Lecture 2

Xiaoping Shi
Department of Statistics, University of Toronto
xpshi@utstat.toronto.edu
http://www.utstat.utoronto.ca/∼xpshi/sta261-2013s.html

July 8, 2013

- Method of Moment
- Method of Maximum Likelihood
- Brief review

Note that we should choose the lowest possible order moment

Example 2: Consider $X_1, \cdots, X_n$ are iid Poisson(λ) with $E(X_i) = Var(X_i) = \lambda$.

Our aim is to show that we should choose the lowest possible order moment in the method of moment.

Two estimators of λ by the method of moment are $\hat{\lambda}_1 = \sum_{i=1}^n X_i/n$ and $\hat{\lambda}_2 = \sum_{i=1}^n (X_i - \sum_{i=1}^n X_i/n)^2/n$.

We use simulations to show that $\hat{\lambda}_1$ is better by applying the lower order moment.
For $n = 20$ and $n = 100$, we take $\lambda = 5$ and $\lambda = 10$. Simulate 1000 times and we obtain $\hat{\lambda}_{1i}, \hat{\lambda}_{2i}$ for $i = 1, \cdots, 1000$. We calculate the sample means and sample variances of them as presented in the following Table

n	λ	$\hat{E}(\hat{\lambda}_1)$	$\hat{V}ar(\hat{\lambda}_1)$	$\hat{E}(\hat{\lambda}_2)$	$\hat{V}ar(\hat{\lambda}_2)$
20	5	5.0025	0.2656	4.7813	2.5178
	10	10.0406	0.5163	9.5677	9.8858
100	5	4.9864	0.0500	4.9440	0.5278
	10	9.9916	0.0964	9.8545	2.0582

The Table indicates $\hat{\lambda}_1$ is more accurate than $\hat{\lambda}_2$, why?

$$\begin{aligned} E(\hat{\lambda}_1) &= \lambda \quad \text{unbiased} \\ E(\hat{\lambda}_2) &< \lambda \quad \text{biased} \\ Var(\hat{\lambda}_1) &< Var(\hat{\lambda}_2) \end{aligned}$$

Conclusion: suppose we wish to estimate two parameters θ_1 and θ_2:

$$\begin{aligned} \theta_1 &= f_1(\mu_1, \mu_2) \\ \theta_2 &= f_2(\mu_1, \mu_2) \end{aligned}$$

The the method of moments estimates are

$$\begin{aligned} \hat{\theta}_1 &= f_1(\hat{\mu}_1, \hat{\mu}_2) \\ \hat{\theta}_2 &= f_2(\hat{\mu}_1, \hat{\mu}_2) \end{aligned}$$

Usually we know $\mu_i = g_i(\theta_1, \theta_2)$ for $i = 1, 2$, then we solve the invert functions g_i and get θ_i. Use low order moments and replace moments by sample moments.

Example 3: Consider $X_1, \cdots, X_n$ is iid Normal distribution $N(\mu, \sigma^2)$ with parameters $\theta_1 = \mu$ and $\theta_2 = \sigma^2$.

$$\begin{aligned} \mu_1 = E(X) &= \mu \\ \mu_2 = E(X^2) &= \sigma^2 + \mu^2 \end{aligned}$$

$$\begin{aligned} \mu &= \mu_1 \\ \sigma^2 &= \mu_2 - \mu_1^2 \end{aligned}$$

So $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = \sum_{i=1}^n X_i^2/n - \bar{X}^2 = \sum(X_i - \bar{X})^2/n$ with $E(\hat{\mu}) = \mu$, $Var(\hat{\mu}) = \sigma^2/n$, $E(\hat{\sigma}^2) = (n-1)\sigma^2/n$ and $Var(\hat{\sigma}^2) = 2(n-1)\sigma^4/(n^2)$

Example 4: Consider $X_1, \cdots, X_n$ are iid Gamma distribution $G(\alpha, \lambda)$ with parameters $\theta_1 = \alpha$ and $\theta_2 = \lambda$.

