Lecture 11: Using Fisher Information, the Cramér-Rao Bound & Conditional Inference

· After going over its theoretical properties, let's use the Fisher Information definition to compute its value on a familiar setting.

$$\text{Let } X_i \overset{iid}{\sim} N(\theta, \sigma^2), \text{ where } \sigma^2 \text{ is known}$$

– We compute log-likelihood $\ell_{\underline{x}}(\theta)$

$$f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$f_\theta(\underline{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$\Downarrow$$

$$\ell_\theta(\underline{x}) = -\frac{1}{2} \cdot \sum_{i=1}^{n} \frac{(x_i-\theta)^2}{\sigma^2} - \frac{n}{2} \cdot \log(2\pi\sigma^2)$$

$$\dot{\ell}_\theta(\underline{x}) = \cancel{-}\frac{1}{\cancel{2}} \cdot \frac{1}{\sigma^2} \cdot \cancel{2} \cdot \cancel{(-1)} \cdot \sum_{i=1}^{n} (x_i - \theta)$$

$$= \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \theta)$$

Now, let's get $\ddot{\ell}_\theta(\underline{x})$

Recall $E[\ddot{\ell}_\theta(\underline{x})] = -I_\theta$

$$\ddot{\ell}_{\underline{x}}(\theta) = \frac{1}{\sigma^2}(-1) \cdot \sum_{i=1}^{n} 1 = -\frac{n}{\sigma^2} \Rightarrow \boxed{I_\theta = \frac{n}{\sigma^2}}$$

• Further, Recall $\boxed{E[\dot{\ell}_{\underline{x}}(\theta)] = 0}$ ✓

$$E[\dot{\ell}_{\underline{x}}(\theta)] = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (x_i - \theta) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i = n \cdot \theta \Rightarrow \boxed{\hat{\theta}^{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}}$$

Finally, we conclude that (for large enough n)

$$\hat{\theta}^{MLE} \overset{a}{\sim} N(\theta, \sigma^2/n) \qquad 1/I_0$$

To finish our discussion about the properties of MLE in a 1-dimensional setting, suppose that $\boxed{\tilde{\theta} = t(\underline{x})}$ is any unbiased estimate of

theta "tilde"

θ, based on iid sample $\underline{x} = (x_1, \dots, x_n)$ from $f_\theta(x)$. That is:

$$E_\theta\{t(\underline{x})\} = \theta$$

$$Bias = \theta - E_\theta\{t(\underline{x})\} = 0$$

Then, the Cramér-Rao lower bound says that the variance of $\tilde{\theta}$ exceeds the Fisher Information Bound.

$$Var_\theta\{\tilde{\theta}\} \geq 1/(n I_\theta)$$

. MLE has (asymptotic) variance at least as small as the best unbiased estimate of θ. Note that MLE is not unbiased in general, but its bias is small (order $1/n$), making the comparison w/ unbiased estimates appropriate.

Conditional Inference

Say we have iid sample.

$$x_i \overset{iid}{\sim} N(\theta, 1)$$

which has produced estimate $\hat{\theta} = \bar{x}$.

However, people conducting the sampling initially disagreed on what the sample size would be, so they flipped a coin to decide:

$$n = \begin{cases} 25 & \text{prob } 1/2 \\ 100 & \text{prob } 1/2 \end{cases}$$

and $n = 25$ won.

Q: what is $\sigma_{\bar{x}}$?

Classic frequentist rationale would have resulted in $\underbrace{\left[\frac{1}{2} \cdot \frac{\sigma^2}{100} + \frac{1}{2} \cdot \frac{\sigma^2}{25}\right]}_{\text{variance}}^{1/2} = 0.158$

However, conditional Inference would lead you to answer $[\sigma^2/25]^{1/2} = 0.2$.

Fisher's arguments for conditional inference:

1. More Relevant Inferences
(Inferences only have to do with what really happened)

2. Simpler inferences (we didn't have to assess any correlation between the result & the sample size selection step).

Example: Observed Fisher Information.

Rather than using $\hat{\theta}^{MLE} \overset{a}{\sim} N(\theta, 1/(n\mathcal{I}_\theta))$

Fisher suggested using $\hat{\theta}^{MLE} \overset{a}{\sim} N(\theta, 1/I(x))$

where $I(x)$ is the observed Fisher Inf.

$$I(\underline{x}) \underbrace{\overset{\Delta}{=}}_{\text{is defined as.}} -\ddot{\ell}_{\underline{x}}(\hat{\theta}^{MLE}) = \frac{-\partial^2}{\partial \theta^2} \ell_{\underline{x}}(\theta) \Big|_{\hat{\theta}^{MLE}}$$

Of course $E[I(\underline{x})] = n\mathcal{I}_\theta$, so in large samples, the observed Fisher Inf. is same as the Fisher Information. However, Fisher suggested that in smaller samples $I(\underline{x})$ gives a better idea of $\hat{\theta}$'s accuracy.

. We can check this by sampling from a distribution.

. For instance, let's use the Cauchy distribution.

$$f_\theta(x) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}$$

we're estimating this.

$$\ell_x(\theta) = \log\left(\frac{1}{\pi}\right) + \overset{0}{\cancel{\log(1)}} - \log(1+(x-\theta)^2)$$

$$\dot{\ell}_x(\theta) = \frac{2(x-\theta)}{1+(x-\theta)^2}$$

$$\ddot{\ell}_x(\theta) = \frac{-2 \cdot [1+(x-\theta)^2] - [2 \cdot (x-\theta) \cdot (-1) \cdot (2)(x-\theta)]}{[1+(x-\theta)^2]^2}$$

$$\ddot{\ell}_x(\theta) = \frac{-2[1+(x-\theta)^2] + 4(x-\theta)^2}{[1+(x-\theta)^2]^2}$$

. 10 000 samples of size = 20 drawn with $\theta = 0$. We computed $1/I(\underline{x})$ for each.

. Grouped 10 000 $\hat{\theta}^{MLE}$ values according to

• [illegible]

quintiles of $1/I(\underline{x})$ and calculated the empirical varince for each group.

Rough estimate of conditional variance of $\hat{\theta}^{MLE}$ given $1/I(x)$.

• Note that for all samples, the unconditional variance $1/n\mathcal{I}_\theta$ is the same.

• Specifically $\mathcal{I}_\theta = \frac{1}{2}$ for a single Cauchy observation. Thus, $\underbrace{\frac{1}{20 \cdot (\frac{1}{2})}}_{1/n\mathcal{I}_\theta} = \frac{1}{10} = 0.1$.