3 The Rao-Blackwell theorem

Variance is what any two statisticians are at.

3.1 Mean squared error

A good estimator should take values close to the true value of the parameter it is attempting to estimate. If $\hat{\theta}$ is an unbiased estimator of θ then $\mathbb{E}(\hat{\theta} - \theta)^2$ is the variance of $\hat{\theta}$. If $\hat{\theta}$ is a biased estimator of θ then $\mathbb{E}(\hat{\theta} - \theta)^2$ is no longer the variance of $\hat{\theta}$, but it is still useful as a measure of the **mean squared error** (**MSE**) of $\hat{\theta}$.

Example 3.1 Consider the estimators in Example 1.3. Each is unbiased, so its MSE is just its variance.

$$\text{var}(\hat{p}) = \text{var}\left[\frac{1}{n}\left(X_1 + \cdots + X_n\right)\right] = \frac{\text{var}(X_1)\cdots + \text{var}(X_n)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

$$\text{var}(\tilde{p}) = \text{var}\left[\frac{1}{3}(X_1 + 2X_2)\right] = \frac{\text{var}(X_1) + 4\,\text{var}(X_2)}{9} = \frac{5p(1-p)}{9}$$

Not surprisingly, $\text{var}(\hat{p}) < \text{var}(\tilde{p})$. In fact, $\text{var}(\hat{p})/\text{var}(\tilde{p}) \to 0$, as $n \to \infty$.

Note that $\hat{p}$ is the MLE of p. Another possible unbiased estimator would be

$$p^* = \frac{1}{\frac{1}{2}n(n+1)}(X_1 + 2X_2 + \cdots + nX_n)$$

with variance

$$\text{var}(p^*) = \frac{1}{\left[\frac{1}{2}n(n+1)\right]^2}\left(1 + 2^2 + \cdots + n^2\right)p(1-p) = \frac{2(2n+1)}{3n(n+1)}p(1-p).$$

Here $\text{var}(\hat{p})/\text{var}(p^*) \to 3/4$.

The next example shows that neither a MLE or an unbiased estimator necessarily minimizes the mean square error.

Example 3.2 Suppose $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, μ and σ^2 unknown and to be estimated. To find the MLEs we consider

$$\log f(x \mid \mu, \sigma^2) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

This is maximized where $\partial(\log f)/\partial\mu = 0$ and $\partial(\log f)/\partial\sigma^2 = 0$. So

$$(1/\hat{\sigma}^2)\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0, \qquad \text{and} \qquad -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = 0,$$

and the MLEs are

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}S_{XX} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

It is easy to check that $\hat{\mu}$ is unbiased. As regards $\hat{\sigma}^2$ note that

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu + \mu - \bar{X})^2\right] = \mathbb{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n\mathbb{E}(\mu -$$
$$= n\sigma^2 - n(\sigma^2/n) = (n-1)\sigma^2$$

so $\hat{\sigma}^2$ is biased. An unbiased estimator is $s^2 = S_{XX}/(n-1)$.

Let us consider an estimator of the form λS_{XX}. Above we see S_{XX} has $(n-1)\sigma^2$ and later we will see that its variance is $2(n-1)\sigma^4$. So

$$\mathbb{E}\left[\lambda S_{XX} - \sigma^2\right]^2 = \left[2(n-1)\sigma^4 + (n-1)^2\sigma^4\right]\lambda^2 - 2(n-1)\sigma^4\lambda + \sigma^4.$$

This is minimized by $\lambda = 1/(n+1)$. Thus the estimator which minimizes the squared error is $S_{XX}/(n+1)$ and this is neither the MLE nor unbiased. Of c there is little difference between any of these estimators when n is large.

Note that $\mathbb{E}[\hat{\sigma}^2] \to \sigma^2$ as $n \to \infty$. So again the MLE is asymptotically unbia

3.2 The Rao-Blackwell theorem

The following theorem says that if we want an estimator with small MSE w confine our search to estimators which are functions of the sufficient statistic.

Theorem 3.3 (Rao-Blackwell Theorem) *Let $\hat{\theta}$ be an estimator of θ with $\mathbb{E}($* ∞ *for all θ. Suppose that T is sufficient for θ, and let $\theta^* = \mathbb{E}(\hat{\theta} \mid T)$. Then for*

$$\mathbb{E}(\theta^* - \theta)^2 \leq \mathbb{E}(\hat{\theta} - \theta)^2.$$

The inequality is strict unless $\hat{\theta}$ is a function of T.

Proof.

$$\mathbb{E}[\theta^* - \theta]^2$$
$$= \mathbb{E}\left[\mathbb{E}(\hat{\theta} \mid T) - \theta\right]^2 = \mathbb{E}\left[\mathbb{E}(\hat{\theta} - \theta \mid T)\right]^2 \leq \mathbb{E}\left[\mathbb{E}((\hat{\theta} - \theta)^2 \mid T)\right] = \mathbb{E}(\hat{\theta} - \theta$$

The outer expectation is being taken with respect to T. The inequality follows the fact that for any RV, W, $\text{var}(W) = \mathbb{E}W^2 - (\mathbb{E}W)^2 \geq 0$. We put $W = (\hat{\theta} - $ and note that there is equality only if $\text{var}(W) = 0$, i.e., $\hat{\theta} - \theta$ can take just one for each value of T, or in other words, $\hat{\theta}$ is a function of T. ■

Note that if $\hat{\theta}$ is unbiased then θ^* is also unbiased, since

$$\mathbb{E}\theta^* = \mathbb{E}\left[\mathbb{E}(\hat{\theta} \mid T)\right] = \mathbb{E}\hat{\theta} = \theta \, .$$

We now have a quantitative rationale for basing estimators on sufficient statistics: if an estimator is not a function of a sufficient statistic, then there is another estimator which is a function of the sufficient statistic and which is at least as good, in the sense of mean squared error of estimation.

