NMAI059 – Probability and Statistics 1

Mykhaylo Tyomkyn

Lecture 11 - More on estimators.

In the previous lecture we saw two criteria of 'goodness' of an estimator: being consistent and being (asymptotically) unbiased. Let us complete this discussion by introducing a third useful criterion, and this time it is a quantitative one. Let us from now on assume that we work in a parametric model $(\Omega, \mathcal{F}, \mathbb{P}_\vartheta : \vartheta \in \Theta)$.

Definition 1 (Mean square error (MSE)) *The* mean square error (MSE) *of an estimator T of $g(\theta)$ is the function $\Theta \to \mathbb{R}$ defined as*

$$MSE_\vartheta(T) = \mathbb{E}_\vartheta((T - g(\vartheta))^2).$$

There is a useful alternative expression.

Lemma 1

$$MSE_\vartheta(T) = Var_\vartheta(T) + bias_\vartheta(T)^2.$$

In particular, if T is unbiased, we have $MSE_\vartheta(T) = Var_\vartheta(T)$.

Proof

$$MSE_\vartheta(T) = \mathbb{E}_\vartheta((T - g(\vartheta))^2) = Var_\vartheta(T - g(\vartheta)) + (\mathbb{E}_\vartheta(T - g(\vartheta)))^2 = Var_\vartheta(T) + bias_\vartheta(T)^2,$$

where in the last step we used that, once ϑ has been fixed, $g(\vartheta)$ is just a constant, and so does not affect the variance. □

The goal, of course, is to try to minimize the mean square error. This provides us with a good criterion for comparing different estimators.

Example 1 *Let us deal again with an unknown coin $X \sim Bern(p)$. That is, $\Theta = [0, 1]$, $p = \vartheta$, and $\mathbb{P}_\vartheta = Bern(p)$. Note that we have $p = \mathbb{E}(X)$, and therefore a natural candidate for an estimator of $p = \vartheta$ is the aforementioned empirical mean $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Since it is unbiased, we have*

$$MSE_p(\bar{X}_n) = \mathrm{Var}_p(\bar{X}_n) = \frac{Var_p(X)}{n} = \frac{p(1-p)}{n}.$$

Now, let us consider a different estimator, namely

$$T_n = T_n(X_1, \ldots, X_n) = \frac{(\sum_{i=1}^n X_i) + 1}{n + 2} = \frac{n\bar{X}_n + 1}{n + 2}.$$

Note that $T_n \geq \bar{X}_n$ if and only if $\bar{X}_n \leq 1/2$, so, qualitatively speaking, "T_n pushes the estimate closer to the middle of $\Theta = [0,1]$". Now, let us compute the mean square error of T_n. We have

$$Var_p(T_n) = Var_p\left(\frac{n\bar{X}_n + 1}{n+2}\right) = \frac{n^2}{(n+2)^2} Var_p(\bar{X}_n) = \frac{np(1-p)}{(n+2)^2},$$

and, by linearity of expectation,

$$bias_p(T_n) = \mathbb{E}_p(T_n - p) = \mathbb{E}_p(T_n) - p = \mathbb{E}_p\left(\frac{n\bar{X}_n + 1}{n+2}\right) - p = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2}.$$

So, we obtain

$$MSE_p(T_n) = Var_p(T_n) + bias_p(T_n)^2 = \frac{np(1-p) + (1-2p)^2}{(n+2)^2}.$$

Comparing the just computed values (and skipping a routine algebraic calculation) it turns out that $MSE_p(T_n) \leq MSE_p(\bar{X}_n)$ for any n if $|p - 1/2| \leq 0.35$. So, if we suspect the true value of p to be close to $1/2$, it might be advised to to use T_n instead of $\bar{X}_n$, even though the former is not unbiased. Ultimately though, all criteria are subjective and should be applied according to the situation and our preferences.

Now that we have formulated some properties and parameters of estimators, it is natural to ask how to actually design a good estimator of ϑ in a parametric model. We present two approaches.

Definition 2 (Method of moments (MoM) estimator) *In a parametric model with k parameters[1] (typically $k = 1$ or $k = 2$) and a sample of size n, the* Method of moments *is the rule to choose the estimator $\hat{\vartheta}$ of ϑ satisfying for all $j = 1, \ldots, k$ the equations*

$$\mathbb{E}_{\hat{\vartheta}}(X^j) = \frac{1}{n}\sum_{i=1}^{n} X_i^j.$$

Example 2 (MoM estimator in the Bernoulli model) *Let $k = 1$, $\Theta = [0,1]$, and for $\vartheta = p$, let $\mathbb{P}_\vartheta = Bern(p)$. Note that p is also the expectation. Then the MoM estimator for p is obtained by setting*

$$\hat{p} = \mathbb{E}_{\hat{p}}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n.$$

In other words, for $k = 1$ we infer ϑ from the (hypothetical) assumption that the sample mean is the true mean. Note that, although $\bar{X}_n$ is an unbiased estimator of μ, the resulting estimator for ϑ could be biased.[2]

Example 3 (MoM estimator in the Gaussian model) *Let $k = 2$, $\Theta = \mathbb{R} \times \mathbb{R}^+$ and for $\vartheta = (\mu, \nu)$, let $\mathbb{P}_\vartheta = \mathcal{N}(\mu, \nu)$. Note that for $X \sim \mathcal{N}(\mu, \nu)$ we have $\mathbb{E}(X) = \mu$ and $\mathbb{E}(X^2) = \nu + \mu^2$. So, we use the ansatz*

$$\bar{X}_n = \hat{\mu} \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\nu} + \hat{\mu}^2.$$

[1] Meaning Θ is k-dimensional, e.g. $\Theta = \mathbb{R} \times \mathbb{R}^+$ and $k = 2$

[2] That said, under very mild additional assumptions the MoM estimators are consistent.

