MTH211A: Theory of Statistics

Module 3: Point Estimation

Recall the (parametric) inference problem: Let $X_1, \cdots, X_n$ be a random sample from some distribution F, which is parameterized by some parameter vector $\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \Theta$, where Θ is the parameter space. Our goal is to infer about F (which is equivalent to infer about $\boldsymbol{\theta}$, or some function of $\boldsymbol{\theta}$) using the random sample $X_1, \cdots, X_n$.

Estimator. The main task of parametric inference is to estimate the parameter $\boldsymbol{\theta}$ with the help of samples $X_1, \cdots, X_n$. For example, suppose it is assumed that the random sample belongs to the `normal` family of distributions, but the parameters of the distribution, μ, σ^2, are not specified. A statistician naturally uses functions of sample observations (statistics) to estimate the parameters. For example, one may use the sample mean, $\bar{X}_n$, to estimate μ, and the sample variance S_n^2 to estimate σ^2. A statistic, which are used to estimate a parameter, are called an estimator. A realization of the estimator based on a sample, which serves as a potential value of the parameter, is called an estimate.

Of course, for any parameter there exists many estimators. For example, one can also put forward the sample median $\tilde{X}_{me}$, instead of sample mean $\bar{X}_n$, to estimate μ. In this module, we will learn some desirable properties that a good estimator should satisfy.

1 Desirable properties of a point estimators

Definition 1 (Mean Squared Error)**.** *The mean squared error (MSE) of an estimator $T(\mathbf{X})$ of the parameter θ is defined as $E_\theta\{T(\mathbf{X}) - \theta\}^2$.*

Remark 1. *In the definition of MSE, the suffix θ in $E_\theta(\cdot)$ is used to emphasize that the data generating process of the sample $\{X_1, \ldots, X_n\}$ involves the parameter θ. As the expectation is taken over $\mathbf{X}$, the MSE is a function of θ only. For example, let $\{X_1, \ldots, X_n\}$ be a random sample from $N(\mu, \sigma^2)$, then the MSE of $\bar{X}_n$ is $E_{\mu,\sigma^2}(\bar{X}_n - \mu)^2 = \mathrm{var}_{\mu,\sigma^2}(\bar{X}_n) = \sigma^2/n$.*

Remark 2. *Instead of MSE one can use other (expected) measures of difference, for example, mean absolute error (MAE) $E_\theta|T(\mathbf{X}) - \theta|$. However, MSE is usually preferred over MAE due to algebraic amenability.*

Remark 3. *The MSE can be decomposed into two parts:*

$$MSE = E_\theta\left[\{T(\mathbf{X}) - E_\theta(T(\mathbf{X})\} + \{E_\theta(T(\mathbf{X}) - \theta\}\right]^2 = \mathrm{var}_\theta\{T(\mathbf{X})\} + \{E_\theta(T(\mathbf{X})) - \theta\}^2 = variance + bias^2.$$

Intuitively, the MSE is the sum of two quantities, one measuring accuracy (bias^2), and the other measuring precision (variance).

Remark 4. *Naturally, one would like to use the point estimator having minimum MSE, among the pool of all possible estimators. However, the minimum MSE estimator do not generally exist. Therefore, one possible way is to focus on a (reasonable) sub-class of all possible estimators, and try to find estimator with minimum MSE in that subclass. The subclass we focus on is the class of unbiased estimators.*

Definition 2 (Unbiased estimator)**.** *The expected difference of the estimator $T(\mathbf{X})$ and the parameter θ, when the data generating process involves θ, is called bias. An estimator $T(\mathbf{X})$ is called unbiased for θ if the bias is zero for all $\theta \in \Theta$, i.e., if $E_\theta\{T(\mathbf{X})\} = \theta$ for all $\theta \in \Theta$.*

Example 1. Let $\{X_1, \ldots, X_n\}$ be a random sample from some population with finite mean μ. Then sample mean $\bar{X}_n$ is an unbiased estimator of μ. Further, if $\{X_1, \ldots, X_n\}$ is a random sample from some population with finite variance σ^2, then the estimator $S_n^{\star 2} = nS_n^2/(n-1)$ is an unbiased estimator of σ^2.

