4 Lecture 4: Classical Estimation Theory

4.1 Cramer-Rao Inequality

Obtaining a point estimator of the parameter of interest is usually the first step in inference. Suppose $\mathbf{X} = (X_1, X_2, .., X_n)$ are i.i.d. from $f(x, \theta), \theta \in R$ and we use a statistic $T_n(\mathbf{X})$ to estimate θ. If $E_\theta(T_n) = \theta + b_n(\theta)$ then the quantity $b_n(\theta)$ is called *bias*. Note that it generally may depend on both θ and the sample size although this dependence may sometimes be suppressed in the notation. We would hope for a zero bias for all θ and n, called *unbiasedness*. When used repeatedly, an unbiased estimator, in the long run, will estimate the true value on average.

Caution: Note, however , that for some families an unbiased estimators may not exist or, even when they exist, may not be very useful. For example, in the case of the geometric distribution $f(x, \theta) = \theta(1-\theta)^{x-1}, x = 1, 2, \ldots$ an unbiased estimator of θ, say, $T(x)$ must satisfy $\sum_{x=1}^{\infty} T(x)\theta(1-\theta)^{x-1} = \theta$ for all $\theta \in [0, 1]$. By a polynomial expansion, the only estimator satisfying this requirement would be $T(1) = 1, T(x) = 0$ if $x \geq 2$. Cancelling θ on both sides, and setting $\tilde{\theta} = 1 - \theta$ we get

$$\sum_{x=1}^{\infty} T(x)\tilde{\theta}^{x-1} = 1 \text{ for all } \tilde{\theta} \in [0, 1]$$

Hence, the only estimator satisfying this requirement would be $T(1) = 1, T(x) = 0$ if $x \geq 2$. Having in mind the interpretation of θ (probability of success in a single trial), such an estimator is neither very reliable, nor very useful.

As an exercise, show that the MLE is $\hat{\theta} = 1/x$. It is biased but makes much more sense!

When looking for an estimator of a "good" quality, we are inclined to analyse the *mean squared error*

$$MSE_\theta(T_n) = E_\theta(T_n - \theta)^2 = Var_\theta T_n + (b_n(\theta))^2.$$

Remark: The following property holds:

$$MSE_\theta(T_n) = Var_\theta T_n + (b_n(\theta))^2$$

Indeed,

$$MSE_\theta(T_n) = E_\theta[(T_n - E_\theta T_n + E_\theta T_n - \theta]^2 =$$

$$E_\theta(T_n - E_\theta T_n)^2 - 2E_\theta[(T_n - E_\theta T_n)(\theta - E_\theta T_n)] + E_\theta(E_\theta T_n - \theta)^2 = Var(T_n) + (b_n(\theta))^2.$$

A small mean squared error as a criterion for choosing a point estimator, is in general more important than unbiasedness. To perform optimally, we would try to find an estimator that minimizes the MSE. Unfortunately, in the class of *all* estimators, an estimator that minimizes the MSE simultaneously *for all θ values*, does not exist!

(Indeed, take *any* estimator $\tilde{\theta}$. Since the parameter $\theta \in \Theta$ is unknown, there will be certain value $\theta_0 \in \Theta$ for which $MSE_{\theta_0}(\tilde{\theta}) > 0$. Then we can consider as a competitor to $\tilde{\theta}$ the estimator $\theta^* \equiv \theta_0$. Note that θ^* is not a very reasonable estimator (it does *not* even use the data (!)) *but* for the *particular* point θ_0 we have $MSE_{\theta_0}(\theta^*) = 0$ and hence, $MSE_{\theta_0}(\tilde{\theta}) > MSE_{\theta_0}(\theta^*)$.)

With other words, when considering the class of *all* estimators, there are so many estimators available to us that to find a single one that is *uniformly* better with respect to the MSE criterion, is just impossible. Way out of this situation is either to restrict the class of estimators considered, or to change the evaluation criterion. We shall be dealing with the first way out right now (the other way was discussed *already* in the Decision theory chapter: Bayes and minimax estimation).

We choose to impose the criterion of unbiasedness. This greatly simplifies the task of minimizing the mean squared error because then, we only have to minimize the variance. In the (smaller subset of unbiased estimators) one can very often find an estimator with the smallest MSE(=Var) *for all θ values*. It is called the uniformly minimum variance unbiased estimator (UMVUE). Let us first look at a well-known result that will help us in our search of the UMVUE (the **Cramer-Rao** theorem).

