3. Maximum likelihood estimators and efficiency

3.1. **Maximum likelihood estimators.** Let $X_1, \ldots, X_n$ be a random sample, drawn from a distribution P_θ that depends on an unknown parameter θ. We are looking for a general method to produce a statistic $T = T(X_1, \ldots, X_n)$ that (we hope) will be a reasonable estimator for θ.

One possible answer is the *maximum likelihood* method. Suppose I observed the values $x_1, \ldots, x_n$. Before the experiment, the probability that exactly these values would occur was $P_\theta(X_1 = x_1, \ldots, X_n = x_n)$, and this will depend on θ. Since I did observe these values, maybe it's a good idea to look for a θ that maximizes this probability (which, to impress the uninitiated, we now call *likelihood*).

Please do not confuse this maximization with the futile attempt to find the θ that is now most likely, given what I just observed. I really maximize over the condition: given that θ has some concrete value, we can work out the probability that what I observed occurred, and this is what I maximize.

Exercise 3.1. Please elaborate. Can you also make it plausible that there are (artificial) examples where the MLE is in fact quite likely to produce an estimate that is hopelessly off target?

Definition 3.1. We call a statistic $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ a *maximum likelihood estimator* for θ if $P_\theta(X_1 = x_1, \ldots, X_n = x_n)$ is maximal at $\theta = \widehat{\theta}(x_1, \ldots, x_n)$.

There is, in general, no guarantee that this maximum exists or (if it does) is unique, but we'll ignore this potential problem and just hope for the best. Also, observe that if we take the definition apart very carefully, we discover a certain amount of juggling around with arguments of functions: the MLE $\widehat{\theta}$ is a *statistic,* that is, a random variable that is a function of the random sample, but the maximizing value of the parameter is obtained by replacing the X_j by their observed values x_j. Alternatively, we could say that we consider the *likelihood function* $L(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$, then plug the random variables X_j into their own likelihood function and finally maximize, which then produces a maximizer that is a random variable itself (and in fact a statistic). None of this matters a whole lot right now; we'll encounter this curious procedure (plug random variables into functions obtained from their own distribution) again in the next section.

Example 3.1. Let's return to the coin flip example: $P(X_1 = 1) = \theta$, $P(X_1 = 0) = 1 - \theta$, and here it's convenient to combine this into one

formula by writing $P(X_1 = x) = \theta^x(1-\theta)^{1-x}$, for $x = 0, 1$. Thus

$$P(X_1 = x_1, \ldots, X_n = x_n) = \theta^{\sum x_j}(1-\theta)^{n-\sum x_j}.$$

We are looking for the θ that maximizes this expression. Take the θ derivative and set this equal to zero. Also, let's abbreviate $S = \sum x_j$.

$$S\theta^{S-1}(1-\theta)^{n-S} - (n-S)\theta^S(1-\theta)^{n-S-1} = 0$$

or $S(1-\theta) - (n-S)\theta = 0$, and this has the solution $\widehat{\theta} = S/n$. (We'd now have to check that this is indeed a maximum, but we skip this part.)

So the MLE for this distribution is given by $\widehat{\theta} = T = \overline{X}$. It is reassuring that this obvious choice now receives some theoretical justification.

We know that this estimator is unbiased. In general, however, MLEs can be biased. To see this, let's return to another example that was discussed earlier.

Example 3.2. Consider again the urn with an unknown number $N = \theta$ of balls in it, labeled $1, \ldots, N$. We form a random sample $X_1, \ldots, X_n$ by drawing n times, with replacement, according to the distribution $P(X_1 = x) = (1/N)\chi_{1,\ldots,N}(x)$. For fixed $x_1, \ldots, x_n \geq 1$, the probability of observing this outcome is then given by

$$P(X_1 = x_1, \ldots, X_n = x_n) = \begin{cases} N^{-n} & \max x_j \leq N \\ 0 & \max x_j > N \end{cases}. \tag{3.1}$$

We want to find the MLE, so we are trying to maximize this over N, for fixed $x_1, \ldots, x_n$. Clearly, entering the second line of (3.1) is no good, so we must take $N \geq \max x_j$. For any such N, the quantity we're trying to maximize equals N^{-n}, so we get the largest possible value by taking the smallest N that is still allowed. In other words, the MLE is given by $\widehat{N} = \max X_j$.

We know that this estimator is not unbiased. Again, it is nice to see some theoretical justification emerging for an estimator that looked reasonable.