Note: For an unbiased estimator $T(\mathbf{X})$, MSE is equal to variance of the estimator.

Remark 5. *A biased estimator may have lower MSE than an unbiased estimator. For example, consider a random sample* $\{X_1, \ldots, X_n\}$ *from* $\texttt{Normal}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$. *The MSE of unbiased sample variance* $S_n^{\star 2}$, $2\sigma^4/(n-1)$, *is higher than the MSE of sample variance* S_n^2, $(2n-1)\sigma^4/n^2$.

Example 2. Let $\{X_1, \ldots, X_n\}$ be a random sample from $\texttt{Bernoulli}(p)$, $0 \le p \le 1$. Then $\bar{X}_n$ is unbiased estimator of p, and the MSE of $\bar{X}_n$ is $p(1-p)/n$.

Remark 6. *An unbiased estimator of* θ *may not exist. For example, in the* $\texttt{Bernoulli}(p)$ *example above, an unbiased estimator of* p^2 *based on a random sample of size* $n = 1$ *does not exist.* (WHY?)

For a parameter θ, *if there exist an unbiased estimator of* θ, *then* θ *is called U-estimable.*

Remark 7. *An unbiased estimator may not always be a good estimator. Consider a random sample* X *of size* $n = 1$ *from* $\texttt{Poisson}(\lambda)$ *distribution, and suppose* $\theta = \exp\{-3\lambda\}$. *It can be shown that* $T = T(X) = (-2)^X$ *is an unbiased estimator of* θ. *However,* T *is an absurd estimator as it may have large negative realization, although* θ *is positive.*

Remark 8. *Usually unbiased estimator of an U-estimable parameter* θ *is not unique. In fact if there exists two different unbiased estimator* T_1 *and* T_2 *of* θ, *then for any* $\alpha \in (0,1)$, $T_\alpha = \alpha T_1 + (1-\alpha)T_2$ *is also unbiased for* θ.

As MSE is equal to variance for an unbiased estimator, the best estimator (in terms of MSE) in the class of unbiased estimators is the one with minimum variance.

Definition 3 (Uniformly minimum variance unbiased estimator)**.** *An estimator of* θ, $T(\mathbf{X})$, *is called a uniformly minimum variance unbiased estimator (UMVUE) if*

(i) $E_\theta\{T(\mathbf{X})\} = \theta$ *for all* $\theta \in \Theta$, *and*

(ii) for any other estimator $T'(\mathbf{X})$ *with* $E_\theta\{T'(\mathbf{X})\} = \theta$, $\mathrm{var}_\theta\{T(\mathbf{X})\} \le \mathrm{var}_\theta\{T'(\mathbf{X})\}$ *for all* $\theta \in \Theta$.

Remark 9. *In the above definition, it is important that (i) and (ii) satisfy for all* $\theta \in \Theta$. *If (i) and (ii) satisfy for a particular choice of* θ, *say* θ_0, *then the corresponding estimator* $T(\mathbf{X})$ *is called locally minimum variance unbiased estimator (LMVUE).*

1.1 Properties of UMVUE

Understanding the properties of UMVUE is necessary for finding UMVUE in practical applications. The first property of UMVUE says that, an UMVUE must be uncorrelated to any unbiased estimator of zero. One can interpret a random variable having zero expectation as a random noise. Ideally, the best unbiased estimator should be independent of an unbiased estimator of zero, as any part of UMVUE should not be explained by a random noise.

