Chapter 7. Statistical Estimation

7.7: Properties of Estimators II

Slides (Google Drive) | Alex Tsun | Video (YouTube)

We'll discuss even more desirable properties of estimators. Last time we talked about bias, variance, and MSE. Bias measured whether or not, in expectation, our estimator was equal to the true value of θ. MSE measured the expected squared difference between our estimator and the true value of θ. If our estimator was unbiased, then the MSE of our estimator was precisely the variance.

7.7.1 Consistency

Definition 7.7.1: Consistency

An estimator $\hat{\theta}_n$ (depending on n iid samples) of θ is said to be **consistent** if it converges (in probability) to θ. That is, for any $\varepsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$$

Basically, as $n \to \infty$, $\hat{\theta}_n$ in the limit will be extremely close to θ.

As usual, we'll do some examples to see how to show this.

Example(s)

Recall that, if $x_1, ..., x_n$ are iid realizations from (continuous) Unif$(0, \theta)$, then

$$\hat{\theta}_n = \hat{\theta}_{n,MoM} = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i$$

Let $\varepsilon > 0$. Show that $\hat{\theta}_n$ is a consistent estimator of θ.

Solution

Since $\hat{\theta}_n$ is unbiased, we have that

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = \mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]| > \varepsilon\right)$$

because we can replace θ with the expected value of the estimator. Now, we can apply Chebyshev's inequality (6.1) to see that

$$\mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]| > \varepsilon\right) \leq \frac{\mathsf{Var}\left(\hat{\theta}_n\right)}{\varepsilon^2}$$

Now, we can take out the 2^2 from the estimator's expression and are left only with the variance of the sample

mean, which is always just $\frac{\sigma^2}{n} = \frac{\mathsf{Var}(x_i)}{n}$.

$$\mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]| > \varepsilon\right) \leq \frac{\mathsf{Var}\left(\hat{\theta}_n\right)}{\varepsilon^2} = \frac{2^2\mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)}{\varepsilon^2} = \frac{4 \cdot \mathsf{Var}\left(x_i\right)/n}{\varepsilon^2}$$

So now we take the limit with this expression.

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \leq \lim_{n\to\infty} \frac{4 \cdot \mathsf{Var}\left(x_i\right)/n}{\varepsilon^2} = 0$$

So, $\hat{\theta}_{n,MoM}$ is a consistent estimator of θ. □

We're also going to show that the MLE estimator is consistent!

Example(s)

Recall that, if $x_1, ..., x_n$ are iid realizations from (continuous) Unif$(0, \theta)$, then

$$\hat{\theta}_n = \hat{\theta}_{n,MLE} = \max\{x_1, ..., x_n\}$$

Let $\varepsilon > 0$. Show that $\hat{\theta}_n$ is a consistent estimator of θ.

Solution

In this case, we cannot use Chebyshev's inequality unfortunately, because the maximum likelihood estimator is not unbiased. The CDF for $\hat{\theta}_n$ is

$$F_{\hat{\theta}_n}(t) = \mathbb{P}\left(\hat{\theta}_n \leq t\right)$$

which is the probability that each individual sample is less than t because only in that case will the max be less than t, and we have independence so we can say

$$\mathbb{P}\left(\hat{\theta}_n \leq t\right) = \mathbb{P}\left(X_1 \leq t\right)\mathbb{P}\left(X_2 \leq t\right)...\mathbb{P}\left(X_n \leq t\right)$$

This is just the CDF of X_i to the n-th power, where the CDF of Unif$(0, \theta)$ is just $\frac{t}{\theta}$ (see the distribution sheet):

$$F_{\hat{\theta}_n}(t) = F_X^n(t) = \begin{cases} 0, & t < 0 \\ (\frac{t}{\theta})^n, & 0 \leq t \leq \theta \\ 1, & t > 0 \end{cases}$$

There are two ways we can have the absolute value from before be greater than epsilon

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = \mathbb{P}\left(\hat{\theta}_n > \theta + \varepsilon\right) + \mathbb{P}\left(\hat{\theta}_n < \theta - \varepsilon\right)$$

