $b_n = \frac{1}{\sqrt{n}}$, then $a_n = o(b_n)$
3. $a_n = O(1)$ means a_n is eventually bounded for all n large enough, $|a_n| < c$ for some $c > 0$. Note that $a_n = o(1)$ implies $a_n = O(1)$
4. $a_n = O(b_n)$ means $\frac{a_n}{b_n} = O(1)$. Likewise, $a_n = o(b_n)$ implies $a_n = O(b_n)$.
   Examples:
   - If $a_n = \frac{n}{2}$, then $a_n = O(n)$

Stochastic analogues:

1. $Y_n = o_p(1)$ if for all $\epsilon > 0$, then $P(|Y_n| > \epsilon) \to 0$
2. We say $Y_n = o_p(a_n)$ if $\frac{Y_n}{a_n} = o_p(1)$
3. $Y_n = O_p(1)$ if for all $\epsilon > 0$, there exists a $c > 0$ such that $P(|Y_n| > c) < \epsilon$
4. We say $Y_n = O_p(a_n)$ if $\frac{Y_n}{a_n} = O_p(1)$
   Examples:
   - $\overline{X}_n - \mu = o_p(1)$ and $S_n - \sigma^2 = o_p(1)$. By the the Law of Large Numbers.
   - $\sqrt{n}(\overline{X}_n - \mu) = O_p(1)$ and $\overline{X}_n - \mu = O_p(\frac{1}{\sqrt{n}})$. By the Central Limit Theorem.

2.2 Statistical Inference

A statistical model $\mathcal{P}$ is a collection of probability distributions or densities. A parametric model has the form

$$\mathcal{P} = \{p(x;\theta) : \theta \in \Theta\} \tag{2.1}$$

where $\Theta \subset \mathbb{R}^d$ in the parametric case.

Examples of nonparametric statistical models:

- $\mathcal{P} = \{$ all continuous CDF's $\}$
- $\mathcal{P} = \{f : \int (f''(x))^2 dx < \infty\}$

2.3 Parametric Point Estimation

Let $X_1, \ldots, X_n$ be independent and identically distributed i.e. *iid* random variables with some distribution $p(x;\theta)$. We want to estimate $\theta = (\theta_1, \ldots, \theta_n)$. An *estimator* is a function of data that does not depend on θ.

Example 2.1. *Suppose* $X \sim N(\mu, 1)$.

- μ *is not an estimator.*
- *Things that are estimators:* X*, any functions of* X*, 3,* $\sqrt{X}$*, etc.*

2.4 Ways to Evaluate Estimators

1. Bias and Variance
2. Mean Squared Error
3. Minimaxity and Decision Theory
4. Large Sample Evaluations

Definition 2.2. *Mean Squared Error (MSE). Suppose* $\theta, \widehat{\theta}$*, define*

$$\mathbb{E}\left[\left(\theta - \widehat{\theta}\right)^2\right] = \int \cdots \int \left[\left(\widehat{\theta}(x_1, \ldots, x_n) - \theta\right) f(x_1;\theta)^2 \cdots f(x_n;\theta)\right] dx_1 \cdots dx_n. \tag{2.2}$$

Definition 2.3. *Bias and Variance The bias is*

$$B = \mathbb{E}\left[\widehat{\theta}\right] - \theta, \tag{2.3}$$

and variance is

$$V = \mathrm{Var}\left[\widehat{\theta}\right]. \tag{2.4}$$

Result 2.4. *Bias-Variance Decomposition*

$$MSE = B^2 + V \tag{2.5}$$

Proof.

$$\begin{aligned} MSE &= \mathbb{E}\left[(\widehat{\theta} - \theta)^2\right] \\ &= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right] + \mathbb{E}\left[\widehat{\theta}\right] - \theta\right)^2\right] \\ &= \mathbb{E}\left[\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right] + \left(\mathbb{E}\left[\widehat{\theta}\right] - \theta\right)^2 + \underbrace{2\mathbb{E}\left[\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right]}_{=0}\left(\mathbb{E}\left[\widehat{\theta}\right] - \theta\right) \\ &= V + B^2 \end{aligned}$$

■

An estimator is unbiased if $B = 0$. Then $MSE = Variance$.

Example 2.5. *Let* $x_1, \ldots, x_n \overset{iid}{\sim} N(\mu, \sigma^2)$.

$$\mathbb{E}[\overline{x}] = \mu, \qquad \mathbb{E}\left[s^2\right] = \sigma^2$$

$$\mathbb{E}\left[(\overline{x} - \mu)^2\right] = \frac{\sigma^2}{n} = O\left(\frac{1}{n}\right) \qquad \mathbb{E}\left[(s^2 - \sigma^2)^2\right] = \frac{2\sigma^4}{n-1} = O\left(\frac{1}{n}\right).$$

2.5 Maximum Likelihood

Definition 2.6. *Let $X_1, \ldots, X_n$ have joint density $p(\vec{x}; \theta)$ where $\theta \in \Theta$. The* likelihood $\mathcal{L} : \Theta \to [0, \infty]$ *is defined by*

$$\mathcal{L}(\theta) := \mathcal{L}(\theta; \vec{x}) = p(\vec{x}; \theta) \tag{2.6}$$

Here $\vec{x}$ is fixed and θ varies in Θ.