$$\begin{aligned} \mu_1 = E(X) &= \alpha/\lambda \\ \mu_2 = E(X^2) &= \mu_1/\lambda + \mu_1^2 \end{aligned}$$

$$\begin{aligned} \lambda &= \mu_1/(\mu_2 - \mu_1^2) \\ \alpha &= \mu_1^2/(\mu_2 - \mu_1^2) \end{aligned}$$

So $\hat{\alpha} = \bar{X}^2/(\sum(X_i - \bar{X})^2/n)$ and $\hat{\lambda} = \bar{X}/(\sum(X_i - \bar{X})^2/n)$. To find the sample distribution of $\hat{\alpha}$ and $\hat{\lambda}$ such as their expectations and variances is difficult.

- **Approximating by simulations**
- **Approximating by central limit theorem (CLT)**

Example 5: Consider $X_1, \cdots, X_n$ are iid Angular distribution $A(\alpha)$ with parameters $\theta_1 = \alpha$ and pdf

$$f(x|\alpha) = \frac{1+\alpha x}{2}, \quad -1 \le x \le 1, \quad -1 \le \alpha \le 1.$$

$$\mu_1 = E(X) \quad = \quad \alpha/3$$

So $\hat{\alpha} = 3\bar{X}$.
The sampling distribution of $\hat{\alpha}$ can be obtained by **central limit theorem**.
Note: $\hat{\alpha}$ is a consistent estimate of α in probability, that is, for any $\varepsilon > 0$,

$$P(|\hat{\alpha} - \alpha| > \varepsilon) \to 0, \quad \text{as} \quad n \to \infty \tag{1.1}$$

by **Chebyschev's Lemma**.

The sample distribution of moment estimates is obtained by CLT or simulation. However, we will later see that Maximum Likelihood estimates (MLE) have nice theoretical properties.

Suppose that r.v.'s $X_1, \cdots, X_n$ have a joint density function

$$f(x_1, x_2, \cdots, x_n|\theta.$$

Then given observed data $X_i = x_i$, the likelihood function is

$$L(\theta) = f(X_1, X_2, \cdots, X_n),$$

where $L(\theta)$ is considered as a function of θ.

The MLE of θ is that value of θ that maximizes the likelihood

$$\hat{\theta} = \arg\max_{\theta} L(\theta),$$

-that is, makes the observed data "most likely".

If X_i is iid,

$$L(\theta = \prod_{i=1}^{n} f(X_i|\theta).$$

Log likelihood function

$$\ell(\theta) = \log\{L(\theta)\} = \sum_{i=1}^{n} \log\{f(X_i|\theta)\},$$

and $\hat{\theta} = \arg\max \ell(\theta)$.

For understanding "most likely", we take an example.

Example 6: Consider $X_1, \cdots, X_n$ are iid Exponential distribution $Exp(\lambda)$ with parameter λ and pdf

$$f(x|\lambda) = \lambda e^{-\lambda x} \quad \text{if} \quad x \geq 0, \quad \text{otherwise} \quad 0.$$

$$\ell(\lambda) = n \log \lambda - \lambda n \bar{X}$$

$$\frac{d\ell(\lambda)}{d\lambda} = n/\lambda - n\bar{X} \begin{cases} > 0 & \text{if } \lambda < 1/\bar{X} \\ = 0 & \text{if } \lambda = 1/\bar{X} \\ < 0 & \text{if } \lambda > 1/\bar{X}. \end{cases}$$

MLE estimate: $\hat{\lambda} = 1/\bar{X}$.

Example 7: Consider $X_1, \cdots, X_n$ are iid Poisson distribution $Pois(\lambda)$ with parameter λ and pdf

$$P(X = x|\lambda) = \lambda^x e^{-\lambda}/x!.$$

$$\ell(\lambda) = \sum_{i=1}^{n} X_i \log \lambda - \lambda n - \sum_{i=1}^{n} \log(X_i!)$$

$$\frac{d\ell(\lambda)}{d\lambda} = \sum_{i=1}^{n} X_i/\lambda - n = 0$$

MLE estimate: $\hat{\lambda} = \bar{X}$.

