About the selection of Variance Estimator

2023.4.7 By Hou Dongyu

When in high school, we know that

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

But at present, we need to change the formulas above into

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

The reason is shown below:

If we use the defination of Variance:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Then we can calculate the expectation of S^2 :

$$
\begin{aligned}
E(S^2) &= \frac{1}{n}\sum_{i=1}^{n} E((X_i - \overline{X})^2) = \frac{1}{n}\sum_{i=1}^{n} E((X_i - \mu + \mu + \overline{X})^2) \\
&= \frac{1}{n} E(\sum_{i=1}^{n} (X_i - \mu + \mu + \overline{X})^2) \\
&= \frac{1}{n} E(\sum_{i=1}^{n} ((X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2) \\
&= \frac{1}{n} E(\sum_{i=1}^{n} (X_i - \mu)^2 - 2\sum_{i=1}^{n} (X_i - \mu)(\overline{X} - \mu) + n(\overline{X} - \mu)^2) \\
&= \frac{1}{n} E(\sum_{i=1}^{n} (X_i - \mu)^2 - 2n(\overline{X} - \mu)(\overline{X} - \mu) + n(\overline{X} - \mu)^2) \\
&= \frac{1}{n} E(\sum_{i=1}^{n} (X_i - \mu)^2 - n(\overline{X} - \mu)^2) \\
&= \frac{1}{n} (\sum_{i=1}^{n} E((X_i - \mu)^2) - nE((\overline{X} - \mu)^2)) \\
&= \frac{1}{n} (nVar(X) - nVar(\overline{X})) \\
&= Var(X) - Var(\overline{X}) \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2
\end{aligned}
$$

And we found that

$$
\frac{n-1}{n}\sigma^2 \neq \sigma^2
$$

To avoid using the estimator with bias, we often use the corrected estimator

$$
S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2
$$