Remark 10 (Role of unbiased estimator of zero)**.** *An unbiased estimator of zero, say* $S(\mathbf{X})$, *is an estimator satisfying* $E_\theta\{S(\mathbf{X})\} = 0$ *for all* $\theta \in \Theta$. *For any unbiased estimator of* θ, *say* $T(\mathbf{X})$, *and an unbiased estimator of zero, say* $S(\mathbf{X})$, *a (uncountable) class of unbiased estimators of* θ *can be obtained by adding a multiple of* $S(\mathbf{X})$ *with* $T(\mathbf{X})$, *i.e.,* $T(\mathbf{X}) + \alpha S(\mathbf{X})$, $\alpha \in \mathbb{R}$. *An UMVUE must have the minimum variance among this class of unbiased estimators, considering all possible* α *and* S.

Theorem 1 (Properties of UMVUE: 1)**.** *Let* $\mathcal{U}$ *be the class of all unbiased estimators of* $\theta \in \Theta$ *with finite variance, i.e.,* $\mathcal{U} = \left\{T = T(\mathbf{X}) : E_\theta(T) = \theta, \text{ and } E_\theta\left(T^2\right) < \infty \text{ for all } \theta \in \Theta\right\}$, *and* $\mathcal{V}_0$ *be the class of unbiased estimators of zero with finite variance, i.e.,* $\mathcal{V}_0 = \left\{S = S(\mathbf{X}) : E_\theta(S) = 0, \text{ and } E_\theta\left(S^2\right) < \infty \text{ for all } \theta \in \Theta\right\}$. *Suppose* $\mathcal{U}$ *is non-empty. Then* $T(\mathbf{X}) \in \mathcal{U}$ *is UMVUE of* $\theta \in \Theta$ *if and only if* $T(\mathbf{X})$ *is uncorrelated with all unbiased estimators of zero.* [Proof]

Remark 11. *The above characterization of UMVUE is of limited application, as the class of unbiased estimators of* zero *is also very large. However, the above theorem is useful in proving that an unbiased estimator of* θ *is not an UMVUE.*

Example 3. Let $X_1, \ldots, X_n$ be a random sample from Poisson(θ) distribution. The estimators $T_1(\mathbf{X}) = \bar{X}_n$, $T_2(\mathbf{X}) = S_n^{\star 2}$ and $T_3(\mathbf{X}) = X_1$ are unbiased for θ. However, we can discard T_3 as a possible candidate of UMVUE as $\mathrm{var}_\theta\left(\bar{X}_n\right) < \mathrm{var}_\theta(X_1)$ for $n > 1$. Now, to check if T_2 can be the UMVUE, consider the covariance of T_2 and $T_1 - T_3$. Verify that the covariance is non-zero. Hence T_2 can not be an UMVUE of θ.

Corollary 1 (Properties of UMVUE: 2)**.** *The UMVUE of θ is unique.* [Proof]

The next theorem provides a way to improve an unbiased estimator of θ using a sufficient statistic.

Theorem 2 (Rao-Blackwell Theorem)**.** *Let $T_1(\mathbf{X})$ be an unbiased estimator of θ and $T(\mathbf{X})$ be a sufficient statistic for θ. Then the conditional expectation $\phi(t) = E(T_1(\mathbf{X}) \mid T(\mathbf{X}) = t)$ defines a statistic $\phi(T)$. This statistic $\phi(T)$ is*

(i) a function of the sufficient statistic $T(\mathbf{X})$ for θ,

(ii) is an unbiased estimator of θ, and

(iii) satisfies $\mathrm{var}_\theta(\phi(T)) \le \mathrm{var}_\theta(T_1(\mathbf{X}))$ for all $\theta \in \Theta$.

Proof of Theorem 2 requires the following result:

Result 1. *Under the existence and finiteness of all the relevant expectations, the following properties of conditional expectation and variance are satisfied:*

(A) $$E_Z\left[E_{Y|Z}\left\{h(Y) \mid Z\right\}\right] = E\left\{h(Y)\right\},$$

(B) $$\mathrm{var}_Z\left[E_{Y|Z}\left\{h(Y) \mid Z\right\}\right] + E_Z\left[\mathrm{var}_{Y|Z}\left\{h(Y) \mid Z\right\}\right] = \mathrm{var}_Y\left\{h(Y)\right\}.$$

[Proof of Theorem 2]

Example 4. Let $X_1, \ldots, X_n$ be n random samples from `Bernoulli`(θ). Then X_1 an unbiased estimator of θ. However, a better estimator can be constructed by considering $T_1(\mathbf{X}) = E(X_1 \mid \sum_{i=1}^n X_i)$ as $\sum_{i=1}^n X_i$ is a sufficient statistic.

