STAT 383C: Statistical Modeling I **Fall 2016**

Lecture 2 — August 30

Lecturer: Purnamrita Sarkar *Scribe: Spencer Woody*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley.

2.1 Review of Slutsky's Theorem and the Delta Method

2.1.1 Slutsky's Theorem

Theorem 2.1.

If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ where c is a constant, then

$$X_n + Y_n \xrightarrow{d} X + c \tag{2.1}$$

and

$$X_n Y_n \xrightarrow{d} Xc. \tag{2.2}$$

2.1.2 Delta method (convergence in density)

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma} \xrightarrow{d} N(0, 1) \tag{2.3}$$

More generally, if $g'(\theta)$ exists and is nonzero, then

$$\sqrt{n}\frac{g(\hat{\theta}) - g(\theta)}{\sigma|g'(\theta)|} \xrightarrow{d} N(0, 1). \tag{2.4}$$

2.2 Maximum Likelihood Estimators (MLEs)

2.2.1 An example with the Bernoulli distribution

Suppose we want to estimate a parameter p from $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, i.e.,

$$P(X = x) = \begin{cases} p & x = 1 \\ 1 - p & x = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{2.5}$$

The likelihood function for p is

$$P(X_1 = x_1, \ldots, X_n = x_n; p) = \prod_{i=1}^{n} p(X_i = x_i; p) \tag{2.6}$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}. \tag{2.7}$$

We must maximize this function in order to find the MLE for p. However, this product equation is messy to maximize as is, so we will apply the logarithm transform to obtain the log likelihood. Maximizing the log likelihood will give us the same MLE because the logarithm is *monotonic* (i.e., it is strictly increasing).

$$\ell(p) = \log \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \tag{2.8}$$

$$= \sum_{i=1}^{n} (x_i \log(p) + (1 - x_i) \log(1-p)) \tag{2.9}$$

We maximize $\ell(p)$ by setting the first derivative to zero.

$$\frac{d}{dp}\ell(p) = \sum_{i=1}^{n} \left(\frac{x_i}{p} - \frac{1 - x_i}{1-p} \right) \tag{2.10}$$

$$= \frac{1}{p} \sum_{i=1}^{n} x_i - \frac{1}{1-p} \sum_{i=1}^{n} (1 - x_i) = 0 \tag{2.11}$$

$$\Rightarrow \hat{p} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \tag{2.12}$$

This is a *closed-form* solution to solving the problem of maximum likelihood. Sometimes, however, there is no closed-form solution to this problem and thus we rely upon the concativity of the likelihood function and use iterative methods (e.g., gradient descent) for optimization.

2.2.2 Concativity

A function f is concave on an interval if for any x_1 and x_2 in the interval and any $\alpha \in [0, 1]$,

$$f((1-\alpha)x_1 + \alpha y_1) \geq (1-\alpha)f(x_1) + \alpha f(y_1). \tag{2.13}$$

When $f'(x)$ and $f''(x)$ both exist, $f''(x) \leq 0$ iff f is concave.

From the Bernoulli example, the second derivative of the likelihood is

$$\frac{d^2}{dp^2}\ell(p) = -\sum_{i=1}^{n}\left(\frac{x_i}{p^2} + \frac{1-x_i}{(1-p)^2}\right) < 0. \tag{2.14}$$

2.2.3 Properties of MLEs

1. Consistency

Under regularity conditions, from the weak law of large numbers,

$$\hat{\theta} \xrightarrow{p} \theta \tag{2.15}$$

From the Bernoulli example,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} p \tag{2.16}$$

2. Asymptotic normality

$$\frac{\hat{\theta} - \theta}{\mathrm{Var}(\hat{\theta})} \xrightarrow{d} N(0, 1) \tag{2.17}$$

From the Bernoulli example,

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \xrightarrow{d} N(0, 1) \tag{2.18}$$

3. Invariance MLEs are invariant under different parameterizations. That is, if $\hat{\theta}$ is an MLE for θ and g is continuous and continuously differentiable, then $g(\hat{\theta})$ is an MLE of $g(\theta)$.

2.2.4 Asymptotic Relative Efficiency (ARE)

$$\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, t^2) \tag{2.19}$$

$$\sqrt{n}(U_n - \theta) \xrightarrow{d} N(0, u^2) \tag{2.20}$$

$$\text{ARE}(T_n, U_n) = \frac{u^2}{t^2} \tag{2.21}$$

Consider $X_1, \dots X_n \sim N(\mu, \sigma^2)$. Let $T_n : \bar{X}$ be the mean and $U_n : \tilde{X}$ be the median of this sequence. Then,

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{2.22}$$

$$\sqrt{n}(\tilde{X} - \mu) \xrightarrow{d} N(0, \frac{\pi}{2}) \tag{2.23}$$

$$\text{ARE}(\bar{X}, \tilde{X}) = \frac{\frac{\pi}{2}\sigma^2}{\sigma^2} = \frac{\pi}{2} > 1, \tag{2.24}$$

Note that the median's asymptotic distribution depends on the distribution function the data are coming from. so $\bar{x}$ has less variance than $\tilde{x}$.

