Probability Theory and UMVUE

Wang Mingxun BEcon&Fin IV

1 Introduction

The target audience of the article are students who have taken probability and statistics courses with a certain level of mathematical maturity. It took me roughly two years to understand what exactly complete and sufficient statistics are. Indeed, it's true that you can memorize the methodology and still succeed in the exam but I want to reveal some big pictures about modern probability and statistics through this article. With the huge effort of former Soviet mathematician Andrey Kolmogorov and American mathematician Joseph Doob, probability has become a rigorous branch of mathematics. However, the rigor of the subject, which mainly comes from the implementation of real analysis has greatly increased the difficulty. Nowadays, research papers are in general written in this kind of 'formal' language, so without some familiarity with measure theory, it's hard to understand modern theoretical statistics research. I hope after reading this article, readers can acquire some basic understanding of real analysis while learning its application in mathematical statistics which is UMVUE here. The notations can be terrifying at the beginning, but don't be afraid. The concepts are not that hard and in fact they are quite naturally developed. The final study tip is that people always get stuck when they learn something new, just remember, grasping the big picture is all you need at first.

2 Measure Theory

I'm sure that all of you have 'learned' some measure theory in elementary school. You are asked to find the length of a ruler, the area of a rectangle, the volume of a cube, etc. That's why we need a concept called 'measure' to measure things. Here in probability, the event is the thing that we need to 'measure' . Think about the classic coin toss example, you have an outcome space $\Omega = \{H, T\}$. Suppose I ask you what the probability of getting a head is, you can answer one-half without a second thought. That's the 'measure' for the 'event' getting a head. Just like the length of a ruler, the area of a rectangle, you need some number to quantify the measure of a certain thing. That is probability. Events are just subsets of our sample space, in our previous example, getting a head can be written as $\{H\}$. We can calculate its probability by $P(\{H\}) = 0.5$. Here we can notice that probability is a function that maps the 'event space' into $[0, 1]$. That's something really important to bear in mind, mathematicians always take certain characteristics out from some specific examples. A certain level of abstraction can help us deal with the intrinsic characteristics that will not be noticed if we only consider some specific examples. Normally we require the 'event space' to be a σ-algebra.

Definition 1. We say that $\mathscr{F} \subset 2^{\Omega}$ is a σ-algebra, if the following are satistied.
(a) $\Omega \in \mathscr{F}$.
(b) If $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$.
(c) If $A_i \in \mathscr{F}$ for $i = 1, 2, 3, \ldots$, then $\cup_i A_i \in \mathscr{F}$.

Now we provide the rigorous definition of a probability measure.

Definition 2. $(\Omega, \mathscr{F})$ with $\mathscr{F}$ being a σ-algebra of subsets of Ω is called a measurable space. A measure μ is any countably additive non-negative set function on this space $\mu : \mathscr{F} \to [0, \infty]$:
(a) $\mu(A) \geq \mu(\phi) = 0$ for all $A \in \mathscr{F}$.
(b) $\mu(\cup_n A_n) = \sum_n \mu(A_n)$ for any countable collection of disjoint sets $A_n \in \mathscr{F}$. If $\mu(\Omega) = 1$, we call it a probability measure.

Definition 3. A measure space is a triplet $(\Omega, \mathscr{F}, \mu)$, with μ being a measure on the measurable space $(\Omega, \mathscr{F})$. A measure space $(\Omega, \mathscr{F}, P)$ with P as a probability measure is called a probability space.

The tools in real analysis help us rigorously define what exactly an event is, and only events in a σ-algebra will be inside our scope of consideration when we apply the measure P to the events.

Recall an elementary probability exercise, let's say you are asked to compute the probability of getting an odd number from 1 to 10, where each number has an equal probability to be obtained. Getting an odd number actually means taking the union of the sets $\{1\}$, $\{3\}$, $\{5\}$, $\{7\}$, $\{9\}$. The final probability can be written as $P(\{1, 3, 5, 7, 9\}) = \frac{5}{10} = 0.5$. Here the σ-algebra is taken to be all the subsets formed by the 10 numbers, implicitly, we are using property (b) in Definition 1 to calculate the probability. The two definitions are general enough for us to cope with all the situations. Below are two important measures, counting, and the Lebesgue measure.