Example 3.3. Recall that the *Poisson distribution* with parameter $\theta > 0$ is given by

$$P(X = x) = \frac{\theta^x}{x!}\, e^{-\theta}, \qquad (x = 0, 1, 2, \ldots).$$

Let's try to find the MLE for θ. A random sample drawn from this distribution has the likelihood function

$$P(X_1 = x_1, \ldots, X_n = x_n) = \frac{\theta^{x_1+\ldots+x_n}}{x_1! \cdots x_n!} e^{-n\theta}.$$

We want to maximize this with respect to θ, so we can ignore the denominator, which does not depend on θ. Let's again write $S = \sum x_j$; we then want to maximize $\theta^S e^{-n\theta}$. This leads to

$$S\theta^{S-1}e^{-n\theta} - n\theta^S e^{-n\theta} = 0$$

or $\widehat{\theta} = S/n$, that is $\widehat{\theta} = \overline{X}$.

Exercise 3.2. Show that $EX = \theta$ if X is Poisson distributed with parameter θ. Conclude that the MLE is unbiased.

For random samples drawn from continuous distributions, the above recipe cannot literally be applied because $P(X_1 = x_1, \ldots, X_n = x_n) = 0$ always in this situation. However, we can modify it as follows: call a statistic $\widehat{\theta}$ a MLE for θ if $\widehat{\theta}(x_1, \ldots, x_n)$ maximizes the (joint) *density*

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) \cdots f(x_n; \theta),$$

for all possible values x_j of the random sample. In analogy to our terminology in the discrete case, we will again refer to this product of the densities as the *likelihood function.*

Example 3.4. Consider the *exponential distribution* with parameter θ; this is the distribution with density

$$f(x) = \frac{e^{-x/\theta}}{\theta} \qquad (x \geq 0), \tag{3.2}$$

and $f(x) = 0$ for $x < 0$. Let's first find EX for an exponentially distributed random variable X:

$$EX = \frac{1}{\theta}\int_0^\infty x e^{-x/\theta}\, dx = -x e^{-x/\theta}\Big|_0^\infty + \int_0^\infty e^{-x/\theta}\, dx = \theta,$$

by an integration by parts in the first step. (So it *is* natural to use θ as the parameter, rather than $1/\theta$.)

To find the MLE for θ, we have to maximize $\theta^{-n}e^{-S/\theta}$ (writing, as usual, $S = \sum x_j$). This gives

$$-n\theta^{-n-1}e^{-S/\theta} + \frac{S}{\theta^2}\theta^{-n}e^{-S/\theta} = 0$$

or $\widehat{\theta} = S/n$, that is, as a statistic, $\widehat{\theta} = \overline{X}$ (again...). This MLE is unbiased.

What would have happened if he had used $\eta = 1/\theta$ in (3.2) instead, to avoid the reciprocals? So $f(x) = \eta e^{-\eta x}$ for $x \geq 0$, and I now want to find the MLE $\widehat{\eta}$ for η. In other words, I want to maximize $\eta^n e^{-\eta S}$, and proceeding as above, we find that this happens at $\widehat{\eta} = n/S$ or $\widehat{\eta} = 1/\overline{X}$.

Now recall that $\eta = 1/\theta$, and the MLE for θ was $\widehat{\theta} = \overline{X}$. This is no coincidence; essentially, we solved the same maximization problem

twice, with slightly changed notation the second time. In general, we have the following (almost tautological) statement:

Theorem 3.2. *Consider parameters η, θ that parametrize the same distribution. Suppose that they are related by $\eta = g(\theta)$, for a bijective g. Then, if $\widehat{\theta}$ is a MLE for θ, then $\widehat{\eta} = g(\widehat{\theta})$ is a MLE for η.*

Exercise 3.3. Give a somewhat more explicit version of the argument suggested above.

Notice, however, that the MLE estimator is no longer unbiased after the transformation. This could be checked rather quickly by an indirect argument, but it is also possible to work things out explicitly.

To get this started, let's first look at the distribution of the sum $S_2 = X_1 + X_2$ two independent exponentially distributed random variables X_1, X_2. We know that the density of S_2 is the convolution of the density from (3.2) with itself:

$$f_2(x) = \frac{1}{\theta^2} \int_0^x e^{-t/\theta} e^{-(x-t)/\theta}\, dt = \frac{1}{\theta^2} x e^{-x/\theta}$$

Next, if we add one more independent random variable with this distribution, that is, if we consider $S_3 = S_2 + X_3$, then the density of S_3 can be obtained as the convolution of f_2 with the density f from (3.2), so

$$f_3(x) = \frac{1}{\theta^3} \int_0^x t e^{-t/\theta} e^{(t-x)/\theta}\, dt = \frac{1}{2\theta^3} x^2 e^{-x/\theta}.$$

Continuing in this style, we find that

$$f_n(x) = \frac{1}{(n-1)!\theta^n}\, x^{n-1} e^{-x/\theta}.$$

Exercise 3.4. Denote the density of $S = S_n$ by f_n. Show that then S/n has density $f(x) = n f_n(nx)$.

