

# STATS300B – Lecture 6

Julia Palacios  
Scribed by Michael Howes

01/20/22

## Contents

|          |   |          |
|----------|---|----------|
| <b>1</b> | <b>The delta method</b>                                 | <b>1</b> |
| <b>2</b> | <b>Maximum likelihood estimation</b>                    | <b>2</b> |
| 2.1      | MLEs for one-dimensional exponential families . . . . . | 2        |
| 2.2      | Asymptotic efficiency . . . . .                         | 3        |
| 2.3      | The ARE of the median . . . . .                         | 3        |
| <b>3</b> | <b>M-estimators and Z-estimators</b>                    | <b>4</b> |
| 3.1      | A weak law for random functions . . . . .               | 4        |

## 1 The delta method

We ended last class with the statement of the higher order delta method. We stated the following.

**Theorem 1** (Delta method 3 (higher order)). *Suppose that  $X_n$  are random  $k$ -vectors such that*

$$r_n(X_n - \theta) \xrightarrow{d} X,$$

*where  $r_n$  is a deterministic function with  $r_n \rightarrow +\infty$ . Let  $\phi$  be a real-valued function that is twice differentiable at  $\theta$  with  $\phi'(\theta) = 0$ . Then,*

$$r_n^2(\phi(X_n) - \phi(\theta)) \xrightarrow{d} \frac{1}{2}X^T \nabla^2 \phi(\theta) X.$$

We will now prove the above.

*Proof.* Note that,

$$\begin{aligned}\phi(X_n) &= \phi(\theta) + \nabla \phi(\theta)^T (X_n - \theta) + (X_n - \theta)^T \nabla^2 \phi(\theta) (X_n - \theta) + o(\|X_n - \theta\|_2^2) \\ &= \phi(\theta) + \frac{1}{2}(X_n - \theta)^T \nabla^2 \phi(\theta) (X_n - \theta) + o(\|X_n - \theta\|_2^2).\end{aligned}$$

□

Note that  $o(\|X_n - \theta\|_2^2) = o_p(r_n^{-2})$ . Thus, by Slutsky's

$$\begin{aligned}r_n^2(\phi(X_n) - \phi(\theta)) &= \frac{1}{2}(r_n(X_n - \theta))^T \nabla^2 \phi(\theta) (r_n(X_n - \theta)) + r_n^2 o_p(r_n^{-2}) \\ &\xrightarrow{d} \frac{1}{2}X^T \nabla^2 \phi(\theta) X.\end{aligned}$$

The central limit theorem gives a special one-dimensional case of the higher-order delta method. If  $X_1, \dots$  are i.i.d. with mean  $\mu$  and variance  $\sigma^2 < \infty$ , then  $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathbf{N}(0, \sigma^2)$ . Thus, if  $g$  is a twice differentiable at  $\mu$ , then

$$n(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} g''(\mu)\sigma^2 Z^2,$$

where  $Z \sim \mathbf{N}(0, 1)$ . So that,  $\frac{n}{g''(\mu)\sigma^2}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} \chi_1^2$ .

## 2 Maximum likelihood estimation

Suppose we have data  $X$  in the form of  $n$  i.i.d. observations  $X_1, \dots, X_n$  drawn from a distribution  $\mathbb{P}_\theta$  with density  $p_\theta$ . The likelihood  $L(\theta) = p_\theta(X)$  is the density evaluated at  $X$  viewed as a function of  $\theta$ . The value  $\hat{\theta}$  that maximizes  $L(\theta)$  is called the maximum likelihood estimator (MLE) of  $\theta$ . Much of this course will focus on properties of MLEs.

### 2.1 MLEs for one-dimensional exponential families

Suppose that  $X$  is generated from a one dimensional exponential family in canonical form. That is,

$$p_\eta(x) = \exp\{\eta T(x) - A(\eta)\}h(x),$$

where  $\eta$  is the natural parameter for the family,  $T(X)$  are the sufficient statistics for the family and  $A(\eta) = \log(\int \exp\{\eta T(x)\}h(x)dx)$  is the log-partition function for the family. The log-partition function is convex and differentiable with,

$$A'(\eta) = \mathbb{E}_\eta[T(X)],$$

and

$$A''(\eta) = \text{Var}_\eta(T(X)).$$

Another name of the  $A(\eta)$  is the *cummulant generating function* of  $T(X)$ . Suppose now that we have an i.i.d. sample of size  $n$  drawn from  $p_\eta$ . Let  $l(\eta) = \log(p_\eta(X_1, \dots, X_n))$  be the log-likelihood function. Then,

$$\begin{aligned} l(\eta) &= \eta \sum_{i=1}^n T(X_i) - nA(\eta) + \log\left(\prod_{i=1}^n h(X_i)\right) \\ \therefore l'(\eta) &= \sum_{i=1}^n T(X_i) - nA'(\eta) = \sum_{i=1}^n T(X_i) - n\mathbb{E}_\eta[T(X)] \\ \therefore l''(\eta) &= -nA''(\eta) = -n\text{Var}_\eta(T(X)) \leq 0. \end{aligned}$$