From this we obtain

$$\hat{\nu} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2,$$

which can be re-written as[3]

$$\hat{\nu} = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \bar{S}_n^2.$$

To summarize, the MoM estimator in the Gaussian model is $(\hat{\mu}, \hat{\nu}) = (\bar{X}_n, \bar{S}_n^2)$.

The second method of designing an estimator plays a fundamental role in Statistics.

Definition 3 (Likelihood function) *In a parametric statistical model the* likelihood-function $L : \mathbb{R}^n \times \Theta \to \mathbb{R}$ *is the joint pmf/pdf of* $X_1, \ldots, X_n$*, assuming* $X \sim \mathbb{P}_\vartheta$*. That is, with* $x = (x_1, \ldots, x_n)$*, when* X *is discrete we have*

$$L(x, \vartheta) = \prod_{i=1}^{n} p_{X,\vartheta}(x_i) = \prod_{i=1}^{n} \mathbb{P}_\vartheta(X_i = x_i),$$

and when X *is continuous we have*

$$L(x, \vartheta) = \prod_{i=1}^{n} f_{X,\vartheta}(x_i).$$

Definition 4 (Maximum likelihood estimator (MLE)) *The* Maximum likelihood estimator (MLE) *for* ϑ *is given by*

$$\hat{\vartheta} = \max_{\vartheta \in \Theta} L((X_1, \ldots, X_n), \vartheta).$$

In other words, based on our sample, we choose ϑ that gives our sample the highest likelihood among all possible choices of ϑ.

Example 4 (MLE in the Bernoulli model) *Consider the Bernoulli model from Example 2. For a vector* $x \in \{0, 1\}$ *We have*

$$L(x, p) = p^\ell (1-p)^{n-\ell},$$

where ℓ *is the number of* 1*-coordinates of* x*. So, with a fixed parameter* $\ell = \sum_{i=1}^n X_i$*, we need to maximize the function* $p^\ell(1-p)^{n-\ell}$ *with respect to* p*. A helpful idea in this situation is to take the logarithm*[4]*. Since* log *is a strictly monotone increasing function, the maxima of* $\log L(x, \vartheta)$ *is attained at exactly the same values of the argument* ϑ *as the maxima of* $L(x, \vartheta)$*. So, applying the logarithm and differentiating,*[5] *gives*

$$\frac{d(\log L)}{dp} = \frac{d}{dp}\left(\ell \log p + (n-\ell)\log(1-p)\right) = \frac{\ell}{p} - \frac{n-\ell}{1-p},$$

which is 0 *when* $p = \ell/n$*. This is indeed the maximum. Thus, here the MLE again coincides with* $\bar{X}_n$.

[3] Exercise.

[4] Throughout these notes log always refers to the natural logarithm.

[5] Exercise: deal with the boundary cases

The function $\log L$ is referred to as the *log-likelihood function* and denoted $\ell_x(\vartheta)$

Example 5 (MLE in the Gaussian model) *Consider the Gaussian model from Example 3, but parametrize it with $\mathcal{N}(\mu, \sigma^2)$ instead of $\mathcal{N}(\mu, \nu)$ (this is purely for computational convenience, the result will not be affected). We have*

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

Which (μ, σ) maximizes this for a fixed x? Taking the logarithm again gives

$$\ell_x(\mu, \sigma) = -\frac{1}{2} \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma} \right)^2 - n \log \sigma - n \log \sqrt{2\pi}. \tag{1}$$

Differentiating with respect to μ:

$$\frac{\partial \ell}{\partial \mu} = -\frac{1}{2} \sum_{i=1}^{n} 2 \left(\frac{x_i - \mu}{\sigma} \right) \left(-\frac{1}{\sigma} \right) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu),$$

which is 0 when

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

thus the MLE for μ is our 'old friend' $\bar{X}_n$. Now, differentiating (1) with respect to σ, we obtain

$$\frac{\partial \ell}{\partial \sigma} = -\frac{1}{2} \sum_{i=1}^{n} (-2) \frac{(x_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma} = \frac{1}{\sigma^3} \left(\sum_{i=1}^{n} (x_i - \mu)^2 - n\sigma^2 \right),$$

which is 0 when

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2,$$

so we again obtain

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2} = \sqrt{\bar{S}_n^2}.$$

So, in both the Bernoulli and the Gaussian model, we got the same estimators using MoM and MLE. They also coincided with the generic estimators for the mean and the variance $\bar{X}_n$ and $\bar{S}_n^2$. We emphasize though, that this is rather a coincidence, and perhaps a testimony to the 'quality' of both methods. In general, however, they may produce different estimators.