Theorem 4.1. *Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ have a distribution that depends on θ and $L(\mathbf{X}, \theta)$ be the joint density. Let $\tau(\theta)$ be a smooth (i.e. differentiable) function of θ that has to be estimated. Consider any unbiased estimator $W(\mathbf{X})$ of $\tau(\theta)$, i.e. $E_\theta W(\mathbf{X}) = \tau(\theta)$. Suppose, in addition, that $L(\mathbf{X}, \theta)$ satisfies:*

$$\frac{\partial}{\partial\theta} \int .. \int h(\mathbf{X}) L(\mathbf{X}, \theta) dX_1..dX_n = \int .. \int h(\mathbf{X}) \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta) dX_1..dX_n \qquad (*)$$

for any function $h(\mathbf{X})$ with $E_\theta |h(\mathbf{X})| < \infty$. Then:

$$Var_\theta(W(\mathbf{X})) \geq \frac{(\frac{\partial}{\partial\theta}\tau(\theta))^2}{I_{\mathbf{X}}(\theta)}$$

for all θ holds.

Proof: The proof is elegantly simple and is a clever application of the Cauchy- Schwartz inequality. In the setting of random variables, this Inequality is equivalent to the following statement:

Cauchy-Schwartz Inequality: If Z and Y are two random variables with finite variances $Var(Z)$ and $Var(Y)$ then

$$[Cov(Z, Y)]^2 = \{E[(Z - E(Z))(Y - E(Y))]\}^2 \leq Var(Z)Var(Y)$$

holds.

To prove the Cramer-Rao Theorem, we choose W to be the Z-variable, and the score V to be the Y-variable in the Cauchy-Schwartz Inequality. Since $E_\theta V(\mathbf{X}, \theta) = 0$ holds for the score, we have that

$$[Cov_\theta(W, V)]^2 = [E_\theta(WV)]^2 \leq Var_\theta(W)Var_\theta V. \qquad (6)$$

Substituting the definition of the score, we get:

$$Cov_\theta(W, V) = E_\theta(WV) = \int ... \int W(\mathbf{X}) \frac{\frac{\partial}{\partial\theta} L(\mathbf{X}, \theta)}{L(\mathbf{X}, \theta)} L(\mathbf{X}, \theta) d\mathbf{X}$$

where $d\mathbf{X} = dX_1 dX_2 \dots dX_n$ is used as shorthand notation. Now if we utilise condition (*), we can continue to get:

$$Cov_\theta(W, V) = \frac{\partial}{\partial \theta} E_\theta W = \frac{\partial}{\partial \theta} \tau(\theta).$$

Then, Inequality (6) implies:

$$Var_\theta(W) \geq \frac{(\frac{\partial}{\partial \theta}\tau(\theta))^2}{I_{\mathbf{X}}(\theta)}. \tag{7}$$

Also, in lectures, we will discuss the *multivariate version of this inequality*, applicable for the case of estimating a multidimensional parameter.

4.1.1. Note: The Cramer-Rao (CR) Inequality was stated for continuous random variables. By an obvious modification of condition (*) requiring the ability to interchange differentiation and summation (instead of differentiation and integration) one can formulate this for discrete random variables, too. Note that in this case, even though $L(\mathbf{X}, \theta)$ may not be differentiable w.r. to x, it has to be assumed to be differentiable w.r. to θ.

4.1.1 Corollary for i.i.d. case.

If $\mathbf{X} = (X_1, X_2, \dots, X_n)$ are i.i.d. with $f(x, \theta)$ then $L(\mathbf{X}, \theta) = \prod_{i=1}^n f(X_i, \theta); I_{\mathbf{X}}(\theta) = nI_{X_1}(\theta)$ and the CR Inequality becomes:

$$\text{Var}_\theta(W(\mathbf{X})) \geq \frac{(\frac{\partial}{\partial \theta}\tau(\theta))^2}{nI_{X_1}(\theta)}$$

4.2 Comments on applying the CR Inequality in the search of the UMVUE

a) In case there exists an unbiased estimator of $\tau(\theta)$ whose variance is equal to the lower bound given by CR Inequality, this will be the UMVUE of $\tau(\theta)$. Such a situation occurs often in the case of observations that come from an exponential family;

b) Let us note the drawback related to the fact that condition (*) in the CR Theorem is a strong one and it often happens that it is not satisfied. A typical situation is when the range of the random variables $X_i, i = 1, 2, .., n$ depends on θ, for example, in the case of a random sample from uniform $[0, \theta)$ observations. According to the general Leibnitz$'$ rule for differentiation of parameter-dependent integrals:

$$\frac{\partial}{\partial \theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) dx = f(b(\theta), \theta) \frac{\partial}{\partial \theta} b(\theta) - f(a(\theta), \theta) \frac{\partial}{\partial \theta} a(\theta) + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x, \theta) dx$$

holds and we see that, if a and b were genuine functions of θ, on the RHS there would be some additional non-zero terms included and condition (*) would not hold.

