Review Session 6 – Point Estimation

References/suggested reading

(i) Casella & Berger: section 6.2, chapter 7.

1 Introduction

In point estimation, we consider a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\theta)$ from a population/distribution with pmf/pdf $f(x|\theta)$. We seek a method of finding a good estimator of the unknown parameter θ. Previously, we've seen two examples where θ is the the mean $\mathbb{E}_{X\sim f(x|\theta)}[X]$ or variance $\mathrm{Var}_{X\sim f(x|\theta)}(X)$ of the population. By *estimator*, we simply mean a statistic, or some function $W(X_1, \ldots, X_n)$ of the sample. For example, we might take $W(X_1, \ldots, X_n) = \overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$, the sample mean, or $W(X_1, \ldots, X_n) = S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, the sample variance, as seen in previous review sessions.

We've seen in the previous review session how simple estimators, such as $\overline{X}_n$ and S^2, behave in the large-sample setting through the law of large numbers. Now, we want to understand how these estimators fare for a fixed sample size n. More generally, outside of these simple cases, we want (1) way(s) of obtaining a reasonable estimator for a general parameter θ and (2) some means of comparing the performance of different estimators in estimating a given parameter θ. If the performance of an estimator can be properly quantified, then we can develop a notion of a "best estimator".

Note that, in the most general case, θ might be a vector here of univariate parameters, or some function of another parameter of interest.

2 Method of Moments

The *method of moments* gives a fairly straightforward way of obtaining an estimator by conflating the population and sample moments. It works best when the parameter θ is something easily related to the moments of the distribution of $f(x|\theta)$ (e.g., when θ is the mean or variance).

Definition 2.1 (method of moments)**.** Let $X_1, \ldots, X_n$ be a sample from a population with pdf or pmf $f(x|\theta_1, \ldots, \theta_k)$. Method of moments estimators are found by equating the first k sample moments to the corresponding k population moments, and solving the resultant system of equations. Define:

$$m_k := \frac{1}{n}\sum_{i=1}^n X_i^k, \quad \mu_k := \mathbb{E}[X^k]$$

The population moment μ_j' is a function of $\theta_1, \ldots, \theta_k$. The method of moments estimator $(\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ of $(\theta_1, \ldots, \theta_k)$ is obtained by solving the system

$$\begin{aligned} m_1 &= \mu_1(\theta_1, \ldots, \theta_k) \\ m_2 &= \mu_2(\theta_1, \ldots, \theta_k) \\ &\vdots \\ m_k &= \mu_k(\theta_1, \ldots, \theta_k) \end{aligned}$$

Example 2.2 (normal method of moments)

Suppose $X_1, \ldots, X_n$ are iid $\mathcal{N}(\theta, \sigma^2)$. Our parameters here are (θ, σ^2). We then have $\mu_1 = \theta$ and $\mu_2 = \theta^2 + \sigma^2$ so that solving the system $\overline{X}_n = \theta$ and $\frac{1}{n}\sum X_i^2 = \theta^2 + \sigma^2$, we get

$$\tilde{\theta} = \overline{X}, \tilde{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2.$$

Example 2.3 (binomial method of moments)

Let $X_1, \ldots, X_n$ be iid binomial(k, p). Our parameters here are (k, p). Equating the sample moments to those of the population gives

$$\overline{X}_n = kp \text{ and } \frac{1}{n}\sum_i X_i^2 = kp(1-p) + k^2p^2 \implies \tilde{k} = \frac{\overline{X}^2}{\overline{X} - (1/n)\sum_i (X_i - \overline{X})^2} \text{ and } \tilde{p} = \frac{\overline{X}}{\tilde{k}}$$

Admittedly, the method of moments estimators are not often the best estimators for the population parameters. In the above example, we see that it is even possible for $\tilde{k}$ and $\tilde{p}$ to be negative, which goes against the ranges of the parameters k and p. Method of moment estimators are consistent under very weak assumptions since the sample moments m_k converge to the population moments μ_k by LLN. However, they tend to be *biased*. In Example 2.2, we see that $\mathbb{E}[\tilde{\sigma}^2] = \left(\frac{n-1}{n}\right) \cdot \sigma^2 \neq \sigma^2$.

3 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is another standard technique for finding estimators. Recall that if $X_1, \ldots, X_n$ are an i.i.d. sample from a population with pdf or pmf $f(x|\theta)$, the *likelihood function* is defined by

$$L(\theta | X_1, \ldots, X_n) := \prod_{i=1}^{n} f(X_i | \theta).$$

Definition 3.1 (MLE). For each sample point $\mathbf{x} = (x_1, \ldots, x_n)$, which we consider as realized values of the random sample $X_1, \ldots, X_n$, the *maximum likelihood estimator* of the parameter θ based on the sample $\mathbf{x}$ is the value of θ which maximizes $L(\theta|\mathbf{x})$.

> **Remark 3.2.** It is often easier to work with the *log-likelihood* $\log(L(\theta|\mathbf{x}))$ (which turns the product over $i \in [n]$ into a sum). Since $\log(\cdot)$ is a monotone transformation, the MLE is always the maximizer of $\log(L(\theta|\mathbf{x}))$.