Since $\overline{X} = S/n$, the Exercise in particular says that $\overline{X}$ has density

$$f(x) = \frac{n}{(n-1)!\theta^n} (nx)^{n-1} e^{-nx/\theta} \qquad (x \geq 0). \tag{3.3}$$

This is already quite interesting, but let's keep going. We were originally interested in $Y = 1/\overline{X}$, the MLE for $\eta = 1/\theta$. We apply the usual technique to transform the densities:

$$P(Y \leq y) = P(\overline{X} \geq 1/y) = \int_{1/y}^{\infty} f(x)\, dx,$$

and since $g = f_Y$ can be obtained as the y derivative of this, we see that

$$g(y) = \frac{1}{y^2} f(1/y) = \frac{n}{(n-1)!\theta^n} y^{-2} (n/y)^{n-1} e^{-n/(\theta y)} \qquad (y > 0). \tag{3.4}$$

This gives

$$\begin{aligned} EY &= \int y g(y)\, dy = \frac{n}{(n-1)!\theta^n} \int_0^\infty y^{-1} \left(\frac{n}{y}\right)^{n-1} e^{-n/(\theta y)}\, dy \\ &= \frac{n}{(n-1)!\theta} \int_0^\infty t^{n-2} e^{-t}\, dt. \end{aligned}$$

We have used the substitution $t = n/(\theta y)$ to pass to the second line. The integral can be evaluated by repeated integration by parts, or, somewhat more elegantly, you recognize it as $\Gamma(n-1) = (n-2)!$. So, putting things together, it follows that

$$E(1/\overline{X}) = \frac{n}{(n-1)\theta} = \frac{n}{n-1}\, \eta.$$

In particular, $Y = 1/\overline{X}$ is not an unbiased estimator for η; we are off by the factor $n/(n-1) > 1$ (which, however, is very close to 1 for large n).

Exercise 3.5. Check one more time that $\overline{X}$ is an unbiased estimator for θ, this time by making use of the density f from (3.3) to compute $E\overline{X}$ (in an admittedly rather clumsy way). You can again use the fact that $\Gamma(k) = (k-1)!$ for $k = 1, 2, \ldots$.

Example 3.5. Consider the uniform distribution on $[0, \theta]$:

$$f(x) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

We would like to find the MLE for θ. We then need to maximize with respect to θ (for given $x_1, \ldots, x_n \ge 0$) the likelihood function

$$f(x_1) \cdots f(x_n) = \begin{cases} \theta^{-n} & \max x_j \le \theta \\ 0 & \max x_j > \theta \end{cases}.$$

This first of all forces us to take $\theta \ge \max x_j$, to enter the first line, and then θ as small as (still) possible, to maximize θ^{-n}. Thus $\widehat{\theta} = \max(X_1, \ldots, X_n)$. This estimator is not unbiased.

Exercise 3.6. Why?

This whole example is an exact (continuous) analog of its discrete version Example 3.2.

Example 3.6. Finally, let's take a look at the normal distribution. Let's first find the MLE for $\theta = \sigma^2$, for a normal distribution with known μ. We then need to maximize

$$\theta^{-n/2} e^{-A/\theta}, \qquad A = \sum \frac{(x_j - \mu)^2}{2}.$$

This gives $-(n/2)/\theta + A/\theta^2 = 0$ or $\theta = 2A/n$, that is,

$$\widehat{\theta} = \frac{1}{n} \sum_{j=1}^{n} (X_j - \mu)^2. \tag{3.5}$$

Exercise 3.7. (a) Show that $n\widehat{\theta}/\sigma^2 \sim \chi^2(n)$.
(b) Conclude that $\widehat{\theta}$ is unbiased.

By Theorem 3.2, the MLE for σ is then given by

$$\widehat{\sigma} = \sqrt{\frac{1}{n} \sum (X_j - \mu)^2}.$$

This estimator is not unbiased.

