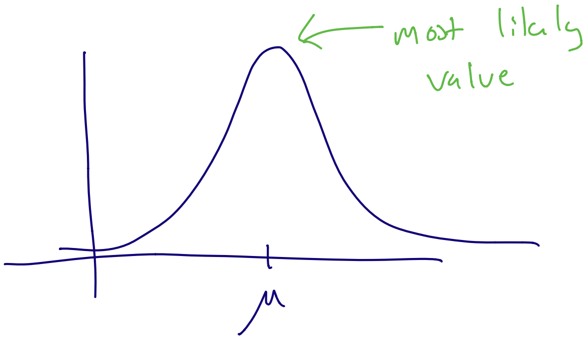
Statistics

Feb 24, 2020

Example: Imagine you have a single piece of data x_1, which you believe is an observation from a normal distribution: $N(\mu, \sigma^2)$.



$\Rightarrow$ best guess at μ is $\hat{\mu} = x_1$.

I.e. we chose the value of μ that maximized

$$f(x_1; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_1-\mu)^2/2\sigma^2}. \Big\} \text{ PDF of } N(\mu,\sigma^2) \text{ evaluated at } x_1.$$

$\Rightarrow$ This is in essence the method of Maximum Likelihood.

If $X_1, \dots, X_n$ are a collection of random variables with joint PDF $f = f(x_1, \dots, x_n; \theta)$, then the Likelihood function is:

$$\mathcal{L}(\theta) = f(x_1, \dots, x_n; \theta)$$

$$= f(x_1;\theta)\, f(x_2;\theta) \dots f(x_n;\theta) \quad \left(\text{if } X_1,\dots,X_n \text{ are IID}\right)$$

$$= \prod_{i=1}^{n} f(x_i; \theta)$$

Log-likelihood: $\log \mathcal{L}(\theta) = \ell(\theta)$

$$= \sum_{i=1}^{n} \log f(x_i;\theta) \quad \text{if IID.}$$

The Maximum Likelihood estimator $\hat{\theta}$ is the value of θ which maximizes $\mathcal{L}(\theta)$ or $\ell(\theta)$.

Notationally : $\mathcal{L}(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x_i-\mu)/2\sigma^2}$

$$\underset{\sim}{\propto} \frac{1}{\sigma} e^{-(x_i-\mu)/2\sigma^2}$$

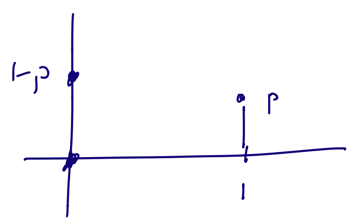
<u>Example</u>: $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ IID r.v.'s

$X_i = 1$ with prob p

$X_i = 0$ with prob $1-p$.

$\Rightarrow$ mass p at $X_i = 1$

mass $1-p$ at $X_i = 0$



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

$$\Rightarrow \mathcal{L}(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

$$= p^{\sum x_i} (1-p)^{n-\sum x_i}$$

$$\ell(p) = \left(\sum x_i\right) \log p + \left(n - \sum x_i\right) \log(1-p).$$

$\Rightarrow$ Solve $\ell'(p) = 0$

$$\ell'(p) = \frac{1}{p} \sum x_i - \frac{1}{1-p}\left(n - \sum x_i\right) = 0.$$

$$\Rightarrow \boxed{\hat{p} = \frac{1}{n} \sum x_i}.$$

Example: Let $X_1, \ldots, X_n \sim U(0, \theta)$ IID.

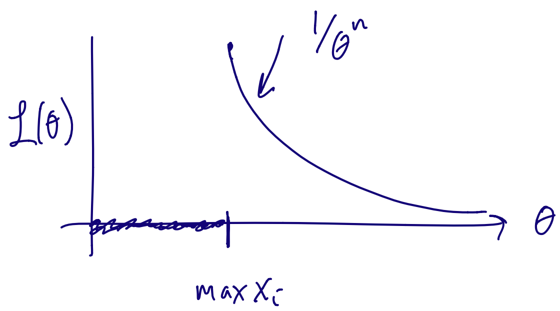
$$f(x;\theta) = \begin{cases} 1/\theta & \text{for } x \in [0,\theta] \\ 0 & \text{otherwise} \end{cases}$$

$$\cancel{\mathcal{L}'(\theta) = \frac{d}{d\theta}\frac{1}{\theta^n} = -n\frac{1}{\theta^{n-1}} = 0}$$

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

$$= \begin{cases} 1/\theta^n & \text{if all } x_i \le \theta \\ 0 & \text{otherwise, i.e. if } \max x_i > \theta. \end{cases}$$

Plot:



$\Rightarrow \hat{\theta} = \max X_i$.