Remark 12. *If $T_1(\mathbf{X})$ is solely a function of a the sufficient statistic $T(\mathbf{X})$ then $\mathrm{var}_\theta(\phi(T)) = \mathrm{var}_\theta(T_1(\mathbf{X}))$ for all $\theta \in \Theta$.* [Proof]

Remark 13. *For any statistic $T(\mathbf{X})$, and an unbiased estimator $T_1(\mathbf{X})$, $E\left\{E(T_1(\mathbf{X}) \mid T(\mathbf{X}) = t)\right\} = \theta$, and conclusion (iii) in Theorem 2 holds. However, if $T(\mathbf{X})$ is not sufficient, then $E(T_1(\mathbf{X}) \mid T(\mathbf{X}) = t)$ may not be a statistic. For instance, consider the function $E(n^{-1}\sum_{i=1}^n X_i \mid X_1)$ in the above example. Observe that*

$$E\left(\frac{1}{n}\sum_{i=1}^n X_i \mid X_1 = x_1\right) = \frac{x_1}{n} + \frac{1}{n}\sum_{i=2}^n E(X_i) = \frac{x_1 + (n-1)\theta}{n},$$

which is not an estimator.

Remark 14. *Rao-Blackwell theorem indicates that the UMVUE must be a function of sufficient statistic. If not, then a better estimate can be obtained by considering the conditional expectation given a sufficient statistic.*

Note: Rao-Blackwell theorem provides a way to improve on an existing estimator. Based on the theorem, we can make a general recommendation of selecting an appropriate (unbiased) function of a sufficient statistic as an estimator of θ. However, the class of sufficient statistics is also uncountable (as one-one function of sufficient statistic is also sufficient). Thus, this theorem does not directly indicate the choice of UMVUE.

Remark 15. *It is not, in general, easy to characterize the class of unbiased estimators of zero, consequently, verifying if $\mathrm{cov}(T(\mathbf{X}), U(\mathbf{X})) = 0$, for all unbiased estimator $U(\mathbf{X})$ of* zero *is difficult. However, if the distribution of $T(\mathbf{X})$ is complete and U is a function of T, then it implies that the only (with probability one) unbiased estimator of* zero *is* zero *itself (with probability one). Thus, correlation of any estimator T of θ and an unbiased estimator of* zero*, must be zero. The Lehman-Scheffe theorem formalizes this idea.*

Theorem 3 (Lehman-Scheffe Theorem)**.** *If T is a complete-sufficient statistic, and there exists an unbiased estimator $T_1(\mathbf{X})$ of θ, then the UMVUE of θ is given by $\phi(T) = E(T_1 \mid T)$.* [Proof]

Remark 16. *The above theorem says that if one can obtain an unbiased estimator, $T(\mathbf{X})$, of θ based on a complete-sufficient statistics, then that much be the UMVUE of θ. For instance, in Example 4, $T_1(\mathbf{X})$ is the UMVUE of θ.*

2 Methods of Finding UMVUE

I. **Rao-Blackwellization:** If an unbiased estimator of θ is available, and the distribution of the complete sufficient statistic in known. Then one might find the UMVUE ba taking the conditional expectation the unbiased estimator given the complete sufficient statistic.

Example 7. Let $X_1, \ldots, X_n$ be a random sample from Poisson(λ) distribution. Consider the problem of estimating $P(X=1) = \theta = \lambda \exp\{-\lambda\}$. Observe that $Y_1 = I_{\{1\}}(X_1)$ is an unbiased estimator of θ, and $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is sufficient for λ, and hence for θ. The conditional distribution of Y_1 given T is a two point distribution with probability mass $\phi(t) = t(n-1)^{t-1}/n^t$ for $Y_1 = 1$ and $1-q$ otherwise. Thus the conditional expectation is $\phi(t)$. Thus the statistic $\phi(T) = T(1-1/n)^T/(n-1)$ is the UMVUE for θ.

