Expected Value of the Sample Variance

– Robert Serfling –

The Setting

Suppose that we have a sample $X_1, \ldots, X_n$ of observations from a population having mean μ and variance σ^2. We do *not* assume that the X_i's are mutually independent, but we do suppose that the pairwise covariances are constant, i.e.,

$$\mathrm{Cov}(X_i, X_j) = \gamma \text{ (constant), all } i \neq j \ .$$

Let us consider estimation of σ^2 by the so-called *sample variance*

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \ ,$$

where $\bar{X}$ is defined as usual. *What is the expected value of s^2?*

The Theorem

Theorem. *Under the above assumption,*

$$E(s^2) = \sigma^2 - \gamma \ .$$

PROOF. First, note that s^2 may be written in "U-statistic" form, as an average of the kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$ over all $n(n-1)$ pairs of observations (X_i, X_j) with $i \neq j$. That is,

$$s^2 = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{(X_i - X_j)^2}{2} \ . \qquad (1)$$

To see this, start with the right-hand side of (1) and perform some routine algebraic reduction:

$$\begin{aligned}
RHS &= \frac{1}{2n(n-1)} \left[(n-1) \sum_1^n X_i^2 + (n-1) \sum_1^n X_j^2 - 2 \sum_{i \neq j} X_i X_j \right] \\
&= \frac{1}{n} \sum_1^n X_i^2 - \frac{1}{n(n-1)} \sum_{i \neq j} X_i X_j \\
&= \frac{1}{n} \sum_1^n X_i^2 - \frac{1}{n(n-1)} \left[\sum_{i,j} X_i X_j - \sum_1^n X_i^2 \right] \\
&= \ldots \\
&= \frac{1}{n}(1 + \frac{1}{n-1}) \sum_1^n X_i^2 - \frac{1}{n(n-1)} \left(\sum_1^n X_i \right)^2 \\
&= \ldots
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n-1}\left[\sum_{1}^{n} X_i^2 - n\bar{X}^2\right] \\
&= \ldots \\
&= s^2 \ .
\end{aligned}
$$

(Isn't this *amazing?*)

Now using (1), it follows that

$$
\begin{aligned}
E(s^2) &= E\left[\frac{(X_1 - X_2)^2}{2}\right] \\
&= \frac{1}{2}[2E(X_1^2) - 2E(X_1X_2)] \\
&= E(X_1^2) - E(X_1X_2) \\
&= \sigma^2 - \mathrm{Cov}(X_1, X_2) \ .
\end{aligned}
$$

The Interesting Special Cases

We consider two important cases.

EXAMPLE 1. *Independent X_i's.* In this case the covariance parameter $\gamma = 0$, and we have

$$E(s^2) = \sigma^2 \ ,$$

i.e., s^2 is *unbiased.*

EXAMPLE 2. *Sampling from a finite population.*

With replacement. In this case, the X_i's are independent and the result of Example 1 applies.

Without replacement. Let N be the population size. In this case, the X_i's have pairwise covariance $\gamma = -\sigma^2/(N-1)$ (as seen in class lectures on sample survey theory). Hence

$$
\begin{aligned}
E(s^2) &= \sigma^2 - (-\frac{\sigma^2}{N-1}) = \left(1 + \frac{1}{N-1}\right)\sigma^2 && (2) \\
&= \frac{N}{N-1}\sigma^2 \ . && (3)
\end{aligned}
$$