Outline

1) Log-likelihood and score
2) Fisher information
3) Cramér–Rao Lower Bound
Log-likelihood score

Assume $\mathcal{Y}$ has densities $p_\theta$ wrt $\mu$, $\theta \in \mathbb{R}^d$

Common support: $\{x : p_\theta(x) > 0\}$ same $\forall \theta$

Recall $l(\theta; x) = \log p_\theta(x)$,

Thought of as random function of $\theta$

**Def** The score is $\nabla l(\theta; x)$; plays a key role in many areas of statistics, esp. asymptotics.

Can think of as "local sufficient statistic":

$$p_{\theta + \varepsilon}(x) = e^{l(\theta + \varepsilon; x)}$$

$$\approx e^{\varepsilon \cdot \nabla l(\theta; x)} p_\theta(x) \quad \text{for } \varepsilon \approx 0$$

**Differential identities:** (assuming enough regularity)

$$1 = \int x e^{l(\theta; x)} \, d\mu(x)$$

$$\frac{\partial}{\partial \theta_j} \Rightarrow 0 = \int \frac{\partial}{\partial \theta_j} l(\theta; x) e^{l(\theta; x)} \, d\mu(x)$$

$$\Rightarrow \quad E_\theta \left[ \nabla l(\theta; x) \right] = 0$$

\[\text{only true if these are the same value of } \theta!\]
\[
\frac{\partial}{\partial \theta_k} = 0 = \int \left( \frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} + \frac{\partial \ell}{\partial \theta_j} \cdot \frac{\partial \ell}{\partial \theta_k} \right) e^\ell \, dm
\]

\[
= \mathbb{E}_{\theta} \left[ \frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} \right] + \mathbb{E}_{\theta} \left[ \frac{\partial \ell}{\partial \theta_j} \frac{\partial \ell}{\partial \theta_k} \right]
\]

\[
\Rightarrow \quad \text{Var}_{\theta} \left[ \nabla \ell (\theta; x) \right] = \mathbb{E}_{\theta} \left[ - \nabla^2 \ell (\theta; x) \right]
\]

\[
J(\theta) = \text{same } \theta \quad \text{same } \theta
\]

Called “Fisher Information”

It is possible to extend this definition to certain cases where \( \ell \) is not even differentiable, e.g. Laplace location family, but for our purposes we can just assume “sufficient regularity.”

Try with another statistic \( S(X) \), let

\[
g(\theta) = \mathbb{E}_{\theta} \left[ S(X) \right] \quad \text{ (“unbiased estimator”)}
\]

\[
g(\theta) = \int S \, e^\ell \, dm
\]

\[
\Rightarrow \quad \nabla g(\theta) = \int \nabla \ell \, e^\ell \, dm = \mathbb{E}_{\theta} \left[ S(X) \nabla \ell (\theta; x) \right]
\]

\[
= \text{Cov}_{\theta} (S(X), \nabla \ell (\theta; x)) \quad \text{why?} \\
\text{Since } \mathbb{E} \nabla \ell = 0
\]
Combining these results with Cauchy–Schwarz gives us the Cramér–Rao Lower Bound or Information Lower Bound:

\[ \text{1-param: } \text{Var}_\theta(\delta) \cdot \text{Var}_\theta(\ell(\theta; x)) \geq \text{Cov}_\theta(\delta, \ell(\theta; x))^2 \]

\[ \Rightarrow \text{Var}_\theta(\delta) = \frac{g(\theta)^2}{J(\theta)} \]

\( \theta \in \mathbb{R}^d, g(\theta) \in \mathbb{R} : \text{Var}_\theta(\delta) \geq g(\theta)'J(\theta)^{-1}g(\theta) \)

Interp: If \( g(\theta) \) is estimand, no unbiased estimator can have smaller variance than \( g(\theta)'J(\theta)^{-1}g(\theta) \)

Ex. (i.i.d. sample)

\[ X_1, \ldots, X_n \overset{iid}{\sim} \rho^{(1)}(x) \quad \theta \in \Theta \]

\[ X \sim \rho^\theta(x) = \prod_i \rho^{(i)}(x_i) \]

