Handout 19: Cramér-Rao Lower Bound

Consider a random variable X with a probability density function $f(\theta; x)$ with one univariate parameter θ. We can define a random variable V, called the **score**, as the derivative of the logarithm of the density of X.[1] We see that this has a nice form by applying the chain rule:

$$\begin{aligned} V &= \frac{\partial}{\partial \theta} \left[\log(f(\theta; X)) \right] \\ &= \frac{1}{f(\theta; X)} \cdot \frac{\partial}{\partial \theta} \left[f(\theta; X) \right]. \end{aligned}$$

[1] The important point is that the score tells us how much the density f changes at a point x with respect to θ. The logarithm is there to make the score measure the relative change rather than the absolute change, which can also be seen through the the application of the chain-rule.

The score measures the sensitivity of the data to the parameter θ. However, because it can be positive or negative, on average it turns out that the score will have an expected value of zero:

$$\begin{aligned} \mathbb{E}V &= \int f(\theta; x) \cdot \frac{1}{f(\theta; x)} \cdot \frac{\partial}{\partial \theta} \left[f(\theta; x) \right] dx \\ &= \int \frac{\partial}{\partial \theta} f(\theta; x) dx = \frac{\partial}{\partial \theta} \int f(\theta; x) dx = \frac{\partial}{\partial \theta} [1] = 0. \end{aligned}$$

This holds for any value of θ.

Because the positive and negative scores cancel each other out, in order to use the score as a measurement of the relationship between the paramter θ and a value of the data X, we need to look at the square of the score. The expected value of this is called the **Fisher information**, commonly denoted by $\mathcal{I}(\theta)$:

$$\mathcal{I}(\theta) = \mathbb{E}[V^2 | \theta] = Var(V^2 | \theta).$$

The Fisher information serves as a measurment of how much information about θ is provided by the data X. The Fisher information can change for different values of θ, but does not depend on the data X, which has been integrated out.

Now, let $T = t(X)$ be an unbiased point estimator for the parameter θ. Let's look at the covariance of T and V, note that this is equal to just $\mathbb{E}[VT]$ since the expected value of V is zero.[2] This has, by construction, a nice form:

[2] Recall that the covariance in general would be $\mathbb{E}[(V - \mathbb{E}V)(T - \mathbb{E}T)]$.

$$\begin{aligned} Cov(V, T) &= \int \left[f(\theta; x) \times t(x) \times \frac{1}{f(\theta; x)} \times \frac{\partial}{\partial \theta} \left[f(\theta; x) \right] \right] dx \\ &= \frac{\partial}{\partial \theta} \left[\int t(x) f(\theta, x) dx \right] = \frac{\partial}{\partial \theta} \mathbb{E}T = 1. \end{aligned}$$

Where the last step comes from the fact that T is unbiased. Next, we need to use the **Cauchy-Schwartz Inequality**, which for probability spaces says that covariance of two random variables is always less in absolute value than the square-root of the product of their variances.[3]

[3] The more general form says that the squared inner product $|\langle u, v \rangle|^2$. is less than $\langle u, u \rangle \cdot \langle v, v \rangle$. Applying this to the integration with density f yields the probabilistic version.

Applying this to T and V shows that:

$$
\begin{aligned}
Var(T) \cdot Var(V) &\geq |Cov(V, T)|^2 \\
Var(T) \cdot \mathcal{I}(\theta) &\geq |1|^2 \\
Var(T) &\geq \frac{1}{\mathcal{I}(\theta)}.
\end{aligned}
$$

So, the variance of T can never be less than the inverse of the Fisher information. This provides a bound on the best that we can hope to do in terms of estimating the parameter θ from the data X. This result is called the **Cramér-Rao** lower bound.

The **efficency** of an unbiased estimator, written $e(\widehat{\theta})$, provides a measurement of how far away the variance of the estimator is away from the Cramèr-Rao bound. Namely, we have:

$$
e(\widehat{\theta}) = \frac{\mathcal{I}(\theta)^{-1}}{Var(\widehat{\theta})}.
$$

We say that an estimator is **efficent** if it has an effiency of 1. Another way to state the Cramér-Rao bound is to simply say that the efficency is never greater than 1.

Under some regularity conditions—in particular, that the logarithm of the density function f is twice-differentiable—the Fisher information can be written in a somewhat simplified form:

$$
\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(\theta; x)\right].
$$

Typically, squaring the log density requires having a number of cross terms, whereas the second derivative removes a number of terms, simplifying the calculation. This is the version that we will use on the worksheet.

It is possible to extend the result above to the case where X and θ are vectors. The extension for a vector X, which includes the important case of a random sample of size n, is fairly trivial. We just replace all of the single integrals above with n-dimensional integrals over $\mathbb{R}^n$. Generalizing to a vector value for θ is a bit more work, requiring some vector calculus that goes beyond the prerequisites for this course. The general idea, however, is very similar.