Example 8: Consider $X_1, \cdots, X_n$ are iid Normal distribution $N(\mu, \sigma^2)$.

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

$$\begin{aligned} \frac{\partial \ell}{\partial \mu} &= \frac{2}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0 \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \sum_{i=1}^{n}(X_i - \mu)^2/(2\sigma^4) = 0 \end{aligned}$$

MLE estimate: $\hat{\mu} = \bar{X}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Question: What's the distribution of $\hat{\sigma}^2$.

Example 9: Consider $X_1, \cdots, X_n$ are iid Gamma distribution $G(\alpha, \lambda)$ with pdf

$$f(x|\alpha, \lambda) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x}, \quad 0 \leq x < \infty.$$

$$\ell(\alpha, \lambda) = n\alpha \log \lambda + (\alpha - 1) \sum \log X_i - \lambda \sum X_i - n \log\{\Gamma(\alpha)\}$$

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha} &= n \log \lambda + \sum \log(X_i) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0 \\ \frac{\partial \ell}{\partial \lambda} &= \frac{n\alpha}{\lambda} - \sum X_i = 0 \end{aligned}$$

MLE estimate: $\hat{\lambda} = \hat{\alpha}/\bar{X}$ and

$$n \log \hat{\alpha} - n \log \bar{X} + \sum \log X_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0.$$

$\hat{\alpha}$ can note be solved in closed form, and then obtaining their exact sampling distribution would appear to be intractable.

- Give initial estimates α_0 and λ_0 by method of moment.
- Use an iterative method to solve $\hat{\alpha}$ and $\hat{\lambda}$ by method of maximum likelihood.
- Use Bootstrap/Simulation method to obtain an approximation of sample distribution of $\hat{\alpha}$ and $\hat{\lambda}$. See Example 2.

Example 10: Consider $X_1, \cdots, X_n$ are iid Laplace distribution $Lap(\theta)$ with pdf

$$f(x|\theta) = \frac{1}{2} e^{-|x-\theta|}, \quad -\infty < x < \infty.$$

$$\ell(\theta) = -n \log 2 - \sum_{i=1}^{n} |X_i - \theta|$$

$$\ell'(\theta) = \sum_{i=1}^{n} \text{sgn}(X_i - \theta)$$

where $\text{sgn}(t) = 1, 0, -1$ depending on whether $t > 0$, $t = 0$, or $t < 0$.
MLE estimate: $\hat{\theta} = \text{median}(X_1, X_2, \cdots, X_n)$ because the median will make half the terms of the sum in $\ell'(\theta)$ nonpositive and half nonnegative.

Example 11: Consider $X_1, \cdots, X_n$ are iid Uniform distribution $U(0, \theta)$ with pdf

$$f(x|\theta) = \frac{1}{\theta}, \quad \text{if} \quad 0 \leq x \leq \theta, \quad \text{otherwise} \quad 0.$$

$$L(\theta) = \theta^{-n} \mathbf{1}(\max\{X_1, X_2, \cdots, X_n\}, \theta)$$

where $\mathbf{1}(a, b)$ is 1 or 0 if $a \leq b$ or $a > b$, respectively. The function is a decreasing function of θ for all $\theta \geq \max\{X_1, X_2, \cdots, X_n\}$ and is 0, otherwise. Hence the maximum occurs at the smallest value of θ; i.e.
MLE estimate: $\hat{\theta} = \max\{X_1, X_2, \cdots, X_n\}$.
Question: What's the pdf of $\max\{X_1, X_2, \cdots, X_n\}$ or $\min\{X_1, X_2, \cdots, X_n\}$.

