4 Best unbiased estimators

1. Notation. In this section we suppose that we have a random sample of size n with sample random variables $X_1, X_2, ..., X_n$ from a population with distribution function $F(x, \theta)$ with associated p.d.f./p.f. $f(x, \theta)$, where $\theta \in \Theta$.

2. Definition. An unbiased estimator $\widehat{g}(\mathbf{X})$ of the real-valued function $g(\theta)$ is a minimum variance unbiased estimator (**MVUE**) if, for any other unbiased estimator $\widetilde{g}(\mathbf{X})$, $\text{var}(\widehat{g}(\mathbf{X})) \leq \text{var}(\widetilde{g}(\mathbf{X}))$ for all $\theta \in \Theta$.

3. Theorem (Rao-Blackwell). Suppose that $\widetilde{g}(\mathbf{X})$ is an unbiased estimator of the real-valued function $g(\theta)$ and that T is a sufficient statistic for θ. Then the conditional expectation, $\text{E}(\widetilde{g}(\mathbf{X})|T) = \widehat{g}(T)$, is an unbiased estimator of $g(\theta)$ and $\text{var}(\widehat{g}) \leq \text{var}(\widetilde{g})$ for all $\theta \in \Theta$.

4. Example. Let $X_1, X_2..., X_n$ be i.i.d. Poisson(θ).

5. Note. It follows that, conditioning any unbiased estimator on a sufficient statistic will result in an improvement. So, in looking for a best unbiased estimator of $g(\theta)$, we need to consider only estimators based on a sufficient statistic. If an estimator is not a function of a sufficient statistic, then there is another estimator which is a function of a sufficient statistic and which is at least as good, in the sense of having smaller variance.

6. Complete statistics. Suppose that $\widehat{g} = \widehat{g}(\mathbf{X})$ is an unbiased estimator which is based on a sufficient statistic T. How do we know that $\widehat{g}$ is best unbiased? If the probability distribution of T has a property called completeness, then $\widehat{g}$ is the unique MVUE of $g(\theta)$.

Definition: Let $\{\psi(t, \theta) : \theta \in \Theta\}$ be a family of p.d.f./p.f. for a statistic T. T is called a **complete** statistic if $\text{E}(\phi(T)) = 0$ for all $\theta \in \Theta$ implies that $\phi(T) = 0$ (a.s.).

7. Theorem. (*Completeness theorem for an exponential family*) Let $X_1, X_2..., X_n$ be i.i.d. observations from a p.d.f./p.f. $f(x, \theta)$ that belongs to the exponential family, that is

$$f(x, \theta) = \exp\{\textstyle\sum_{i=1}^{k} w(\theta_i)K_i(x) + S(x) + q(\theta)\}$$

for all $\{x : f(x, \theta) \neq 0\}$, where $\theta = (\theta_1, ..., \theta_k)$ and the set $\{x : f(x, \theta) \neq 0\}$ does not depend on θ. Suppose also that the parameter space contains a k dimensional open rectangle. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} K_1(X_j), \sum_{j=1}^{n} K_2(X_j), \ldots, \sum_{j=1}^{n} K_m(X_j) \right)$$

is a complete sufficient statistic for θ.

8. Theorem. Let T be a complete sufficient statistic for a parameter θ, and let $\hat{g}(T)$ be any estimator based only on T. Then $\hat{g}(T)$ is the unique MVUE of its expected value.

9. Remark. Given that there is a complete sufficient statistic T for θ, the search for a **MVUE** for $g(\theta)$ requires us to find either an unbiased estimator $\widehat{g}(T)$ or to find any unbiased estimator $\widetilde{g}(\mathbf{X})$ and then calculate $\text{E}(\widetilde{g}(\mathbf{X})|T) = \widehat{g}(T)$.

10. Example. Let $X_1, X_2..., X_n$ be i.i.d. (a) N($0, \sigma^2$), (b) Bin(N, θ).

11. Notes. (i) If a minimal sufficient statistic exists, any complete sufficient statistic

is also a minimal sufficient statistic. However, it is *not* true that any minimal sufficient statistic is necessarily complete.

(ii) In general, lack of completeness of a minimal sufficient statistic prevents our using Theorem 8 to establish the existence of an **MVUE** of a real parameter. In searching for an optimal estimator, we might ask whether there is a lower bound for the variance of an unbiased estimator. If such a lower bound existed, it would function as a benchmark against which estimators could be compared. If an estimator achieved this lower bound, we would know that it could not be improved upon. In the case of unbiased estimators, the following theorem provides such a lower bound.