The first term is 0, because there's no way our estimator is greater than $\theta + \varepsilon$, as it's never going to be greater than θ by definition (the samples are between 0 and θ so there's no way the max of the samples is greater than θ). So, now we can just use the CDF on the right term, and just plug in for t:

$$\mathbb{P}\left(\hat{\theta}_n > \theta + \varepsilon\right) + \mathbb{P}\left(\hat{\theta}_n < \theta - \varepsilon\right) = \mathbb{P}\left(\hat{\theta}_n < \theta - \varepsilon\right) = \begin{cases} (\frac{\theta-\varepsilon}{\theta})^n, & \varepsilon < \theta \\ 0, & \varepsilon \geq \theta \end{cases}$$

We can assume that ε is less than θ because we really only care when ε is very very small, so we have that

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = \left(\frac{\theta - \varepsilon}{\theta}\right)^n$$

Thus, when we take the limit as n approaches infinity, we see that in the parenthesis, we have a number less than 1, and we raise it to the n-th power, so it goes to 0

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$$

So, $\hat{\theta}_{n,MLE}$ is also a consistent estimator of θ.

□

Now we've seen that, even though the MLE and MoM estimators of θ given iid samples from Unif$(0, \theta)$ are different, they are both consistent! That means, as $n \to \infty$, they will both converge to the true parameter θ. This is clearly a good property of an estimator.

7.7.2 Consistency vs Unbiasedness

You may be wondering, what's the difference between consistency and unbiasedness? I, for one, was very confused about the difference for a while as well. There is, in fact, a subtle difference, which we'll see by comparing estimators for θ in the continuous Unif$(0, \theta)$ distribution.

Unbiased?	Consistent?	Example
Yes	Yes	$\hat{\theta}_{MoM}$
Yes	No	$2X_1$
No	Yes	$\hat{\theta}_{MLE}$
No	No	$1/X_1^2$

1. For instance, an unbiased and consistent estimator was the MoM for the uniform distribution: $\hat{\theta}_{n,MoM} = 2\bar{x}$. We proved it was unbiased in 7.6, meaning it is correct in expectation. It converges to the true parameter (consistent) since the variance goes to 0.

2. However, if you ignore all the samples and just take the first one and multiply it by 2, $\hat{\theta} = 2X_1$, it is unbiased (as it is $2 \cdot \frac{\theta}{2}$), but it's not consistent; our estimator doesn't get better and better with more n because we're not using all n samples. Consistency requires that as we get more samples, we approach the true parameter.

3. Biased but consistent, on the other hand, was the MLE estimator. We showed its expectation was $\frac{n}{n+1}\theta$, which is actually "asymptotically unbiased" since $\mathbb{E}\left[\hat{\theta}_{n,MLE}\right] = \frac{n}{n+1}\theta \to \theta$ as $n \to \infty$. It does get better and better as $n \to \infty$.

4. Neither unbiased nor consistent would just be some random expression, such as $\hat{\theta} = \frac{1}{X_1^2}$.

7.7.3 Efficiency

To take about our last topic, efficiency, we first have to define Fisher Information. Efficiency says that our estimator has as low variance as possible. This property combined with consistency and unbiasedness mean that our estimator is on target (unbiased), converges to the true parameter (consistent), and does so as fast as possible (efficient).

7.7.3.1 Fisher Information

Definition 7.7.2: Fisher Information

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). The **Fisher Information** of the parameter θ is defined to be:

$$I(\theta) = n \cdot \mathbb{E}\left[\left(\frac{\partial \ln L(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ln L(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]$$

where $L(\mathbf{x} \mid \theta)$ denotes the likelihood of the data given parameter θ (defined in 7.1). From Wikipedia, it "is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability X depends".

That written definition is definitely a mouthful, but if you stop and parse it, you'll see it's not too bad to compute. We always take the second derivative of the log-likelihood to confirm that our MLE was a maximizer; now all you have to do is take the expectation to get the Fisher Information. There's no way though that I can interpret the negative expected value of the second derivative of the log-likelihood, it's just too gross and messy.

7.7.3.2 The Cramer-Rao Lower Bound (CRLB) and Efficiency

Why did we define that nasty Fisher information? (Actually, it's much worse when θ is a vector instead of a single number, as the second derivative becomes a matrix of second partial derivatives). It would be great if the mean squared error of an estimator $\hat{\theta}$ was a low as possible. The Cramer-Rao Lower Bound actually gives a lower bound on the variance on any unbiased estimator $\hat{\theta}$ for θ. That is, if $\hat{\theta}$ is any unbiased estimator for θ, there is a minimum possible variance (variance = MSE for unbiased estimators). And if your estimator achieves this lowest possible variance, it is said to be **efficient**. This is also a highly desirable property of estimators. The bound is called the Cramer-Rao Lower Bound.

