STA 522, Spring 2021
Introduction to Theoretical Statistics II

Lecture 6

Department of Biostatistics
University at Buffalo

8 March, 2021

AGENDA

- Method of maximum likelihood
- Bayesian approach to Statistics

Review: Method of Estimation

- **Method of Moments:** Equate population moments with the sample moments, then solve for parameters.

- **Method of Maximum Likelihood:** For each sample point $\underline{x}$, let $\hat{\theta}(\underline{x})$ be a parameter value at which the likelihood function $L(\theta \mid \underline{x})$ attains its maximum as a function of θ, with $\underline{x}$ held fixed. A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\underline{X}$ is $\hat{\theta}(\underline{X})$.

- **Note:** since the logarithm function is strictly increasing on $(0, \infty)$ (and so one-to-one), the value which maximizes $\log L(\theta \mid \underline{x})$ is the same value that maximizes $L(\theta \mid \underline{x})$.

- **Example:** $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli (p), for $0 \leq p \leq 1$. The MLE of p is $\hat{p} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$

Example: Let $X_1, X_2, \ldots, X_n \sim$ iid Poisson(θ), where θ. Find the MLE of θ.

The likelihood of θ is

$$L(\theta \mid \underline{x}) = \prod_{i=1}^{n} \exp(-\theta) \frac{\theta^{x_i}}{x_i!} \;\; = \exp(-n\theta) \; \frac{\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}$$

The log likelihood is:

$$\log L(\theta \mid \underline{x}) = -n\theta + \left(\sum_{i=1}^{n} x_i \right) \log \theta + \log \left(\prod_{i=1}^{n} x_i! \right)$$

Therefore,

$$\frac{d \;\; \log L(\theta \mid \underline{x})}{d\theta} = -n + \left(\sum_{i=1}^{n} x_i \right) \frac{1}{\theta} \gtreqless 0 \;\; \text{according as} \;\; \theta \lesseqgtr \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}$$

Therefore, $\hat{\theta} = \overline{x}$ is the MLE of θ.

Example: Let $X_1, X_2, \ldots, X_n \sim$ iid $\mathrm{N}(\theta, 1)$ for $-\infty < \theta < \infty$. Find the MLE of θ.

The likelihood function for θ is given by

$$L(\theta \mid \underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \; \exp\left[-\frac{1}{2}(x_i - \theta)^2\right] = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2\right]$$

Therefore the log likelihood is:

$$\log L(\theta \mid \underline{x}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2 = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(\theta - x_i)^2$$

which implies

$$\frac{d \log L(\theta \mid \underline{x})}{d\theta} = -\frac{1}{2}\sum_{i=1}^{n} 2\;(x_i - \theta) \gtreqless 0 \;\; \text{according as} \;\; \theta \lesseqgtr \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

Thus the MLE of θ is $\hat{\theta} = \overline{x}$.

Example (Restricted Range MLE): Let $X_1, X_2, \ldots, X_n \sim$ iid $\mathrm{N}(\theta, 1)$, where $\theta \geq 0$. Find the MLE of θ.

With no restrictions on θ the MLE of θ is $\overline{X}$.

However, if $\overline{X} < 0$, it will be outside the range of the parameter.

log likelihood:

$$\begin{aligned} \log L(\theta \mid \underline{x}) &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \theta)^2 \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \overline{x})^2 - \frac{n}{2} (\theta - \overline{x})^2 \end{aligned}$$

If $\overline{x} < 0$ then $L(\theta \mid \underline{x}) \leq L(0 \mid \underline{x})$ for all $\theta \in [0, \infty)$.

Therefore, the MLE of θ is

$$\hat{\theta} = \begin{cases} \overline{x} & \text{if } \overline{x} \geq 0 \\ 0 & \text{if } \overline{x} < 0 \end{cases}$$

Example (MLE where the likelihood function is non-differentiable): Consider $X_1, X_2, \dots, X_n \sim$ iid Uniform$(0, \theta)$. Find the MLE of θ.

