Law of Large Numbers

A. Consider iid r.v.'s $X_1, X_2, X_3, \cdots$, where
$E[X_i] = \mu < \infty$ and $Var(X_i) = \sigma^2 < \infty$.
Define sample mean $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.
$E[\overline{X}_n] = E[\frac{1}{n}(X_1 + \cdots + X_n)] = \frac{1}{n} \cdot n \mu = \mu$.
$Var(\overline{X}_n) = Var(\frac{1}{n}(X_1 + \cdots + X_n)) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{1}{n}\sigma^2$.
As n increases, $\overline{X}_n$ more likely to be close to μ.

B. Theorem (Weak Law of Large Numbers): For any $\varepsilon > 0$,
as $n \to \infty$, $P(|\overline{X}_n - \mu| > \varepsilon) \to 0$.
Proof: By Chebyshev,
$P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0$ as $n \to \infty$.

C. Strong law of large numbers:
$P(\lim_{n\to\infty} \overline{X}_n = \mu) = 1$.
Strong implies weak, but not vice versa.

D. Demos: 10limits.pdf, slides 14–19, ~~21–25~~

E. Note that laws of large numbers just talk about
$\overline{X}_n \to \mu$, whereas CLT gives entire distribution
of $\overline{X}_n$ as normal.

Maximum Likelihood Estimators

A. Parameter estimation

Given independent samples $x_1, x_2, \ldots, x_n$ from a distribution $f(x|\theta)$, estimate θ.

Ex: Given samples HHTHH of flips of a coin, estimate $\theta = P(\text{heads})$. Recall also discussion of estimating μ, σ^2 for Chebyshev

B. $P(x|\theta)$: Prob of event x given model θ.

Viewed as a function of x (θ fixed), it's a probability.

Viewed as a function of θ (x fixed), it's a likelihood, and often written $L(x|\theta)$.

Ex What θ makes HHTHH most likely? I.e., what θ maximizes $L(HHTHH|\theta)$?

C. Maximum likelihood estimation:

What θ maximizes $L(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$, where $x_1, x_2, \ldots, x_n$ are independent samples from $f(x|\theta)$?

Approach: $\frac{\partial}{\partial\theta} L(\vec{x}|\theta) = 0$ or $\frac{\partial}{\partial\theta} \ln L(\vec{x}|\theta) = 0$

D. Ex: θ = prob of heads, n ind flips $x_1, x_2, \ldots, x_n$ yielding n_0 tails, n_1 heads, $n_0 + n_1 = n$.

$L(x_1, \ldots, x_n|\theta) = (1-\theta)^{n_0}\theta^{n_1}$ (Include factor $\binom{n}{n_0}$?)

$\ln L(x_1, \ldots, x_n|\theta) = n_0 \ln(1-\theta) + n_1 \ln\theta$

$\frac{\partial}{\partial\theta} \ln L(x_1, \ldots, x_n|\theta) = -\frac{n_0}{1-\theta} + \frac{n_1}{\theta} = 0$

$-n_0\hat{\theta} + n_1(1-\hat{\theta}) = 0$

$n_1 = \hat{\theta}(n_0 + n_1)$, $\hat{\theta} = \frac{n_1}{n}$

$\frac{\partial^2}{\partial\theta^2} \ln L(x_1, \ldots, x_n|\theta) = \frac{-n_0}{(1-\theta)^2} - \frac{n_1}{\theta^2} < 0$,

so concave downward everywhere; $\hat{\theta}$ is max and boundary values are less.

Note that $\frac{n_1}{n}$ is the fraction of heads in the sample, and ~~the~~ a good estimate of the prob of heads in the distribution

E. Ex.: $x_i \sim N(\mu, \sigma^2)$, both $\underset{\theta_1 =}{\mu}$ and $\underset{\theta_2 =}{\sigma^2}$ unknown,

$$L(x_1, x_2, \cdots, x_n | \theta_1, \theta_2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} e^{-(x_i-\theta_1)^2/2\theta_2}$$

$$\ln L(x_1, x_2 \cdots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n}\left(-\frac{1}{2}\ln(2\pi\theta_2) - \frac{(x_i-\theta_1)^2}{2\theta_2}\right)$$

$$\frac{\partial}{\partial\theta_1} \ln L(x_1, x_2, \cdots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n} \frac{-2(x_i-\theta_1)}{2\theta_2} = 0$$

$$\sum_{i=1}^{n} x_i = n\hat{\theta}_1, \quad \hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Sample mean is MLE of population mean μ.

But: density is not probability, so why can we take product of densities in $L(\vec{x}|\theta)$?

Answer: for small δ, $P(x_i - \frac{\delta}{2} < Y < x_i + \frac{\delta}{2}) \approx \delta f(x_i)$.

When we take ln and $\frac{\partial}{\partial\theta}$, δ drops out.

Check that $\hat{\theta}_1$ is global maximum:

$$\frac{\partial^2}{\partial\theta_1^2} \ln L(x_1, \cdots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n} \frac{-2}{2\theta_2} < 0,$$ so conc. downward everywhere

E. $$\frac{\partial}{\partial\theta_2} \ln L(x_1, x_2, \cdots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n}\left(-\frac{1}{2}\cdot\frac{2\pi}{2\pi\theta_2} + \frac{(x_i-\theta_1)^2}{2\theta_2^2}\right) = 0$$

$$\sum_{i=1}^{n}\left(-\hat{\theta}_2 + (x_i - \hat{\theta}_1)^2\right) = 0$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_1)^2 \cancel{= \frac{1}{n}\sum_{i=1}(x_i - \hat{\theta}_1)^2}$$ (Setting these 2 eqns to 0 gives 2 eqns in 2 variables to solve.)

Sample variance is MLE of population variance.

$$\frac{\partial^2}{\partial\theta_2^2} \ln L(x_1, x_2, \cdots, x_n | \theta_1, \theta_2) = \sum_{i=1}^{n}\left(\frac{1}{2\theta_2^2} - \frac{(x_i-\theta_1)^2}{\theta_2^3}\right)$$

$$= \frac{n}{2\theta_2^2} - \frac{1}{\theta_2^3}\sum_{i=1}^{n}(x_i - \theta_1)^2$$

At $\theta_2 = \hat{\theta}_2$, this is negative.