Definition 4. If Ω is countable, define

$$\mu(A) = |A| = \text{number of points in } A.$$

Definition 5. Let $\Omega = \mathbb{R}^n$, define

$$\mu(A) = \int \cdots \int_A dx_1 \cdots dx_n.$$

With $n = 1, 2, 3$, $\mu(A)$ is called the length, area, and volume of A. That's all I want to say about measures, let's move on to Lebesgue integration. In fact, you don't have to read Section 3 to understand the rest of the material, but the Lebesgue integral notations will constantly appear in your later studies, so it's beneficial for you to take a look at them. Notice that Lebesgue measure and Lebesgue integrals are two different things.

3 Integration

I think the readers should be familiar with the Riemann integration. Roughly speaking, it partitions the x-axis and uses some rectangles to approximate the area under the curve. In Lebesgue integration, we try to partition the y-axis to achieve the same goal. I believe the readers should be more familiar with random variables and expectations instead of measurable functions and Lebesgue integration. Therefore, we will go with the former track, but actually, they are the same because random variables are just some measurable functions. Actually, you can guess that Lebesgue and Riemann integration are identical for some smooth and 'well-behaved' functions. That's true, the purpose of Lebesgue integration is to deal with some really 'bad' functions like a Dirichlet function. (Search it on the internet if you have no idea what it is.) So after all, statistics students can just treat it as a conceptual/notation difference. You can still calculate the integrals using the technique you have learned in calculus.

Definition 6. $(\Omega_1, \mathscr{A})$ and $(\Omega_2, \mathscr{B})$ are measurable spaces. $X : \Omega_1 \to \Omega_2$ is a measurable mapping if

$$X^{-1}(B) \coloneqq \{\omega : \omega \in \Omega_1, X(\omega) \in B\} \in \mathscr{A}, \ \ \forall B \in \mathscr{B}.$$

A random variable is just a measurable function from $(\Omega, \mathscr{A})$ to $(\mathbb{R}, \mathscr{B})$. The symbol $\mathscr{B}$ here is the Borel σ-algebra. Roughly speaking, the Borel σ-algebra consists of subsets of the real numbers. Therefore open, closed, semi-open, and semi-closed subsets of $\mathbb{R}$ are all inside the Borel σ-algebra. However, it turns out that the Borel σ-algebra does not contain all the subsets of $\mathbb{R}$. If you are interested in all the details of this special kind of σ-algebra, you can consult classic real analysis textbooks.

Now we can start with our construction of the 'expectation' , which is Lebesgue integral in disguise. It normally consists of three to four steps, from simple functions to bounded non-negative functions to general measurable functions.

3.1 Step 1: Simple r.v.

Definition 7. Our probability space is $(\Omega, \mathscr{A}, P)$. The expectation of a simple r.v. $X = \sum_{i=1}^{n} a_i I_{A_i}$, with $\sum_{i=1}^{n} A_i = \Omega$, $A_i \in \mathscr{A}$ is:

$$E(X) = \sum_{i=1}^{n} a_i P(A_i).$$

Here I_{A_i} is just the indicator function or you call it a characteristic function in real analysis. You can see that the calculation of discrete random variables is similar to this kind of simple random variable.

3.2 Step 2: Non-negative r.v.

Theorem 1. Given a non-negative r.v. X, there exists a sequence of increasing simple random variables such that $X_n(\omega) \uparrow X(\omega)$ for every ω. (The upper arrow here means approaching from below.)

I don't want to put down the exact proof of the theorem because the construction of the function is kind of tedious. I believe the diagram from Folland [Fol99] should give you enough intuition why we can find such a sequence of simple functions.

The proof is left to the reader (Exercise 10).
On the other hand, the following result shows that one is unlikely to commit any serious blunders by forgetting to worry about completeness of the measure.

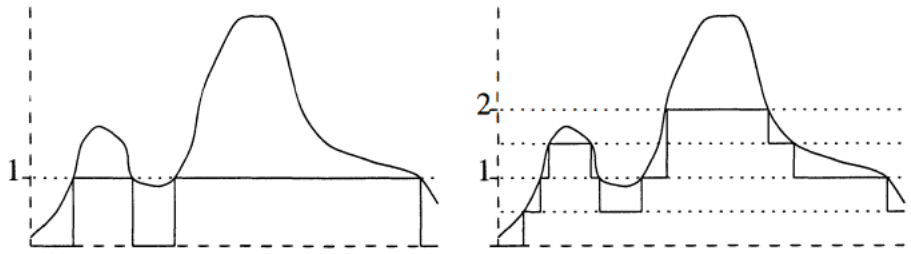

Fig. 2.1 The functions ϕ_0 (left) and ϕ_1 (right) in the proof of Theorem 2.10a.