The likelihood function is given by:

$$L(\theta \mid \underline{x}) = \prod_{i=1}^{n} \frac{1}{\theta} \; I(0 \le x_i \le \theta) = \frac{1}{\theta^n} \; I\left(\theta \ge x_{(n)}\right) \;\; I\left(0 \le x_{(1)}\right)$$

Clearly, $L(\theta \mid \underline{x})$ is not continuous (and hence non-differentiable) because of the indicator function.

Note that $L(\theta \mid \underline{x})$ is zero at $\theta < x_{(n)}$, jumps to $\frac{1}{\theta^n}$ at $\theta = x_{(n)}$ and then steadily declines.

Hence the MLE for θ is $\hat{\theta} = X_{(n)}$.

Example (Problem 7.6): Let $X_1, X_2, \ldots, X_n \sim$ iid Pareto$(\theta, 1)$ with pdf

$$f(x \mid \theta) = \theta x^{-2}; \quad 0 < \theta \leq x < \infty$$

Find (a) a sufficient statistic for θ, (b) the MLE of θ and (c) the method of moments estimator of θ.

(a) The joint pdf is $f(\underline{x} \mid \theta) = \underbrace{\theta^n I(x_{(1)} \geq \theta)}_{=g(T(\underline{x}|\theta)} \prod_{i=1}^{n} x_i^{-2}$. Hence by Factorization theorem, $T(\underline{X}) = X_{(1)}$ is sufficient for θ.

(b) The likelihood function for θ is

$$L(\theta \mid \underline{x}) = \theta^n I\left(\theta \leq x_{(1)}\right) \prod_{i=1}^{n} x_i^{-2}$$

This is maximum when $\theta = x_{(1)}$. Hence the MLE for θ is $\hat{\theta} = X_{(1)}$.

(c) Note that here $\mu_1' = \mathrm{E}_\theta(X_1) = \int_\theta^\infty \theta \frac{dx}{x} = \infty$. Hence method of moment estimator for θ does not exist.

Example (Binomial with unknown number of trials): Let $X_1, X_2, \ldots, X_n \sim$ iid Binomial(k, p), where p is known and k is unknown. The likelihood function is:

$$L(k \mid \underline{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{x_i}$$

Maximizing with differentiation is difficult because of factorials and because k is integer.

Note on the outset that $L(k \mid \underline{x}, p) = 0$ if $k < \max_i x_i$. So the MLE must be an integer $\hat{k} \geq \max_i x_i$ such that

$$\frac{L(\hat{k} \mid \underline{x}, p)}{L(\hat{k}-1 \mid \underline{x}, p)} \geq 1 \;\text{ and }\; \frac{L(\hat{k} \mid \underline{x}, p)}{L(\hat{k}+1 \mid \underline{x}, p)} > 1$$

Note that

$$\frac{L(k \mid \underline{x}, p)}{L(k-1 \mid \underline{x}, p)} = \frac{(k(1-p))^n}{\prod_{i=1}^{n}(k - x_i)}$$

Condition for maximum is

$$(k(1-p))^n \geq \prod_{i=1}^{n}(k-x_i) \quad \text{and} \quad ((k+1)(1-p))^n < \prod_{i=1}^{n}(k+1-x_i)$$

Divide by k^n and set $z = 1/k$. We want to solve

$$(1-p)^n = \prod_{i=1}^{n}(1-x_i z)$$

The RHS is strictly decreasing in z and RHS $= 1$ if $z = 0$ and RHS $= 0$ if $z = 1/\max_i x_i$.

Thus there is a unique z, say $\hat{z}$ that solves the equation. The unique solution is not analytically tractable. Must be approximated using numeric methods in practice.

The quantity $1/\hat{z}$ may not be an integer. The MLE $\hat{k}$ of k, is the largest integer $\leq 1/\hat{z}$.

Invariance Property of Maximum Likelihood

Consider a distribution indexed by a parameter θ. Interest is in finding an estimator for some function of θ, say $\tau(\theta)$.

Invariance property of MLEs says that if $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$.

For example, if θ is the mean of a normal distribution then the MLE of $\sin(\theta)$ is $\sin(\overline{X})$.

Need to be careful when τ is not one-to-one.