Note that, unlike the method of moments estimator, the range of the MLE coincides with the range of the parameter by construction (i.e., the maximization of $L(\theta|\mathbf{x})$ should be treated as a *constrained* maximization over the known range of θ).

Intuitively, the MLE is a reasonable choice of estimator since it is the parameter which is most likely to have produced the observed sample. We'll see later that the MLE will also benefit from some other optimality properties. The main drawback of the MLE is the potential difficulty in maximizing $L(\theta|\mathbf{x})$. If the likelihood function is twice differentiable in θ, the go-to approach is to use calculus (i.e., the second derivative test or its variants). When θ is a vector, one has to be careful in optimizing $L(\theta|\mathbf{x})$ over all the dimensions of θ. In the case of two variables, $\theta = (\theta_1, \theta_2)$, the "second derivative test" for bivariate maximization gives us a way of finding optima in both θ_1 and θ_2.

Lemma 3.3

To verify a function $H(\theta_1, \theta_2)$ has a local maximum at $(\hat{\theta}_1, \hat{\theta}_2)$, it must be shown that

1. The first-order partials $\frac{\partial}{\partial \theta_1} H|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0$ and $\frac{\partial}{\partial \theta_2} H|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} = 0$.
2. At least one second order partial is negative: $\frac{\partial^2}{\partial \theta_1^2} H|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0$ or $\frac{\partial^2}{\partial \theta_2^2} H|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} < 0$.
3. The Jacobian of second-order partials is positive.

$$\frac{\partial^2}{\partial \theta_1^2} H(\theta_1, \theta_2) \frac{\partial^2}{\partial \theta_2^2} H(\theta_1, \theta_2) - \left(\frac{\partial^2}{\partial \theta_1 \partial \theta_2} H(\theta_1, \theta_2) \right)^2 \Bigg|_{\theta_1=\hat{\theta}_1, \theta_2=\hat{\theta}_2} > 0$$

However, it will often be the case that we do not need to check the last condition above. This may occur, for instance, if we can show that for any fixed θ_1, the likelihood can be maximized in θ_2 with a maximizer $\hat{\theta}_2$ not depending on θ_1. Then, it remains to maximize $L(\theta_1, \hat{\theta}_2 | \mathbf{x})$ over θ_1. Let's see an example of this.

Example 3.4 (normal MLE, mean and variance unknown)

Let $X_1, \dots, X_n$ be iid $\mathcal{N}(\theta, \sigma^2)$, with both θ, σ^2 unknown. Then

$$L(\theta, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_{i=1}^n (x_i - \theta)^2 / \sigma^2}$$

and

$$\log L(\theta, \sigma^2 | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log \sigma^2 - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2 / \sigma^2$$

The partials are then

$$\frac{\partial}{\partial \theta} \log L(\theta, \sigma^2 | \mathbf{x}) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta)$$
$$\frac{\partial}{\partial \sigma^2} \log L(\theta, \sigma^2 | \mathbf{x}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \theta)^2$$

Setting both partials equal to 0 gives solution $(\hat{\theta}, \widehat{\sigma^2}) = (\overline{x}_n, n^{-1} \sum (x_i - \overline{x}_n)^2)$. We show this is in fact a global maximum. Recall that if $\theta \neq \overline{x}_n$, then

$$\sum_i (x_i - \theta)^2 > \sum_i (x_i - \overline{x}_n)^2$$

Hence, for any σ^2,

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_i (x_i - \overline{x}_n)^2 / \sigma^2} \geq \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\sum_i (x_i - \theta)^2 / \sigma^2}$$

It suffices to show $(\sigma^2)^{-n/2} \exp(-(1/2) \sum (x_i - \overline{x}_n)^2 / \sigma^2)$ achieves its global maximum at $\widehat{\sigma^2}$. This is straightforward with univariate calculus.

Another consideration that often arises in tricky MLE calculations is to be careful about constraining the maximization to the range of θ. Let's see an example of this.

Example 3.5 (restricted range MLE)

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\theta, 1)$ where is it known that $\theta \geq 0$. With no restrictions on θ, we saw that the MLE of θ is $\overline{X}_n$; however, if $\overline{X}_n$ is negative, it will be outside the range of the parameter which means it cannot be the MLE in this setting with $\theta \geq 0$. However, if $\overline{X}_n$ is negative, then the likelihood function $L(\theta|X_1, \dots, X_n)$ is decreasing in θ for $\theta \geq 0$. Thus, it is maximized at $\hat{\theta} = 0$. If $\overline{X}_n > 0$ on the other hand, the likelihood is maximized at $\hat{\theta} = \overline{X}_n$ as our earlier calculations show. Thus, in this case, the MLE is

$$\hat{\theta} = \begin{cases} \overline{X}_n & \overline{X}_n \geq 0 \\ 0 & \overline{X}_n < 0. \end{cases}$$

Sometimes it is difficult to differentiate the likelihood or log-likelihood, and we have to instead make careful inferences about where the likelihood's maximum can be located. This occurs, for instance, when our parameter θ takes on a discrete range of values.

