

Lecture 13 – Fisher Information

Last Modified: October 11, 2026

Outline:

13.1 Fisher Information	1
13.2 Cramer-Rao Lower Bound	4

13.1 Fisher Information

- In this lecture, we introduce a way to quantify the amount of “information” that a sample contains about an unknown parameter.
- We will assume throughout that the statistical models $\{f_\theta(x) \mid \theta \in \Theta\}$ is **regular**, i.e.,
 - (1) $\Theta = \mathbb{R}$ or $\Theta = (a, b)$.
 - (2) f_θ is differentiable and finite
 - (3) We can swap integration and differentiation

$$\frac{\partial}{\partial \theta} \int f_\theta(x) \, dx = \int \frac{\partial}{\partial \theta} f_\theta(x) \, dx$$

and other nice integrability conditions also hold.

- **Remark:** The pdfs we have seen previously in this course for which the support does *not* depend on θ are examples of regular statistical models.
- **Definition:** The **score function** of a regular statistical model is

$$S_\theta(x) := \frac{\partial}{\partial \theta} \log f_\theta(x) = \frac{f'_\theta(x)}{f_\theta(x)}$$

where $f'_\theta(x) = \frac{\partial}{\partial \theta} f_\theta(x)$. This is the relative rate at which the log-density of X changes at a given x as θ varies.

- **Theorem:** For a regular model we have

- (1) Score has mean zero:

$$\mathbb{E}_\theta[S_\theta(X)] = 0 \quad \text{for all } \theta$$

- (2) Variance of score satisfies

$$\text{Var}(S_\theta(X)) = \mathbb{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log f_\theta(X) \right)^2 \right] = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right]$$

- **Proof (sketch):**

- We know a pdf must integrate to one:

$$\int f_\theta(x) \, dx = 1 \quad \forall \theta$$

- Thus the derivative w.r.t. θ is zero:

$$\begin{aligned}
 0 &= \frac{\partial}{\partial \theta} \int f_{\theta}(x) \, dx \\
 &= \int \frac{\partial}{\partial \theta} f_{\theta}(x) \, dx \\
 &= \int \frac{\left(\frac{\partial}{\partial \theta} f_{\theta}(x)\right)}{f_{\theta}(x)} f_{\theta}(x) \, dx \\
 &= \int S_{\theta}(x) f_{\theta}(x) \, dx \\
 &= \mathbb{E}[S_{\theta}(X)]
 \end{aligned}$$

- For the variance, we write

$$\text{Var}(S_{\theta}(X)) = \mathbb{E}[S_{\theta}^2(X)] = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X)\right)^2\right]$$

It remains to show that $\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X)\right)^2\right] = -\mathbb{E}_{\theta}\left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta}(X)\right]$. To see this, take the derivative of $0 = \mathbb{E}_{\theta}[S_{\theta}(X)]$ w.r.t. θ :

$$\begin{aligned}
 0 &= \frac{\partial}{\partial \theta} \mathbb{E}_{\theta}[S_{\theta}(X)] = \frac{\partial}{\partial \theta} \int S_{\theta}(x) f_{\theta}(x) \, dx \\
 &= \int \frac{\partial}{\partial \theta} [S_{\theta}(x) f_{\theta}(x)] \, dx \\
 &= \int \left(\frac{\partial}{\partial \theta} S_{\theta}(x)\right) f_{\theta}(x) \, dx + \int S_{\theta}(x) \left(\frac{\partial}{\partial \theta} f_{\theta}(x)\right) \, dx \\
 &= \int \left(\frac{\partial^2}{\partial \theta^2} \log f_{\theta}(x)\right) f_{\theta}(x) \, dx + \int \left(\frac{\partial}{\partial \theta} \log f_{\theta}(x)\right)^2 f_{\theta}(x) \, dx \\
 &= \mathbb{E}_{\theta}\left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta}(X)\right] + \mathbb{E}_{\theta}\left[\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X)\right)^2\right].
 \end{aligned}$$

- **Definition:** The (expected) **Fisher Information** is the variance of the score function, i.e.

$$I_X(\theta) := \text{Var}_{\theta}(S_{\theta}(X)) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f_{\theta}(X)\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta}(X)\right]$$

- More generally, let Y be any random variable or random vector whose distribution depends on θ . For example, we could have $Y = \underline{X} = (X_1, \dots, X_n)$, or $Y = T(\underline{X})$, or $Y = X_1$. Write $f_{\theta,Y}(y)$ for its probability density or probability mass function, and assume that this family is regular. The score function and the **Fisher information about θ contained in Y** are

$$S_{\theta,Y}(y) := \frac{\partial}{\partial \theta} \log f_{\theta,Y}(y), \quad I_Y(\theta) := \text{Var}_{\theta}(S_{\theta,Y}(Y)).$$

As above, regularity gives $\mathbb{E}_{\theta}[S_{\theta,Y}(Y)] = 0$, and

$$I_Y(\theta) = \mathbb{E}_{\theta}[S_{\theta,Y}(Y)^2] = -\mathbb{E}_{\theta}\left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta,Y}(Y)\right].$$

- **Theorem:** Consider a regular statistical model and $X_i \stackrel{\text{iid}}{\sim} f_{\theta}$. Then

$$I_{\underline{X}}(\theta) = nI_{X_1}(\theta).$$

The Fisher information increases with the sample size!

