Chapter 2: Point Estimation

Part II: Best Unbiased Estimator (UMVUE)

1 INTRODUCTION

Consider a class M of all estimators for θ, if there exists an estimator $\hat{\theta}^{**}$ in M that is uniformly better than any other estimators in M, then $\hat{\theta}^{**}$ is said to be a uniform minimum MSE estimator for θ in M. However, such an estimator $\hat{\theta}^{**}$ in general does not exist partly because we are too ***G***enerous to consider too many (all) estimators for θ, even some of them are poor or not reasonable (like $\hat{\theta} = 3423$).

Thus, we can restrict us to consider a particular class of estimators. Is it reasonable?
Yes, it is because sometimes a "very poor" estimator can be a (locally) best estimator. For instance, $\hat{\theta} = 3423$ is undoubtedly a poor estimator because no information of data is used, i.e. 3423 is always used to estimate an unknown parameter θ no matter what the observed data are. However, it is the best if the true value of θ is really equal to 3423. Thus, at least, we have to shrink a class of estimator to kick such a poor estimator out.

In this course, we keep mean-unbiased estimators.

Definition (Unbiasedness): If an estimator $\hat{\theta}$ satisfies $E(\hat{\theta}) = \theta$ for all $\theta \in \Theta$, then it is said to be mean-unbiased or unbiased for θ; otherwise, it is biased.

Note that $E\left(\hat{\theta}(\boldsymbol{X}) - \theta\right)^2 = Var\left(\hat{\theta}(\boldsymbol{X})\right) + bias^2$, where $bias = E\left(\hat{\theta}(\boldsymbol{X})\right) - \theta$.
So, if $\hat{\theta}$ is unbiased for θ, then its MSE is just its variance! In other words, we fix the bias to be zero, and then look for an estimator with the smallest variance (or the most efficient!!). Such an estimator is called a *UMVUE* --- ***uniform minimum variance unbiased estimator***.

More specifically, the *UMVUE* $\hat{\theta}^*$ for θ is an unbiased estimator such that for any other unbiased estimator $\hat{\theta}$ for θ, $Var\left(\hat{\theta}^*\right) \leq Var(\hat{\theta})$, for all $\theta \in \Theta$.

From now, we would discuss how to catch *UMVUE* $\hat{\theta}^*$. Thus, unless otherwise specified, we assume that a UMVUE exists for our target parameter θ.

Question: Would it be possible to have another unbiased estimator for θ with the same variance as $\hat{\theta}^*$for all $\theta \in \Theta$? That is, is the UMVUE $\hat{\theta}^*$ unique?

Lemma 1 (Uniqueness of UMVUE).

Assume that a UMVUE $\hat{\theta}^*$ for θ exists. Then, $\hat{\theta}^*$ is unique.

Proof:

Suppose that $\hat{\theta}^{**}$ ($\neq \hat{\theta}^*$) is another UMVUE, where

$$Var(\hat{\theta}^*) = Var(\hat{\theta}^{**}) \leq Var(\hat{\theta}), \text{ for all } \theta \in \Theta,$$

and $\hat{\theta}$ is any unbiased estimator for θ. Now, let $\hat{\theta}' = \frac{1}{2}(\hat{\theta}^* + \hat{\theta}^{**})$. It is easy to verify that $\hat{\theta}'$ is also unbiased for θ, and

$$\begin{aligned} Var(\hat{\theta}') &= \frac{1}{4}Var(\hat{\theta}^*) + \frac{1}{4}Var(\hat{\theta}^{**}) + \frac{1}{2}Cov(\hat{\theta}^*, \hat{\theta}^{**}) \\ &\leq \frac{1}{4}Var(\hat{\theta}^*) + \frac{1}{4}Var(\hat{\theta}^{**}) + \frac{1}{2}\sqrt{Var(\hat{\theta}^*)Var(\hat{\theta}^{**})} = Var(\hat{\theta}^*) \leq Var(\hat{\theta}') \end{aligned}$$

Thus, $Var(\hat{\theta}') = Var(\hat{\theta}^*)$, for all $\theta \in \Theta$.

The equality holds if and only if $Cov(\hat{\theta}^*, \hat{\theta}^{**}) = \sqrt{Var(\hat{\theta}^*)Var(\hat{\theta}^{**})}$, i.e. $\hat{\theta}^* = a\hat{\theta}^{**} + b$, where $a \neq 0$. It can be shown that $a = 1$ and $b = 0$, i.e. $\hat{\theta}^* = \hat{\theta}^{**}$. Therefore, $\hat{\theta}^*$ is unique.