$X_1, \cdots, X_m$ the counts in cells $1, \cdots, m$, follow a multinomial distribution with a total count of n, cell prob. $p_1, \cdots, p_m$ (we wish to estimate) and pdf

$$f(x_1, \cdots, x_m | p_1, \cdots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i},$$

where $\sum_{i=1}^m X_i = n, \sum_{i=1}^m p_i = 1$.
Multinomial distribution is a generalization of the Binomial distribution $B(p_1)$ $(m = 2)$ with

$$f(x_1, x_2 | p_1, p_2) = \frac{n!}{x_1!(n - x_1)!} p_1^{x_i} (1 - p_1)^{n - x_1}.$$

$$\ell(p_1, \cdots, p_m) = \log(n!) - \sum_{i=1}^m \log(X_i!) + \sum_{i=1}^m X_i \log(p_i).$$

To solve $p_1, \cdots, p_m$, we introduce a Lagrange multiplier. Since there is a constraint on $p_1, \cdots, p_m$: $\sum p_i = 1$,

$$La(p_1, \cdots, p_m, \lambda) = \ell(p_1, \cdots, p_m) + \lambda(\sum p_i - 1).$$

$$\frac{\partial La}{\partial p_j} = X_j/p_j + \lambda = 0$$

We have $p_j = -X_j/\lambda$ for $j = 1, \cdots, m$ and sum both sides of this equation, then $1 = \sum p_j = \sum X_j/\lambda$, i.e. $\lambda = -n$.
MLE estimate: $\hat{p}_j = X_j/n$.

Example 12:

	Blood Type			
	M	MN	N	Total
Frequency	342	500	187	1029
	X_1	X_2	X_3	n
Suppose	$p_1 = (1-\theta)^2$	$p_2 = 2\theta(1-\theta)$	$p_3 = \theta^2$	1

What's the MLE of θ?

$$\ell(\theta) = \ell(p_1, p_2, p_3) = \log n! - \sum \log X_i! + X_1 \log(1-\theta)^2 + X_2 \log\{2\theta(1-\theta)\} + X_3 \log \theta^2.$$

$$\ell'(\theta) = \frac{-2X_1}{1-\theta} + \frac{2X_2}{\theta} + \frac{-X_2}{1-\theta} + \frac{2X_3}{\theta} = 0$$

MLE estimate: $\hat{\theta} = \frac{X_2+2X_3}{2X_1+2X_2+2X_3} = \frac{X_2+2X_3}{2n} = \frac{500+2\times187}{2\times1029} = 0.4247.$

$X_1, \cdots, X_n$ are iid with pdf $f(X_i|\theta_0)$ and $\ell(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$.

MLE estimate: $\hat{\theta} = \arg\min_\theta \ell(\theta)$.

- $\hat{\theta}$ converges to θ_0 in probability.
- $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ converges to standard normal distribution $N(0,1)$ in distribution, where

$$\begin{aligned} I(\theta) &= E\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right] \\ &= Var\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] \end{aligned}$$

  is called as **Fisher information**.
- since θ_0 is unknown, we will use $I(\hat{\theta})$ in place of $I(\theta_0)$.

Proof of Part I: Show that $\ell(\theta_0) > \ell(\theta)$ in probability for all $\theta \neq \theta_0$. Since $\ell(\hat{\theta})$ reaches the maximum of $\ell(\theta)$, $\hat{\theta} \to \theta_0$ in probability.

Proof of Part II: From a Taylor series expansion,
$0 = \ell'(\theta) \approx \ell'(\theta_0) + (\hat{\theta} - \theta_0)\ell''(\theta_0)$ by Part I.

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \approx \frac{\sqrt{I(\theta_0)/n}}{-\ell''(\theta_0)/n} \sum_{i=1}^{n} \frac{f'(X_i|\theta_0)}{f(X_i|\theta_0)}$$

where $-\ell''(\theta_0)/n \to I(\theta_0)$ by **the law of large number**. The remaining part is proved by **CLT**.

We will discuss three methods for forming CI for MLE:

- Exact method for sample distribution of MLE.
- Large Sample Theory for MLE.
- Bootstrap method. See Example 2 for simulations. After simulations, the empirical distribution can be used to approximate the sample distribution. Hence, CI will be constructed based on empirical quantiles.