Properties of the Maximum Likelihood Estimator

Consistency

true value of θ

$\Rightarrow \hat{\theta} \xrightarrow{\mathbb{P}} \theta_*$

i.e. $\mathbb{P}(|\hat{\theta} - \theta_*| > \epsilon) \to 0$ for any $\epsilon > 0$ as $n \to \infty$.

Sketch of Proof

Consider first the Kullback-Leibler "distance"

$$D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx.$$

"distance" between pdfs f and g.

$$\Rightarrow \quad D(f,g) \geq 0$$
$$D(f,f) = 0$$

Write $D(\theta,\psi) = D\left(f(x;\theta), f(x;\psi)\right)$.

We say that the statistical model $\mathcal{F}$ is identifiable if $\theta \neq \psi \Rightarrow D(\theta,\psi) > 0$.

Now, maximizing $\ell(\theta)$ is equivalent to maximizing

$$M(\theta) = \frac{1}{n}\sum_i \log \frac{f(x_i;\theta)}{f(x_i;\theta_*)}$$
$$= \frac{1}{n}\sum\left(\log f(x_i;\theta) - \log f(x_i;\theta_*)\right)$$
$$= \frac{1}{n}\left(\ell(\theta) - \ell(\theta_*)\right)$$

constant with respect to θ.

If $\mathbb{E}\left(\log \frac{f(x_i;\theta)}{f(x_i;\theta_*)}\right)$ exists, then by the Law of Large Numbers, as $n \to \infty$ $M(\theta)$ converges to

$$\mathbb{E}_{\theta_*}\left(\log \frac{f(x_i;\theta)}{f(x_i;\theta_*)}\right) = \int \log \frac{f(x;\theta)}{f(x;\theta_*)} f(x;\theta_*)\, dx$$
$$= -\int \log \frac{f(x;\theta_*)}{f(x;\theta)} f(x;\theta_*)\, dx$$
$$= -D(\theta_*, \theta).$$

So for large n, $M(\theta) \simeq -D(\theta_*, \theta)$, which is maximized at $\theta = \theta_*$ since $-D(\theta_*, \theta_*) = 0$ and $-D(\theta_*, \theta) < 0$ for $\theta \neq \theta_*$.

$\Rightarrow$ The maximizer of $M(\theta)$ tends to θ_*.

Equivariance

Thm: Let $\tau = g(\theta)$ be a function of θ. Let $\hat{\theta}$ be the MLE of θ. Then $\hat{\tau} = g(\hat{\theta})$ is the MLE of τ.

Asymptotic Normality

Goal Show that $\hat{\theta} \to N(\theta_*, ?)$

Def: Score function

$$s(x;\theta) = \frac{\partial}{\partial\theta} \log f(x;\theta)$$

Fisher Information:

$$I(\theta) = \mathrm{Var}(s(X;\theta))$$

$$I_n(\theta) = \mathrm{Var}\left(\sum_i s(X_i;\theta)\right)$$

$$= \sum_i \mathrm{Var}(s(X_i;\theta)).$$

and since X_i are IID

Compute the expected value of score function.

$$\mathbb{E}\left(s(X;\theta)\right) = \int \frac{\partial}{\partial\theta} \log f(x;\theta) \; f(x;\theta) \; dx$$

$$= \int \frac{1}{\cancel{f(x;\theta)}} \left(\frac{\partial}{\partial\theta} f(x;\theta) \right) \cancel{f(x;\theta)} \; dx$$

$$= \frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} f(x;\theta) \; dx$$

$$= \frac{\partial}{\partial\theta} 1 = 0$$

$$\Rightarrow \quad \mathrm{Var}\left(s(x;\theta) \right) = \mathbb{E}\left(s(x;\theta)^2 \right) - \left(\mathbb{E}\left(s(x;\theta) \right) \right)^2$$

$$= \mathbb{E}\left(s(x;\theta)^2 \right)$$

Thm: $I_n(\theta) = n\, I(\theta)$ and furthermore,

$$I(\theta) = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial\theta^2} \log f(x;\theta) \right)$$

$$= -\mathbb{E}_\theta\left(\frac{\partial}{\partial\theta} s(x;\theta) \right).$$

$$= \mathrm{Var}\left(s(x;\theta) \right)$$

$$= \mathbb{E}\left(s(X;\theta)^2 \right)$$

Thm: Let $se = \sqrt{\mathrm{Var}(\hat\theta)}$, "under appropriate regularity conditions"

① $se \approx \sqrt{1/I_n(\theta)}$ and $\dfrac{\hat\theta - \theta}{se} \rightsquigarrow N(0,1)$

② Let $\hat{se} \approx \sqrt{1/I_n(\hat\theta)}$, then $\dfrac{\hat\theta - \theta}{\hat{se}} \rightsquigarrow N(0,1)$.