From the diagram, we can imagine that with a finer partition of the image of the function, the simple function can approximate the target function in a better way. Therefore, it gives us the motivation to put down the next definition.

Definition 8. The expectation of X is $E(X) = \lim_{x\to\infty} E(X_n)$, X_n's are simple, non-negative and $X_n \uparrow X$

One thing you have to notice that is the outcome space of the probability space or the domain of our random variable might not be the real line. So the graph only serves as an intuitive reference. However, that reflects in another way how powerful the new integration is. Once you have the σ-algebra and a measure, you can integrate abstract objects like an outcome space. Let's go back to our fair coin example, suppose we define the random variable X to be:

$$X(\omega) = I_{\{\omega=H\}} = \begin{cases} 1 & \omega = \mathrm{H} \\ 0 & \omega = \mathrm{T} \end{cases}$$

Using Step 1 from above,

$$E(X) = 1 \times P(\{H\}) + 0 \times P(\{T\}) = 1 \times 0.5 = 0.5$$

3.3 Step 3: General r.v.

Define $X^+ = \max\{X,0\}$, $X^- = \max\{-X,0\}$

Definition 9. For general r.v. X, if either $E(X^+) < \infty$ or $E(X^-) < \infty$, but not both, then the expectation of X is

$$E(X) = E(X^+) - E(X^-)$$

Keep in mind that all the functions above are non-negative, we have to add one more definition to deal with the negative situations. We say that X is integrable if $E(|X|) < \infty$. So we have properly defined our expectations or Lebesgue integrals. In general, we denote it as:

$$E(X) = \int_\Omega X(\omega)P(d\omega) = \int X(t)dF_X(t) = \int_\Omega XdP = \int XdP$$

These are just symbols to denote the integral. They confuses me a lot when I first encountered the first two, I thought they are different in some sense. But in the end, I figure out they can all be treated as integrating the random variable with respect to some measure. Here is another tip for you, normally we are not lucky enough to get a simple function (except for a discrete random variable) to integrate, so we don't have an explicit way to compute the integral. The above section just introduces a conceptual framework instead of a new way to compute integrals. The functions/random variables you are going to see are Riemann integrable almost surely. (P.S. almost surely or almost everywhere is very important in real analysis)

4 Sufficient Statistics

I believe this example is one of the best ones I've ever seen from Keener's Theoretical Statistics. [Kee10] The example can be tedious, but please be patient and follow carefully the procedures.

Suppose X and Y are independent with common probability density

$$f_\theta(x) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

and let U be independent of X and Y and uniformly distributed on (0,1). Take $T = X + Y$, and define

$$\tilde{X} = UT \text{ and } \tilde{Y} = (1-U)T$$

Now we try to calculate the joint density of $\tilde{X}, \tilde{Y}$. Because X and Y are independent

$$\begin{aligned} P(T \leq t|Y = y) &= P(X + Y \leq t|Y = y) \\ &= E[I\{X + Y \leq t\}|Y = y] \\ &= \int I\{x + y \leq t\}dP_X(x) \\ &= F_X(t - y). \end{aligned}$$

(Here the above integral is just the usual $\int I\{x+y \le t\} f_X(x)dx$.)
So $P(T \le t|Y) = F_X(t-Y)$ and

$$F_T(t) = P(T \le t) = E[F_X(t-Y)].$$

Look at the distribution again, we know that $F_X(t-Y)$ is $1-e^{\theta(t-Y)}$ on $Y < t$ and 0 on $Y \ge t$. For $t \ge 0$:

$$F_T(t) = \int_0^t (1-e^{-\theta(t-y)})\theta e^{-\theta y} dy = 1 - e^{-\theta t} - t\theta e^{-\theta t}.$$

