Maximum likelihood estimation (part 1)

Lecture 13b (STAT 24400 F24)

The likelihood & log likelihood

Setting: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\cdot \mid \theta)$ for an unknown parameter θ

The joint density or PMF of $(X_1, \ldots, X_n)$ is:

$$\prod_{i=1}^{n} f(X_i \mid \theta) = f(X_1 \mid \theta) \cdot \ \ldots \ \cdot f(X_n \mid \theta)$$

called the **likelihood** of θ given the data

Sometimes it's convenient to work with the log likelihood:

$$\log \left(\prod_{i=1}^{n} f(X_i \mid \theta) \right) = \sum_{i=1}^{n} \log \left(f(X_i \mid \theta) \right)$$

The likelihood & log likelihood (notations)

Switching notation:

- $\theta_0 \in \Theta$ is the unknown true value of the parameter
- $\theta \in \Theta$ represents any possible value of the parameter
  (so that we can study the function Likelihood(θ), over all $\theta \in \Theta$,
  even though θ_0 is fixed)

The maximum likelihood estimator (MLE)

The maximum likelihood estimator (MLE) is the value of θ that maximizes the likelihood:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} f(X_i \mid \theta) \right\}$$

$\max_x h(x)$ = maximum value of $h(x)$

$\operatorname{argmax}_x h(x)$ = value of x that yields maximum value of $h(x)$

Often more convenient to work with log likelihood:

$$\widehat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \log \left(f(X_i \mid \theta) \right) \right\}$$

Example: MLE of Normal mean (σ^2 known)

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ for unknown $\mu \in \mathbb{R}$ (σ^2 is known)

- The density is $f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
- Likelihood function $= \prod_{i=1}^n f(x_i \mid \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2}$
- Log likelihood:

$$\sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}\right) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2$$

- Solve for MLE:

$$0 = \frac{\partial}{\partial\mu}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2\right) = \frac{1}{\sigma^2}\sum_i (X_i - \mu)$$

$$\Rightarrow \quad \widehat{\mu} = \bar{X} \text{ (same as MoM)}$$

Check: it is a global maximum point (by 2nd derivative test, 1st derivative test, etc.)

Example: MLE of Normal μ and σ^2

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ for unknown $\mu \in \mathbb{R}$, $\sigma^2 > 0$

- The density is $f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
- Log likelihood:

$$\sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(X_i-\mu)^2/2\sigma^2}\right) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2$$

- Solve for MLE $(\hat{\mu}, \hat{\sigma}^2)$:

$$0 = \frac{\partial}{\partial\mu}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2\right) = \frac{1}{\sigma^2}\sum_i (X_i - \mu)$$

$$0 = \frac{\partial}{\partial(\sigma^2)}\left(-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (X_i-\mu)^2\right) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (X_i-\mu)^2$$

$$\Rightarrow \quad \widehat{\mu} = \bar{X}, \ \widehat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2 \text{ (same as MoM)}$$

Example: MLE for Exponential rate

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Exponential}(\lambda)$ for unknown $\lambda > 0$

- The density is (using a new, useful notation)

$$f(x \mid \lambda) = \lambda e^{-\lambda x} \cdot \mathbb{1}_{x>0}$$

- Log likelihood:

$$\sum_{i=1}^n \log\left(\lambda e^{-\lambda X_i} \cdot \mathbb{1}_{X_i > 0}\right) \overset{\text{since } X_i > 0 \text{ for all } i}{=} \sum_{i=1}^n \log\left(\lambda e^{-\lambda X_i}\right) = n\log(\lambda) - \lambda \sum_i X_i$$

- Solve for MLE:

$$0 = \frac{\partial}{\partial\lambda}\left(n\log(\lambda) - \lambda\sum_i X_i\right) = \frac{n}{\lambda} - \sum_i X_i$$

$$\Rightarrow \quad \widehat{\lambda} = 1/\bar{X} \text{ (same as MoM)}$$

Check: it is a global max.

