18.05 Lecture 25
April 11, 2005

Maximum Likelihood Estimators
$X_1, ..., X_n$ have distribution $\mathbb{P}_{\theta_0} \in \{\mathbb{P}_\theta : \theta \in \Theta\}$
Joint p.f. or p.d.f.: $f(x_1, ..., x_n) = f(x_1|\theta) \times ... \times f(x_n|\theta) = \psi(\theta)$ - likelihood function.
If $\mathbb{P}_\theta$ - discrete, then $f(x|\theta) = \mathbb{P}_\theta(X = x)$,
and $\psi(\theta)$ - the probability to observe $X_1, ..., X_n$

Definition: A Maximum likelihood estimator (M.L.E.):
$\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$ such that $\psi(\hat{\theta}) = \max_\theta \psi(\theta)$
Suppose that there are two possible values of the parameter, $\theta = 1, \theta = 2$
p.f./p.d.f. - $f(x|1), f(x|2)$
Then observe points $x_1, ..., x_n$
view probability with first parameter and second parameter:
$\psi(1) = f(x_1, ..., x_n|1) = 0.1, \psi(2) = f(x_1, ..., x_n|2) = 0.001$,
The parameter is much more likely to be 1 than 2.

Example: Bernoulli Distribution B(p), $p \in [0.1]$,
$\psi(p) = f(x_1, ..., x_n|p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$
$\psi(\theta) \to \max \leftrightarrow \log \psi(\theta) \to \max$ (log-likelihood)
$\log \psi(p) = \sum x_i \log p + (n - \sum x_i)\log(1-p)$, maximize over $[0, 1]$
Find the critical point:

$$\frac{\partial}{\partial p} \log \psi(p) = 0$$

$$\frac{\sum x_i}{p} - \frac{n - \sum x_i}{1-p} = 0$$

$$\sum x_i(1-p) - p(n - \sum x_i) = \sum x_i - p\sum x_i - np + p\sum x_i = 0$$

$$\hat{p} = \frac{\sum x_i}{n} = \overline{x} \to \mathbb{E}(X) = p$$

For Bernoulli distribution, the MLE converges to the actual parameter of the distribution, p.

Example: Normal Distribution: $N(\mu, \sigma^2)$,

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\psi(\mu, \sigma^2) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$$

$$\log \psi(\mu, \sigma^2) = n \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2 \to \max : \mu, \sigma^2$$

Note that the two parameters are decoupled.

First, for a fixed σ, we minimize $\sum_{i=1}^n (x_i - \mu)^2$ over μ

$$\frac{\partial}{\partial \mu}\sum_{i=1}^{n}(x_i-\mu)^2=-\sum_{i=1}^{n}2(x_i-\mu)=0,$$

$$\sum_{i=1}^{n}x_i-n\mu=0,\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i=\overline{x}\to\mathbb{E}(X)=\mu_0$$

To summarize, the estimator of μ for a Normal distribution is the sample mean.

To find the estimator of the variance:

$$-n\log(\sqrt{2\pi}\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\overline{x})^2\to \text{ maximize over } \sigma$$

$$\frac{\partial}{\partial \sigma}=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\overline{x})^2=0$$

$$\hat{\sigma}^2=\frac{1}{n}\sum(x_i-\overline{x})^2\text{- MLE of } \sigma_0^2; \hat{\sigma}^2- \text{ a sample variance}$$

Find $\hat{\sigma}^2$

$$\hat{\sigma}^2=\frac{1}{n}\sum(x_i^2-2x_i\overline{x}+(\overline{x})^2)=\frac{1}{n}\sum x_i^2-2\overline{x}\frac{1}{n}\sum x_i+(\overline{x})^2=\frac{1}{n}\sum x_i^2-2(\overline{x})^2+(\overline{x})^2=$$

$$=\frac{1}{n}\sum x_i^2-(\overline{x})^2=\overline{x^2}-(\overline{x})^2\to\mathbb{E}(x_1^2)-\mathbb{E}(x_1)^2=\sigma_0^2$$

To summarize, the estimator of σ_0^2 for a Normal distribution is the sample variance.