Definition: Let $\eta = \tau(\theta)$ be any function of θ. The **induced likelihood function** L^* is given by

$$L^*(\eta \mid \underline{x}) = \sup_{\{\theta\,:\,\tau(\theta)=\eta\}} L(\theta \mid \underline{x}).$$

The value $\hat{\eta}$ that maximizes $L^*(\eta \mid \underline{x})$ will be called the MLE of $\eta = \tau(\theta)$. Note that the maxima of L^* and L coincide.

Theorem (7.2.10)

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Proof: Define the induced likelihood L^* as before. Let $\hat{\eta}$ denote the value that maximizes $L^*(\eta \mid \underline{x})$.

Need to show that $L^*(\hat{\eta} \mid \underline{x}) = L^*(\tau(\hat{\theta}) \mid \underline{x})$.

Since the maxima of L and L^* coincide, therefore,

$$\begin{aligned}
L^*(\hat{\eta} \mid \underline{x}) &= \sup_{\{\theta : \tau(\theta)=\eta\}} L(\theta \mid \underline{x}) \\
&= \sup_{\theta} L(\theta \mid \underline{x}) \\
&= L(\hat{\theta} \mid \underline{x}) \\
&= \sup_{\{\theta : \tau(\theta)=\tau(\hat{\theta})\}} L(\theta \mid \underline{x}) \\
&= L^*(\tau(\hat{\theta}) \mid \underline{x})
\end{aligned}$$

Hence, $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$

Examples of the invariance property of MLE:

- If $X_1, X_2, \ldots, X_n$ iid $\mathrm{N}(\theta, 1)$ then the MLE of θ^2 is $\overline{X}^2$.

- If $X_1, X_2, \ldots, X_n$ iid Bernoulli(p) then the MLE of $\sqrt{p(1-p)}$ is $\sqrt{\hat{p}(1-\hat{p})}$ where $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

MLE of multiple parameters

Using calculus is tedious. In two parameter case, finding Local Maxima of a function $H(\theta_1, \theta_2)$ involves:

(a) Compute the first-order partial derivatives of $H(\theta_1, \theta_2)$, set them equal to 0, and solve for θ_1 and θ_2. Denote the solution by $(\hat{\theta}_1, \hat{\theta}_2)$.

(b) Show that the Jacobian of the second-order partial derivatives, evaluated at $(\hat{\theta}_1, \hat{\theta}_2)$, is positive (recall the Jacobian is $H_{11}H_{22} - H_{12}H_{21}$, where H_1 means $\frac{\partial H}{\partial \theta_1}$, and so on).

(c) Show that at least one of H_{11} or H_{22}, evaluated at $(\hat{\theta}_1, \hat{\theta}_2)$, is negative.

Instead, successive maximizations, if possible, usually makes the problem easier.

Example: Suppose $X_1, X_2, \ldots, X_n \sim$ iid $\mathrm{N}(\mu, \sigma^2)$. Find the MLEs for μ and σ^2.

The likelihood function is

$$L(\mu, \sigma^2 \mid \underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \ \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

First fix σ. The log likelihood is

$$\log L(\mu, \sigma^2 \mid \underline{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\frac{\partial \log L(\mu, \sigma^2 \mid \underline{x})}{\partial \mu} = -\frac{1}{2}\sum_{i=1}^{n} 2\,(x_i - \mu) \gtreqless 0 \ \text{ according as } \ \mu \lesseqgtr \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

So, for each σ, $\hat{\mu} = \overline{x}$ is the MLE of μ.

Plug in $\hat{\mu}$ into $\log L(\mu, \sigma^2 \mid \underline{x})$ to obtain the profile log-likelihood of σ:

$$\log \tilde{L}(\sigma^2 \mid \underline{x}) = \log L(\hat{\mu}, \sigma^2 \mid \underline{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\frac{\partial \log \tilde{L}(\sigma^2 \mid \underline{x})}{\partial\left(\sigma^2\right)} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 \gtreqless 0$$

according as

$$\sigma^2 \lesseqgtr \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

which means the MLE for σ^2 is $\hat{\sigma}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Therefore, the MLE for the (μ, σ^2) is $(\hat{\mu}, \hat{\sigma}^2) = \left(\overline{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)$.