III. **Method of Solving:** Let $T(\mathbf{X})$ be a complete sufficient statistic. As it is known that any function of T which is an unbiased estimator of θ is the UMVUE of θ, one can obtain the UMVUE directly by solving for the function $g(T)$ such that $E_\theta(g(T)) = \theta$.

Example 8. In the same problem of estimating $P(X=1) = \theta = \lambda \exp\{-\lambda\}$, we can apply the direct solving method as follows:

$$E_\lambda \{g(T)\} = \theta \quad \Leftrightarrow \quad g(0)+g(1)n\lambda+\cdots+g(t)\frac{n^t\lambda^t}{t!}+\cdots = \lambda+(n-1)\lambda^2+\cdots+\frac{(n-1)^{t-1}\lambda^t}{t!}+\cdots,$$

which arrives at the same choice of UMVUE.

Remark 17. *From Lehman-Scheffe theorem, it is intuitive that if T is a complete sufficient statistic, and T' is any other sufficient statistic, then $\phi(T')$ (ϕ as described in Rao-Blackwell theorem) must be a function of $\phi(T)$, which in turn implies that T must be minimal sufficient. We conclude this section with the proof of Theorem 6, stated in Module 2.*

We now prove the theorem stated in the last module.

Theorem 4. *A complete-sufficient statistic is minimal sufficient.* [Proof]

3 Variance Inequalities

In many situations it is not possible to obtain a complete sufficient statistic, or it is difficult to verify if a given estimator is the UMVUE. Further, in many situations an UMVUE does not exist. In such cases, if one can obtain a tight (achievable) lower-bound for the class of variances of all the unbiased estimators, then it would be possible evaluate the performance of any unbiased estimator.

The Rao-Cramer lower bound serves this purpose.

Theorem 5 (Cramer-Rao Lowerbound, CRLB)**.** *Let $X_1, \cdots, X_n$ be a sample with pdf $f_{\mathbf{X}}(\cdot;\theta)$, $\theta \in \Theta$, satisfying the following regularity conditions:*

(i) Θ is an open interval in $\mathbb{R}$, and the support $S_{\mathbf{X}}$ does not depend on θ.

(ii) For each $\mathbf{x} \in S_{\mathbf{X}}$ and $\theta \in \Theta$, the derivative $\partial \log f_{\mathbf{X}}(\mathbf{x};\theta)/\partial\theta$ exists and is finite.

(iii) For any statistic $S(\mathbf{X})$ with $E(|S(\mathbf{X})|) < \infty$ for all θ, we have

$$\frac{\partial}{\partial\theta} E_\theta\{S(\mathbf{X})\} = \int S(\mathbf{x}) \frac{\partial}{\partial\theta} f_{\mathbf{X}}(\mathbf{x};\theta) d\mathbf{x}.$$

Let $T(\mathbf{X})$ be a statistic satisfying $\text{var}_\theta\{T(\mathbf{X})\} < \infty$.

Define $E_\theta\{T(\mathbf{X})\} = \psi(\theta)$, $\psi'(\theta) = \dfrac{\partial}{\partial\theta}\psi(\theta)$ and $I(\theta) = E_\theta\left[\left\{\dfrac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x};\theta)\right\}^2\right]$.