2.2.5 Bias, Variance, & Mean squared error (MSE) of estimators

$\text{Bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta$

$\text{Var}(\hat{\theta}) = \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2]$

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \text{E}[(\hat{\theta} - \theta)^2] \\
&= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}] + \text{E}[\hat{\theta}] - \theta)^2] \\
&= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \text{E}[\hat{\theta}])(\text{E}[\hat{\theta}] - \theta) + (\text{E}[\hat{\theta}] - \theta)^2] \\
&= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] + \text{E}[2(\hat{\theta} - \text{E}[\hat{\theta}])(\text{E}[\hat{\theta}] - \theta)] + \text{E}[(\text{E}[\hat{\theta}] - \theta)^2] \\
&= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] + \{2(\text{E}[\hat{\theta}] - \theta)\}\text{E}[\hat{\theta} - \text{E}[\hat{\theta}]] + (\text{E}[\hat{\theta}] - \theta)^2 \\
&= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] + 2(\text{E}[\hat{\theta}] - \theta)\{\text{E}[\hat{\theta}] - \text{E}[\hat{\theta}]\} + (\text{E}[\hat{\theta}] - \theta)^2 \\
&= \text{E}[(\hat{\theta} - \text{E}[\hat{\theta}])^2] + (\text{E}[\hat{\theta}] - \theta)^2 \\
&= \text{Var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}))^2
\end{aligned}$$

2.3 Fisher Information

Fisher information tells us the asymptotic variance of an estimator.

$$\sqrt{I_n(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} N(0, 1), \tag{2.25}$$

where

$$I(\theta) = \mathrm{E}\left[\left(\frac{d}{d\theta}\log f(x;\theta)\right)\right] \tag{2.26}$$

$$= -\mathrm{E}\left(\frac{d^2}{d\theta^2}\log(f(x;\theta))\right) \tag{2.27}$$

From the Bernoulli example,

$$I_n(p) = -\mathrm{E}\left(\frac{d^2}{d\theta^2}\log(\ell(x;p))\right) \tag{2.28}$$

$$= -\mathrm{E}\left(-\sum_{i=1}^{n}\left(\frac{x_i}{p^2} + \frac{1-x_i}{(1-p)^2}\right)\right) \tag{2.29}$$

$$= \frac{1}{p^2}\sum_{i=1}^{n}\mathrm{E}(x_i) + \frac{1}{(1-p)^2}\sum_{i=1}^{n}\mathrm{E}(1-x_i) \tag{2.30}$$

$$= \frac{np}{p^2} + \frac{n(p-1)}{(1-p)^2} \tag{2.31}$$

$$= \frac{n}{p(1-p)} \tag{2.32}$$

$$\therefore \sqrt{\frac{n}{p(1-p)}}(\hat{p} - p) \xrightarrow{d} N(0, 1) \tag{2.33}$$

2.3.1 Cramér-Rao Lower Bound

IF $\tilde{\theta}$ is an unbiased estimator of θ and $\hat{\theta}$ is the MLE of θ, then

$$\mathrm{Var}(\tilde{\theta}) \geq \frac{1}{I_n(\theta)} \approx \mathrm{Var}(\hat{\theta}) \tag{2.34}$$

2.3.2 Confidence Invervals

The use of Fisher's Information is useful for finding confidence intervals for $\hat{\theta}$. Suppose we want to find t such that $P(|\hat{\theta} - \theta| \leq t) \approx 1 - \alpha$.

$$P(|\hat{\theta} - \theta| \leq t) = P(\sqrt{I_n(\theta)}|\hat{\theta} - \theta| \leq t\sqrt{I_n(\theta)}) \tag{2.35}$$

$$= P(|z| \leq t\sqrt{I_n(\theta)}), \quad Z \sim N(0, 1) \tag{2.36}$$

Since z is from the standard normal distribution,

$$t\sqrt{I_n(\theta)} = z^*_{1-\frac{\alpha}{2}} \quad \text{where} \quad z^*_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{2.37}$$

$$\therefore t = \frac{z^*_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\theta)}} \tag{2.38}$$

so our α-level confidence interval for $\hat{\theta}$ is

$$\left[\hat{\theta} - \frac{z^*_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\theta)}},\ \ \hat{\theta} + \frac{z^*_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\theta)}}\right] \tag{2.39}$$

We rarely know the true value of $I_n(\theta)$ so we approximate it using $I_n(\hat{\theta})$ and so our confidence interval becomes

$$\left[\hat{\theta} - \frac{z^*_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\hat{\theta})}},\ \ \hat{\theta} + \frac{z^*_{1-\frac{\alpha}{2}}}{\sqrt{I_n(\hat{\theta})}}\right]. \tag{2.40}$$