In order to get the density:

$$p_T(t) = F_T'(t) = \begin{cases} t\theta^2 e^{-\theta t} & t \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Because T and U are independent, the joint density is:

$$p_\theta(t,u) = \begin{cases} t\theta^2 e^{-\theta t} & t \ge 0, u \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

We know that,

$$P((\tilde{X},\tilde{Y}) \in B) = \int\int I_B(tu, t(1-u)) p_\theta(t,u) du dt.$$

Changing variables to $x = ut$, $du = dx/t$ and applying Fubini's theorem (reverse the order of integration),

$$P((\tilde{X},\tilde{Y}) \in B) = \int\int I_B(x, t-x) t^{-1} p_\theta(t, x/t) dt dx.$$

Now change our y to $t-x$,

$$P((\tilde{X},\tilde{Y}) \in B) = \int\int I_B(x,y)(x+y)^{-1} p_\theta\left(x+y, \frac{x}{x+y}\right) dy dx$$

Thus $\tilde{X}$,$\tilde{Y}$ have joint density

$$f_{\tilde{X},\tilde{Y}}(\tilde{x},\tilde{y}) = \frac{p_\theta(x+y, \frac{x}{x+y})}{x+y} = \begin{cases} \theta^2 e^{-\theta(x+y)} & x \ge 0, y \ge 0 \\ 0 & \text{otherwise} \end{cases}$$

Observing the above double integral, ignoring the indicator part, the remaining terms are just the density of the joint random variables, so we get the above conclusion.

Notice that it's just the joint density of X and Y, so we have found that the joint distribution of $\tilde{X}$ and $\tilde{Y}$ and the joint distribution of X and Y are the same. Consider some fake data sets generated from $\tilde{X}$ and $\tilde{Y}$, the pair $(\tilde{X},\tilde{Y})$ should be just as informative as (X,Y). But we can compute $(\tilde{X},\tilde{Y})$ from $T = X+Y$ and U. The distribution of U does not depend on θ, so we can obtain it from

a table of random numbers. Thus T by itself also provides as much information about θ as the pair (X,Y) because we could construct fake data $(\tilde{X},\tilde{Y})$ equivalent to (X,Y) using any uniformly distributed U on (0,1). The sum $T = X+Y$ is called a sufficient statistic. The construction of fake data works because the conditional distribution Q_t for X and Y given $T = t$, is given explicitly by

$$Q_t(B) = P_\theta[(X,Y) \in B)|T = t] = P[(Ut,(1-U)t) \in B]$$

which does not depend on θ. So it motivates the following definition.

Definition 10. Suppose X follows a distribution from a family $\mathscr{P} = \{P_\theta : \theta \in \Omega\}$. Then $T = T(X)$ is a sufficient statistic for $\mathscr{P}$ (or for X, or for θ), if the conditional distribution of X under P_θ given $T = t$ does not depend on θ.

The distribution family is a collection of distribution functions depending on the parameter θ from the parameter space Ω. Notice that the parameter space is completely different from our sample space in the beginning.

Suppose T is sufficient, we let

$$Q_t(B) = P_\theta(X \in B|T = t)$$

Then $P_\theta(X \in B|T) = Q_T(B)$ and

$$P_\theta(X \in B) = E_\theta[P_\theta(X \in B|T)] = E_\theta[Q_T(B)].$$

Suppose we use a random number generator to construct 'fake' data $\tilde{X}$ from T taking $\tilde{X} \sim Q_t$ when $T = t$. Then

$$\tilde{X}|(T = t) \sim Q_t$$

and we can consider an expectation

$$P_\theta(\tilde{X} \in B) = E_\theta[P_\theta(\tilde{X} \in B|T)] = E_\theta[Q_T(B)].$$

So X and $\tilde{X}$ have the same distribution.

Hope the above example gives you a deeper understanding of what exactly a sufficient statistic is. However, the definition does not help us in terms of finding sufficient statistics. In fact, we can use the factorization theorem, which is much easier to compute than the definition. You can find examples of how to apply the factorization theorem on the internet.

5 Minimal Sufficiency

Definition 11. A statistic T is minimal sufficient if T is sufficient, and for every sufficient statistic $\tilde{T}$ there exists a function f such that $T = f(\tilde{T})$ a.e.