How do we catch UMVUE?

In general, it is not east to find a UMVUE for a parameter being estimated. However, in some cases, we can still get it easily. We would study two approaches by:

1. **Cramér-Rao inequality**
2. **Complete and Sufficient statistics**

2 THE CRAMéR-RAO INEQUALITY (OR THE C-R INEQUALITY)

The C-R inequality provides a lower bound (usually called a ***C-R lower bound***) of the variance of an **UNBIASED** estimator for a parameter being estimated. So, if we can find an unbiased estimator whose variance achieves the C-R lower bound, then it must be a UMVUE of the parameter.

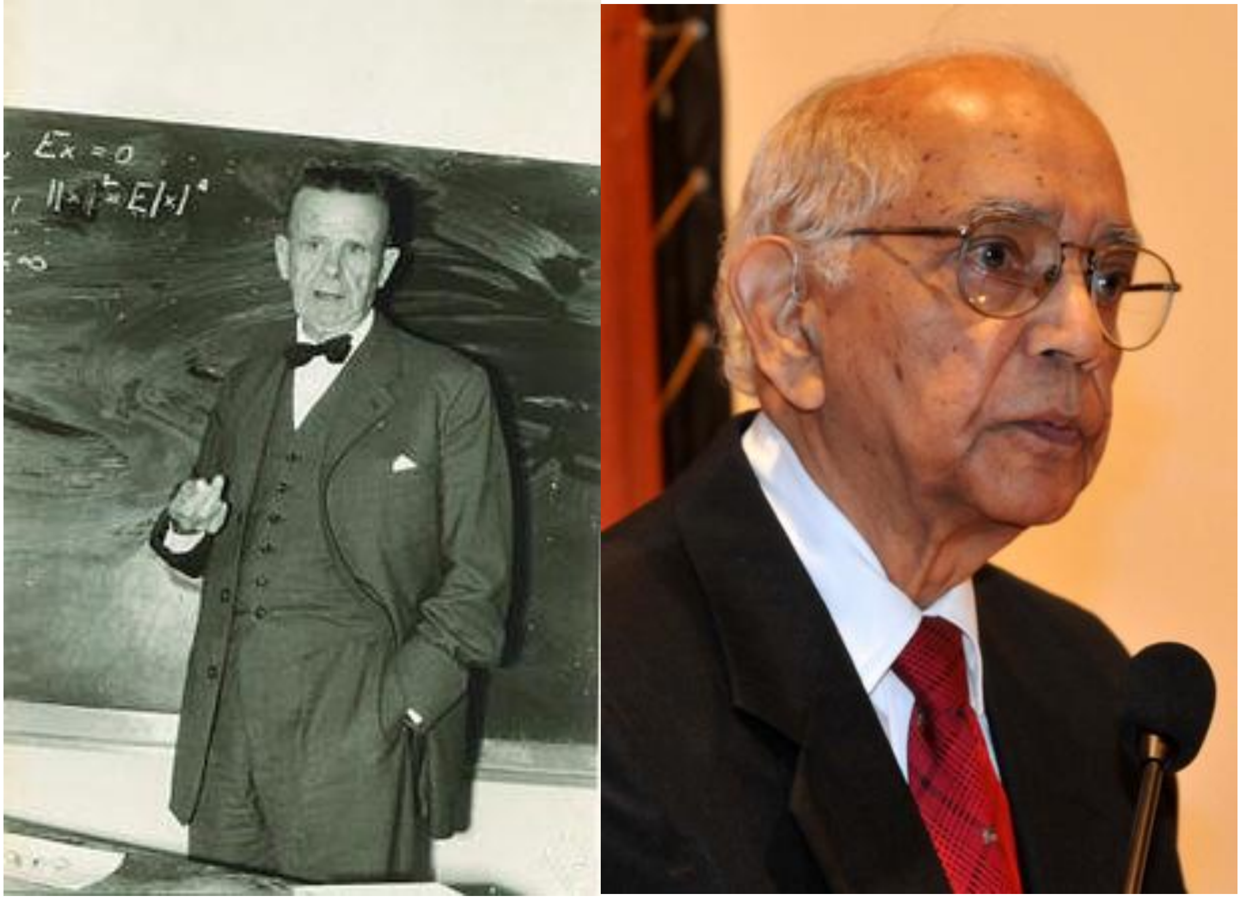


Figure 1: Harald Cramér (left) and Calyampudi Radhakrishna Rao (right)