Example: $U(0,\theta),\theta>0$ - parameter.

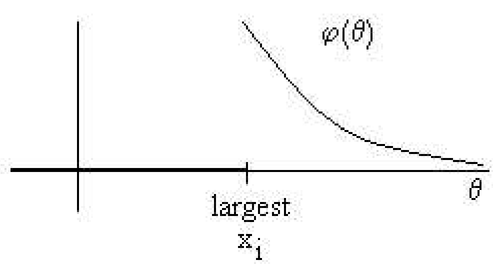
$$f(x|\theta)=\{\frac{1}{\theta},0\le x\le\theta;0, \text{ otherwise }\}$$

Here, when finding the maximum we need to take into account that the distribution is supported on a finite interval $[0,\theta]$.

$$\psi(\theta)=\prod_{i=1}^{n}\frac{1}{\theta}I(0\le x_i\le\theta)=\frac{1}{\theta^n}I(0\le x_1,x_2,...,x_n\le\theta)$$

The likelihood function will be 0 if any points fall outside of the interval.
If θ will be the correct parameter with $\mathbb{P}=0$,
you chose the wrong θ for your distribution.

$\psi(\theta)\to$ maximize over $\theta>0$



If you graph the p.d.f., notice that it drops off when θ drops below the maximum data point.

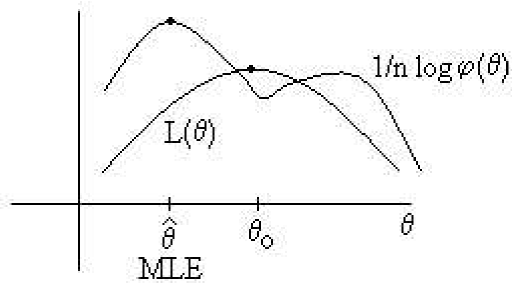
$$\hat{\theta} = \max(X_1, ..., X_n)$$

The estimator converges to the actual parameter θ_0:
As you keep choosing points, the maximum gets closer and closer to θ_0

Sketch of the consisteny of MLE.

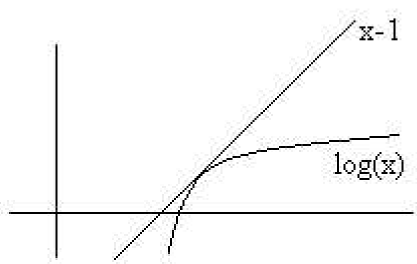
$$\psi(\theta) \to \max \Leftrightarrow \frac{1}{n} \log \psi(\theta) \to \max$$

$$L_n(\theta) = \frac{1}{n} \log \psi(\theta) = \frac{1}{n} \log \prod f(x_i|\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(x_i|\theta) \to L(\theta) = \mathbb{E}_{\theta_0} \log f(x_1|\theta).$$



$L_n(\theta)$ is maximized at $\hat{\theta}$, by definition of MLE. Let us show that $L(\theta)$ is maximized at θ_0.
Then, evidently, $\hat{\theta} \to \theta_0$. $L(\theta) \le L(\theta_0)$:
Expand the inequality:

$$\begin{aligned} L(\theta) - L(\theta_0) &= \int \log \frac{f(x|\theta)}{f(x|\theta_0)} f(x|\theta_0) dx \le \int \left(\frac{f(x|\theta)}{f(x|\theta_0)} - 1 \right) f(x|\theta_0) dx \\ &= \int \left(f(x|\theta) - f(x|\theta_0) \right) dx = 1 - 1 = 0. \end{aligned}$$



Here, we used that the graph of the logarithm will be less than the line y = x - 1 except at the tangent point.

** End of Lecture 25