Example 3.6 (binomial MLE, unknown number of trials)

Let $X_1, \dots, X_n$ be a random sample from a binomial(k, p) population, where p is known and k is unknown. The likelihood is then

$$L(k|\mathbf{x}, p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i} (1-p)^{k-x_i}$$

Maximizing L by differentiation is difficult since k has to be an integer. Observe $L(k|\mathbf{x}, p) = 0$ if $k < \max(x_i)$. Thus, our maximizing k must satisfy $k \geq \max(x_i)$. Since the function $k \mapsto L(k|\mathbf{x}, p)$ is increasing and then decreasing in k, we can instead find the MLE k by mandating that it satisfies:

$$\frac{L(k|\mathbf{x}, p)}{L(k-1|\mathbf{x}, p)} \geq 1, \frac{L(k+1|\mathbf{x}, p)}{L(k|\mathbf{x}, p)} < 1.$$

These conditions become

$$(k(1-p))^n \geq \prod_{i=1}^{n} (k - x_i) \text{ and } ((k+1)(1-p))^n < \prod_{i=1}^{n} (k+1-x_i)$$

Dividing by k^n, we get:

$$(1-p)^n \geq \prod_{i=1}^{n} (1 - x_i/k) \text{ and } (1-p)^n < \prod_{i=1}^{n} (1 - x_i/(k+1)).$$

In fact, it will suffice to solve for $\prod_{i=1}^{n}(1 - x_i/z) = (1-p)^n$ on the domain of real numbers $z \in [\max\{x_i\}, \infty)$. Note that $f(z) := \prod_{i=1}^{n}(1 - x_i/z)$ is a continuous strictly increasing function in increasing z whose boundary values are $f(\max\{x_i\}) = 0$ and $\lim_{z \to \infty} f(z) = 1$. Thus, there is some value of $\hat{z}$ such that $f(\hat{z}) = (1-p)^n$, whence $\lceil \hat{z} \rceil$ will be the MLE.

Finally, sometimes, we may be interested in the MLE of a function of another parameter η which can be written as the function of θ, $\eta = \tau(\theta)$. Then, it turns out the MLE of η is just $\tau(\cdot)$ applied to the MLE of θ. Thus, MLE is invariant under transformations of the parameter. We make this formal as follows:

Theorem 3.7 (functional invariance of MLE)

Suppose that a distribution is indexed by a parameter θ, but the interest is in finding an esimtator for some function of θ, say $\eta := \tau(\theta)$. If $\tau(\cdot)$ is one-to-one, then it is clear that if $\hat{\theta}$ is the MLE of θ, then $\tau(\hat{\theta})$ should be the MLE of $\tau(\theta)$. This is evident from the fact that we can write the likelihood of η as

$$L^*(\eta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\tau^{-1}(\eta)) = L(\tau^{-1}(\eta)|\mathbf{x}),$$

so that if $\theta = \tau^{-1}(\eta)$ maximizes $L(\cdot|\mathbf{x})$, then η maximizes $L^*(\cdot|\mathbf{x})$. If $\tau(\cdot)$ is not one-to-one, then we need a more general notion of the likelihood of η since there is no longer a unique agreed-upon value of θ such that $\tau(\theta) = \eta$. In this case, we consider the *induced likelihood function*:

$$L^*(\eta|\mathbf{x}) = \sup_{\theta:\tau(\theta)=\eta} L(\theta|\mathbf{x}).$$

The value $\hat{\eta}$ that maximizes $L^*(\eta|\mathbf{x})$ is what we call the MLE of $\eta = \tau(\theta)$. Then, similar to before, we have $\hat{\eta} = \tau(\hat{\theta})$ where $\hat{\theta}$ is the MLE of θ.

4 Bayes Estimators

In the Bayesian approach, a parameter θ is thought to itself arise from a probability distribution, called the *prior distribution*, which captures an experimenter's subjective and prior belief about the value of θ. This is so-called for being determined prior to observing the random sample $X_1, \dots, X_n \sim f(x|\theta)$. Upon observing the sample, the prior distribution on θ is updated to the so-called *posterior distribution*. The update procedure is rooted in Bayes' rule, which tells us how to relate the conditional distribution $\theta|\mathbf{x}$ to the distribution $\mathbf{x}|\theta$.

In particular, let $\pi(\theta)$ be a prior distribution and let $f(\mathbf{x}|\theta)$ be the sampling distribution. Then, the posterior distribution, i.e. the conditional distribution of θ given the sample $\mathbf{x}$, is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})},$$

where $m(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)\, d\theta$. Typically, it will suffice to compute only the part of the RHS which depends on θ, $f(\mathbf{x}|\theta)\pi(\theta)$, as this will often identify the posterior distribution. From the posterior distribution, we can concoct point estimates of θ. For example, we know the mean $\mathbb{E}[\theta|\mathbf{x}]$ is a fairly "representative" deterministic value of the distribution $\pi(\theta|\mathbf{x})$. So, we can consider the point estimate $\delta(\mathbf{x}) := \mathbb{E}[\theta|\mathbf{x}]$ of θ. We can also consider similar measures of central tendency such as the median median$(\theta|\mathbf{x})$ of the posterior. The *maximum a posteriori (MAP)* estimator is the mode of the posterior distribution $\mathrm{argmax}_\theta\, \pi(\theta|\mathbf{x})$.