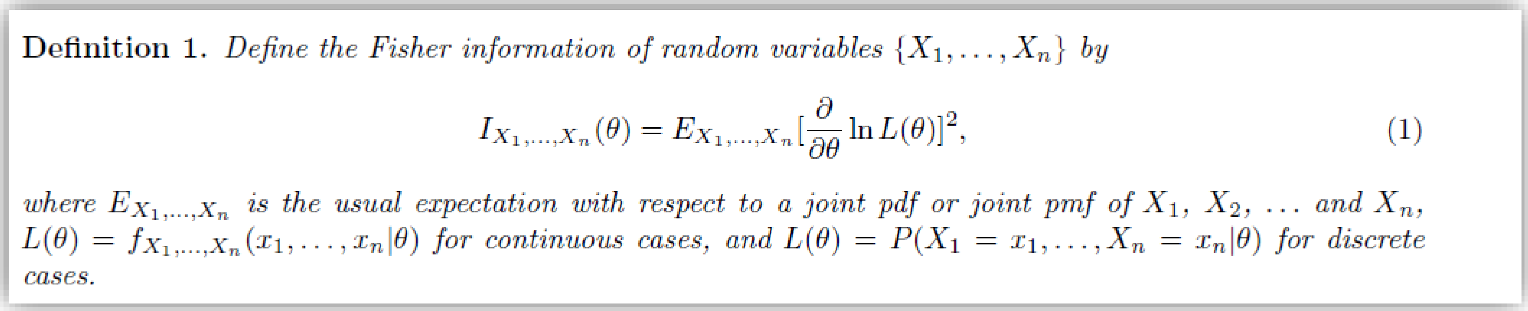
Note that a UMVUE may have a variance greater than the C-R lower bound. We will have an example to illustrate this later.

Before we study the details of the C-R inequality, let's turn to talk about Fisher Information first because the C-R lower bound is based on it.

2.1 FISHER INFORMATION

The ***Fisher information*** proposed by R.A. Fisher is a measure of the amount of information about an unknown parameter θ that a random variable or data carries.

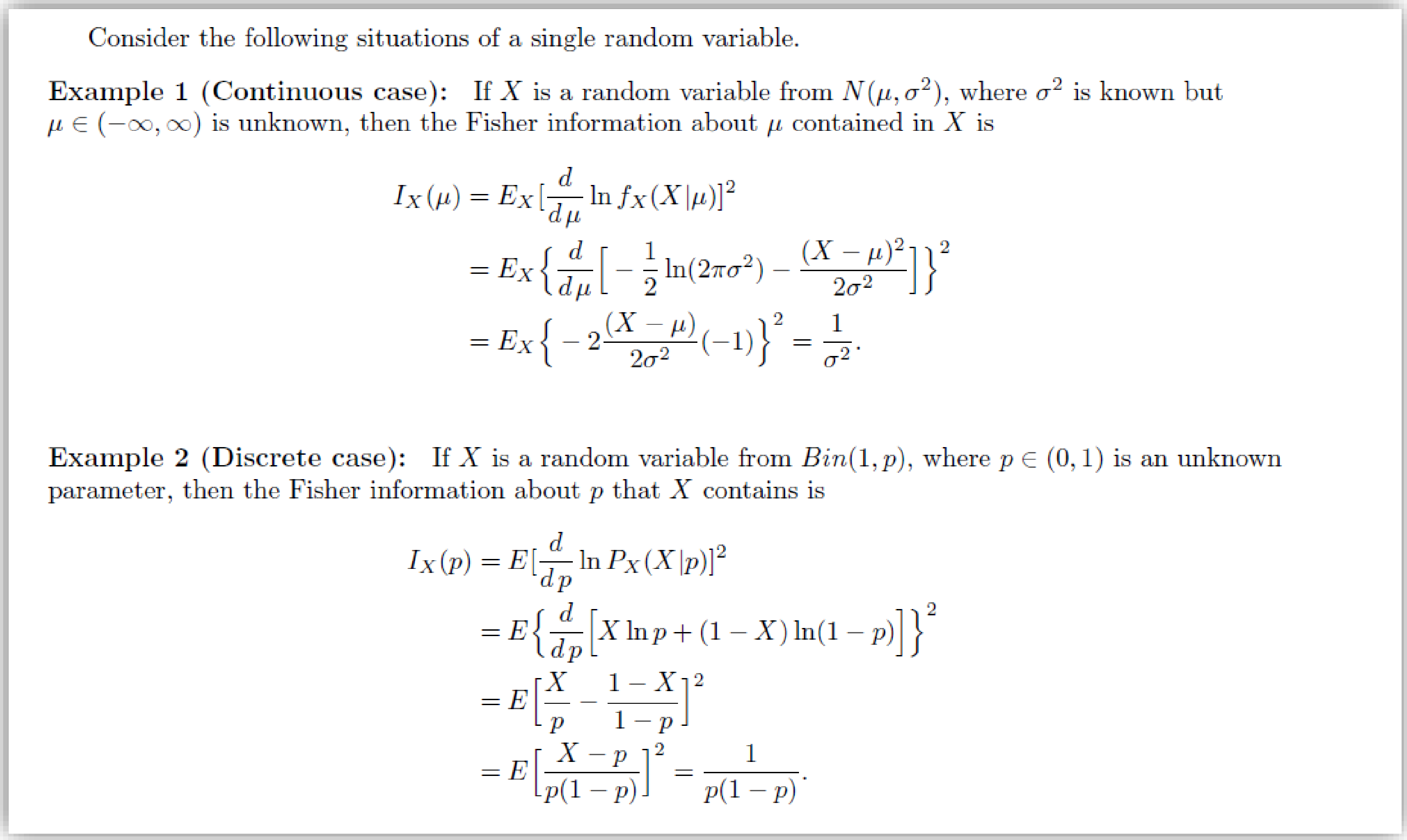
It is very important for us to have such a measure because data contain a certain amount of information about θ and we do want to know how to quantify this amount appropriately. Fisher information is commonly used because it has a lot of nice properties such as additivity.