Let \( l_i(\theta; x_i) = \log \rho^{(i)}(x_i) \)

\[ \ell(\theta; x) = \sum_i l_i(\theta; x_i) \]

\[ J(\theta) = \text{Var}_\theta(\nabla \ell(\theta; x)) \]

\[ = \text{Var}_\theta(\sum_i \nabla l_i(\theta; x_i)) \]

\[ = nJ_1(\theta) \quad \text{where } J_1(\theta) \text{ is Fisher info in single observation} \]

\( \Rightarrow \) Lower bound scales like \( n^{-1} \) (SD = \( n^{-1/2} \) for “regular” families).
Efficiency

CRLB is not nec. attainable.

We define the efficiency of an unbiased estimator as:
\[
\text{eff}_\theta(\delta) = \frac{\text{CRLB}_\theta}{\text{Var}_\theta(\delta)} \quad (= \frac{1/\ell(\theta)}{\text{Var}_\theta(\delta)} \quad \text{if } g(\theta) = \theta \in \mathbb{R})
\]
\[
\text{eff}_\theta(\delta) \leq 1
\]

We say \( \delta(X) \) is efficient if \( \text{eff}_\theta(\delta) = 1 \quad \forall \theta \)

Depends on \( \text{Corr}_\theta(\delta(X), \ell(\theta; x)) \):
\[
\text{eff}_\theta(\delta) = \frac{\text{Cov}_\theta(\delta(X), \ell(\theta; x))}{\text{Var}_\theta(\delta) \cdot \text{Var}_\theta(\ell(\theta))}
\]
\[
= \text{Corr}_\theta(\delta, \ell(\theta))
\]
\[
\leq 1
\]

\( \delta(X) \) is efficient \( \iff \text{Corr}_\theta^2(\delta, \ell(\theta)) = 1 \quad \forall \theta \)

Rarely achieved in finite samples but we can approach it asymptotically as \( n \to \infty \)
CRLB requires differentiation under integral

Can make more general statement if we replace $\nabla l(\theta; x)$ with finite-difference:

$$\frac{\rho_{\theta+\varepsilon}(x)}{\rho_{\theta}(x)} - 1 = \epsilon \quad \left( 2 \varepsilon' \nabla l(\theta; x) \text{ small } \varepsilon \right)$$

$$\mathbb{E}_{\theta} \left[ \frac{\rho_{\theta+\varepsilon} - 1}{\rho_{\theta}} \right] = \int \delta (\frac{\rho_{\theta+\varepsilon} - 1}{\rho_{\theta}}) \rho_{\theta} \, dm = 1 - 1 = 0$$

(assuming common support, or $\rho_{\theta+\varepsilon} \ll \rho_{\theta}$)

$$\text{Cov}_{\theta} (\delta, \frac{\rho_{\theta+\varepsilon} - 1}{\rho_{\theta}}) = \int \delta (\frac{\rho_{\theta+\varepsilon} - 1}{\rho_{\theta}}) \rho_{\theta} \, dm$$

$$= \mathbb{E}_{\theta+\varepsilon} (\delta) - \mathbb{E}_{\theta} (\delta)$$

$$= g(\theta+\varepsilon) - g(\theta)$$

$$\Rightarrow \text{Var}_{\theta} (\delta) \geq \frac{(g(\theta+\varepsilon) - g(\theta))^2}{\mathbb{E}_{\theta} \left[ \left( \frac{\rho_{\theta+\varepsilon} - 1}{\rho_{\theta}} \right)^2 \right]}$$

CRLB follows from $\varepsilon \to 0$, but sup gives better bound
Ex. Exponential Families

\[ \rho_\eta(x) = e^{\eta^T T(x) - A(\eta)} h(x) \]

\[ l(\eta; x) = \eta^T T(x) - A(\eta) + \log h(x) \]

\[ \nabla l(\eta; x) = T(x) - \nabla A(\eta) \]

\[ = T(x) - \mathbb{E}_\eta T(x) \]

\[ \text{Var}_\eta(\nabla l(\eta)) = \text{Var}_\eta(T(x)) = \nabla^2 A(\eta) \]

\[ \nabla^2 l(\eta; x) = -\nabla^2 A(\eta) \]

\[ \mathbb{E}_\eta \left[ -\nabla^2 l(\eta; x) \right] = \nabla^2 A(\eta) \quad \checkmark \]