What if μ and σ are both unknown? There is an obvious way to adapt our procedure: we can maximize over both parameters simultaneously to obtain *two* statistics that can serve as MLE style estimators. So we now want to maximize

$$\theta^{-n/2} \exp\left(-\frac{1}{2\theta} \sum_{j=1}^{n} (x_j - \mu)^2 \right)$$

over both μ and θ. We set the partial derivatives equal to zero and obtain the two conditions

$$-\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum_{j=1}^{n} (x_j - \mu)^2 = 0, \quad \sum_{j=1}^{n} (x_j - \mu) = 0.$$

The second equation says that $\mu = (1/n) \sum x_j =: \overline{x}$, and then, by repeating the calculation from above, we see from this and the first equation that $\theta = (1/n) \sum (x_j - \overline{x})^2$. In other words,

$$\widehat{\mu} = \overline{X}, \qquad \widehat{\theta} = \frac{1}{n} \sum_{j=1}^{n} (X_j - \overline{X})^2 = \frac{n-1}{n} S^2.$$

So $\widehat{\mu}$ is unbiased, but $\widehat{\theta}$ is not since $ES^2 = \sigma^2 = \theta$, so $E\widehat{\theta} = ((n-1)/n)\theta$.

Exercise 3.8. Find the MLE for θ for the following densities: (a) $f(x) = \theta x^{\theta - 1}$ for $0 < x < 1$, and $f(x) = 0$ otherwise, and $\theta > 0$;
(b) $f(x) = e^{\theta - x}$ for $x \geq \theta$ and $f(x) = 0$ otherwise

Exercise 3.9. Here's an example where the maximization does not produce a unique value. Consider the density $f(x) = (1/2)e^{-|x-\theta|}$. Assume for convenience that $n = 2k$ is even and consider data $x_1 < x_2 < \ldots < x_n$. Then show that any $\widehat{\theta}$ in the interval $x_k < \widehat{\theta} < x_{k+1}$ maximizes the likelihood function.

Exercise 3.10. (a) Show that

$$f(x, \theta) = \frac{1}{\theta^2} x e^{-x/\theta} \qquad (x \geq 0)$$

(and $f(x) = 0$ for $x < 0$) is a density for $\theta > 0$.
(b) Find the MLE $\widehat{\theta}$ for θ.
(c) Show that $\widehat{\theta}$ is unbiased.

3.2. **Cramer-Rao bounds.** If an estimator is unbiased, it delivers the correct value at least on average. It would then be nice if this estimator showed only little variation about this correct value (of course, if T is biased, it is less clear if little variation about the incorrect value is a good thing).

Let's take another look at our favorite example from this point of view. So $P(X_1 = 1) = \theta$, $P(X_1 = 0) = 1 - \theta$, and we are going to use the MLE $T = \widehat{\theta} = \overline{X}$. Since the X_j are independent, the variances add up and thus

$$\operatorname{Var}(T) = \frac{1}{n^2} n \operatorname{Var}(X_1) = \frac{\theta(1-\theta)}{n}$$

and $\sigma_T = \sqrt{\theta(1-\theta)/n} \leq 1/(2\sqrt{n})$. This doesn't look too bad. In particular, for large random samples, it gets small; it decays at the rate $\sigma_T \sim 1/\sqrt{n}$.

Could we perhaps do better than this with a different unbiased estimator? It turns out that this is not the case. The statistic $T = \overline{X}$ is optimal in this example in the sense that it has the smallest possible variance among all unbiased estimators. We now derive such a result in a general setting.

Let $f(x, \theta)$ be a density that depends on the parameter θ. We will assume throughout this section that f is sufficiently well behaved so that the following manipulations are justified, without actually making explicit a precise version of such assumptions. We will certainly need f to be twice differentiable with respect to θ since we will take this second derivative, but this on its own is not sufficient to justify some of the other steps (such as differentiating under the integral sign).

We have $\int_{-\infty}^{\infty} f \, dx = 1$, so by taking the θ derivative (and interchanging differentiation and integral), we obtain $\int \partial f / \partial \theta \, dx = 0$. This we

may rewrite as

$$\int_{-\infty}^{\infty} f(x,\theta)\frac{\partial}{\partial\theta}\ln f(x,\theta)\,dx = 0. \tag{3.6}$$

There are potential problems here with regions where $f = 0$; to avoid these, I will simply interpret (3.6) as an integral over only those parts of the real line where $f > 0$. (To make sure that the argument leading to (3.6) is still justified in this setting, we should really make the additional assumption that $\{x : f(x,\theta) > 0\}$ does not depend on θ, but we'll ignore purely technical points of this kind.)

