Chapter 7. Statistical Estimation

7.6: Properties of Estimators I

(From "Probability & Statistics with Applications to Computing" by Alex Tsun)

Now that we have all these techniques to compute estimators, you might be wondering which one is the "best". Actually, a better question would be: how can we determine which *estimator* is "better" (rather than the technique)? There are even more different ways to estimate besides MLE/MoM/MAP, and in different scenarios, different techniques may work better. In these notes, we will consider some properties of estimators that allow us to compare their "goodness".

7.6.1 Bias

The first estimator property we'll cover is Bias. The bias of an estimator measures whether or not in expectation, the estimator will be equal to the true parameter.

Definition 7.6.1: Bias

Let $\hat{\theta}$ be an estimator for θ. The **bias** of $\hat{\theta}$ as an estimator for θ is

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

If

- $\text{Bias}(\hat{\theta}, \theta) = 0$, or equivalently $\mathbb{E}\left[\hat{\theta}\right] = \theta$, then we say $\hat{\theta}$ is an **unbiased** estimator of $\hat{\theta}$.
- $\text{Bias}(\hat{\theta}, \theta) > 0$, then $\hat{\theta}$ typically overestimates θ.
- $\text{Bias}(\hat{\theta}, \theta) < 0$, then $\hat{\theta}$ typically underestimates θ.

Let's go through some examples!

Example(s)

First, recall that, if $x_1, ..., x_n$ are iid realizations from $\text{Poi}(\theta)$, then the MLE and MoM were both the sample mean.

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

Show that $\hat{\theta}$ is an unbiased estimator of θ.

Solution

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[x_i\right] && \text{[LoE]} \\
&= \frac{1}{n}\sum_{i=1}^{n} \theta && \left[\mathbb{E}\left[\text{Poi}(\theta)\right] = \theta\right] \\
&= \frac{1}{n} n\theta \\
&= \theta
\end{aligned}$$

This makes sense: the average of your samples should be "on-target" for the true average! □

Example(s)

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

$$\hat{\theta}_{MLE} = x_{\max} \qquad \hat{\theta}_{MoM} = 2 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sure, $\hat{\theta}_{MLE}$ maximizes the likelihood, so in a way $\hat{\theta}_{MLE}$ is better than $\hat{\theta}_{MOM}$. But, what are the biases of these estimators? Before doing any computation: do you think $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MoM}$ are overestimates, underestimates, or unbiased?

Solution I actually think $\hat{\theta}_{MoM}$ is spot-on since the average of the samples should be close to $\theta/2$, and multiplying by 2 would seem to give the true θ. On the other hand, $\hat{\theta}_{MLE}$ might be a bit of an underestimate, since we probably wouldn't have θ be exactly the largest (maybe a little larger).

- **Bias of the maximum likelihood estimator.**
  Recall from 5.10 that the density of the largest order statistic (i.e. the maximum of the sample) is

  $$f_{X_{\max}}(y) = n\, F_X^{n-1}(y)\, f_X(y) = n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta}$$

  You could also instead first find the CDF of $X_{\max}$ as

  $$F_{X_{\max}}(y) = \mathbb{P}\left(X_{\max} \le y\right) = \mathbb{P}\left(X_i \le y\right)^n = F_X(y)^n = \left(\frac{y}{\theta}\right)^n$$

  since the max is less than or equal to a value if and only if each of them is, then take the derivative. Using this density function we can compute the expected value of the $\hat{\theta_{MLE}}$ as follows:

  $$\mathbb{E}\left[\hat{\theta}_{MLE}\right] = \mathbb{E}\left[X_{\max}\right] = \int_0^\theta y\left(n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta}\right)dy = \frac{n}{\theta^n}\int_0^\theta y^n dy = \frac{n}{\theta^n}\left[\frac{1}{n+1}y^{n+1}\right]_0^\theta = \frac{n}{n+1}\,\theta$$

  This makes sense because if I had 3 samples from Unif$(0, 1)$ for example, I would expect them at $1/4, 2/4, 3/4$, and so it would be $\dfrac{n}{n+1}$ as my expected max. Similarly, if I had 4 samples, then I would expect them at $1/5, 2/5, 3/5, 4/5$, and so it would again be $\dfrac{n}{n+1}$ as my expected max.

Finally,

$$\text{Bias}(\hat{\theta}_{MLE}, \theta) = \mathbb{E}\left[\hat{\theta}_{MLE}\right] - \theta = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta$$

- **Bias of the method of moments estimator.**

$$\mathbb{E}\left[\hat{\theta}_{MOM}\right] = \mathbb{E}\left[2 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i\right] = \frac{2}{n}\, n\, \frac{\theta}{2} = \theta$$

$$\text{Bias}(\hat{\theta}_{MoM}, \theta) = \mathbb{E}\left[\hat{\theta}_{MoM}\right] - \theta = \theta - \theta = 0$$

- **Analysis of Results**

  This means that $\hat{\theta}_{MLE}$ typically underestimates θ and $\hat{\theta}_{MOM}$ is an unbiased estimator of θ. But something isn't quite right...

  Suppose the samples are $x_1 = 1, x_2 = 9, x_3 = 2$. Then, we would have

$$\hat{\theta}_{MLE} = \max\{1, 9, 2\} = 9 \qquad \hat{\theta}_{MOM} = \frac{2}{3}(1 + 9 + 2) = 8$$

  However, based on our sample, the MoM estimator is impossible. If the actual parameter were 8, then that means that the distribution we pulled the sample from is Unif$(0, 8)$, in which case the likelihood that we get a 9 is 0. But we did see a 9 in our sample. So, even though $\hat{\theta}_{MOM}$ is unbiased, it still yields an impossible estimate. This just goes to show that finding the right estimator is actually quite tricky.