Example: MLE for Binomial

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(p)$ for unknown $p \in (0,1)$ $\rightsquigarrow \sum_i X_i \sim \mathrm{Binomial}(n, p)$

- The PMF is (using a new expression)

$$f(x \mid p) = p^x (1-p)^{1-x}$$

- Log likelihood:

$$\sum_{i=1}^n \log\left(p^{X_i}(1-p)^{1-X_i}\right) = \sum_i X_i \cdot \log(p) + \sum_i (1 - X_i) \cdot \log(1-p)$$

- Solve for MLE:

$$0 = \frac{\partial}{\partial p}\left(\sum_i X_i \cdot \log(p) + \sum_i (1-X_i)\cdot\log(1-p)\right) = \frac{\sum_i X_i}{p} - \frac{\sum_i (1-X_i)}{1-p}$$

$$\Rightarrow \quad \widehat{p} = \bar{X} \text{ (= the proportion of successes in the sample)}$$

Example: MLE for Uniform

$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Uniform}[0,\theta]$ for unknown $\theta > 0$

- The density is
$$f(x \mid \theta) = \frac{1}{\theta} \cdot \mathbb{1}_{0 \le x \le \theta}$$
- Likelihood:
$$\prod_{i=1}^n \left(\frac{1}{\theta} \cdot \mathbb{1}_{0 \le X_i \le \theta} \right) \overset{\substack{\text{since } X_i \ge 0 \text{ for all } i \\ \downarrow}}{=} \theta^{-n} \cdot \mathbb{1}_{\theta \ge \max_i X_i}$$
- Solve for MLE:
$$\widehat{\theta} = \max_i X_i = X_{(n)} \ (\textit{not} \text{ same as MoM})$$

MoM vs MLE (Example for Uniform)

Compare MoM and MLE estimators for Uniform$[0,\theta]$:

MoM estimator $\widehat{\theta} = 2\bar{X}$:

- For each X_i, $\mathbb{E}(X_i) = \frac{\theta}{2}$, $\text{Var}(X_i) = \frac{\theta^2}{12}$
- Bias:
$$\mathbb{E}(\widehat{\theta}) = 2\mathbb{E}(\bar{X}) = 2 \cdot \frac{\theta}{2} = \theta \quad \Rightarrow \quad \text{bias} = 0$$
- Variance:
$$\text{Var}(\widehat{\theta}) = \text{Var}\left(\frac{2}{n} \sum_i X_i \right) = \frac{4}{n^2} \text{Var}(\sum_i X_i) = \frac{4}{n^2} \sum_i \text{Var}(X_i) = \frac{\theta^2}{3n}$$
- MSE $= \text{bias}^2 + \text{Var}(\widehat{\theta}) = \frac{\theta^2}{3n}$

MoM vs MLE (Example for Uniform)

Compare MoM and MLE estimators for Uniform$[0,\theta]$:

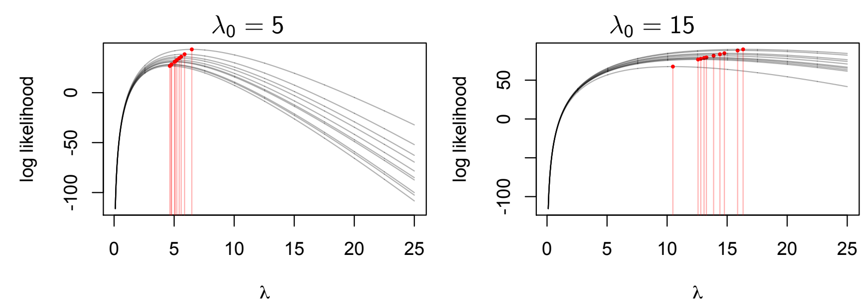
MLE estimator $\widehat{\theta} = X_{(n)}$:

- The sampling distribution has density $\quad f(x) = \frac{nx^{n-1}}{\theta^n} \cdot \mathbb{1}_{0 \le x \le \theta}$
- Bias:
$$\mathbb{E}(\widehat{\theta}) = \int_{x=0}^{\theta} x \cdot \frac{nx^{n-1}}{\theta^n} \, \mathrm{d}x = \frac{n\theta}{n+1} \quad \Rightarrow \quad \text{bias} = -\frac{\theta}{n+1}$$
- Variance:
$$\mathbb{E}(\widehat{\theta}^2) = \int_{x=0}^{\theta} x^2 \cdot \frac{nx^{n-1}}{\theta^n} \, \mathrm{d}x = \frac{n\theta^2}{n+2}$$
$$\Rightarrow \quad \text{Var}(\widehat{\theta}) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1} \right)^2 = \frac{n\theta^2}{(n+1)^2(n+2)}$$
- MSE $= \text{bias}^2 + \text{Var}(\widehat{\theta}) = \frac{2\theta^2}{(n+1)(n+2)}$ (comparison: By MoM, MSE $= \frac{\theta^2}{3n}$)

Accuracy of the MLE

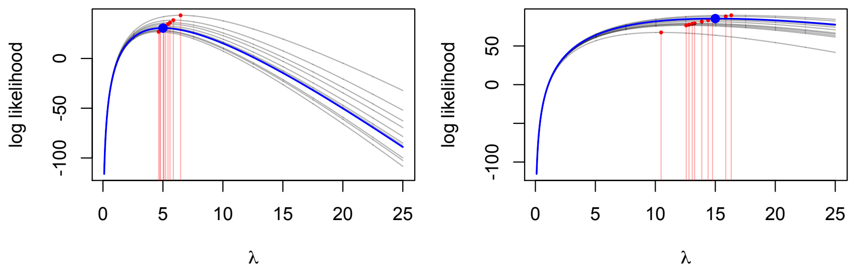
Example: suppose $X_1, \ldots, X_{50} \overset{\text{iid}}{\sim} \text{Exponential}(\lambda)$

Here is a plot of the log likelihood function, and the MLE, over 10 trials:



Accuracy of the MLE

Add in the expected log likelihood curve (with λ_0 highlighted):



$\Rightarrow$ higher curvature leads to a more accurate estimate

Fisher information (Example: Exponential rate)

(Now we are going to work with one-dimensional θ.)

The **Fisher information** is defined as:

$$\mathcal{I}(\theta) = \mathbb{E}\left(\left(\frac{\partial}{\partial\theta}\log\left(f(X \mid \theta)\right)\right)^2\right) = \mathbb{E}\left(-\frac{\partial^2}{\partial\theta^2}\log\left(f(X \mid \theta)\right)\right)$$

with some regularity conditions

(smoothness of $\log(f)$ as a function of θ)

Example:

- Exponential(λ):

$$-\frac{\partial^2}{\partial\lambda^2}\log\left(f(X \mid \lambda)\right) = -\frac{\partial^2}{\partial\lambda^2}\log(\lambda e^{-\lambda X}) = -\frac{\partial^2}{\partial\lambda^2}\left[\log(\lambda) - \lambda X\right] = \frac{1}{\lambda^2}$$

$$\mathcal{I}(\lambda) = \mathbb{E}\left(-\frac{\partial^2}{\partial\lambda^2}\log\left(f(X \mid \lambda)\right)\right) = \mathbb{E}\left(\frac{1}{\lambda^2}\right) = \frac{1}{\lambda^2}$$

Fisher information example (Normal mean)

Example:

- $N(\mu, \sigma^2)$ with σ^2 known:

$$-\frac{\partial^2}{\partial\mu^2}\log\left(f(X \mid \mu)\right) = -\frac{\partial^2}{\partial\mu^2}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(X-\mu)^2/2\sigma^2}\right)$$
$$= -\frac{\partial^2}{\partial\mu^2}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}\right] = \frac{1}{\sigma^2}$$

$$\mathcal{I}(\mu) = \mathbb{E}\left(\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}$$

Fisher information example (Bernoulli)

Example:

- Bernoulli(p):

$$-\frac{\partial^2}{\partial p^2}\log\left(f(X \mid p)\right) = -\frac{\partial^2}{\partial p^2}\log(p^X(1-p)^{1-X})$$
$$= -\frac{\partial^2}{\partial p^2}\left[X\log(p) + (1-X)\log(1-p)\right] = \frac{X}{p^2} + \frac{1-X}{(1-p)^2}$$

$$\mathcal{I}(p) = \mathbb{E}\left(-\frac{\partial^2}{\partial p^2}\log\left(f(X \mid p)\right)\right) = \frac{\mathbb{E}(X)}{p^2} + \frac{1-\mathbb{E}(X)}{(1-p)^2} = \frac{1}{p(1-p)}$$

Asymptotic distribution of the MLE — Fisher's Theorem

The Fisher information determines the (approximate) variance of the MLE.

Informally: if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(\cdot \mid \theta_0)$ and $\widehat{\theta}$ is the MLE,

$$\text{the distribution of } \widehat{\theta} \text{ is } \approx \mathrm{N}\big(\theta_0, \tfrac{1}{n\mathcal{I}(\theta_0)}\big)$$

More formally: under some regularity conditions,

$$\sqrt{n\mathcal{I}(\theta_0)} \cdot (\widehat{\theta} - \theta_0) \to \mathrm{N}(0,1)$$

"converges in distribution"

This means that the CDF converges — i.e., for all fixed $x \in \mathbb{R}$,

$$\mathbb{P}\left(\sqrt{n\mathcal{I}(\theta_0)} \cdot (\widehat{\theta} - \theta_0) \leq x\right) \to \Phi(x)$$

The same holds with $\mathcal{I}(\widehat{\theta})$ in place of $\mathcal{I}(\theta_0)$: (very useful in practice)

$$\sqrt{n\mathcal{I}(\widehat{\theta})} \cdot (\widehat{\theta} - \theta_0) \to \mathrm{N}(0,1)$$

Asymptotic distribution of the MLE (examples)

Examples:

- Exponential(λ): $\widehat{\lambda} = 1/\bar{X}$ and $\mathcal{I}(\lambda) = 1/\lambda^2$, so:

$$\widehat{\lambda} \approx \mathrm{N}\big(\lambda_0, \tfrac{\lambda_0^2}{n}\big) \text{ or } \approx \mathrm{N}\big(\lambda_0, \tfrac{\widehat{\lambda}^2}{n}\big)$$

- $\mathrm{N}(\mu, \sigma^2)$ with σ^2 known: $\widehat{\mu} = \bar{X}$ and $\mathcal{I}(\mu) = 1/\sigma^2$ so:

$$\widehat{\mu} \approx \mathrm{N}\big(\mu_0, \frac{\sigma^2}{n}\big)$$

  (In fact, in this case we know this is the *exact* distribution!)

- Bernoulli(p): $\widehat{p} = \bar{X}$ and $\mathcal{I}(p) = \frac{1}{p(1-p)}$, so:

$$\widehat{p} \approx \mathrm{N}\big(p_0, \tfrac{p_0(1-p_0)}{n}\big) \text{ or } \approx \mathrm{N}\big(p_0, \tfrac{\widehat{p}(1-\widehat{p})}{n}\big)$$

Asymptotic distribution of the MLE (counterexamples)

Examples:

- Uniform$[0, \theta]$: in this case the regularity conditions do not hold.

  We need $\log(f(X \mid \theta))$ to be smooth as a function of θ,
  but $\log(f(X \mid \theta)) = \log(0) = -\infty$ if $X > \theta$.

  To confirm the theorem doesn't hold:

  We've calculated $\mathrm{Var}(\widehat{\theta}) = \frac{n\theta^2}{(n+1)^2(n+2)} = \mathcal{O}(\frac{1}{n^2})$,
  while asymptotic normality of the MLE would yield $\mathrm{Var}(\widehat{\theta}) = \mathcal{O}(\frac{1}{n})$

  In fact, no approximation is needed here, since we actually know
  the *exact* distribution of the MLE in this case (via order statistics)