Definition 1. *Define the Fisher information of random variables* $\{X_1, \dots, X_n\}$ *by*

$$I_{X_1,\dots,X_n}(\theta) = E_{X_1,\dots,X_n}[\frac{\partial}{\partial\theta}\ln L(\theta)]^2, \tag{1}$$

where $E_{X_1,\dots,X_n}$ *is the usual expectation with respect to a joint pdf or joint pmf of* X_1, X_2, $\dots$ *and* X_n, $L(\theta) = f_{X_1,\dots,X_n}(x_1, \dots, x_n|\theta)$ *for continuous cases, and* $L(\theta) = P(X_1 = x_1, \dots, X_n = x_n|\theta)$ *for discrete cases.*



Consider the following situations of a single random variable.

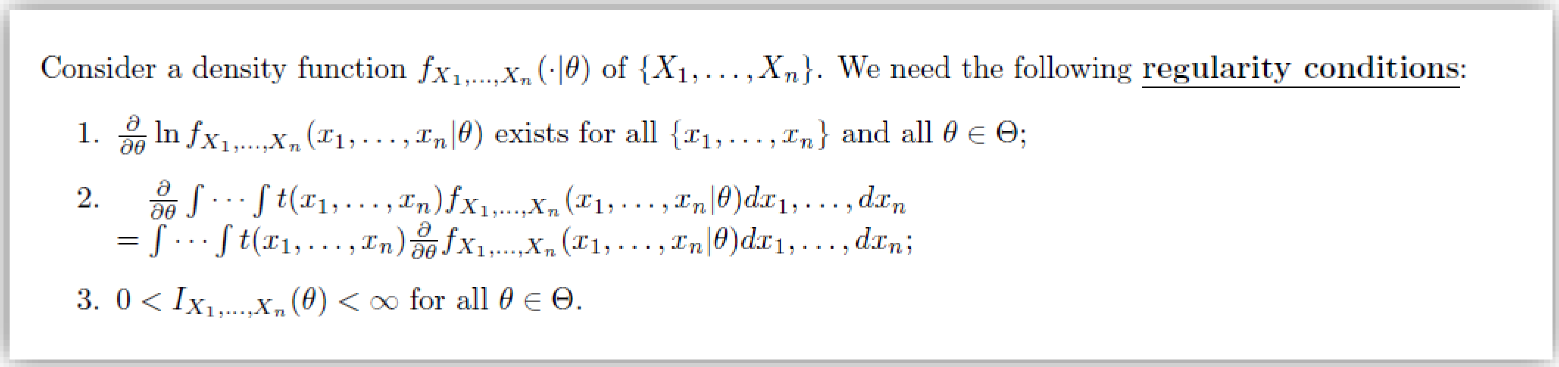
Example 1 (Continuous case): If X is a random variable from $N(\mu, \sigma^2)$, where σ^2 is known but $\mu \in (-\infty, \infty)$ is unknown, then the Fisher information about μ contained in X is