Thus, any solution to  $l'(\eta) = 0$  is a local maximum of  $l$ . It follows that the MLE  $\hat{\eta}$  solves the equation,

$$\mathbb{E}_{\hat{\eta}}[T(X)] = \frac{1}{n} \sum_{i=1}^n T(X_i) = \bar{T}_n$$

Thus, the MLE is a type of method of moments estimator in this example. In particular, we have  $\hat{\eta} = \phi(\bar{T})$  where  $\phi$  can be thought of as  $(A')^{-1}$ . By the central limit theorem,

$$\sqrt{n}(\bar{T} - \mathbb{E}_\eta[T(X)]) \xrightarrow{d} \mathbf{N}(0, \text{Var}_\eta(T(X))) = \mathbf{N}(0, A''(\eta)).$$

Thus, by the delta method,

$$\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} \mathbf{N}(0, \phi'(\eta)^2 A''(\eta)).$$

We know that  $\phi'(\theta) = \frac{1}{A''(\eta)}$  and so,  $\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{d} \mathbf{N}(0, A''(\eta)^{-1})$ . We will see that similar results hold for distributions that are not exponential families.

## 2.2 Asymptotic efficiency

Recall that the Fisher information of a parameter  $\theta$  is defined to be,

$$I(\eta) = \mathbb{E}_\eta[l'(\eta)^2] = -\mathbb{E}_\eta[l''(\eta)].$$

We have seen that for exponential families,  $I(\eta) = -\mathbb{E}[-A''(\eta)] = A''(\eta)$ . We also have the Cramer–Rao lower bound that states if  $\hat{\eta}$  is an unbiased estimator for  $\eta$  based on an i.i.d. sample of size  $n$ , then

$$\text{Var}_\eta(\hat{\eta}) \geq \frac{1}{nI(\eta)}.$$

For exponentially families,  $\text{Var}(\hat{\eta}_{MLE}) \sim \frac{1}{nA''(\eta)} = \frac{1}{nI(\eta)}$ . Thus, the MLE estimator asymptotically reaches the Cramer–Rao lower bound. Estimators with this property are called *asymptotically efficient*.

**Definition 1.** An estimator  $\hat{\eta}$  is *asymptotically efficient* if  $\text{Var}(\hat{\eta}) \sim \frac{1}{nI(\eta)}$ .

We can also compare the asymptotic efficiency of two estimators.

**Definition 2.** Let  $\hat{\eta}_1$  and  $\hat{\eta}_2$  be two estimators. The asymptotic relative efficiency (ARE) of  $\hat{\eta}_2$  compared to  $\hat{\eta}_1$  is,

$$\lim_{n \rightarrow \infty} \frac{\text{Var}_\eta(\hat{\eta}_2)}{\text{Var}_\eta(\hat{\eta}_1)}.$$

## 2.3 The ARE of the median

Let  $X_1, X_2, \dots$  be i.i.d. with common CDF  $F$ . Let  $\gamma \in (0, 1)$  and let  $\tilde{\theta}_n$  be the  $[\gamma n]^{th}$  order statistic for  $X_1, \dots, X_n$ , where  $[y]$  denotes the ceiling of  $y$ . If  $F(\theta) = \gamma$  and if  $F'(\theta)$  exists and is strictly positive, then

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{\gamma(1-\gamma)}{[F'(\theta)]^2}\right).$$

The idea is that the event  $\sqrt{n}(\tilde{\theta}_n - \theta) \leq a$  is equivalent to  $\tilde{\theta}_n \leq \theta + \frac{a}{\sqrt{n}}$ . Which is in turn to at least  $[\gamma n]$  of  $X_i$ 's being less than  $b = \theta + \frac{a}{\sqrt{n}}$ . Thus,

$$\{\tilde{\theta}_n \leq a\} = \left\{ \sum_{i=1}^n \mathbf{1}_{\{X_i \leq b\}} \geq [\gamma n] \right\}.$$

Furthermore, the random variables  $\mathbf{1}_{\{X_i \leq b\}}$  are i.i.d. and with mean  $F(\theta + a/\sqrt{n})$  and variance  $F(\theta + a/\sqrt{n})(1 - F(\theta + a/\sqrt{n}))$ . Thus, we can apply the central limit theorem to and use the fact the CDF is differentiable at  $\theta$ .