The following illustrative example will be discussed in detail at the lecture: Assume that the right-hand limit θ of the interval $[0, \theta)$ is to be estimated and n i.i.d. observations from the uniform in $[0, \theta)$ distribution are given. It is easy to show that the density of $Y = X_{(n)}$ is given by

$$f_Y(y, \theta) = \begin{cases} ny^{n-1}/\theta^n, & \text{if } 0 < y < \theta \\ 0 \text{ else} \end{cases}$$

Then we can calculate easily that $E[\frac{n+1}{n}X_{(n)}] = \theta$ holds, that is, $\frac{n+1}{n}X_{(n)}$ is an unbiased estimator of $\tau(\theta) = \theta$. It is also easy to calculate its variance: $Var(\frac{n+1}{n}X_{(n)}) = \frac{1}{n(n+2)}\theta^2$ holds. The latter value is **less** than the value $\frac{\theta^2}{n}$ we would get if we recklessly calculated the CR bound ignoring the fact that the regularity condition (*) is in fact **violated** in this example (since the support of the density depends on the unknown parameter).

c) Even in cases where the CR Theorem is applicable, there is no guarantee that the lower bound on the variance is attainable. In fact, looking at the proof of the theorem, we can see that the bound is achievable iff one has an equality in the Cauchy-Schwartz Inequality which means that the score $V(\mathbf{X}, \theta)$ must have a representation of the form $V(\mathbf{X}, \theta) = k_n(\theta)[W(\mathbf{X}) - \tau(\theta)]$. If, alternatively, $V(\mathbf{X}, \theta)$ can not be written in this form then no unbiased estimator of $\tau(\theta)$ would have a variance equal to the one given by the CR bound and in this case the CR Inequality would be of no use when searching for UMVUE. There still might be an UMVUE (with a variance slightly higher than the value given by the CR lower bound) but one would need to develop a method to find it in such situations.

4.3 Examples

(to be discussed at lectures)

a) Estimating the parameter θ in a Poisson(θ) distribution. In this case the CR bound is achievable and the unbiased estimator that achieves it, is $\hat{\theta} = \bar{X}$.

b) Estimating the function $\tau(\theta) = \exp(-\theta)$ from a sample of Poisson(θ) distribution. In this case no unbiased estimator of $\tau(\theta)$ has variance equal to the bound. Nevertheless, UMVUE of $\tau(\theta)$ exists.

Some details:

$$\log L(X; \theta) = \log \left\{ \frac{\theta^{\sum_{i=1}^n x_i} e^{(-n\theta)}}{\prod_{i=1}^n x_i!} \right\} = -n\theta + (\sum_{i=1}^n x_i) \log \theta - \sum_{i=1}^n \log (x_i!) \tag{8}$$

Taking derivatives with respect to θ in (3):

$$\frac{\partial}{\partial \theta} \log L(X; \theta) = -n + \sum_{i=1}^n x_i/\theta = n \exp(\theta)[\frac{1}{\theta} \exp(-\theta)\bar{x} - \exp(-\theta)]$$

and this can not be represented as $k(\theta, n)[\text{ statistic } - \exp(-\theta)]$ Formal calculation of the Cramer-Rao bound gives $\frac{\theta}{n}e^{(-2\theta)}$ (check (!)) but this bound is not attainable by any unbiased estimator of $\tau(\theta) = \exp(-\theta)$.

Nevertheless UMVUE does exist and is given (as we shall see later) by $T = (1 - \frac{1}{n})^{n\bar{X}}$. Another method needs to be applied to finding this UMVUE (see Theorem of Lehmann-Scheffe below).

4.4 Which are the estimators that could attain the bound?

Theorem 4.2. *If under the regularity conditions of CR Theorem there is an estimator of $\tau(\theta)$ which attains the lower bound, it should be the MLE of $\tau(\theta)$.*

Proof: At lectures.