$$\begin{aligned} I_X(\mu) &= E_X[\frac{d}{d\mu}\ln f_X(X|\mu)]^2 \\ &= E_X\Big\{\frac{d}{d\mu}\Big[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(X-\mu)^2}{2\sigma^2}\Big]\Big\}^2 \\ &= E_X\Big\{-2\frac{(X-\mu)}{2\sigma^2}(-1)\Big\}^2 = \frac{1}{\sigma^2}. \end{aligned}$$

Example 2 (Discrete case): If X is a random variable from $Bin(1, p)$, where $p \in (0, 1)$ is an unknown parameter, then the Fisher information about p that X contains is

$$\begin{aligned} I_X(p) &= E[\frac{d}{dp}\ln P_X(X|p)]^2 \\ &= E\Big\{\frac{d}{dp}\Big[X\ln p + (1-X)\ln(1-p)\Big]\Big\}^2 \\ &= E\Big[\frac{X}{p} - \frac{1-X}{1-p}\Big]^2 \\ &= E\Big[\frac{X-p}{p(1-p)}\Big]^2 = \frac{1}{p(1-p)}. \end{aligned}$$

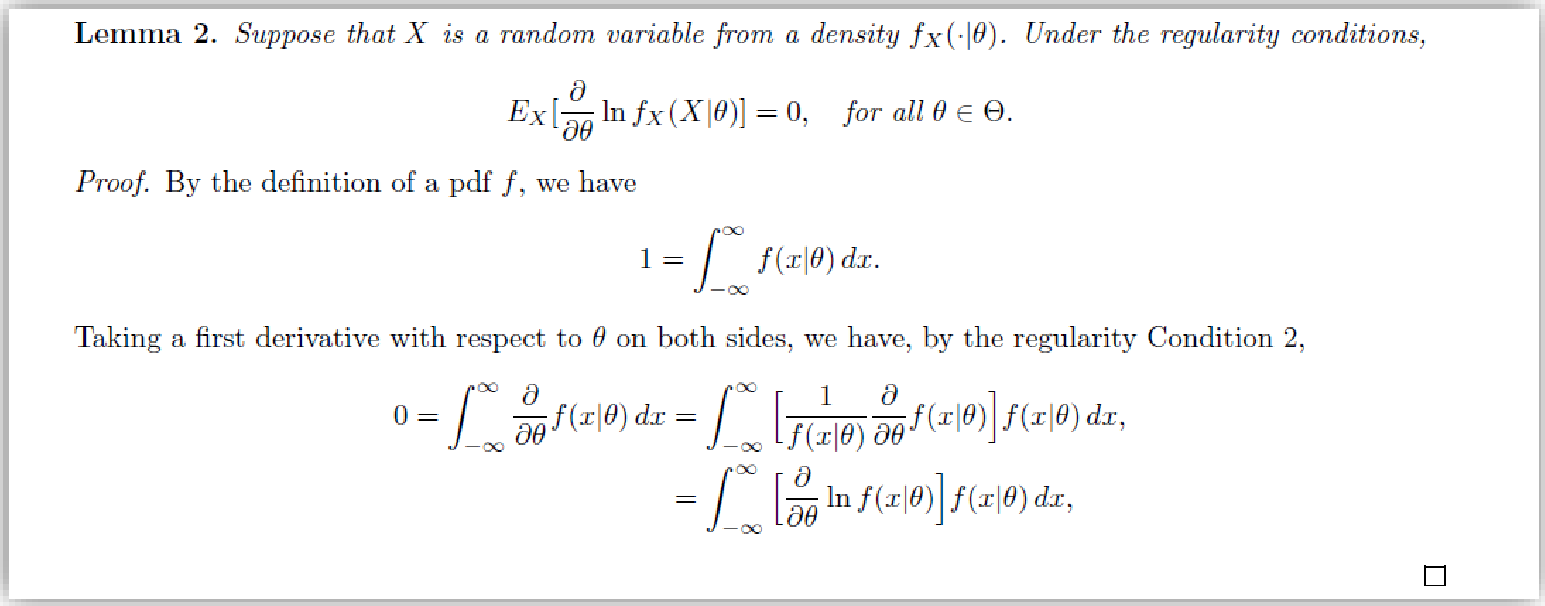
In the following, we will see some properties of the Fisher information. For simplicity and similarity, I only discuss the situations of continuous random variables. However, all of the properties below still hold true in discrete cases. Please keep it in mind.