Asymptotic Confidence Intervals

Let $C = \left(\hat{\theta} - z_{\alpha/2}\, \hat{se} \,,\ \hat{\theta} + z_{\alpha/2}\, \hat{se} \right)$

then $\mathbb{P}_\theta(\theta \in C) \to 1-\alpha$ as $n \to \infty$.

(same exact proof as before, just rearrange terms).

Optimality

Suppose we want to estimate μ from $X_1, \dots, X_n \sim N(\mu, \sigma^2)$ IID.

Let $\hat{\mu} = \text{MLE} = \frac{1}{n} \sum X_i$.

We could alternatively estimate μ using $\tilde{\mu} = \text{median}(X_1, \dots, X_n)$.

It can be shown that

$$\sqrt{n}(\hat{\mu} - \mu) \rightsquigarrow N(0, \sigma^2)$$

$$\sqrt{n}(\tilde{\mu} - \mu) \rightsquigarrow N(0, \sigma^2 \pi/2).$$

In general, let T, U be two estimators of θ, each of which is asymptotically normal:

$$\sqrt{n}(T - \theta) \rightsquigarrow N(0, t^2)$$

$$\sqrt{n}(U - \theta) \rightsquigarrow N(0, u^2)$$

$\text{ARE}(U, T)$ = asymptotic relative efficiency of U to T

$= t^2/u^2$ = ratio of variances.

Back to our example:

$$ARE(\tilde{\mu}, \hat{\mu}) = \frac{Var(\hat{\mu})}{Var(\tilde{\mu})} = \frac{\sigma^2}{\frac{\pi}{2}\sigma^2} = \frac{2}{\pi} \approx .63.$$

<u>Thm</u>: If $\hat{\theta}$ is the MLE of θ, and $\tilde{\theta}$ is any other asymptotically normal estimator, then

$$ARE(\tilde{\theta}, \hat{\theta}) \leq 1.$$

$\Rightarrow$ the MLE is <u>efficient</u> or <u>asymptotically optimal</u>

$\Rightarrow$ MLE has the smallest asymptotic variance.

<u>Multiparameter Models</u>

Extend to models with several parameters:

Let $\theta = (\theta_1, \ldots, \theta_k)$, and let
$\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$ be the MLE, i.e. the solution to the system of equations

$$\left.\begin{aligned} \frac{\partial}{\partial\theta_1} \ell(\theta) &= 0 \\ \frac{\partial}{\partial\theta_2} \ell(\theta) &= 0 \\ &\vdots \\ \frac{\partial}{\partial\theta_k} \ell(\theta) &= 0 \end{aligned}\right\} \text{ system of } k \text{ equations in the } k \text{ unknowns } \theta_1, \ldots, \theta_k.$$

Let us also define $H_{jk} = \dfrac{\partial^2 \ell}{\partial\theta_j \partial\theta_k}$

Fisher Information Matrix:

$$I_n(\theta) = -\begin{pmatrix} \mathbb{E}(H_{11}) & \cdots & \mathbb{E}(H_{1k}) \\ \mathbb{E}(H_{21}) & & \vdots \\ \vdots & & \vdots \\ \mathbb{E}(H_{k1}) & & \mathbb{E}(H_{kk}) \end{pmatrix}$$

and $J_n = I_n^{-1}$. (Question for home: why does J_n exist?).

Thm Under the same regularity conditions on f as before,

$$\hat{\theta} - \theta \approx N(0, J_n).$$

this is a k-dimensional vector.

And furthermore if $\hat{\theta}_j$ is the j^{th} component of $\hat{\theta}$, then

$$\frac{\hat{\theta}_j - \theta_j}{\hat{se}_j} \rightsquigarrow N(0,1)$$

where $\hat{se}_j^2 = J_n(j,j)$

and $\mathrm{Cov}(\hat{\theta}_j, \hat{\theta}_k) \approx J_n(j,k)$.