The Maximum Likelihood Estimation For The Variance Is Biased

BIO210 Biostatistics

Extra Reading Material for Lecture 18

Xi Chen
School of Life Sciences
Southern University of Science and Technology

Fall 2024

The random variable $\boldsymbol{X}$ denotes certain metric (*e.g.* height, weight) we are interested in from a population, and $\boldsymbol{X} \sim \mathcal{N}(\mu, \sigma^2)$. We draw a random sample of size n from the population. Like we discussed during the lecture, a random sample of size n can be thought as n **i.i.d.** random variables. That is:

$$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \cdots, \boldsymbol{X}_n \sim \mathcal{N}(\mu, \sigma^2)$$

We have seen that the maximum likelihood estimator for σ^2 is:

$$\hat{\boldsymbol{\sigma^2}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^2$$

Then, what is $\mathbb{E}\left[\hat{\boldsymbol{\sigma^2}}\right]$? If $\mathbb{E}\left[\hat{\boldsymbol{\sigma^2}}\right] = \sigma^2$, it is an unbiased estimator. Otherwise, it is a biased one.

Now let's have a look.

$$\mathbb{E}\left[\hat{\boldsymbol{\sigma^2}}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^2\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\boldsymbol{X}_i^2 - 2\bar{\boldsymbol{X}}\boldsymbol{X}_i + \bar{\boldsymbol{X}}^2)\right]$$
$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^2 - 2\bar{\boldsymbol{X}}\sum_{i=1}^{n}\boldsymbol{X}_i + \sum_{i=1}^{n}\bar{\boldsymbol{X}}^2\right]$$

Note that: $\sum_{i=1}^{n}\boldsymbol{X}_i = n\bar{\boldsymbol{X}}$. Since $\bar{\boldsymbol{X}}$ remains the same for each i, we have $\sum_{i=1}^{n}\bar{\boldsymbol{X}}^2 = n\bar{\boldsymbol{X}}^2$. Replacing the blue terms above, we have:

$$\mathbb{E}\left[\hat{\boldsymbol{\sigma^2}}\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^2 - 2\bar{\boldsymbol{X}}\cdot n\bar{\boldsymbol{X}} + n\bar{\boldsymbol{X}}^2\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^2 - n\bar{\boldsymbol{X}}^2\right]$$
$$= \frac{1}{n}\left(\mathbb{E}\left[\sum_{i=1}^{n}\boldsymbol{X}_i^2\right] - \mathbb{E}\left[n\bar{\boldsymbol{X}}^2\right]\right) \tag{1}$$

Since $\mathbb{V}\text{ar}(\boldsymbol{X}) = \mathbb{E}\left[\boldsymbol{X}^2\right] - (\mathbb{E}[\boldsymbol{X}])^2$, so we have $\mathbb{E}\left[\boldsymbol{X}^2\right] = \mathbb{V}\text{ar}(\boldsymbol{X}) + (\mathbb{E}[\boldsymbol{X}])^2$, then,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} \boldsymbol{X}_i^2\right] &= \mathbb{E}\left[\boldsymbol{X}_1^2\right] + \mathbb{E}\left[\boldsymbol{X}_2^2\right] + \mathbb{E}\left[\boldsymbol{X}_3^2\right] + \cdots + \mathbb{E}\left[\boldsymbol{X}_n^2\right] \\
&= \mathbb{V}\mathrm{ar}\left(\boldsymbol{X}_1\right) + \left(\mathbb{E}\left[\boldsymbol{X}_1\right]\right)^2 + \mathbb{V}\mathrm{ar}\left(\boldsymbol{X}_2\right) + \left(\mathbb{E}\left[\boldsymbol{X}_2\right]\right)^2 + \cdots \\
&\quad + \mathbb{V}\mathrm{ar}\left(\boldsymbol{X}_n\right) + \left(\mathbb{E}\left[\boldsymbol{X}_n\right]\right)^2 \\
&= \sigma^2 + \mu^2 + \sigma^2 + \mu^2 + \cdots + \sigma^2 + \mu^2 \\
&= n\sigma^2 + n\mu^2 \qquad (2)
\end{aligned}$$

Putting equation (2) into equation (1), we have:

$$\begin{aligned}
\mathbb{E}\left[\hat{\boldsymbol{\sigma}^2}\right] &= \sigma^2 + \mu^2 - \frac{1}{n} \cdot \mathbb{E}\left[n\bar{\boldsymbol{X}}^2\right] = \sigma^2 + \mu^2 - \mathbb{E}\left[\bar{\boldsymbol{X}}^2\right] \\
&= \sigma^2 + \mu^2 - (\sigma_{\bar{X}}^2 + \mu_{\bar{X}}^2) \qquad (3)
\end{aligned}$$

According to the central limit theorem, we have $\mu_{\bar{\boldsymbol{X}}} = \mu$ and $\sigma_{\bar{\boldsymbol{X}}}^2 = \dfrac{\sigma^2}{n}$. Therefore, equation (3) becomes:

$$\mathbb{E}\left[\hat{\boldsymbol{\sigma}^2}\right] = \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

Hence, it is not an unbiased estimator.