Consider a density function $f_{X_1,\ldots,X_n}(\cdot|\theta)$ of $\{X_1,\ldots,X_n\}$. We need the following <u>**regularity conditions**</u>:

1. $\frac{\partial}{\partial\theta}\ln f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta)$ exists for all $\{x_1,\ldots,x_n\}$ and all $\theta\in\Theta$;
2. $\frac{\partial}{\partial\theta}\int\cdots\int t(x_1,\ldots,x_n)f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta)dx_1,\ldots,dx_n$
   $=\int\cdots\int t(x_1,\ldots,x_n)\frac{\partial}{\partial\theta}f_{X_1,\ldots,X_n}(x_1,\ldots,x_n|\theta)dx_1,\ldots,dx_n;$
3. $0<I_{X_1,\ldots,X_n}(\theta)<\infty$ for all $\theta\in\Theta$.

Condition 2 can be satisfied when the unknown parameter θ does not appear in an endpoint of all intervals in which the pdf is positive. Note that a uniform distribution over an interval from 0 to θ violates this condition.



Lemma 2. *Suppose that X is a random variable from a density $f_X(\cdot|\theta)$. Under the regularity conditions,*

$$E_X[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)]=0,\quad \textit{for all } \theta\in\Theta.$$

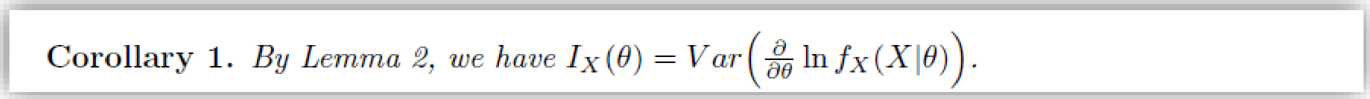
Proof. By the definition of a pdf f, we have

$$1=\int_{-\infty}^{\infty}f(x|\theta)\,dx.$$

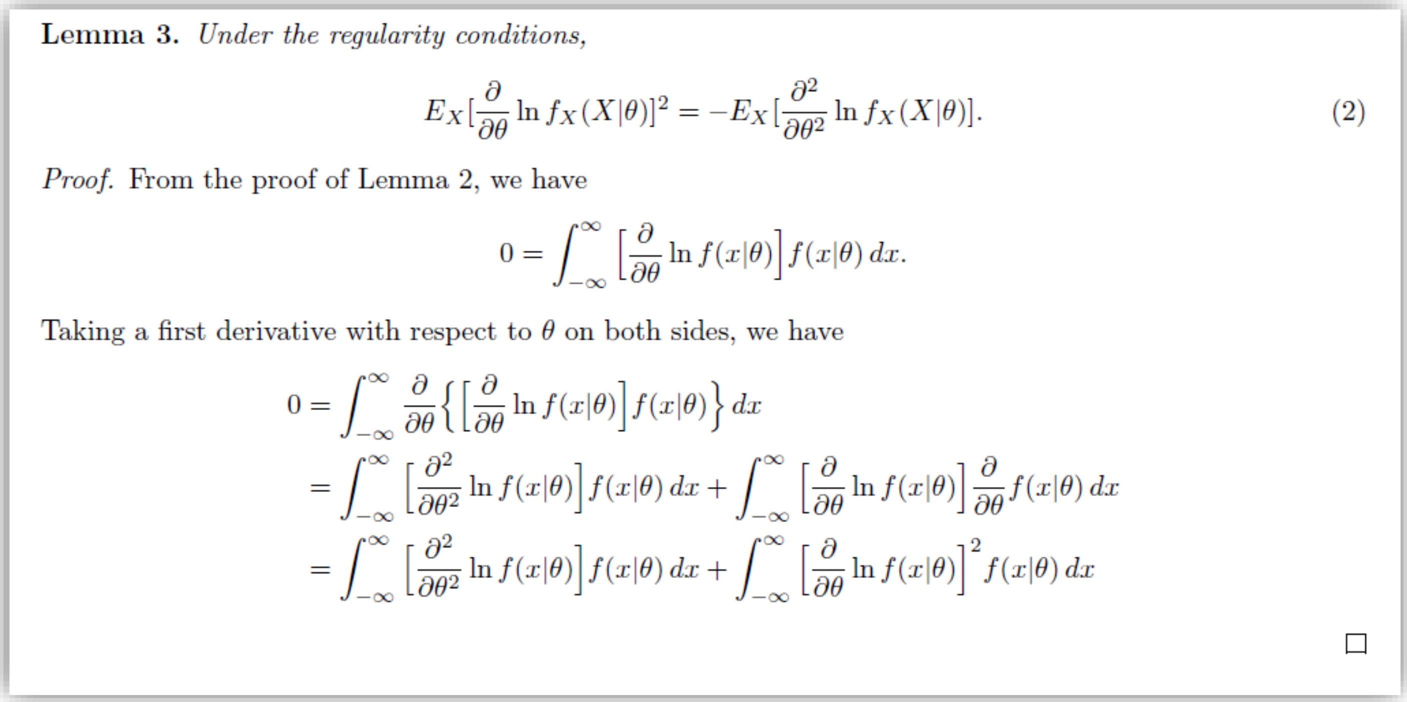
Taking a first derivative with respect to θ on both sides, we have, by the regularity Condition 2,