An alternative reading of (3.6) is $E(\partial/\partial\theta)\ln f(X,\theta) = 0$. Here (and below) I use the general fact that $Eg(X) = \int g(x)f(x)\,dx$ for any function g.

Also note the somewhat curious construction here: we plug the random variable X into its own density (and then take the logarithm) to produce the new random variable $\ln f(X)$ (which also depends on θ). If we take one more derivative, then (3.6) becomes

$$\int_{-\infty}^{\infty} f(x,\theta)\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta)\,dx + \int_{-\infty}^{\infty} f(x,\theta)\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)^2 dx = 0. \tag{3.7}$$

Definition 3.3. The *Fisher information* is defined as

$$I(\theta) = E\left(\frac{\partial}{\partial\theta}\ln f(X,\theta)\right)^2.$$

This assumes that X is a continuous random variable; in the discrete case, we replace f by $P(X = x,\theta)$ (and again plug X into its own distribution). From (3.7), we obtain the alternative formula

$$I(\theta) = -E\frac{\partial^2}{\partial\theta^2}\ln f(X,\theta); \tag{3.8}$$

moreover, it is also true that

$$I(\theta) = \mathrm{Var}((\partial/\partial\theta)\ln f(X,\theta)). \tag{3.9}$$

Example 3.7. Let's return one more time to the coin flip example: $P(X = x) = \theta^x(1-\theta)^{1-x}$ $(x = 0, 1)$, so $\ln P = x\ln\theta + (1-x)\ln(1-\theta)$ and

$$\frac{\partial}{\partial\theta}\ln P = \frac{x}{\theta} - \frac{1-x}{1-\theta}. \tag{3.10}$$

To find the Fisher information, we plug X into this function and take the square. This produces

$$\frac{X^2}{\theta^2} + \frac{(1-X)^2}{(1-\theta)^2} - 2\frac{X(1-X)}{\theta(1-\theta)} =$$
$$X^2\left(\frac{1}{\theta^2} + \frac{1}{(1-\theta)^2} + \frac{2}{\theta(1-\theta)}\right) - 2X\left(\frac{1}{(1-\theta)^2} + \frac{1}{\theta(1-\theta)}\right) + \frac{1}{(1-\theta)^2}.$$

Now recall that $EX = EX^2 = \theta$, and take the expectation. We find that

$$\begin{aligned} I(\theta) &= \frac{\theta(1-\theta)^2 + \theta^3 + 2\theta^2(1-\theta) - 2\theta^3 - 2\theta^2(1-\theta) + \theta^2}{\theta^2(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)}. \end{aligned}$$

Alternatively, we could have obtained the same result more quickly from (3.8). Take one more derivative in (3.10), plug X into the resulting function and take the expectation:

$$I(\theta) = -E\left(-\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2}\right) = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$$

Example 3.8. Consider the $N(\theta, 1)$ distribution. Its density is given by $f = (2\pi)^{-1/2}e^{-(x-\theta)^2/2}$, so $\ln f = -(x-\theta)^2/2 + C$. Two differentiations produce $(\partial^2/\partial\theta^2)\ln f = -1$, so $I = 1$.

When dealing with a random sample $X_1, \ldots, X_n$, Definition 3.3 can be adapted by replacing f by what we called the likelihood function in the previous section. More precisely, we could replace (3.9) with

$$\operatorname{Var}\left(\frac{\partial}{\partial\theta}\ln L(X_1, \ldots, X_n; \theta)\right),$$

where $L(x_1, \ldots, x_n) = f(x_1)\cdots f(x_n)$ (continuous case) or $L(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$ (discrete case). Then, however, we can use the product structure of L and independence to evaluate (in the continuous case, say)

$$\operatorname{Var}\left(\sum_{j=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_j, \theta)\right) = \sum_{j=1}^{n}\operatorname{Var}\left(\frac{\partial}{\partial\theta}\ln f(X_j, \theta)\right) = nI(\theta),$$

where now I is the Fisher information of an individual random variable X. An analogous calculation works in the discrete case.