- **Proof:** By independence, $f_{\theta, \underline{X}}(\underline{x}) = \prod_{i=1}^n f_{\theta}(x_i)$. Taking logarithms and differentiating gives

$$S_{\theta, \underline{X}}(\underline{X}) = \frac{\partial}{\partial \theta} \sum_{i=1}^n \log f_{\theta}(X_i) = \sum_{i=1}^n S_{\theta}(X_i).$$

Since $X_1, \dots, X_n$ are i.i.d., their scores $S_{\theta}(X_1), \dots, S_{\theta}(X_n)$ are also i.i.d. Hence

$$I_{\underline{X}}(\theta) = \text{Var}_{\theta} \left(\sum_{i=1}^n S_{\theta}(X_i) \right) = \sum_{i=1}^n \text{Var}_{\theta} (S_{\theta}(X_i)) = n I_{X_1}(\theta).$$

- **Example:** $X_i \stackrel{\text{iid}}{\sim} \text{Exp}(\theta)$. Compute the Fisher Information of $\underline{X} = (X_1, \dots, X_n)$.

- We'll use that $I_{\underline{X}}(\theta) = n I_{X_1}(\theta)$. The pdf of X_1 is

$$f_{\theta}(x) = \theta e^{-\theta x} \mathbb{1}_{\{x \geq 0\}}$$

and so

$$\log f_{\theta}(x) = n \log \theta - \theta x$$

- The first and second derivatives w.r.t. θ are

$$\begin{aligned} \frac{\partial}{\partial \theta} \log f_{\theta}(x) &= \frac{1}{\theta} - x, \\ \frac{\partial^2}{\partial \theta^2} \log f_{\theta}(x) &= -\frac{1}{\theta^2} \end{aligned}$$

- Thus

$$I_{\underline{X}}(\theta) = n I_{X_1}(\theta) = -n \mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log f_{\theta}(X_1) \right] = \frac{n}{\theta^2}$$

- Why did we introduce this quantity? The Fisher information allows us to quantify the loss of information when using a statistic which is not sufficient.
- **Theorem:** Consider a regular statistical model and let T be a statistic. Then

$$I_{\underline{X}}(\theta) \geq I_T(\theta) \quad \forall \theta$$

and the equality holds if and only if T is **sufficient**.

- **Remark:** The higher the Fisher Information, the closer we are to the scenario with full information.
- **Example:** $X_i \stackrel{\text{iid}}{\sim} \text{Exp}(\theta)$.

- We know that $T(\underline{X}) = \sum_{i=1}^n X_i \sim \text{Gamma}(n, \theta)$ is sufficient.
- Let's manually compute $I_T(\theta)$ and show it equals $I_{\underline{X}}(\theta) = n/\theta^2$, as just computed above. We have

$$\begin{aligned} f_{T, \theta}(t) &= \frac{\theta^n}{\Gamma(n)} t^{n-1} e^{-\theta t} \\ \implies \log f_{T, \theta}(t) &= n \log(\theta) + (n-1) \log(t) - \log(\Gamma(n)) - \theta t \end{aligned}$$

- The first and second derivatives w.r.t. θ are

$$\frac{\partial}{\partial \theta} \log f_{T, \theta}(t) = \frac{n}{\theta} - t \quad \text{and} \quad \frac{\partial^2}{\partial \theta^2} \log f_{T, \theta}(t) = -\frac{n}{\theta^2}$$

- Thus

$$I_T(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log f_{T, \theta}(t) \right] = \frac{n}{\theta^2}$$

This is equal to $I_{\underline{X}}(\theta) = n/\theta^2$ as expected, since T is sufficient.

13.2 Cramer-Rao Lower Bound

- We can use Fisher information to find a lower bound on the variance of our estimators.
- **Lemma (Cauchy-Schwarz Inequality):** For any random variables (X, Y) ,

$$\text{Cov}^2(X, Y) \leq \text{Var}(X) \text{Var}(Y)$$

where

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Equality holds if and only if X and Y are linearly dependent.