Usually, unless stated otherwise, the *Bayes estimator* is understood to be the posterior mean $\mathbb{E}[\theta|\mathbf{x}]$.

Example 4.1 (binomial Bayes estimation)

Let $X_1, \dots, X_n$ be iid Bernoulli(p). Then $Y = \sum_i X_i$ is binomial(n, p). Assume the prior on p is beta(α, β). The joint distribution of Y and p is

$$f(y,p) = \left(\binom{n}{y} p^y (1-p)^{n-y}\right) \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right) = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

The marginal of Y is then (by recognizing the integral contains the kernel of a beta pdf)

$$f(y) = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}.$$

The posterior is then

$$f(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}$$

which is beta$(y+\alpha, n-y+\beta)$. The Bayes estimator is then:

$$\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}$$

When estimating a binomial parameter, as in the example above, it was not absolutely necessary to choose a prior distribution from the beta family. However, there was a certain advantage to choosing the beta family in that the estimator ended up having a nice closed-form expression. Moreover, this was made possible by the fact that the posterior was in the same family as the prior. There is a broad class of examples for which this phenomenon holds.

Definition 4.2 (conjugate family)**.** Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$. Examples of conjugate families can be found here.

Note: by class Π we mean a collection of distributions or pdf's/pmf's, typically parametrized by one or two real numbers, much like how the class $\mathcal{F}$ is indexed by θ. We've seen many examples of such classes already, e.g. the beta family, the normal family, the gamma family, etc.

Example 4.3 (normal Bayes estimators)

Let $X \sim \mathcal{N}(\theta, \sigma^2)$ and suppose the prior on θ is $\mathcal{N}(\mu, \tau^2)$. The posterior of θ then is also normal with mean and variance

$$\mathbb{E}[\theta|x] = \frac{\tau^2}{\tau^2+\sigma^2}x + \frac{\sigma^2}{\sigma^2+\tau^2}\mu \text{ and } \mathrm{Var}(\theta|x) = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$$

Thus, the normal family is its own conjugate family.

5 Methods of Evaluating Estimators

5.1 Mean Squared Error

We want some way of measuring the quality of an estimator. A natural approach is to first consider some loss function $L(\theta, W)$ of the true value of the parameter θ and the estimator $W = W(X_1, \dots, X_n)$. For instance, we might take $L(\theta, W)$ to simply be the Euclidean distance between W and θ. Since $X_1, \dots, X_n$ are random, we want to consider the average error $\mathbb{E}_\theta[L(\theta, W)]$. This running standard we will use in this section is the *mean squared error* where the loss $L(\theta, W) := |\theta - W|^2$ (for $\theta \in \mathbb{R}$).

Definition 5.1 (mean squared error)**.** The *mean squared error* (MSE) of an estimator W of a parameter θ is the function of θ defined by $\mathbb{E}_\theta[|W - \theta|^2]$.

Remark 5.2 (bias-variance decomposition)**.** The MSE can be rewritten as:

$$\mathbb{E}_\theta[W-\theta]^2 = \operatorname{Var}_\theta W + (\mathbb{E}_\theta W - \theta)^2 =: \operatorname{Var}_\theta W + (\text{Bias}_\theta W)^2$$

Let's look at how some of the familiar estimators we've established so far fare in terms of MSE.

Example 5.3 (normal MSE)

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. The statistics $\overline{X}_n$ and S^2 are both unbiased estimators of their population analogues μ and σ^2:

$$\mathbb{E}[\overline{X}_n] = \mu, \mathbb{E}[S^2] = \sigma^2$$

for all μ, σ^2. In fact, both of the above are always true without the normality assumption. The MSE's of these estimators are, respectively,

$$\mathbb{E}[\overline{X}_n - \mu]^2 = \operatorname{Var} \overline{X}_n = \frac{\sigma^2}{n} \text{ and } \mathbb{E}[S^2 - \sigma^2]^2 = \operatorname{Var} S^2 = \frac{2\sigma^4}{n-1}$$

Example 5.4

We've seen before that an alternative estimator for σ^2 is the MLE $\hat{\sigma}^2 := \frac{1}{n}\sum(X_i - \overline{X})^2 = \frac{n-1}{n} \cdot S^2$. We have

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$$

so that $\hat{\sigma}^2$ is a biased estimator of σ^2. The variance of $\hat{\sigma}^2$ is then

$$\operatorname{Var}(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2}$$

and, hence, its MSE is

$$\mathbb{E}[\hat{\sigma}^2 - \sigma^2]^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4$$

Thus,

$$\mathbb{E}[\hat{\sigma}^2 - \sigma^2]^2 = \left(\frac{2n-1}{n^2}\right)\sigma^4 < \left(\frac{2}{n-1}\right)\sigma^4 = \mathbb{E}[S^2 - \sigma^2]^2$$

meaning $\hat{\sigma}^2$ has a smaller MSE than S^2. Thus, by trading off variance for bias, the MSE is improved by using $\hat{\sigma}^2$ instead of S^2. This does not necessarily mean that $\hat{\sigma}^2$ is a better estimator than S^2: it is still biased and, on average, will underestimate σ^2. Moreover, there is the question of whether the MSE is the right notion of error for scale parameters such as σ^2.