Bayesian Approach to Statistics

(a) In the classical approach, the parameter θ is thought to be an unknown, but fixed, quantity.

(b) In the Bayesian approach, θ is considered to be a quantity whose variation can be described by a probability distribution (called the **prior distribution**).

(c) The prior distribution is subjective and is based on the experimenter's belief. It is formulated before the data are seen.

(d) A sample is then taken from a population indexed by θ, and the prior distribution is updated (using Bayes' Rule) with the sample information. The updated prior is called the **posterior distribution**.

(e) Denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(\underline{x} \mid \theta)$.

(f) The posterior distribution is the conditional distribution of θ, given the sample $\underline{x}$:

$$\begin{aligned} \pi(\theta \mid \underline{x}) &= \frac{f(\underline{x} \mid \theta)\pi(\theta)}{m(\underline{x})} \\ &= \frac{f(\underline{x}, \theta)}{m(\underline{x})}, \end{aligned}$$

where $m(\underline{x})$ is the marginal distribution of $\underline{X}$:

$$m(\underline{x}) = \int f(\underline{x} \mid \theta)\pi(\theta)\, d\theta.$$

Example (Binomial Bayes estimation): Let $X_1, X_2, \ldots, X_n \sim$ iid Bernoulli$\,(p)$, and let $= \sum_{i=1}^{n} X_i$. Then $Y \sim$ binomial(n, p).

Assume the prior distribution on p to be beta(α, β). Determine the Bayes estimator of p.

The joint distribution of Y and p is

$$\begin{aligned} f_{Y,p}(y,p) &= \left[\binom{n}{y} p^y (1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] \\ &= \binom{n}{y}\ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\ p^{y+\alpha-1}(1-p)^{n-y+\beta-1};\ y = 0, 1, \ldots, n;\ 0 \le p \le 1 \end{aligned}$$

The marginal pmf of Y is:

$$f_Y(y) = \int_0^1 f(y,p)\ dp = \binom{n}{y}\ \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\ \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}$$

The posterior pdf of θ is

$$f_{Y|p}(y \mid p) = \frac{f_{Y,P}(y,p)}{f_P(p)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\, p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

which is Beta$(y+\alpha, n-y+\beta)$.

A natural Bayesian (point) estimator is the mean of the posterior distribution, given by

$$\hat{p}_B = \mathrm{E}(p \mid Y) = \frac{y+\alpha}{n+\alpha+\beta}$$

Note that

$$\hat{p}_B = \left(\frac{n}{n+\alpha+\beta}\right) \underbrace{\left(\frac{y}{n}\right)}_{=\text{sample mean}} + \left(\frac{\alpha+\beta}{n+\alpha+\beta}\right) \underbrace{\left(\frac{\alpha}{\alpha+\beta}\right)}_{=\text{prior mean}}$$

Conjugate Family

Definition: Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x \mid \theta)$, indexed by θ. A class Π of prior distributions is a **conjugate family** for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$.

- The beta family is conjugate for the binomial family, which is why it was chosen as the prior distribution in the previous example.
- the gamma family is conjugate for the Poisson family.
- the normal family is its own conjugate.

Example (Normal Bayes Estimator) Let $X \sim \mathsf{N}(\theta, \sigma^2)$, and suppose that the prior distribution on θ is $\mathsf{N}(\mu, \tau^2)$ where σ^2, μ and τ^2 are all known.

The posterior distribution of θ is also normal (Exercise 7.22; Homework) with

$$\mathsf{E}(\theta \mid x) = \frac{\tau^2}{\sigma^2 + \tau^2} x + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu = \frac{1/\sigma^2}{1/\sigma^2 + 1/\tau^2} x + \frac{1/\tau^2}{1/\sigma^2 + 1/\tau^2} \mu$$

and

$$\mathsf{Var}(\theta \mid x) = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2} = \frac{1}{1/\sigma^2 + 1/\tau^2}$$

Using the posterior mean, a Bayes point estimator is given by $E(\theta \mid X)$

Note that the Bayes estimator is again a linear combination of prior and sample means.

Homework

- Read p. 316 – 326.
- Exercises: TBA.