- **Theorem (Cramer-Rao Lower Bound):** Consider a regular statistical model and let T be any estimator. Then

$$\text{Var}_\theta(T) \geq \frac{\left[\frac{\partial}{\partial \theta} \mathbb{E}_\theta[T]\right]^2}{I_{\underline{X}}(\theta)}$$

- **Proof Sketch:**

- The derivative of $\mathbb{E}_\theta[T]$ w.r.t. θ satisfies

$$\begin{aligned} \frac{\partial}{\partial \theta} \mathbb{E}_\theta[T(\underline{X})] &= \frac{\partial}{\partial \theta} \int T(\underline{x}) f_\theta(\underline{x}) \, d\mathbf{x} \\ &= \int T(\underline{x}) \frac{\partial}{\partial \theta} f_\theta(\underline{x}) \, d\mathbf{x} \\ &= \int T(\underline{x}) S_\theta(\underline{x}) f_\theta(\underline{x}) \, d\mathbf{x} \\ &= \mathbb{E}_\theta[T(\underline{X}) S_\theta(\underline{X})] \end{aligned}$$

- Since the score function has zero mean, this expression is equal to the covariance between T and S_θ ,

$$\text{Cov}_\theta(T(\underline{X}), S_\theta(\underline{X})) = \mathbb{E}_\theta[T(\underline{X}) S_\theta(\underline{X})] - \underbrace{\mathbb{E}_\theta[T(\underline{X})] \mathbb{E}_\theta[S_\theta(\underline{X})]}_{=0}$$

- Applying the Cauchy-Schwartz inequality, we conclude that

$$\left[\frac{\partial}{\partial \theta} \mathbb{E}_\theta[T(\underline{X})]\right]^2 = \text{Cov}_\theta^2(T(\underline{X}), S_\theta(\underline{X})) \leq \text{Var}_\theta(T(\underline{X})) \underbrace{\text{Var}_\theta(S_\theta(\underline{X}))}_{=I_{\underline{X}}(\theta)}$$

- Rearranging the inequality completes the proof:

$$\text{Var}_\theta(T) \geq \frac{\left[\frac{\partial}{\partial \theta} \mathbb{E}_\theta[T]\right]^2}{I_{\underline{X}}(\theta)}$$

- **Remark:** If T is unbiased for $g(\theta)$, then

$$\text{Var}(T) \geq \frac{(g'(\theta))^2}{n I_{X_1}(\theta)}$$

In particular, if $g(\theta) = \theta$ then

$$\text{Var}(T) \geq \frac{1}{n I_{X_1}(\theta)}$$

- **Remark:** If an unbiased estimator T for $g(\theta)$ is such that

$$\text{Var}(T) = \frac{(g'(\theta))^2}{nI_{X_1}(\theta)}$$

then T is the UMVUE!

- **Example:** $X_i \stackrel{\text{iid}}{\sim} \text{Poisson}(\theta)$. Show that $T(\underline{X}) = \frac{1}{n} \sum_{i=1}^n X_i$ achieves the Cramer-Rao lower bound, thus is UMVUE for θ .

- The pdf is

$$f_X(x) = \frac{\theta^x e^{-\theta}}{x!} \quad \implies \quad \log f_X(x) = x \log(\theta) - \theta - \log(x!)$$

- The first and second derivatives w.r.t. θ are

$$\frac{\partial}{\partial \theta} \log f_X(x) = \frac{x}{\theta} - 1 \quad \text{and} \quad \frac{\partial^2}{\partial \theta^2} \log f_X(x) = -\frac{x}{\theta^2}$$

- Thus

$$I_{X_1}(\theta) = -\mathbb{E}\left[-\frac{X_1}{\theta^2}\right] = \frac{1}{\theta^2} \mathbb{E}[X_1] = \frac{1}{\theta}$$

$$\text{Var}(T) = \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{\theta}{n} = \frac{1}{n \cdot \frac{1}{\theta}} = \frac{1}{nI_{X_1}(\theta)}$$

- Since T achieves the Cramer-Rao lower bound, we conclude that T is the UMVUE

- **Remark:** It may happen that *no* estimator achieves the Cramer-Rao lower bound! Based on the Cramer-Rao lower bound, we can define the notion of **efficiency**.
- **Definition:** Given two unbiased estimators T and S for $g(\theta)$, we say that T is **more efficient** than S if

$$\text{Var}(T) \leq \text{Var}(S) \quad \forall \theta \in \Theta$$

- Let ℓ_{CR} be the value of the CR lower bound so $\ell_{CR} = \frac{(g'(\theta))^2}{nI_{X_1}(\theta)}$, i.e. $\text{Var}(T) \geq \ell_{CR}$ for all θ .
- **Definition:** Let T be unbiased for $g(\theta)$. Then the **efficiency** of T is defined as

$$\text{eff}(T) = \frac{\ell_{CR}}{\text{Var}(T)}$$

- $\text{eff}(T) \leq 1$ for all $\theta \in \Theta$.
- If $\text{Var}(T) = \ell_{CR}$ for all θ , then T is **efficient**.
- If $\lim_{n \rightarrow \infty} \frac{\ell_{CR}}{\text{Var}(T)} = 1$ for all θ , then T is **asymptotically efficient**.

- **Remark:** If T is efficient, then T is the UMVUE. However, the converse is not necessarily true!