Theorem 3.4 (Cramer-Rao)**.** *Let $T = T(X_1, \dots, X_n)$ be a statistic and write $k(\theta) = ET$. Then, under suitable (smoothness) assumptions,*

$$\mathrm{Var}(T) \geq \frac{(k'(\theta))^2}{nI(\theta)}.$$

Corollary 3.5. *If the statistic T in Theorem 3.4 is unbiased, then*

$$\mathrm{Var}(T) \geq \frac{1}{nI(\theta)}.$$

As an illustration, let's again look at the coin flip example with its MLE $T = \widehat{\theta} = \overline{X}$. We saw earlier that $\mathrm{Var}(T) = \theta(1-\theta)/n$, and this equals $1/(nI)$ by our calculation from Example 3.7. Since T is also unbiased, this means that this estimator achieves the Cramer-Rao bound from Corollary 3.5. We give a special name to estimators that are optimal, in this sense:

Definition 3.6. Let T be an unbiased estimator for θ. We call T *efficient* if T achieves the CR bound:

$$\mathrm{Var}(T) = \frac{1}{nI(\theta)}$$

So we can summarize by saying that $\overline{X}$ is an efficient estimator for θ.

Let's now try to derive the CR bound. I'll do this for continuous random variables, with density $f(x,\theta)$. Then

$$k(\theta) = \int dx_1 \int dx_2 \dots \int dx_n T(x_1, \dots, x_n) f(x_1,\theta) \cdots f(x_n,\theta)$$

and thus (at least if we are allowed to freely interchange differentiations and integrals)

$$\begin{aligned} k'(\theta) &= \sum_{j=1}^{n} \int dx_1 \int dx_2 \dots \int dx_n T(x_1, \dots, x_n) \times \\ &\qquad\qquad f(x_1,\theta) \cdots \frac{\partial f(x_j,\theta)}{\partial \theta} \cdots f(x_n,\theta) \\ &= \int dx_1 \int dx_2 \dots \int dx_n T(x_1, \dots, x_n) \times \\ &\qquad \left(\sum_{j=1}^{n} \frac{\partial}{\partial \theta} \ln f(x_j,\theta) \right) f(x_1,\theta) \cdots f(x_n,\theta) \\ &= ETZ, \end{aligned}$$

where we have abbreviated $Z = \sum(\partial/\partial\theta)\ln f(X_j, \theta)$. We know that $EZ = 0$ (compare (3.6)) and $\mathrm{Var}(Z) = nI$, by independence of the X_j. We will now need the following tool (which has many other uses):

Exercise 3.11. Establish the *Cauchy-Schwarz inequality:* For any two random variables X, Y,

$$|EXY| \le \left(EX^2\right)^{1/2} \left(EY^2\right)^{1/2}.$$

Suggestion: Consider the parabola $f(t) = E(X + tY)^2 \ge 0$ and find its minimum.

Exercise 3.12. Can you also show that we have equality in the CSI precisely if $X = cY$ or $Y = cX$ for some $c \in \mathbb{R}$?

Exercise 3.13. Define the *correlation coefficient* of two random variables X, Y as

$$\rho_{X,Y} = \frac{E(X - EX)(Y - EY)}{\sigma_X \sigma_Y}.$$

Deduce from the CSI that $-1 \le \rho \le 1$. Also, show that $\rho = 0$ if X, Y are independent. (The converse of this statement is not true, in general.)

Since $EZ = 0$, we can write

$$k'(\theta) = ETZ = E(T - ET)Z = E(T - ET)(Z - EZ),$$

and now the CSI shows that

$$k'^2 \le \mathrm{Var}(T)\mathrm{Var}(Z) = nI(\theta)\mathrm{Var}(T),$$

as claimed. □

Exercise 3.14. Observe that the inequality was only introduced in the very last step. Thus, by Exercise 3.12, we have equality in the CR bound precisely if $T - ET$ and Z are multiples of one another. In particular, this must hold for the efficient statistic $T = \overline{X}$ from the coin flip example. Confirm directly that indeed $\overline{X} - \theta = cZ$.

Example 3.9. We saw in Example 3.4 that the MLE for the exponential distribution $f(x) = e^{-x/\theta}/\theta$ $(x \ge 0)$ is given by $T = \widehat{\theta} = \overline{X}$ and that T is unbiased. Is T also efficient? To answer this, we compute the Fisher information: $\ln f = -\ln\theta - x/\theta$, so $-\partial^2 \ln f/\partial\theta^2 = -1/\theta^2 + 2X/\theta^3$, and, taking expectations, we see that $I = 1/\theta^2$. On the other hand, $\mathrm{Var}(T) = (1/n)\mathrm{Var}(X_1)$ and

$$EX_1^2 = \frac{1}{\theta}\int_0^\infty x^2 e^{-x/\theta}\,dx = \theta^2 \int_0^\infty t^2 e^{-t}\,dt = 2\theta^2,$$

by two integrations by parts. This implies that $\mathrm{Var}(X_1) = EX_1^2 - (EX_1)^2 = \theta^2$, and thus $\mathrm{Var}(T) = \theta^2/n = 1/(nI)$, and T is indeed efficient.