Examples 3.4

(a) $X_1, \ldots, X_n \sim P(\lambda)$, λ *to be estimated.*
In Example 2.3 (a) we saw that a sufficient statistic is $\sum_i x_i$. Suppose we start with the unbiased estimator $\tilde{\lambda} = X_1$. Then 'Rao–Blackwellization' gives

$$\lambda^* = \mathbb{E}\left[X_1 \mid \sum_i X_i = t\right] \, .$$

But

$$\sum_i \mathbb{E}\left[X_i \mid \sum_i X_i = t\right] = \mathbb{E}\left[\sum_i X_i \mid \sum_i X_i = t\right] = t \, .$$

By the fact that $X_1, \ldots, X_n$ are IID, every term within the sum on the l.h.s. must be the same, and hence equal to t/n. Thus we recover the estimator $\lambda^* = \hat{\lambda} = \bar{X}$.

(b) $X_1, \ldots, X_n \sim P(\lambda)$, $\theta = e^{-\lambda}$ *to be estimated.*
Now $\theta = \mathbb{P}(X_1 = 0)$. So a simple unbiased estimator is $\hat{\theta} = 1\{X_1 = 0\}$. Then

$$\theta^* = \mathbb{E}\left[1\{X_1 = 0\} \,\Big|\, \sum_{i=1}^n X_i = t\right] = \mathbb{P}\left(X_1 = 0 \,\Big|\, \sum_{i=1}^n X_i = t\right)$$

$$= \mathbb{P}\left(X_1 = 0; \sum_{i=2}^n X_i = t\right) \Big/ \mathbb{P}\left(\sum_{i=1}^n X_i = t\right)$$

$$= e^{-\lambda}\frac{((n-1)\lambda)^t e^{-(n-1)\lambda}}{t!} \Big/ \frac{(n\lambda)^t e^{-n\lambda}}{t!} = \left(\frac{n-1}{n}\right)^t$$

Since $\hat{\theta}$ is unbiased, so is θ^*. As it should be, θ^* is only a function of t. If you do Rao-Blackwellization and you do not get just a function of t then you have made a mistake.

(c) $X_1, \ldots, X_n \sim U[0, \theta]$, θ *to be estimated.*
In Example 2.3 (c) we saw that a sufficient statistic is $\max_i x_i$. Suppose we start with the unbiased estimator $\tilde{\theta} = 2X_1$. Rao–Blackwellization gives

$$\theta^* = \mathbb{E}\left[2X_1 \mid \max_i X_i = t\right] = 2\left(\frac{1}{n}t + \frac{n-1}{n}(t/2)\right) = \frac{n+1}{n}t \, .$$

This is an unbiased estimator of θ. In the above calculation we use the idea that $X_1 = \max_i X_i$ with probability $1/n$, and if X_1 is not the maximum then its expected value is half the maximum. Note that the MLE $\hat{\theta} = \max_i X_i$ is biased.

3.3 Consistency and asymptotic efficiency*

Two further properties of maximum likelihood estimators are consistency and as- totic efficiency. Suppose $\hat{\theta}$ is the MLE of θ.
To say that $\hat{\theta}$ is **consistent** means that

$$\mathbb{P}(|\hat{\theta} - \theta| > \epsilon) \to 0 \quad \text{as } n \to \infty \, .$$

In Example 3.1 this is just the weak law of large numbers:

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - p\right| > \epsilon\right) \to 0 \, .$$

It can be shown that $\mathrm{var}(\tilde{\theta}) \geq 1/nI(\theta)$ for any unbiased estimate $\tilde{\theta}$, where $1/$ is called the *Cramer-Rao lower bound*. To say that $\hat{\theta}$ is **asymptotically effi** means that

$$\lim_{n\to\infty} \mathrm{var}(\hat{\theta})/[1/nI(\theta)] = 1 \, .$$

The MLE is asymptotically efficient and so asymptotically of minimum varian

3.4 Maximum likelihood and decision-making

We have seen that the MLE is a function of the sufficient statistic, asymptot unbiased, consistent and asymptotically efficient. These are nice properties. consider the following example.

Example 3.5 You and a friend have agreed to meet sometime just after 12 You have arrived at noon, have waited 5 minutes and your friend has not s up. You believe that either your friend will arrive at X minutes past 12, wher believe X is exponentially distributed with an unknown parameter λ, $\lambda > 0$, or she has completely forgotten and will not show up at all. We can associate the event with the parameter value $\lambda = 0$. Then

$$\mathbb{P}(\text{data} \mid \lambda) = \mathbb{P}(\text{you wait at least 5 minutes} \mid \lambda) = \int_5^\infty \lambda e^{-\lambda t}\, dt = e^{-5\lambda} \, .$$

Thus the maximum likelihood estimator for λ is $\hat{\lambda} = 0$. If you base your decisi to whether or not you should wait a bit longer only upon the maximum likel estimator of λ, then you will estimate that your friend will never arrive and c not to wait. This argument holds even if you have only waited 1 second.

The above analysis is unsatisfactory because we have not modelled the co either waiting in vain, or deciding not to wait but then having the friend turn