So any unbiased est. of \( \eta \) has

\[ \text{Var}_\eta(\delta) \succeq \nabla^2 A(\eta)^{-1} \]
Curved family: 

\[ p_\theta(x) = e^{\gamma(\theta)'T(x) - B(\theta)} h(x), \quad \theta \in \mathbb{R} \]

\[ B(\theta) = A(\gamma(\theta)) \]

\[ l(\theta; x) = \gamma(\theta)'T(x) - B(\theta) + \log h(x) \]

\[ \dot{l}(\theta; x) = \dot{\gamma}(\theta)'T(x) - \dot{\gamma}(\theta)' \nabla \gamma A(\gamma(\theta)) \]

\[ = \dot{\gamma}(\theta)'(T(x) - \mathbb{E}_\theta T(x)) \]

\[ \Rightarrow \dot{\gamma}(\theta)'T(x) \text{ is "locally complete suff. stat."} \]
Doubts about unbiasedness

The UMVUE might be very inefficient, or inadmissible, or just dumb, in cases where another approach makes much more sense.

Ex. \( X \sim \text{Bin}(1000, \theta) \)

Estimate \( g(\theta) = \mathbb{P}_\theta(X \geq 500) \)

UMVUE is \( 1\{X \geq 500\} \) \( \text{(why?)} \)

\( \Rightarrow X = 500 \) \( \Rightarrow \) Conclude \( g(\theta) = 100\% \)

\( X = 499 \) \( \Rightarrow \) Conclude \( g(\theta) = 0\% \)

This is not epistemically reasonable!!

Could do much better with e.g. MLE or a Bayes estimator.

In fact, our theorem should make us suspicious of UMVUE's: every idiotic function of \( T \) is a UMVUE (of its own expectation)
Gaussian Sequence Model

\[ X_i \sim N(m_i, 1), \quad i = 1, \ldots, d \text{ indep.} \]

or \( X \sim N_d(m, I_d) \quad m \in \mathbb{R}^d \), estimate \( \phi = \|m\|^2 \)

\( X \) is complete sufficient

\[
\mathbb{E}_m \|X\|^2 = \mathbb{E}_0 \left[ \|m + X\|^2 \right]
\]
\[
= \|m\|^2 + \mathbb{E}_0 \|X\|^2 + 2\mathbb{E}_0 [m^T X]
\]
\[
= \|m\|^2 + d
\]

\[ \Rightarrow \delta(x) = \|x\|^2 - d \]

If \( m = 0 \), \( \delta(x) < 0 \) about half the time!

\[ (\|x\|^2 - d)_+ = \max(0, \|x\|^2 - d) \]

strictly dominates UMVU
Gets worse: Ex 4.7 in Keener

\[ X \sim \text{Truncated Poisson}(\theta) \]

\[ p_\theta(x) = \frac{\theta^x e^{-\theta}}{x! (1-e^{-\theta})}, \quad x = 1, 2, \ldots \]

\[ \theta > 0 \]

Estimate \( \hat{\theta}(x) = e^{-\theta} \) (mass lost to truncation)

Keener shows UMVUE is \( S(X) = (-1)^{x+1} \)

\[ \frac{e^{-\theta}}{1-e^{-\theta}} \left( \theta - \frac{\theta^2}{2} + \frac{\theta^3}{3!} - \ldots \right) = \frac{e^{-\theta}}{1-e^{-\theta}} \left[ 1 - (1 + (\theta^2) + (\theta^3) + \ldots) \right] \]

[Idiotic but we cannot improve using any unbiased estimator]

Sometimes insisting on unbiasedness leads us to absurd results.

Unbiasedness has bad reputation, but other methods have their problems too.]