Let's now take another look at the uniform distribution from Example 3.5. Its density equals

$$f(x,\theta) = \begin{cases} 1/\theta & 0 < x < \theta \\ 0 & \text{otherwise} \end{cases};$$

recall that the MLE is given by $\widehat{\theta} = \max(X_1, \ldots, X_n)$. We know that $T = \widehat{\theta}$ is not unbiased. Let's try to be more precise here. Since $P(T \leq t) = (t/\theta)^n$, the statistic T has density $f(t) = nt^{n-1}/\theta^n$ $(0 < t < \theta)$. It follows that

$$ET = \frac{n}{\theta^n} \int_0^\theta t^n \, dt = \frac{n}{n+1}\, \theta.$$

Exercise 3.15. Show by a similar calculation that $ET^2 = (n/(n+2))\theta^2$.

In particular, if we introduce

$$U = \frac{n+1}{n}\, T = \frac{n+1}{n} \max(X_1, \ldots, X_n),$$

then this new statistic is unbiased (though it is no longer the MLE for θ). By the exercise,

$$EU^2 = \left(\frac{n+1}{n}\right)^2 ET^2 = \frac{(n+1)^2}{n(n+2)}\, \theta^2,$$

so

$$\mathrm{Var}(U) = EU^2 - (EU)^2 = \left(\frac{(n+1)^2}{n(n+2)} - 1\right)\theta^2 = \frac{\theta^2}{n(n+2)}. \tag{3.11}$$

This looks great: In our previous examples, the variance decayed only at the rate $\sim 1/n$, and here we now have $\mathrm{Var}(U) \lesssim 1/n^2$. Come to think of it, is this consistent with the CR bound? Doesn't Corollary 3.5 say that $\mathrm{Var}(T) \gtrsim 1/n$ for *any* unbiased statistic T? The answer to this is that the whole theory doesn't apply here. The density $f(x,\theta)$ is not continuous (let alone differentiable) as a function of θ; it jumps at $\theta = x$. In fact, the problems can be pinpointed more precisely: (3.6) fails, the integrand equals $-1/\theta^2$, and (3.6) was used to deduce that $EZ = 0$, so the whole argument breaks down. Recall that by our discussion following (3.6), the integration in (3.6) is really only extended over $0 < x < \theta$, so problems with the jump of f are temporarily avoided. (However, I also remarked parenthetically that I would like the set $\{x : f(x,\theta) > 0\}$ to be independent of θ, and this clearly fails here.)

Let's compare U with another unbiased estimator. Let $V = 2\overline{X}$. Since $E\overline{X} = EX_1 = \theta/2$, this is indeed unbiased. It is a continuous analog of the unbiased estimator that we suggested (not very seriously, though) in the urn example from Chapter 2; see pg. 10. We have $\mathrm{Var}(\overline{X}) = \mathrm{Var}(X_1)/n$ and

$$EX_1^2 = \frac{1}{\theta}\int_0^\theta t^2\,dt = \frac{\theta^2}{3},$$

so $\mathrm{Var}(X_1) = \theta^2(1/3 - 1/4) = \theta^2/12$, thus

$$\mathrm{Var}(V) = \frac{\theta^2}{3n}.$$

This is markedly inferior to (3.11). We right away had a bad feeling about V (in Chapter 2); this now receives precise theoretical confirmation.

Exercise 3.16. However, if $n = 1$, then $\mathrm{Var}(V) = \mathrm{Var}(U)$. Can you explain this?

Exercise 3.17. Consider the density

$$f(x,\theta) = \begin{cases} 2x/\theta^2 & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}.$$

(a) Find the MLE $\widehat{\theta}$.
(b) Show that $T = \frac{2n+1}{2n}\widehat{\theta}$ is unbiased.
(c) Find $\mathrm{Var}(T)$.
Suggestion: Proceed as in the discussion above.

Example 3.10. Let's return to the MLE $T = \widehat{\theta} = \overline{X}$ for the Poisson distribution; compare Example 3.3. We saw earlier that this is unbiased. Is T also efficient?