Here equal almost everywhere (a.e.) or equal almost surely (a.s.) means that the two functions differ from only a P-null set which can be written as $P(T = f(\tilde{T})) = 1$. The definition rises from the fact that sometimes letting two things be exactly equal is too strong, so we introduce the concept that two things are almost equal except for some 'ignorable' parts.

A P-null set A is just a set satisfying $P(A) = 0$ which is in some sense negligible. That's the reason why we call it equal almost everywhere. Intuitively, a 'point' has a measure of 0 which is a P-null set. Recall that the continuous probability density function has the property of $f_X(x) = 0$. Because a single point is a P-null set, so the probability of that point is 0.

6 Completeness

Definition 12. A statistic T is complete for a family $\mathscr{P} = \{P_\theta : \theta \in \Omega\}$ if there exists a function f s.t.

$$E_\theta[f(T)] = c, \text{ for all } \theta,$$

implies $f(T) = c$ a.s.

I shall provide an example here to illustrate how we check the completeness. Suppose $X_1, ..., X_n$ are i.i.d. from a uniform distribution on $(0, \theta)$. Using indicator functions, the joint density is $I(\min x_i > 0)I(\max x_i < \theta)/\theta^n$, and so $T = \max\{X_1, ..., X_n\}$ is sufficient by the factorization theorem. By independence, for $t \in (0, \theta)$,

$$\begin{aligned} P_\theta(T \le t) &= P_\theta(X_1 \le t, \ldots, X_n \le t) \\ &= P_\theta(X_1 \le t) \times \cdots \times P_\theta(X_n \le t) \\ &= (t/\theta)^n, \end{aligned}$$

Then, T has density nt^{n-1}/θ^n, $t \in (0, \theta)$. Suppose $E_\theta[f(T)] = c$ for all $\theta > 0$, then

$$E_\theta[f(T) - c] = \frac{n}{\theta^n} \int_0^\theta [f(t) - c] t^{n-1} dt = 0$$

We know that $[f(t) - c]t^{n-1} = 0$ almost surely for $t > 0$. So $P(f(T) = c) = 1$, and T is complete. The final step is a classic result in integration theory. If the integral of non-negative integrand has value 0, then the integrand has to be 0 almost surely. That makes perfect sense because having some bad points (equal to something else rather than 0) will not affect the value of the integral. However if your function has even a tiny little range equal to something else, then your integral will not be 0.

Theorem 2. If T is complete and sufficient, then T is minimal sufficient.

The theorem is called the Bahadur's theorem, you can take a look at the technical proof by searching it on the internet. Before we discuss the Rao-Blackwell theorem, which is involved in the proof

of the main theorem related to the properties of the UMVUE, we shall discuss the risk and loss functions first. In decision theory, we use a loss function $L(\theta, d)$ to measure the loss generated by estimating θ with a value d. You can simply treat them as a tool to measure the error. Because a random variable X would be involved, in general, we are dealing with $L(\theta, \delta(X))$ which is also random. Therefore, we come up with the risk function which is just the average loss of the estimator, defined as:

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X)].$$

Theorem 3 (Rao-Blackwell)**.** Let T be a sufficient statistic for $\mathscr{P} = \{P_\theta : \theta \in \Omega\}$, let δ be an estimator of $g(\theta)$, and define $\eta(T) = E[\delta(X)|T]$. If $\theta \in \Omega$, $R(\theta, \delta) < \infty$, and $L(\theta, \cdot)$ is convex, then

$$R(\theta, \eta) \leq R(\theta, \delta).$$

Furthermore, if $L(\theta, \cdot)$ is strictly convex, the inequality will be strict unless $P(\delta(X) = \eta(T)) = 1$.

Proof. Applying Jensen's inequality with expectations against the conditional distribution of $\delta(X)$ given T gives

$$L(\theta, \eta(T)) \leq E_\theta[L(\theta, \delta(X))|T].$$

Taking expectations, $R(\theta, \eta) \leq R(\theta, \delta)$. □

7 UMVUE

Definition 13. An estimator δ is called unbiased for $g(\theta)$ if

$$E_\theta[\delta(X)] = g(\theta), \ \forall \theta \in \Omega$$

If an unbiased estimator exists, g is called U-estimable.

Definition 14. An unbiased estimator δ is uniformly minimum variance unbiased estimator (UMVUE) if

$$Var_\theta(\delta) \leq Var_\theta(\delta^*), \ \forall \theta \in \Omega,$$

for any competing unbiased estimator δ^*.