If $0 < I(\theta) < \infty$, then $T(\mathbf{X})$ satisfies

$$\text{var}_\theta\{T(\mathbf{X})\} \geq \frac{[\psi'(\theta)]^2}{I(\theta)}. \tag{1}$$

[Proof of Theorem 5]

Remark 18. *From the proof of Theorem 5 it follows that equality holds in (1) if* $T(\mathbf{X}) - \psi(\theta) = k(\theta)\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)$ *with probability* one. *Integrating both sides with respect to* θ, *we observe that equality holds for one-parameter exponential family with* $T(\mathbf{X})$ *being a sufficient statistic.*

Remark 19. *When* $\psi(\theta) = \theta$, *then the CRLB reduces to* $var_\theta\{T(\mathbf{X})\} \geq [I(\theta)]^{-1}$.

Remark 20. *When* $X_1, \cdots, X_n$ *are i.i.d., then* $I(\theta) = nI_1(\theta)$, *where* $I_1(\theta) = E_\theta\left[\frac{\partial}{\partial\theta}\log f_{X_1}(x;\theta)\right]^2$.

Remark 21. *The quantity* $I(\theta)$ *is called the* Fisher information *of the sample* $X_1, \cdots, X_n$. *As* n *increases,* $I(\theta)$ *increases and consequently the variance of the UMVUE decreases. Thus the estimator becomes more concentrated around* θ, *i.e., has more information about* θ. *To understand the intuition behind the term* information *further, read this.*

Remark 22. *If* $f_{\mathbf{X}}(\mathbf{x};\theta)$ *satisfies* $\frac{\partial}{\partial\theta}E_\theta\left[\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right] = \int \frac{\partial}{\partial\theta}\left[\left\{\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{x};\theta)\right\} f_{\mathbf{X}}(\mathbf{x};\theta)\right] dx$, *then*

$$E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right\}^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f_{\mathbf{X}}(\mathbf{X};\theta)\right]$$

Remark 23. *CRLB provides another way to verify if an unbiased estimator of* θ *is UMVUE or not. We can simply compute the variance of the given unbiased estimator, and check if the variance matches the CRLB. However, one must be cautious about applying this method. The regularity conditions (i)-(iii) must be satisfied by the underlying class of distributions, and the estimators under consideration. For example,* `Uniform`$(0, \theta)$, *location exponential distributions do not satisfy criteria (i).*

Remark 24. *Even when the family of distributions satisfies the regularity conditions, CRLB may not be achieved by the UMVUE of* θ. *The following is one such example.*

Example 9. Let $X_1, \cdots, X_n$ be a random sample from `Normal`(μ, σ^2). It is not difficult to see that the CRLB is $2\sigma^4/n$, and the variance of the UMVUE $S_n^{\star 2}$ (being an unbiased estimator based on a complete sufficient statistic) is $2\sigma^4/(n-1)$.

Example 10. Let $X_1, \ldots, X_n$ be a random sample from a location family of distributions, i.e., $X_i = \theta + W_i$, $i = 1, \ldots, n$, where W_is are i.i.d. from a distribution with p.d.f. f_W, free of θ. Then $I(\theta) = E\left[\{f_W'(W)/f_W(W)\}^2\right]$ is free of θ.

We conclude this section with some definitions.

Definition 4 (Efficiency of two estimators)**.** *Let* T_1 *and* T_2 *be two unbiased estimators for the parameter* $\psi(\theta)$. *Suppose that* $E_\theta\left(T_i^2\right) < \infty$, *for* $i = 1, 2$. *Then the efficiency of* T_1 *relative to* T_2 *is defined as*

$$\text{eff}_\theta(T_1 \mid T_2) = \frac{\text{var}_\theta(T_2)}{\text{var}_\theta(T_1)}.$$

We say that T_1 is more efficient than T_2 if $\text{eff}_\theta(T_1 \mid T_2) > 1$.

The following definition assumes that the underlying family of distributions satisfy the regularity conditions stated in Theorem 5.

Definition 5 (Efficiency of an Estimator)**.** *The ratio of the CRLB to the actual variance of an unbiased estimator* $T(\mathbf{X})$ *is called efficiency of* T.

Definition 6 (Efficient Estimator)**.** *An unbiased estimator of* θ, $T(\mathbf{X})$ *is said to be* efficient *or* most efficient *if variance of* T *attains the CRLB.*