5.2 Best Unbiased Estimators

It is not always obvious how to compare two estimators even based on mean squared error. Namely, the MSE $\mathbb{E}_\theta[[W-\theta|^2]$ is a function of θ and thus will vary in value for different values of θ. As a trivial example, the constant estimator $W(X_1, \dots, X_n) \equiv 0$ would have an MSE of 0 at $\theta = 0$, but would be a very unsuitable estimator for any other value of θ. One way to make this task of finding a "best" estimator more tractable is to limit the class of estimators. In particular, we consider the class of unbiased estimators W, i.e. such that $\mathbb{E}_\theta[W] = \theta$. From the bias-variance decomposition of the MSE, it then suffices to find an unbiased estimator with smallest variance, as this will also have smallest MSE.

Definition 5.5 (best unbiased estimator, UMVUE)**.** An estimator W^* is a *best unbiased estimator* of θ if it is unbiased for all θ and, for any other unbiased estimator W, we have

$$\operatorname{Var}_\theta(W^*) \le \operatorname{Var}_\theta(W)$$

for all θ. W^* is also called a *uniform minimum variance unbiased estimator* (UMVUE) of θ. This is also the estimator with the smallest MSE in this class of unbiased estimators.

It is often not difficult to come up with examples of unbiased estimators. For starters, if we can even come upon two unbiased estimators W_1, W_2, then any linear combination $c \cdot W_1 + (1-c) \cdot W_2$ for $c \in [0,1]$ will also be an unbiased estimator. But, it might be difficult to determine which unbiased estimator W^* truly minimizes the variance $\mathrm{Var}_\theta(W^*)$. However, it turns out the minimum variance has an exact formula, given by the Cramér-Rao inequality/bound.

Theorem 5.6 (Cramér-Rao Inequality)

Let $X_1, \ldots, X_n$ be a sample (not necessarily iid) with joint pdf $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying

$$\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} W(\mathbf{x}) f(\mathbf{x}|\theta)\, d\mathbf{x} \text{ and } \mathrm{Var}_\theta\, W(\mathbf{X}) < \infty \tag{1}$$

Then,

$$\mathrm{Var}_\theta\, W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X})\right)^2}{\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]}$$

Proof. Recall by Cauchy-Schwarz that

$$|\mathrm{Cov}(X,Y)|^2 \leq (\mathrm{Var}\, X)(\mathrm{Var}\, Y) \implies \mathrm{Var}(X) \geq \frac{|\mathrm{Cov}(X,Y)|^2}{\mathrm{Var}(Y)}.$$

Let $X = W(\mathbf{X})$ and let $Y = \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)$. Then, because we can switch the order of differentiation and integration by (1),

$$\mathbb{E}[Y] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right] = \mathbb{E}\left[\frac{\frac{\partial}{\partial\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} f(\mathbf{X}|\theta)\, d\mathbf{X} = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(\mathbf{X}|\theta)\, d\mathbf{X} = \frac{\partial}{\partial\theta}1 = 0.$$

Thus, $\mathrm{Var}(Y) = \mathbb{E}[Y^2]$ and, by a similar computation as above,

$$\mathrm{Cov}(X,Y) = \mathbb{E}[X\cdot Y] - \mathbb{E}[X]\cdot\mathbb{E}[Y] = \mathbb{E}\left[W(\mathbf{X})\cdot\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right] = \int_{\mathcal{X}} W(\mathbf{X})\cdot\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\, d\mathbf{X} = \frac{\partial}{\partial\theta}\mathbb{E}[W(\mathbf{X})].$$

■

Note that (1) is a fairly reasonable condition: that we can switch the order of differentiation and integration. It will hold for many standard pdf's and estimators, and is ensured, for instance, if the integrand $W(\mathbf{x})\cdot f(\mathbf{x}|\theta)$ and its derivative (w.r.t. θ) are uniformly bounded.

Corollary 5.7 (Cramér-Rao Inequality or Information Inequality, i.i.d. case)

If the assumptions of the previous theorem are satisfied and, additionally, if $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\theta)$, then

$$\mathrm{Var}_\theta(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}_\theta W(\mathbf{X})\right)^2}{n\cdot\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}$$

For unbiased estimators $\mathbb{E}_\theta[W(\mathbf{X})] = \theta$, the RHS numerator will be 1 and thus the Cramér-Rao lower bound looks like:

$$\mathrm{Var}_\theta(W(\mathbf{X})) \geq \frac{1}{n\cdot\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]}.$$

Proof. Use the fact that the joint pdf factors: $f(\mathbf{X}|\theta) = \prod_{i=1}^n f(X_i|\theta)$ and expand the square in the RHS denominator after converting the log of a product to a sum of logs. The cross-terms will vanish. ■

The Cramér-Rao inequality for discrete distributions/pmf's is analogous with the only modification being in (1), where the integral changes to a sum.