We finally achieve the main theorem of the article:

Theorem 4. Suppose g is U-estimable and T is complete sufficient. Then there is an essentially unique unbiased estimator based on T that is the UMVUE.

Proof. Let $\delta = \delta(X)$ be any unbiased estimator and define

$$\eta(T) = E[\delta|T]$$

as in the Rao-Blackwell theorem. Taking expectation,

$$g(\theta) = E_\theta[\delta] = E_\theta[E_\theta[\delta|T]] = E_\theta[\eta(T)]$$

and thus $\eta(T)$ is unbiased. (We apply the law of iterated expectation here, it's useful.) Suppose $\eta^*(T)$ is also unbiased. Then

$$E_\theta[\eta(T) - \eta^*(T)] = 0 \ \forall \theta \in \Omega$$

and by completeness, $P(\eta(T) - \eta^*(T) = 0) = 1$. This shows that the estimator $\eta(T)$ is essentially unique; any other unbiased estimator based on T will equal $\eta(T)$ except on a P-null set. The estimator $\eta(T)$ has minimum variance by the Rao-Blackwell theorem with squared error loss. Specifically, if δ^* is any unbiased estimator, then $\eta^*(T) = E_\theta[\delta^*|T]$ is unbiased by the calculation above. With squared error loss, the risk of δ^* on $\eta^*(T)$ is the variance, and so

$$Var_\theta(\delta^*) \geq Var_\theta(\eta^*(T)) = Var_\theta(\eta(T)), \ \forall \theta \in \Omega.$$

Thus, $\eta(T)$ is the UMVUE. □

You can come up with the choice of $\eta(T)$ by yourself if you are highly sensitive to the application of the law of iterated expectation. The proof might not be easy to understand first, but I want to draw your attention to how we apply the definition of the completeness to show the uniqueness of the estimator $\eta(T)$.

By the theorem above, we know that once we have a complete and sufficient statistic T, it is automatically the UMVUE given that our estimand is U-estimable. Now, we introduce another method to find the UMVUE that is relevant to the Cramer-Rao lower bound (CRLB). The proof will be lengthy but you can benefit a lot by going through it. It involves a lot of common techniques.

Theorem 5 (Cramer-Rao lower bound (CRLB))**.** Let $X = \{X_1, \ldots, X_n\} \sim f_\theta(x)$, where $\theta \in \Omega \subset \mathbb{R}^k$. Let $E[\delta(X)] = g(\theta)$. Assume for $h(x) = 1$ and $h(x) = \delta(x)$,

$$\frac{\partial}{\partial \theta} \int h(x) f_\theta(x) dx = \int h(x) \frac{\partial}{\partial \theta} f_\theta(x) dx, \ \text{ for } \ \theta \in \Omega$$

Let $l(\theta) = \log f_\theta(X)$. We have

$$Var(\delta(X)) \geq (g'(\theta))_{1\times k} [I_X(\theta)]^{-1}_{k\times k} (g'(\theta))^T_{k\times 1},$$

where the equality holds iff $\delta(X)$ and $l'(\theta)$ are linearly dependent, which means $\delta(X) - E[\delta(X)] = \lambda_0 (l'(\theta))^T$ for some constant λ_0.

The following proof requires two theorems about matrix differentiation, they are important in computations and can be easily found on the internet. The first one is $\frac{\partial}{\partial x} x^T A x = 2Ax$. The second one is $\frac{\partial}{\partial x} b^T x = b$. Notice that $\frac{\partial}{\partial x^T} x^T b = b$.

Proof. The main lemma we are proving is: For any random variable δ and random vector $Y = (Y_1, \ldots, Y_k)$, we have

$$Var(\delta) \geq Cov(\delta, Y)[Var(Y)]^{-1}[Cov(\delta, Y)]^T$$

where the equality holds iff δ and Y are linearly dependent. Here,

$$\begin{aligned} Cov(\delta, Y) &= (Cov(\delta, Y_1), \ldots, Cov(\delta, Y_k))_{1\times k} \\ Var(Y) &= (Cov(Y_i, Y_j))_{k\times k} \end{aligned}$$

