Unbiased estimator for Cov(X, Y)

This is Exercise 7 of Lecture 10 in MITx: 18.6501x Fundamentals of Statistics.

Let (X_1, Y_1), $(X_2, Y_2), \ldots, (X_n, Y_n) \stackrel{iid}{\sim} (X, Y)$, with$\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_X Y$ and $\mathbb{E}[XY] = \mu_X Y$. That is, each random variable pair (X_1, Y_1) has the same distribution as the random variable pair (X, Y) and the pairs are independent of one another.

Estimating the covariance between X and Y based on observed sequences is useful because non-zero covariance implies dependence between X and Y. In this problem, we study one way to obtain an unbiased estimator for $\mathsf{Cov}(X, Y)$.

Consider the following estimator for the covariance:

$$\widetilde{S}_{XY} = \frac{1}{n}\left(\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)\right),$$

where $\overline{X}_n$ and $\overline{Y}_n$ are the sample mean estimators of μ_X and μ_Y.

First, we note that

$$\begin{aligned}
\mathbb{E}\left[\frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} Y_j\right)}{n}\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i Y_i + \sum_{i=1}^{n}\sum_{i \neq j=1}^{n} X_i Y_j\right] \\
&= \mu_{XY} + (n-1)\mu_X \mu_Y
\end{aligned}$$

where we have used the property that X_i and Y_j are independent whenever $i \neq j$. (In the first of the product of sums, we need to divide by n, and we are, but in the second, we need to divide by $(n-1)$, but we are not, so this multiplier needs to appear in the final answer.)

Then,

$$
\begin{aligned}
\mathbb{E}\left[\widetilde{S}_{XY}\right] &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i}{n}\sum_{j=1}^{n} Y_j - \frac{\sum_{i=1}^{n} Y_i}{n}\sum_{j=1}^{n} X_j + \frac{\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j}{n}\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{j=1}^{n} Y_j}{n}\right] \\
&= \frac{1}{n}\left(n\mu_{XY} - \mu_{XY} + (n-1)\mu_X\mu_Y\right) \\
&= \frac{n-1}{n}\left(\mu_{XY} - \mu_X\mu_Y\right) \\
&= \frac{n-1}{n}\mathsf{Cov}(X, Y)
\end{aligned}
$$

Hence, the estimator is biased, since $\mathbb{E}[\widetilde{S}_{XY}] \neq \mathsf{Cov}(X, Y)$.

We can fix this by multiplying $\widetilde{S}_{XY}$ by $\frac{n}{n-1}$ to obtain the unbiased estimator of $\mathsf{Cov}(X, Y)$:

$$
\widehat{S}_{XY} = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)\right).
$$