The quantity $\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right]$ in the lower bound is called the *Fisher information*. It is so-called since the larger it is, i.e. the more information we have, the more possible it is to better estimate θ (by decreasing the variance and hence the MSE). The Fisher information can in fact be computed with two different formulas.

Theorem 5.8

If $f(x|\theta)$ satisfies

$$\frac{d}{d\theta}\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X|\theta)\right] = \int \frac{\partial}{\partial\theta}\left(\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right) f(x|\theta)\right) dx$$

(again, a mild condition that we can exchange differentiation and integration; this is true for most common families of distributions), then

$$\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)\right]$$

Theorem 5.9 (multivariate Fisher information and multivariate Cramér-Rao)

Suppose $\theta = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$. Then, the *Fisher information matrix* of θ with respect to sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\theta)$ is the $p \times p$ matrix $I(\theta)$ with (i,j)-th entry:

$$I_{i,j} := \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log(f(x|\theta)\right)\cdot\left(\frac{\partial}{\partial\theta_j}\log(f(x|\theta)\right)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(f(x|\theta))\right].$$

Let $T(X) = (T_1(X), \ldots, T_p(X))$ be an estimator of θ and denote by its expectation $\psi(\theta) := \mathbb{E}_\theta[T(X)] \in \mathbb{R}^p$. The multivariate Cramér-Rao bound then states

$$\mathrm{Cov}_\theta(T(X)) \succeq \left(\frac{\partial\psi(\theta)}{\partial\theta}\right)\cdot[I(\theta)]^{-1}\left(\frac{\partial\psi(\theta)}{\partial\theta}\right)^T,$$

where $\partial\psi(\theta)/\partial\theta$ is the Jacobian matrix of $\psi(\theta)$ with respect to θ, and where the ordering on matrices "$A \succeq B$" means that $A - B$ is p.s.d. or $\lambda_{\min}(A-B) \geq 0$.

Theorem 5.10 (Fisher information of transformation)

If we are interested in a function of a parameter $\tau = \tau(\theta)$, then the Fisher information $I(\tau) := \mathbb{E}_\tau\left[\left(\frac{\partial}{\partial\tau}\log f(X|\theta)\right)^2\right]$ of τ can be obtained from the Fisher information $I(\theta)$ of θ, via chain rule:

$$I(\theta) = I(\tau(\theta))\cdot\left(\frac{\partial\tau}{\partial\theta}\right)^2.$$

If $\theta, \tau \in \mathbb{R}^p$, then we have

$$I(\theta) = J^T I(\tau(\theta))J,$$

where J is the $p \times p$ Jacobian matrix with (i,j)-th coordinate $J_{ij} = \frac{\partial\tau_i}{\partial\theta_j}$.

The question remains, however, as to which estimator W attains the Cramér-Rao lower bound. The answer turns out to be surprisingly simple.

Corollary 5.11 (attainment of Cramer-Rao bound)

Let $X_1, \ldots, X_n$ be iid $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramer-Rao inequality. Let $L(\theta|x) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ is any unbiased estimator of θ, then $W(\mathbf{X})$ attains the Cramer-Rao lower bound iff

$$a(\theta)(W(\mathbf{x}) - \theta) = \frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x})$$

for some function $a(\theta)$, i.e. if the log-likelihood and $W(\mathbf{x})$ are proportional to each other as in the equality case of Cauchy-Schwarz.

Proof. This follows from the equality case of Cauchy-Schwarz $\mathrm{Cov}(X,Y)^2 \leq \mathrm{Var}(X)\cdot\mathrm{Var}(Y)$ which occurs when X and Y are linear transformations of each other. ∎

5.3 Bayes risk

Definition 5.12 (risk function)**.** Recall we considered a loss function $L(\theta, W)$ and assessed the quality of an estimator by considering the average loss $R(\theta, W) := \mathbb{E}_\theta[L(\theta, W)]$. This is also called the *risk function*, and is a function of θ.

We discussed previously how it might be difficult to compare two estimators based on their risk functions $R(\theta, \cdot)$ since this varies with θ. However, in the Bayesian setup, we can further average out by the prior distribution $\pi(\theta)$ to obtain an "average risk" of an estimator W over all θ.

Definition 5.13 (Bayes risk)**.** For a prior distribution $\pi(\theta)$, we define the *Bayes risk* to be

$$\int_\Theta R(\theta, W)\pi(\theta)\, d\theta$$

An estimator that yields the smallest value of the Bayes risk is called the *Bayes rule with respect to a prior* π, and is denoted W^π.