$$\begin{aligned}0=\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x|\theta)\,dx&=\int_{-\infty}^{\infty}\Big[\frac{1}{f(x|\theta)}\frac{\partial}{\partial\theta}f(x|\theta)\Big]f(x|\theta)\,dx,\\&=\int_{-\infty}^{\infty}\Big[\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big]f(x|\theta)\,dx,\end{aligned}$$

□



Corollary 1. *By Lemma 2, we have $I_X(\theta)=Var\Big(\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\Big)$.*



Lemma 3. *Under the regularity conditions,*

$$E_X[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)]^2 = -E_X[\frac{\partial^2}{\partial\theta^2}\ln f_X(X|\theta)]. \tag{2}$$

Proof. From the proof of Lemma 2, we have

$$0 = \int_{-\infty}^{\infty}\Big[\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big]f(x|\theta)\,dx.$$

Taking a first derivative with respect to θ on both sides, we have

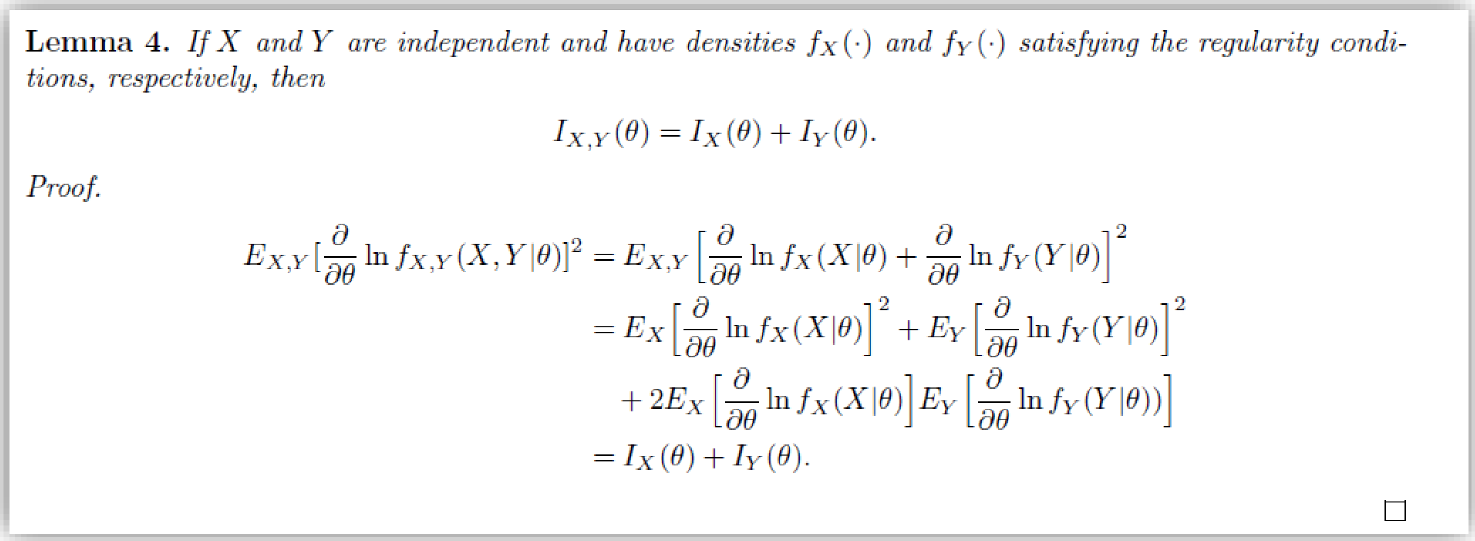
$$\begin{aligned}
0 &= \int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\Big\{\Big[\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big]f(x|\theta)\Big\}\,dx \\
&= \int_{-\infty}^{\infty}\Big[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\Big]f(x|\theta)\,dx + \int_{-\infty}^{\infty}\Big[\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big]\frac{\partial}{\partial\theta}f(x|\theta)\,dx \\
&= \int_{-\infty}^{\infty}\Big[\frac{\partial^2}{\partial\theta^2}\ln f(x|\theta)\Big]f(x|\theta)\,dx + \int_{-\infty}^{\infty}\Big[\frac{\partial}{\partial\theta}\ln f(x|\theta)\Big]^2 f(x|\theta)\,dx
\end{aligned}$$