Example 3 continued: We wish to find $1-\alpha$ CI for parameters μ and σ^2.

Exact Method: Let $S^2 = \sum(X_i - \bar{X})^2/(n-1)$.

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t_{n-1}.$$

$$P\left(-t_{n-1}(\alpha/2) \leq \frac{\sqrt{n}(\bar{X}-\mu)}{S} \leq t_{n-1}(\alpha/2)\right) = 1-\alpha,$$

$1-\alpha$ CI for μ: $\bar{X} \pm t_{n-1}(\alpha/2)S/\sqrt{n}$.

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}.$$

$$P\left(\chi^2_{n-1}(1-\alpha/2) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq \chi^2_{n-1}(\alpha/2)\right) = 1-\alpha,$$

$1-\alpha$ CI for σ^2: $\left(\frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1-\alpha/2)}\right)$.

Large Sample Theory for MLE:

$$\sqrt{nI(\bar{X}, \hat{\sigma}^2)}(\bar{X} - \mu) \to N(0, 1).$$

$$P\left(-z(\alpha/2) \leq \sqrt{nI(\bar{X}, \hat{\sigma}^2)}(\bar{X} - \mu) \leq z(\alpha/2)\right) = 1 - \alpha,$$

$1 - \alpha$ CI for μ: $\bar{X} \pm z(\alpha/2)/\sqrt{nI(\bar{X}, \hat{\sigma}^2)}$. where

$$\begin{aligned} I(\mu, \sigma^2) &= -E\left[\frac{\partial^2}{\partial \mu^2} \log f(X|\theta)\right] \\ &= 1/\sigma^2. \end{aligned} \tag{2.2}$$

Compare it with CI by Exact Method: $S \sim \hat{\sigma}$ and $t_{n-1}(\alpha/2) \sim z(\alpha/2)$ when n is large; for example $z(\alpha/2) = 1.96$ with $\alpha = 0.05$ and $t_{n-1}(\alpha/2)$ is 2.093 or 1.980 if $n = 20$ or $n = 121$, respectively.

Question: What's $1 - \alpha$ CI for σ^2?

Let's take another two examples to show how to calculate Fisher information.

Example 7 continued: $Pois(\lambda)$

$$\ell(\lambda) = X \log \lambda - \lambda - \log(X!)$$

$$\begin{aligned} \frac{\partial^2}{\partial \lambda^2} \ell(\lambda) &= -X/\lambda^2 \\ I(\lambda) &= -E(-X/\lambda^2) = 1/\lambda. \end{aligned} \tag{2.3}$$

Since $\hat{\lambda} = \bar{X}$, $I(\hat{\lambda}) = 1/\bar{X}$.

Example 13: Fisher information for a Bernoulli r.v. Let X be Bernoulli $B(1, \theta)$. Thus

$$\begin{aligned} \ell(\theta) &= X \log \theta + (1 - X) \log(1 - \theta) \\ \frac{\partial^2}{\partial \theta^2} \ell(\theta) &= \frac{-X}{\theta^2} + \frac{X - 1}{(1 - \theta)^2} \\ I(\theta) &= -\frac{-\theta}{\theta^2} - \frac{\theta - 1}{(1 - \theta)^2} = \frac{1}{\theta(1 - \theta)}. \end{aligned}$$

MLE estimate: $\hat{\theta} = 0.4247$.

STA261H1S: First quiz

The first quiz will be held in tutorial sessions on this Wednesday, 6:30-7:00pm. Make sure you show up to write the quiz, which weights 10% of this course.
(Time: **30 minutes**, *Wednesday, July 10, 2013, 6:30 - 7:00pm*)

Three questions will be contained.

The last names A-Ke go to room SS2135 TA: Tadeu. The last names Ki-Por go to room RW229 TA: Qin. The last names Qiu-Zou go to room MP137 TA: Becky. You must write your quiz in your tutorial session.

Since I have not received the confirmation of the TA Qin, if there is no TA in room RW229, please go to room SS2135.