  A good solution would be to "de-bias" the MLE by scaling it appropriately. If you decided to have a new estimator based on the MLE:

$$\hat{\theta} = \frac{n+1}{n}\hat{\theta}_{MLE}$$

  you would now get an unbiased estimator that can't be wrong! But now it does not maximize the likelihood anymore...
  Actually, the MLE is what we say to be "**asymptotically unbiased**", meaning unbiased in the limit. This is because

$$\text{Bias}(\hat{\theta}_{MLE}, \theta) = -\frac{1}{n+1}\theta \to 0$$

  as $n \to \infty$. So usually we might just leave it because we can't seem to win...

□

Example(s)

Recall that if $x_1, \ldots, x_n \sim \text{Exp}(\theta)$ are iid, our MLE and MoM estimates were both the inverse sample mean:

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

What can you say about the bias of this estimator?

Solution

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}\right] &= \mathbb{E}\left[\frac{n}{\sum_{i=1}^{n} x_i}\right] \\
&\geq \frac{n}{\sum_{i=1}^{n} \mathbb{E}\left[x_i\right]} && \text{[Jensen's inequality]} \\
&= \frac{n}{\sum_{i=1}^{n} \frac{1}{\theta}} && \left[\mathbb{E}\left[\text{Exp}(\theta)\right] = \frac{1}{\theta}\right] \\
&= \frac{n}{n\frac{1}{\theta}} \\
&= \theta
\end{aligned}$$

The inequality comes from Jensen's (section 6.3): since $g(x_1, \ldots, x_n) = \frac{1}{\sum_{i=1}^{n} x_i}$ is convex (at least in the positive octant when all $x_i \geq 0$), we have that $\mathbb{E}\left[g(x_1, \ldots, x_n)\right] \geq g(\mathbb{E}\left[x_1\right], \mathbb{E}\left[x_2\right], \ldots, \mathbb{E}\left[x_n\right])$. It is convex for a reason similar to that $\frac{1}{x}$ is a convex function. So $\mathbb{E}\left[\hat{\theta}\right] \geq \theta$ systematically, and we typically have an overestimate. □

7.6.2 Variance and Mean Squared Error

We are often also interested in how much a estimator varies (we would like it to be unbiased and have small variance to that it is more accurate). One metric that captures this property of estimators is an estimators variance.

The variance of an estimator $\hat{\theta}$ is

$$\mathsf{Var}\left(\hat{\theta}\right) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right])^2\right]$$

This is just the definition of variance applied to the random variable $\hat{\theta}$ and isn't actually a new definition.

But maybe instead of just computing the variance, we want a slightly different metric which instead measures the squared difference from the *actual* estimator and not just its expectation:

$$\mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

We call this property the mean squared error (MSE), and it is related to both bias and variance! Look closely at the difference: if $\hat{\theta}$ is unbiased, then $\mathbb{E}\left[\hat{\theta}\right] = \theta$ and the MSE and variance are actually equal!

Definition 7.6.2: Mean Squared Error

The mean squared error of an estimator $\hat{\theta}$ of θ is

$$\text{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

If $\hat{\theta}$ is an unbiased estimator of θ (i.e. $\mathbb{E}\left[\hat{\theta}\right] = \theta$), then you can see that $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right)$. In fact, in general $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2$.

This leads to what is known as the "Bias-Variance Tradeoff" in machine learning and statistics. Usually, we want to minimize MSE, and these two quantities are often inversely related. That is, decreasing one leads to an increase in the other, and finding the balance will minimize the MSE. It's hard to see why that might be the case since we aren't working with as complex of estimators (we're just learning the basics!).

Proof of Alternate MSE Formula. We will prove that $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2$.

$$\begin{aligned}
\text{MSE}(\hat{\theta}, \theta) &= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] && \text{[def of MSE])} \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] + \mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] && \text{[add and subtract } \mathbb{E}\left[\hat{\theta}\right]\text{]} \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right] + 2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)\right] + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] && [(a+b)^2 = a^2 + 2ab + b^2] \\
&= \mathsf{Var}\left(\hat{\theta}\right) + 0 + \mathbb{E}\left[\text{Bias}(\hat{\theta}, \theta)^2\right] && \text{[def of var, bias, } \mathbb{E}\left[\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right] = 0\text{]} \\
&= \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2
\end{aligned}$$

□

It is highly desirable that the MSE of an estimator is low! We want a small difference between $\hat{\theta}$ and θ. Use the formula above to compute MSE: $\mathsf{Var}\left(\hat{\theta}\right)$ is something we learned how to compute a long time ago, and there are several examples of bias computations above.

Example(s)

First, recall that, if $x_1, ..., x_n$ are iid realizations from $\text{Poi}(\theta)$, then the MLE and MoM were both the sample mean.

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

Compute the MSE of $\hat{\theta}$ as an estimator of θ.

Solution To compute the MSE, let's compute the bias and variance separately. Earlier, we showed that

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}\left[\hat{\theta}\right] - \theta = \theta - \theta = 0$$

Now for the variance:

$$\begin{aligned}
\mathsf{Var}\left(\hat{\theta}\right) &= \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{Var}\left(x_i\right) && \text{[variance adds if independent]} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \theta && [\mathsf{Var}\left(\text{Poi}(\theta)\right) = \theta] \\
&= \frac{1}{n^2} n\theta \\
&= \frac{\theta}{n}
\end{aligned}$$

Finally, using both of those results:

$$\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2 = \frac{\theta}{n} + 0^2 = \frac{\theta}{n}$$

□