Remark 5.14. For $\mathbf{X} \sim f(\mathbf{x}|\theta)$ and $\theta \sim \pi$, the Bayes risk of an estimator W can be written as

$$\int_\Theta R(\theta, W)\pi(\theta)\, d\theta = \int_\Theta \left(\int_\mathcal{X} L(\theta, W(\mathbf{x})) f(\mathbf{x}|\theta)\, d\mathbf{x} \right) \pi(\theta)\, d\theta$$

Now, write $f(\mathbf{x}|\theta)\pi(\theta) = \pi(\theta|\mathbf{x})m(\mathbf{x})$ where $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ and $m(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$ so that

$$\int_\Theta R(\theta, W)\pi(\theta)\, d\theta = \int_\mathcal{X} \left[\int_\Theta L(\theta, W(\mathbf{x}))\pi(\theta|\mathbf{x})\, d\theta \right] m(\mathbf{x})\, d\mathbf{x}$$

The quantity in square brackets is called the *posterior expected loss* and is a function only of $\mathbf{x}$ and not of θ. Thus, for each $\mathbf{x}$, if we choose $W(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

Example 5.15 (two Bayes rules)

Consider a point estimation problem for a real-valued parameter θ.

1. For squared error loss, the posterior expected loss is

$$\int_\Theta (\theta - a)^2 \pi(\theta|\mathbf{x})\, d\theta = \mathbb{E}[(\theta - a)^2 | \mathbf{X} = \mathbf{x}]$$

   Here θ is the random variable with distribution $\pi(\theta|\mathbf{x})$. This expected value is minimized by $W^\pi(\mathbf{x}) = \mathbb{E}[\theta|\mathbf{x}]$. So the Bayes rule is the mean of the posterior distribution.

2. For absolute error loss, the posterior expected loss is $\mathbb{E}[|\theta - a||\mathbf{X} = \mathbf{x}]$. We can see that this is minimized by choosing $W^\pi(\mathbf{x})$ to be the median of $\pi(\theta|\mathbf{x})$.

6 Problems

6.1 Previous Core Competency Problems

Problem 1 (May 2018, # 3). Let $W_1, W_2, \ldots, W_k$ be unbiased estimators of a parameter θ with $\mathrm{Var}(W_i) = \sigma_i^2$ and $\mathrm{Cov}(W_i, W_j) = 0$ if $i \neq j$.

(a) Show that among all estimators of the form $\sum_{i=1}^k a_i W_i$, where a_i's are constants and $\mathbb{E}_\theta(\sum_i a_i W_i) = \theta$, the estimator $W^* = \frac{\sum_i W_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$ has minimum variance.

(b) Show that $\mathrm{Var}(W^*) = \frac{1}{\sum_i 1/\sigma_i^2}$.

Problem 2 (May 2018, # 6). Consider observed response variables $Y_1, \ldots, Y_n \in \mathbb{R}$ that depend linearly on covariates $x_1, \ldots, x_n$ as follows:

$$Y_i = \beta x_i + \epsilon_i, \text{ for } i = 1, \ldots, n.$$

Here, the ϵ_i's are independent Gaussian noise variables, but we do not assume they have the same variance. Instead, they are distributed as $\epsilon_i \sim N(0, \sigma_i^2)$ for possibly different variances $\sigma_1^2, \ldots, \sigma_n^2$. The unknown parameter of interest is β.

(a) Suppose that the error variances $\sigma_1^2, \ldots, \sigma_n^2$ are all known. Find the MLE $\hat{\beta}$ for β in this case and derive an explicit formula for $\hat{\beta}$. Show that $\hat{\beta}$ minimizes a certain weighted least-squares criterion.

(b) Show that the estimate $\hat{\beta}$ in part (a) is unbiased, and derive a formula for the variance of $\hat{\beta}$ in terms of $\sigma_1^2, \ldots, \sigma_n^2$ and $x_1, \ldots, x_n$.

(c) Compute the Fisher information $I(\beta)$ in this model (still assuming $\sigma_1^2, \ldots, \sigma_n^2$ are known constants). Compare this value with the variance of $\hat{\beta}$ derived in part (b).

Problem 3 (May 2018, # 7). Suppose that $X \sim$ Poisson(λ) and its parameter $\lambda > 0$ has a prior distribution Gamma(α, β) given by density

$$f(y|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-y\beta} y^{\alpha-1}, \text{ for } y \geq 0, \text{ (and 0 otherwise)}.$$

(a) Find the posterior distribution of λ given the observation X, and identify the distribution with its parameters.

(b) Find the mean of this posterior distribution.

Problem 4 (May 2018, # 8). Suppose $X_1, X_2 \overset{i.i.d.}{\sim} Ber(p)$ for some unknown parameter $p \in (0, 1)$. Find an unbiased estimator for the following functions of p, if there exists one.

(a) $g(p) = 2p$.

(b) $g(p) = p(1-p)$.

(c) $g(p) = p^2$.

(d) $g(p) = p^3$.