Conclusion: When looking for UMVUE, it is a good idea to calculate the MLE first. If the MLE turns out to be unbiased and its variance equals the one given by the CR bound, the UMVUE has been constructed. Otherwise if either the MLE is biased or does not attain the bound then it is sure that the bound is not attainable at all. Very often in such situations the UMVUE (which necessarily will have variance larger than the one given by the bound) turns out to be a bias-corrected MLE.

To outline a more specific way to construct UMVUE in such more delicate situations, let us first formulate the following famous theorem:

4.5 Rao-Blackwell Theorem

Theorem 4.3. *Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\hat{\tau}(T) = E(W \mid T)$. Then $E_\theta \hat{\tau}(T) = \tau(\theta)$ and $Var_\theta \hat{\tau}(T) \leq Var_\theta W$ for all $\theta \in \Theta$, i.e. $\hat{\tau}(T)$ is uniformly better than W as an estimator of $\tau(\theta)$.*

Proof: at lecture.

Note: $\hat{\tau}(T)$ is a function of a sufficient statistic and has uniformly smaller variance than W. This theorem underlines the importance of sufficient statistics. We can therefore only consider functions of sufficient statistics when looking for UMVUE.

4.6 Uniqueness of UMVUE

Theorem 4.4. *If an estimator W is UMVUE for $\tau(\theta)$, then W is unique. Moreover, W is UMVUE iff W is uncorrelated with all unbiased estimators of zero.*

Proof: at lecture.

Note: Characterization of the estimators that are uncorrelated with any unbiased estimator of zero is therefore important. If it turned out that the family $f(T, \theta)$ of distributions *does have* the property that there are *no unbiased estimators of zero* (except the constant zero itself) then our search of UMVUE will be successfully finalized. For such type of characterization , we need the notion of **completeness.**

4.7 Completeness of a family of distributions

Definition 4. Let $\tilde{f}(t,\theta), \theta \in \Theta$ be a family of distributions for a statistic $T(\mathbf{X})$. The family is called *complete* if $E_\theta g(T) = 0$ for all $\theta \in \Theta$ implies $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$. Equivalently, $T(X)$ is called a *complete statistic* for θ.

Note that completeness is a property of the *whole* family of distributions, *not* a property of a particular distribution.

Now we finally can formulate the theorem that allows us to find the UMVUE even in situations when the CR bound is not achievable.

4.8 Theorem of Lehmann-Scheffe

Theorem 4.5. *Let T be a complete sufficient statistic for a parameter θ and W be any unbiased estimator of $\tau(\theta)$. Then $\hat{\tau}(T) = E(W \mid T)$ is the unique UMVUE of $\tau(\theta)$.*

Proof: This theorem is a direct consequence and compilation of the statements we already made. Indeed, according to the Rao-Blackwell Theorem, it suffices to consider as candidates for UMVUE only unbiased functions of the *sufficient* statistic T (otherwise by conditioning T we can improve any other unbiased competitor) and (again due to Rao-Blackwell) $\hat{\tau}(T)$ is such an unbiased estimator of $\tau(\theta)$ (and, of course $\hat{\tau}(T) = E(W \mid T)$ is a function of T.) Because of the completeness of T, no further improvement of $\hat{\tau}(T)$ is possible-hence it is the *unique* UMVUE.

The theorem has very useful applications. In many situations there will be no obvious candidate for the UMVUE of $\tau(\theta)$. But the theorem suggests that if we have found *any* (even if very poor) unbiased estimator of $\tau(\theta)$ and we know a statistic T that is a *complete and sufficient* statistic for θ then $E(W \mid T)$ is the *uniformly best unbiased estimator* of $\tau(\theta)$.

4.9 Examples-

see lectures.

- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. $N(0,\theta)$, $T = \bar{X}$ is *not complete* for θ (note that θ denotes the variance in this example).
- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. Bernoulli with $\theta \in (0,1)$ denoting the probability of success, $T = \sum_{i=1}^n X_i$ is complete for θ and $\bar{X}$ is UMVUE for the expected value parameter θ. Hovewer, for the variance parameter $\theta(1-\theta)$, the UMVUE turns out to be $\bar{X}(1-\bar{X})\frac{n}{n-1}$.

- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. uniform in $[0, \theta)$, the statistic $T = X_{(n)}$ is complete and $\frac{n+1}{n} X_{(n)}$ is UMVUE for θ.
- For $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ i.i.d. Poisson (θ), the statistic $\sum_{i=1}^{n} X_i$ is complete. For $\tau(\theta) = \exp(-\theta)$, the UMVUE turns out to be $(1 - \frac{1}{n})^{n\bar{X}}$.