A special case is when  $\gamma = 1/2$  and  $\tilde{\theta}_n$  is the median of the sample  $X_1, \dots, X_n$ . For a symmetric distribution we can look at the ARE of  $\tilde{\theta}_n$  compared to  $\bar{X}_n$ . If  $X_i \sim N(\theta, \sigma^2)$ , then  $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$  and  $\text{Var}(\tilde{\theta}_n) \sim \frac{1}{n} \cdot \frac{1}{4F'(\theta)}$ . Note that,

$$F'(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\theta - \theta)^2\right\} = \frac{1}{\sqrt{2\pi}\sigma^2}.$$

Thus, the ARE of the sample median compared to the sample mean is

$$\frac{2\pi\sigma^2}{4\sigma^2} = \frac{\pi}{2} \approx 1.577 > 1.$$

Thus, for Gaussian data, the sample mean has lower asymptotic variance than the sample median.

### 3 M-estimators and Z-estimators

The MLE is defined to be,

$$\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} L(\theta|X).$$

We would like to know the following about  $\hat{\theta}_n$ ,

1. Consistency,
2. Asymptotic distribution,
3. Optimality.

Since the MLE is defined as a maximizer it is an M-estimator. In many cases the MLE is also defined by solving the equation  $l'(\hat{\theta}) = 0$ , this makes  $\hat{\theta}$  a Z-estimator as well. To prove consistency for M-estimators and Z-estimators, it helps to think of the log likelihood as a random function. We would like to say that the sample average of the log likelihood converges to the population log likelihood for all  $\theta$ . Thus, we need to understand the convergence of random functions.

#### 3.1 A weak law for random functions

**Definition 3.** Let  $K$  be a compact set and let  $\mu : K \rightarrow \mathbb{R}$  be a continuous function. Define  $\|\mu\|_\infty = \sup_{t \in K} |\mu(t)|$ .

**Lemma 1.** Let  $X$  be a random variable taking values in a compact set  $K$ . Let  $h(t, x)$  be a function such that  $h(\cdot, x)$  is a continuous function from  $K$  to  $\mathbb{R}$ . Define  $W(t) = h(t, X)$  for  $t \in K$ . Thus,  $W$  is a random continuous function on  $K$ . Suppose that  $\mathbb{E}\|W\|_\infty < \infty$ , then

1. The function  $\mu(t) = \mathbb{E}[W(t)]$  is continuous on  $K$ .
2. As  $\varepsilon \searrow 0$ ,

$$\sup_{t \in K} \mathbb{E} \left[ \sup_{s: \|s-t\| \leq \varepsilon} |W(s) - W(t)| \right] \rightarrow 0.$$

*Proof.* Suppose that  $t_n \rightarrow t$ , then  $W(t_n) \rightarrow W(t)$ . Furthermore,  $|W(t_n)| \leq \|W\|_\infty$  and  $\|W\|_\infty$  is integrable. Thus, by the dominated convergence theorem,

$$\mu(t_n) = \mathbb{E}[W(t_n)] \rightarrow \mathbb{E}[W(t)] = \mu(t).$$

Thus,  $\mu$  is continuous. For each  $x$  in our sample space, define

$$M_\varepsilon(t) = \sup_{\|s-t\| \leq \varepsilon} |W(t) - W(s)|.$$

For fixed  $t$ ,  $M_\varepsilon(t) \rightarrow 0$  as  $\varepsilon \rightarrow 0$  and  $|M_\varepsilon(t)| \leq 2\|W\|_\infty$ . Thus, by the dominated convergence theorem for each  $t \in K$  we have,

$$\lim_{\varepsilon \rightarrow 0} \mathbb{E} \left[ \sup_{\|s-t\| \leq \varepsilon} |W(t) - W(s)| \right] = 0.$$

The uniform result of the lemma is derived by the above point-wise result combined with the compactness of  $K$  □

We can use the above lemma to prove our weak law for random functions. We know from the regular weak law that  $|\bar{W}_n(t) - \mu(t)| \xrightarrow{P} 0$ , but we want a result that is uniform in  $t$ .

**Theorem 2.** Let  $X_1, X_2, \dots$  be i.i.d. random variables and let  $W_i(t) = h(t, X_i)$  where  $h$  is a function such that  $h(\cdot, x) : K \rightarrow \mathbb{R}$  is a continuous function on a compact set  $K$  for all  $x$ . Suppose that  $\mathbb{E}[\|W_1\|_\infty] < \infty$ . Let  $\mu(t) = \mathbb{E}[W_1(t)]$ , then

$$\|W_n - \mu\|_\infty \xrightarrow{p} 0,$$

as  $n \rightarrow \infty$ .