To answer this, we first work out the Fisher information: $-\ln P(X = x,\theta) = -x\ln\theta + \theta + \ln x!$, so by taking two derivatives and then the expectation, we find that $I(\theta) = EX/\theta^2 = 1/\theta$. On the other hand,

$$EX_1^2 = \sum_{k=0}^{\infty} k^2 \frac{\theta^k}{k!}e^{-\theta} = \theta^2 \sum_{k=0}^{\infty} \frac{\theta^k}{k!}e^{-\theta} + EX_1 = \theta^2 + \theta;$$

the first step follows by writing $k^2 = k(k-1) + k$. Thus $\mathrm{Var}(X_1) = \theta$, hence $\mathrm{Var}(T) = \theta/n$, and T is efficient.

Exercise 3.18. In this problem, you should frequently refer to results and calculations from Example 3.4. Consider the density $f(x,\theta) =$

$\theta e^{-\theta x}$ $(x \geq 0)$ and $f(x) = 0$ for $x < 0$. Recall that

$$T = \frac{n-1}{n} Y, \qquad Y = 1/\overline{X}$$

is an unbiased estimator for θ.
(a) Find the Fisher information $I(\theta)$ for this density.
(b) Compute $\mathrm{Var}(T)$; conclude that T is not efficient. (Later we will see that T nevertheless has the smallest possible variance among all unbiased estimators.)

Suggestion: Use the density of Y from (3.4) to work out EY^2, and then ET^2 and $\mathrm{Var}(T)$. Avoid the trap of forgetting that the θ of the present exercise corresponds to $1/\theta$ in (3.4).

Example 3.11. Let's now try to estimate the variance of an $N(0, \sigma)$ distribution. We take $\theta = \sigma^2$ as the parameter labeling this family of densities. Two unbiased estimators come to mind:

$$T_1 = \frac{1}{n} \sum_{j=1}^{n} X_j^2, \qquad T_2 = S^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2$$

We know from Example 3.6 that T_1 is the MLE for θ; see (3.5).

We start out by computing the Fisher information. We have $-\ln f = (1/2)\ln\theta + X^2/(2\theta) + C$, so

$$I(\theta) = -\frac{1}{2\theta^2} + \frac{1}{\theta^3} EX^2 = \frac{1}{2\theta^2}.$$

Next, independence gives $\mathrm{Var}(T_1) = (1/n)\mathrm{Var}(X_1^2)$, and this latter variance we compute as $EX_1^4 - (EX_1^2)^2$.

Exercise 3.19. Show that $EX_1^4 = 3\theta^2$.

Suggestion: Use integration by parts in the resulting integral.

Since $EX_1^2 = \mathrm{Var}(X_1) = \theta$, this shows that $\mathrm{Var}(X_1^2) = 2\theta^2$ and thus $\mathrm{Var}(T_1) = 2\theta^2/n$. So T_1 is efficient.

As for T_2, we recall that $(n-1)S^2/\theta \sim \chi^2(n-1)$ and also that $n-1$ iid $N(0,1)$-distributed random variables have this same distribution. More explicitly, $(n-1)S^2/\theta$ has the same distribution as $Z = \sum_{j=1}^{n-1} Y_j^2$, with Y_j iid and $Y_j \sim N(0,1)$. In particular, the variances agree, and $\mathrm{Var}(Z) = (n-1)\mathrm{Var}(Y_1^2) = 2(n-1)$, by the calculation we just did. Thus

$$\mathrm{Var}(S^2) = 2(n-1)\frac{\theta^2}{(n-1)^2} = \frac{2\theta^2}{n-1},$$

and this estimator is not efficient (it comes very close though).

If we had used n instead of the slightly unexpected $n-1$ in the denominator of the formula defining S^2, the resulting estimator $Y_3 = \frac{n-1}{n}S^2$ has variance

$$\operatorname{Var}(Y_3) = \frac{2(n-1)\theta^2}{n^2} = \frac{n-1}{n}\,\frac{1}{nI(\theta)}. \tag{3.12}$$

This, of course, does not contradict the CR bound from Corollary 3.5: this estimator is not unbiased. On the contrary, everything is in perfect order, we only need to refer to Theorem 3.4, which handles this situation. Since $k(\theta) = EY_3 = (n-1)\theta/n$, we have $k'^2 = ((n-1)/n)^2$, and the variance from (3.12) is in fact slightly larger (by a factor of $n/(n-1)$) than the lower bound provided by the theorem.

Exercise 3.20. Consider a random sample drawn from an $N(\theta, 1)$ distribution. Show that (the MLE) $\overline{X}$ is an efficient estimator for θ.