□

Note that if we replace the single random variable X in Lemma 2 and Lemma 3 by a set of random variables, the results still hold true.

2.2 Fisher Information of Independent RV

In particular, if we consider independent rvs, say X and Y, then we have the following result of the relationship between the Fisher information about θ contained in (X, Y) and the Fisher Information about θ in X alone and in Y alone.



Lemma 4. *If X and Y are independent and have densities $f_X(\cdot)$ and $f_Y(\cdot)$ satisfying the regularity conditions, respectively, then*

$$I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta).$$

Proof.

$$\begin{aligned} E_{X,Y}[\frac{\partial}{\partial\theta}\ln f_{X,Y}(X,Y|\theta)]^2 &= E_{X,Y}\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta) + \frac{\partial}{\partial\theta}\ln f_Y(Y|\theta)\right]^2 \\ &= E_X\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]^2 + E_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(Y|\theta)\right]^2 \\ &\quad + 2E_X\left[\frac{\partial}{\partial\theta}\ln f_X(X|\theta)\right]E_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(Y|\theta))\right] \\ &= I_X(\theta) + I_Y(\theta). \end{aligned}$$

□

Lemma 4 indeed tells us ***the additive property of the Fisher information***.

2.3 Fisher Information of a R.S.

For a rs $\{X_1, \dots, X_n\}$ of size n from a distribution with a density function $f(\cdot\,|\theta)$, by Lemma 4, we can see that the Fisher information about θ contained in the random sample is

$$I_{X_1,\dots,X_n}(\theta) = I_{X_1}(\theta) + \cdots + I_{X_n}(\theta) = nI_{X_1}(\theta).$$

This results shows us that the Fisher information about θ in a rs is n times the Fisher information about θ in ONE observation, say X_1. Thus, the Fisher information about θ in a rs of size n is $\frac{n}{\sigma^2}$ in Example 1 ($\theta = \mu$) and $\frac{n}{p(1-p)}$ in Example 2 ($\theta = p$).

Remark that for different i and j, $I_{X_i}(\theta) = I_{X_j}(\theta)$ only means that X_i and X_j carry the same amount of the information about θ. It does not mean that they carry an identical information about θ.

2.4 FISHER INFORMATION OF A STATISTIC OR AN ESTIMATOR

Note that a statistic or an estimator can be regarded as a function for data condensation because we condense a r.s. of size n --- an n-dimensional object $\boldsymbol{X} = (X_1, \dots, X_n)^T$ --- into a lower-dimensional quantity, say a real-valued statistic $T(\boldsymbol{X})$.

In this condensation process, we may or may not lose some information about an unknown parameter θ. Lemma 5 shows us that the Fisher information can clearly reflect this situation.

Lemma 5. Under the regularity conditions, for any statistic $T(\boldsymbol{X})$ for θ, we have

$$I_{T(\boldsymbol{X})}(\theta) \leq I_{\boldsymbol{X}}(\theta),$$

where $I_{T(\boldsymbol{X})}(\theta) = E_{X_1,\dots X_n}\left[\frac{d}{d\theta}\log f_T(T(\boldsymbol{X})|\theta)\right]^2$ and $f_T(\cdot\,|\theta)$ is the density function of $T(\boldsymbol{X})$.

The above inequality means that none of the statistics can carry more information about θ than the information contained in a r.s.