Estimating the population variance from a random sample

Nguyen Duc Thanh

(Introduction to Probability - Spring 2021)

1 Problem

Let $X_1, X_2...$ be uncorrelated random variables, each having mean μ and variance σ^2. If $\overline{X} = n^{-1}(X_1 + X_2 + ... + X_n)$, show that

$$\mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \sigma^2$$

This fact is of importance in statistics and is used when estimating the population variance from knowledge of a random sample.

2 Proof

Since $\mathbb{E}(X_i) = \mu$ for any $i = 1, 2...$, one has

$$\begin{aligned}\mathbb{E}(\overline{X}) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)\\ &= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)\\ &= \frac{n\mu}{n}\\ &= \mu\end{aligned}$$

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

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \overline{X})^2 &= \sum_{i=1}^{n}\left[(X_i - \mu) - (\overline{X} - \mu)\right]^2\\ &= \sum_{i=1}^{n}\left[(X_i - \mu)^2 - 2(X_i - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2\right]\\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\overline{X} - \mu)^2\\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\overline{X} - \mu)(n\overline{X} - n\mu) + n(\overline{X} - \mu)^2\\ &= \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\end{aligned}$$

Because $\mathbb{E}(X_i) = \mathbb{E}(\overline{X}) = \mu$

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\left[X_i - \mathbb{E}(X_i)\right]^2 - n\left[\overline{X} - \mathbb{E}(\overline{X})\right]^2$$

Then

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) &= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{i=1}^{n}\left[X_i-\mathbb{E}(X_i)\right]^2 - n\left[\overline{X}-\mathbb{E}(\overline{X})\right]^2\right) \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n}\mathbb{E}\left(\left[X_i-\mathbb{E}(X_i)\right]^2\right) - n\mathbb{E}\left(\left[\overline{X}-\mathbb{E}(\overline{X})\right]^2\right)\right]
\end{aligned}$$

Notice that $\mathbb{E}\left(\left[X_i-\mathbb{E}(X_i)\right]^2\right) = var(X_i) = \sigma^2$ and $\mathbb{E}\left(\left[\overline{X}-\mathbb{E}(\overline{X})\right]^2\right) = var(\overline{X})$, the equation can then be simplified as

$$\mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \frac{n}{n-1}\left[\sigma^2 - var(\overline{X})\right]$$

Consider $var(\overline{X})$

$$\begin{aligned}
var(\overline{X}) &= var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) \\
&= \frac{1}{n^2}var\left(\sum_{i=1}^{n}X_i\right) \\
&= \frac{1}{n^2}\left[\mathbb{E}\left(\left[\sum_{i=1}^{n}X_i\right]^2\right) - \mathbb{E}\left(\sum_{i=1}^{n}X_i\right)^2\right] \\
&= \frac{1}{n^2}\left[\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}X_iX_j\right) - \left[\sum_{i=1}^{n}\mathbb{E}(X_i)\right]^2\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(X_iX_j) - \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(X_i)\mathbb{E}(X_j)\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j)\right]
\end{aligned}$$

Given that $X_1, X_2 ...$ are uncorrelated, one has

$$cov(X_i, X_j) = \mathbb{E}(X_iX_j) - \mathbb{E}(X_i)\mathbb{E}(X_j) = 0 \quad \text{whenever} \quad i \neq j$$

Then

$$\begin{aligned}
var(\overline{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\left[\mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}var(X_i) \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Hence

$$\mathbb{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) = \frac{n}{n-1}\left(1-\frac{1}{n}\right)\sigma^2 = \sigma^2$$

□