12. Theorem (Cramér-Rao). Subject to regularity conditions (see 13), for any unbiased estimator $\widehat{g}(\mathbf{X})$ of the real function $g(\theta)$,

$$\operatorname{var}(\widehat{g}(\mathbf{X})) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

where

$$I(\theta) = \operatorname{E}\left(\left[\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)\right]^2\right).$$

13. Regularity conditions for Cramér-Rao. Written for a continuous distribution, these are:

(i) $\{x : f(x, \theta) \neq 0\}$ is independent of $\theta \in \Theta$ and Θ contains an open interval in $\mathbb{R}^1$;

(ii) $\frac{\partial f(\mathbf{x}, \theta)}{\partial \theta}$ exists for all $\mathbf{x}$ and $\frac{d}{d\theta} \int \ldots \int_{R^n} f(\mathbf{x}, \theta) d\mathbf{x} = \int \ldots \int_{R^n} [\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta)] d\mathbf{x} = 0;$

(iii) $\operatorname{E}([\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)]^2)$ exists for all $\theta \in \Theta$; and,

(iv) $\frac{d}{d\theta} \int \ldots \int_{R^n} \widehat{g}(\mathbf{x}) f(\mathbf{x}, \theta) d\mathbf{x} = \int \ldots \int_{R^n} \widehat{g}(\mathbf{x}) [\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta)] d\mathbf{x},$ where $\widehat{g}(\mathbf{x})$ is an unbiased estimator of the differentiable function $g(\theta)$.

For a discrete distribution just replace the $\int$ signs by $\sum$ signs.

14. Note. Suppose that the regularity conditions are satisfied.

(i) If $V = \frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)$ then $\operatorname{E}(V) = 0$.

(Indeed, $\operatorname{E}(V) = \int \ldots \int_{R^n} [\frac{\partial}{\partial \theta} \log f(\mathbf{x}, \theta)] f(\mathbf{x}, \theta) d\mathbf{x} = \int \ldots \int_{R^n} [\frac{\partial}{\partial \theta} f(\mathbf{x}, \theta)] d\mathbf{x} = 0.)$

(ii) Write $I(\theta) = \operatorname{var}(V) = \operatorname{E}(V^2)$. Then $I(\theta) = n\operatorname{E}([\frac{\partial}{\partial \theta} \log f(X_1, \theta)]^2)$.

(Indeed, $V = \frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta) = \sum_{k=1}^{n} \frac{\partial}{\partial \theta} \log f(X_k, \theta)$ and, as in (i), $\operatorname{E}(\frac{\partial}{\partial \theta} \log f(X_k, \theta)) = 0$ for $k = 1, 2, \ldots, n$. Also, since $X_1, X_2, \ldots, X_n$ are i.i.d., the random variables

$$\frac{\partial}{\partial \theta} \log f(X_1, \theta), \frac{\partial}{\partial \theta} \log f(X_2, \theta) , \ldots, \quad \frac{\partial}{\partial \theta} \log f(X_n, \theta)$$

are also i.i.d. Hence,

$\operatorname{E}([\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)]^2) = \operatorname{var}(\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)) = \sum_{k=1}^{n} \operatorname{var}(\frac{\partial}{\partial \theta} \log f(X_k, \theta)) = n\operatorname{E}([\frac{\partial}{\partial \theta} \log f(X_1, \theta)]^2).)$

15. The Fisher information. The quantity

$$I(\theta) = E([\frac{\partial}{\partial \theta} \log f(\mathbf{X}, \theta)]^2)$$

is called the (total) Fisher information. This terminology reflects the fact that the amount of information gives a bound on the variance of the best unbiased estimator of $g(\theta)$. As

the information gets bigger and we have more information about θ, we have a smaller bound on the variance of the best unbiased estimator. By 14-(ii),

$$I(\theta) = ni(\theta)$$

where

$$i(\theta) = \mathrm{E}([\frac{\partial}{\partial\theta}\log f(X_1,\theta)]^2)$$

is called the one step Fisher information. So, the information in n i.i.d. observations is n-times that in one of the observations.

16. Note. Subject to the regularity conditions,

$$i(\theta) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\log f(X_1,\theta)\right).$$

17. Examples. Let $X_1, X_2..., X_n$ be i.i.d. (a) N(μ, σ^2), where σ is known, (b) Poisson(θ).

18. Note. The Cramér-Rao Theorem establishes the lower bound for the variance of an unbiased estimator. If one finds an estimator which attains the bound then it is a MVUE. But there is no guarantee that the bound is sharp. The value of the Cramér-Rao lower bound may be strictly smaller than the variance of any unbiased estimator. The following theorem shows that only in exceptional circumstances can the lower bound be attained.