1. The likelihood is a function of θ
2. The likelihood is not a pdf
3. If the data are *iid*, then $\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i, \theta)$
4. The likelihood is only defined up to some constant of proportionality

Definition 2.7. *Let*

$$\widehat{\theta}(\vec{x}) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}}\, p(\vec{x}; \theta).$$

We call $\widehat{\theta}(\vec{x})$ the Maximum Likelihood Estimator (MLE) *for θ. Note that this estimator may not be unique or may not exist.*

We may apply any monotone increasing function, and still achieve maximization. Usually, we solve by taking the log of the likelihood function.

$$\ell(\theta) = \ln \mathcal{L}(\theta) \tag{2.7}$$

Example 2.8. *Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Bernoulli(p).*

$$\begin{aligned}
\mathcal{L}(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{[0,1]}(x_i) \\
&= p^S (1-p)^{n-S} && \text{letting } S = \sum_{i=1}^{n} x_i \\
\ell(p) &= S \log(p) + (n-S)\log(1-p) \\
\frac{\partial}{\partial p} &= \frac{S}{p} - \frac{n-S}{1-p} \overset{set}{=} 0 \\
&\Rightarrow \widehat{p} = \frac{S}{n} = \overline{X}_n
\end{aligned}$$

Note: One should also use second derivative test to ensure critical point is a maximum.

Example 2.9. *Let* $X_1, \ldots, X_n \sim^{iid} N(0, \theta)$.

$$\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} e^{-x_i^2/2\theta} \\
\ell(\theta) &\propto -\frac{n}{2}\log(\theta) - \frac{1}{2\theta}\sum_{i=1}^{n} x_i^2 \\
&= -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2 \overset{=}{set} 0 \\
&\Rightarrow \theta \sum_{i=1}^{n} x_i^2 = n\theta^2 \\
&\Rightarrow \widehat{\theta} = 0 \text{ or } \frac{1}{n}\sum_{i=1}^{n} x_i^2.
\end{aligned}$$

The maximum likelihood estimator (MLE) is equivariant. If $\eta = g(\theta)$, and we want to estimate η, we use $\widehat{\eta} = g(\widehat{\theta})$. It is the MLE for η (g is 1-to-1, η and θ live in same space).

2.6 Bayes Estimator

Definition 2.10. *Start with prior on* θ, $\pi(\theta)$. *Compute posterior by using Bayes Theorem. Note* $p(\vec{x}|\theta)\pi(\theta) = p(x, \theta)$. *Then the* posterior distribution *on* θ *conditional on* $\vec{x}$ *is*

$$\pi(\theta|\vec{x}) = \frac{p(x|\theta)\pi(\theta)}{m(x)} \tag{2.8}$$

here $m(x) = \int (p(x|\theta)\pi(\theta))d\theta$.

Next, use $\pi(\theta|\vec{x})$ to find an estimator.

One Way:

$$\widehat{\theta} = \mathbb{E}[\theta|x] = \int (\theta)(\pi(\theta|x))d\theta \tag{2.9}$$

This is often called the *Bayes Estimator*. It estimates θ by averaging over the posterior.

Example 2.11. *Let* $X_1, \ldots, X_n \sim Bernoulli(p)$ *with prior on* $p \sim Beta(a, b)$.

$$\begin{aligned}
\pi(p) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} I_{[0,1]}(p) \\
\mathcal{L}(p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^S(1-p)^{n-S} \qquad \text{where } S = \sum_{i=1}^{n} x_i \\
\pi(p|x) &\propto p^S(1-p)^{n-S} p^{a-1}(1-p)^{b-1} I_{[0,1]}(p) \\
&= p^{S+a-1}(1-p)^{n-S+b-1} I_{[0,1]}(p) \\
&\propto Beta(S+a, n-S+b)
\end{aligned}$$

Example 2.12 (Maximum a posteriori estimation)**.** *Let* $X_1, \ldots, X_n \sim^{iid} N(0, 1)$. *If*

$$\pi(\mu) = N(0, \frac{1}{\lambda})$$

then

$$\pi(\mu|x) \propto e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} e^{-\frac{\lambda}{2}\mu^2}.$$

We can use the posterior mode as an estimator for μ.

$$\begin{aligned}
\tilde{\mu} &= \underset{\mu}{\operatorname{argmax}}\, e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2} e^{-\frac{\lambda}{2}\mu^2} \\
&= \underset{\mu}{\operatorname{argmax}} -\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{\lambda}{2}\mu^2 \\
&= \underset{\mu}{\operatorname{argmin}} \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2 + \frac{\lambda}{2}\mu^2 \\
&= \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n}(x_i-\mu)^2 + \lambda\mu^2 \\
&= \underset{\mu}{\operatorname{argmin}}\, \mu^2 - 2\overline{x}\mu + \lambda\mu^2 \\
&= \frac{\overline{x}}{1+\lambda}
\end{aligned}$$

This looks similar to Ridge Regression.