Definition 7.7.3: Cramer-Rao Lower Bound (CRLB)

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). If $\hat{\theta}$ is an *unbiased* estimator for θ, then

$$\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is the Fisher information defined earlier. What this is saying is, for any unbiased estimator $\hat{\theta}$ for θ, the variance (=MSE) is at least $\frac{1}{I(\theta)}$. If we achieve this lower bound, meaning our variance is exactly equal to $\frac{1}{I(\theta)}$, then we have the best variance possible for our estimate. That is, we have the **minimum variance unbiased estimator (MVUE)** for θ.

Since we want to find the lowest variance possible, we can look at this through the frame of finding the estimator's efficiency.

Definition 7.7.4: Efficiency

Let $\hat{\theta}$ be an unbiased estimator of θ. The **efficiency** of $\hat{\theta}$ is

$$e(\hat{\theta}, \theta) = \frac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)} \leq 1$$

This will always be between 0 and 1 because if your variance is equal to the CRLB, then it equals 1, and anything greater will result in a smaller value. A larger variance will result in a smaller efficiency, and we want our efficiency to be as high as possible (1).
An *unbiased* estimator is said to be **efficient** if it achieves the CRLB - meaning $e(\hat{\theta}, \theta) = 1$. That is, it could not possibly have a lower variance. Again, the CRLB is not guaranteed for biased estimators.

That was super complicated - let's see how to verify the MLE of Poi(θ) is efficient. It looks scary - but it's just messy algebra!

Example(s)

Recall that, if $x_1, ..., x_n$ are iid realizations from $X \sim \text{Poi}(\theta)$ (recall $\mathbb{E}[X] = \mathsf{Var}(X) = \theta$), then

$$\hat{\theta} = \hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Is $\hat{\theta}$ efficient?

Solution

First, you have to check that it's unbiased, as the CRLB only holds for unbiased estimators...

$$\mathbb{E}\left[\hat{\theta}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \mathbb{E}[x_i] = \theta$$

...which it is! Otherwise, we wouldn't be able to use this bound. We also need to compute the variance. The variance of the sample mean (the estimator) is just $\frac{\sigma^2}{n}$, and the variance of a Poisson is just θ.

$$\mathsf{Var}\left(\hat{\theta}\right) = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{\mathsf{Var}(x_i)}{n} = \frac{\theta}{n}$$

Then, we're going to compute that weird Fisher Information, which gives us the CRLB, and see if our variance matches. Remember, we take the second derivative of the log-likelihood, which we did earlier in 7.2

so we're just going to copy over the answer.

$$\frac{\partial^2}{\partial\theta^2}\ln L(\mathbf{x}\mid\theta)=\sum_{i=1}^{n}-\frac{x_i}{\theta^2}$$

Then, we need to take the expected value of this. It turns out, with some algebra, you get $-\frac{n}{\theta}$.

$$\mathbb{E}\left[\frac{\partial^2\ln L(x\mid\theta)}{\partial\theta^2}\right]=\mathbb{E}\left[\sum_{i=1}^{n}-\frac{x_i}{\theta^2}\right]=-\frac{1}{\theta^2}\sum_{i=1}^{n}\mathbb{E}\left[x_i\right]=-\frac{1}{\theta^2}n\theta=-\frac{n}{\theta}$$

Our Fisher Information was the **negative** expected value of the second derivative of the log-likelihood, so we just flip the sign to get $\frac{n}{\theta}$.

$$I(\theta)=-\mathbb{E}\left[\frac{\partial^2\ln L(\mathbf{x}\mid\theta)}{\partial\theta^2}\right]=\frac{n}{\theta}$$

.

Finally, our efficiency is the inverse of the Fisher Information over the variance:

$$e(\hat{\theta},\theta)=\frac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)}=\frac{(\frac{n}{\theta})^{-1}}{\frac{\theta}{n}}=1$$

Thus, we've shown that, since our efficiency is 1, our estimator is efficient. That is, it has the best possible variance among all unbiased estimators of θ. This, again, is a really good property that we want to have.

To reiterate, this means we cannot possibly do better in terms of mean squared error. Our bias is 0, and our variance is as low as it can possibly go. The sample mean is the unequivocally best estimator for a Poisson distribution, in terms of efficiency, in terms of bias, and MSE (it also happens to be consistent, so there are a lot of good things).

As you can see, showing efficiency is just a bunch of tedious calculations!

□