19. Theorem (Attainment) Subject to the regularity conditions of the Cramér-Rao Theorem, there exists an unbiased estimator $\hat{g}(\mathbf{X})$ of $g(\theta)$ whose variance attains the Cramér-Rao lower bound if and only if

$$\tfrac{\partial}{\partial\theta}\log f(\mathbf{X},\theta) = a(\theta)(\hat{g}(\mathbf{X}) - g(\theta))$$

for some function $a(\theta)$. Furthermore, in this case, var$(\hat{g}) = g'(\theta)/a(\theta)$.

20. Note. If $X_1, X_2, ..., X_n$ are i.i.d. with the common p.d.f./p.f. $f(x, \theta)$, then

$$\frac{\partial}{\partial\theta}\log f(\mathbf{X}, \theta) = \frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f(X_i, \theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i, \theta).$$

21. Examples. Let $X_1, X_2..., X_n$ be i.i.d. (a) N$(0, \sigma^2)$, (b) Poisson(θ) and $g(\theta) = \mathrm{e}^{-\theta}$.

22. Definition. The efficiency (eff) of an unbiased estimator $\hat{g}(\mathbf{X})$ of $g(\theta)$ is defined to be C-R lower bound $\div$ var$(\hat{g}(\mathbf{X}))$, i.e.,

$$\mathrm{eff}\,(\hat{g}(\mathbf{X})) = \frac{[g'(\theta)]^2}{I(\theta)\mathrm{var}(\hat{g}(\mathbf{X}))}.$$

Naturally, we look for estimators with efficiency close to 1. We say that an unbiased estimator $\hat{g}(\mathbf{X})$ of $g(\theta)$ is efficient if eff$(\hat{g}(\mathbf{X})) = 1$ (that is, if the variance of $\hat{g}(\mathbf{X})$ attains the C-R lower bound). Clearly, if an estimator is efficient, it is a MVUE, but not vice versa. In Example 21 (a), the MVUE of σ^2 is efficient, but in (b), the the MVUE of $\hat{g}$ of $g(\theta) = \mathrm{e}^{-\theta}$ is not efficient. Nevertheless, it can be shown that in Example 21 (b), eff$(\hat{g}) \approx 1$ for large n's.

23. Remarks.

(i) **When equality holds in (19).**

We have $\frac{\partial}{\partial\theta}\log f(\mathbf{X},\theta) = a(\theta)(\widehat{g}(\mathbf{X}) - g(\theta))$. Replacing $\mathbf{X}$ by $\mathbf{x}$ and integrating w.r.t. θ we have

$$\log f(\mathbf{x},\theta) = \widehat{g}(\mathbf{x})\int a(\theta)d\theta - \int a(\theta)g(\theta)d\theta + S(\mathbf{x})$$

or (provided that all the integrals exist and are finite),

$$f(\mathbf{x},\theta) = \exp[\widehat{g}(\mathbf{x})p(\theta) + S(\mathbf{x}) + q(\theta)] \text{ (say)}$$

so that the distribution must be in the exponential class if the C-R gives the **MVUE**.

(ii) **Cramér-Rao and MLE.** If, under the conditions of the Cramér-Rao inequality, $\widehat{\theta}$ is efficient for θ, then $\widehat{\theta}$ is the maximum likelihood estimator of θ.

Now, maximizing $f(\mathbf{x},\theta)$ is equivalent to maximizing $\log f(\mathbf{x},\theta)$ and we have

$$0 = \frac{\partial}{\partial\theta}\log f(\mathbf{X},\theta) = a(\theta)(\widehat{\theta}(\mathbf{X}) - \theta),$$

so that the estimator giving the Cramér-Rao lower bound is the MLE.

24. Asymptotic properties of MLE's. In general, MLE's are neither unbiased nor efficient. The main justification of the method of maximum likelihood in terms of the criterion of minimum variance unbiasedness is that it is possible to show that for large samples, subject to certain regularity conditions, maximum likelihood estimators are nearly unbiased and have variances nearly equal to the C-R lower bound.

In particular, it can be shown that, for large values of n, the MLE $\hat{\theta}$ of θ has asymptotically the normal distribution $N(\theta, \frac{1}{I(\theta)})$.

17