So now we start to prove the lemma, for any $\lambda = (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k$, we have

$$\begin{aligned} 0 &\leq Var(\delta - \lambda Y^T) = Cov(\delta - \lambda Y^T, \delta - \lambda Y^T) \\ &= Var(\delta) - 2Cov(\delta, Y\lambda^T) + Cov(\lambda Y^T, \lambda Y^T) \\ &= Var(\delta) - 2Cov(\delta, Y)\lambda^T + \lambda Var(Y)\lambda^T \\ &:= g(\lambda). \end{aligned}$$

Setting $g'(\lambda) = 2\lambda Var(Y) - 2Cov(\delta, Y) = 0$, we get

$$\lambda_0 = Cov(\delta, Y)[Var(Y)]^{-1}.$$

Plugging it back to g, we have

$$g(\lambda_0) = Var(\delta) - Cov(\delta, Y)[Var(Y)]^{-1}[Cov(\delta, Y)]^T \geq 0.$$

where equality holds iff $g(\lambda_0) = 0 = Var(\delta - \lambda_0 Y^T)$ iff $\delta - \lambda_0 Y^T = C$.

Here comes the second part of the proof, taking $Y = \frac{\partial}{\partial\theta} \log f_\theta(X) = l'(\theta)$, we have two more small results to prove. To prove the two small results, we need another observation first.

$$l'_\theta(x) = \frac{1}{f_\theta(x)} \frac{\partial}{\partial\theta} f_\theta(x).$$

Now, the first one is an observation by differentiating $1 = \int f_\theta(x)dx$

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int f_\theta(x)dx \\ &= \int \frac{\partial}{\partial\theta} f_\theta(x)dx \\ &= \int l'_\theta(x) f_\theta(x)dx \\ &= E[l'_\theta(x)]. \end{aligned}$$

Notice that we have applied the regularity condition in the assumption to exchange the integral and the partial derivative sign. The second one comes from differentiating $g(\theta) = E[\delta(X)] = \int \delta(x) f_\theta(x) dx$ w.r.t. θ:

$$\begin{aligned} g'(\theta) &= \int \delta(x) f'_\theta(x) dx \\ &= \int \delta(x) l'_\theta(x) f_\theta(x) dx \\ &= E[\delta(x) l'_\theta(x)]. \end{aligned}$$

So we get

$$\begin{aligned} 0 = E[l'(\theta)] = E[Y], \\ g'(\theta) = Cov(\delta(X), l'(\theta)) = Cov(\delta(X), Y). \\ I_X(\theta) = E[(l'(\theta)^T_{k\times 1} l'(\theta)_{1\times k}] = E[Y^T Y] = Var(Y). \end{aligned}$$

Applying the main lemma, we get

$$Var(\delta) \geq Cov(\delta, Y)[Var(Y)]^{-1}[Cov(\delta, Y)]^T = g'(\theta))_{1\times k}[I_X(\theta)]^{-1}_{k\times k}(g'(\theta))^T_{k\times 1}$$

where the equality holds iff $\delta - \lambda_0 l'(\theta)^T = C$ which implies that $E[\delta(X)] = C$, so $\delta(X) - E[\delta(X)] = \lambda_0 l'(\theta)^T$ □

Corollary 1. If θ is a scalar, then the CRLB becomes

$$Var(\delta(X)) \geq \frac{[g'(\theta)]^2}{I_X(\theta)} \text{ where } I_X(\theta) = E(l'(\theta))^2 = -E(l''(\theta))$$

The last equality of the Fisher Information matrix $I_X(\theta)$ can be proved although not that easily. Readers can give it a try. The expectation of the second derivative of the log-likelihood is easier to compute, so I include it in the corollary to inform the readers that we don't have to always use the definition.

The name of the uniformly minimum variance unbiased estimator reveals that it's related to the variance of the estimator. CRLB is the bound for the variance of an unbiased estimator. So if the CRLB is sharp for some unbiased estimator T, then T must be the UMVUE. However, it's possible that the CRLB can't be achieved. That ends the article. I really hope that my understanding can help you gain a deeper understanding of the topic and provide you with enough motivation and intuition to learn more advanced probability and statistics theory.

References

[Fol99] Gerald B Folland. *Real analysis: Modern techniques and their applications*. Vol. 40. John Wiley & Sons, 1999.

[Kee10] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2010.