Problem 5 (September 2019, # 7). Suppose that $X_1, \ldots, X_n$ are i.i.d. uniform random variables on $[0, \theta]$ for some $\theta \in [1, 2]$.

(i) What is the MLE of θ?

(ii) Suppose that, instead of X_i's, we only observe, for all $i = 1, \ldots, n$,

$$Y_i = \begin{cases} X_i & \text{if } X_i \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

What is the MLE of θ based on $\{Y_1, \ldots, Y_n\}$?

Problem 6 (September 2019, # 8). Suppose that a measurement Y is recorded with a $N(\theta, \sigma^2)$ sampling distribution, with σ known and θ known to lie in the interval $[0,1]$ (but otherwise unknown). Consider two point estimators of θ: (a) the posterior mean $\hat{\theta}_B$ based on the assumption of a uniform prior distribution on θ on $[0,1]$, and (b) the maximum likelihood estimate $\hat{\theta}_M$, restricted to the range $[0,1]$.

(i) Show that, as $\sigma \to \infty$, $\hat{\theta}_B$ converges in distribution (to Y_1, say). Identify the limit Y_1. [**Hint:** You may first find the distribution of $\Theta|Y = y$ and then take limits.]

(ii) Show that, as $\sigma \to \infty$, $\hat{\theta}_M$ converges in distribution (to Y_2, say). Identify the limit Y_2.

(iii) If σ is large enough, which estimator $\hat{\theta}_M$ or $\hat{\theta}_B$ has a higher mean squared error, for any value of θ in $[0,1]$. You may answer this question by comparing the mean squared errors of Y_1 and Y_2 for estimating θ.

Problem 7 (May 2020, # 6). Suppose that we have single observation from X from the exponential distribution with parameter λ. Define $T(X) = I(X > 1)$, where I is the indicator function. Set $\psi(\lambda) := e^{-\lambda}$.

(i) Show that $T(X)$ is unbiased for $\psi(\lambda)$.

(ii) Find the (Fisher) information bound for unbiased estimators of $\psi(\lambda)$.

(iii) Show that the variance of $T(X)$ is strictly larger than the information bound.

Problem 8 (September 2020, # 5). Consider the following Bayesian model

$$Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \text{Uniform}([0,\theta]) \text{ and } \theta \sim \text{Pareto}(\beta, \lambda,)$$

where the pdf of the Pareto distribution is given by

$$\pi(\theta; \beta, \lambda) = \frac{\beta \lambda^\beta}{\theta^{(\beta+1)}}, \quad \theta > \lambda, \quad \beta, \lambda > 0.$$

Moreover, for this exercise you may assume $\beta > 1$.

(a) Use the Bayes formula to derive the posterior density of θ as explicitly as possible.

(b) Compute the prior and posterior means of θ.

Problem 9 (May 2021, # 1). Let $X_1, \dots, X_n$ be an i.i.d. random sample with common density function

$$f(x) = \begin{cases} 3\theta^3 x^{-4} & \text{for } x \geq \theta \\ 0 & \text{otherwise} \end{cases},$$

where $\theta > 0$ is an unknown parameter.

(i) Apply the method of moments to obtain an unbiased estimator of θ.

(ii) Find the maximum likelihood estimator (MLE) of θ and show that it is biased.

(iii) Which of the above two estimators has a smaller mean squared error (MSE)?

Problem 10 (September 2021, # 1). Let $X_1, \dots, X_n$ be an i.i.d. sample with common density

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & \text{otherwise}, \end{cases}$$

where $\theta > 0$ is an unknown parameter.

(i) Find a one dimensional sufficient statistic T_n.

(ii) Derive the cumulative distribution function F_n of T_n.

(iii) Give an exact $(1-\alpha)$-confidence interval for θ. (Hint: What is the distribution of $F_n(T_n)$?).

Problem 11 (September 2021, # 2)**.** Let X and Y be two independent exponential random variables with parameters λ and μ, respectively, i.e. $\mathbb{P}(X \geq x, Y \geq y) = e^{-\lambda x - \mu y}$, $x \geq 0$, $y \geq 0$. Define random variables

$$T = \min(X, Y) \text{ and } \Delta = \begin{cases} 1 & X < Y \\ 0 & \text{otherwise.} \end{cases}$$

(i) Find the probability density function of T and the probability mass function of Δ.

(ii) Find the joint distribution function of (T, Δ).

(iii) Suppose we have a random sample (T_i, Δ_i), $i = 1, \ldots, n$, i.e. i.i.d. copies of (T, Δ). Write down the likelihood function and find the MLE of λ.

6.2 Additional Practice

Problem 12 (Casella & Berger, Exercise 7.12)**.** Let $X_1, \ldots, X_n$ be a random sample from a population with pmf

$$\mathbb{P}_\theta(X = x) = \theta^x (1-\theta)^{1-x}, x = 0 \text{ or } 1, 0 \leq \theta \leq \frac{1}{2}.$$

(i) Find the method of moments estimator and MLE of θ.

(ii) Find the mean squared errors of each of the estimators.

(iii) Which estimator is preferred? Justify your choice.