Activity Solution: Properties of the Sample Mean

Suppose we would like to estimate the mean μ of our process $X(t)$ using some data $x(1), \ldots, x(N)$. We would like to know to what extent the mean of a sample,

$$\bar{x} = \frac{1}{N} \sum_{t=1}^{N} x(t)$$

constitutes a useful estimator of μ. There are several theorems showing how this estimator behaves asymptotically as $N \to \infty$, but these do not tell us the variance of our estimator for a finite amount of data. We derive here the variance of the sample mean when the data are correlated.

1. *Clickers question:* For i.i.d. data of size N from a random variable X with variance σ_X^2 what is $\mathrm{Var}(\bar{x})$?
   We recall
   $$\mathrm{Var}(\bar{x}) = \frac{\sigma_X^2}{N}.$$

2. *Clickers question:* Consider the variance of the sum of three random variables X_1, X_2, and X_3. Find an expression for $\mathrm{Var}(X_1 + X_2 + X_3)$ in variance and covariance terms.
   We have
   $$\begin{aligned} \mathrm{Var}(X_1 + X_2 + X_3) &= E\left([(X_1 + X_2 + X_3) - E(X_1 + X_2 + X_3)]^2 \right) \\ &= E\left([(X_1 - E(X_1)) + (X_2 - E(X_2)) + (X_3 - E(X_3))]^2 \right) \\ &= \sum_{i=1}^{3} \mathrm{Var}(X_i) + 2 \sum_{i,j\,:\,i<j} \mathrm{Cov}(X_i, X_j). \end{aligned}$$
   and the result follows on expansion of the brackets and application of the expectation operator.

3. What is the general result for the variance of $X_1 + X_2 + \cdots + X_n$?
   In general
   $$\begin{aligned} \mathrm{Var}\left(\sum_{i=1}^{n} X_i \right) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i,j\,:\,i<j} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j). \end{aligned}$$

4. *Clickers question:* Using your result from above, find an expression for $\mathrm{Var}(\bar{x})$ where the data are three observations, $x(1)$, $x(2)$, and $x(3)$, from a stationary process with variance σ_X^2 and acf $\rho(\cdot)$.
Recall that for a stationary process, $Cov(X(i), X(j)) = \sigma_X^2 \rho(i-j)$, where $i-j$ is the lag between times i and j. Hence

$$\begin{aligned}
Var(\bar{x}) =& Var\left(\frac{1}{3}\sum_{t=1}^{3} X(t)\right) \\
=& \frac{1}{9} Var\left(\sum_{t=1}^{9} X(t)\right) \\
=& \frac{1}{9}\left(\sum_{t=1}^{3} Var(X(t)) + 2\sum_{i,j\,:\,i<j} Cov(X(i), X(j))\right) \\
=& \frac{1}{9}\left(\sum_{i=1}^{3} \sigma_X^2 + 2\sum_{i,j\,:\,i<j} \sigma_X^2 \rho(i-j)\right) \\
=& \frac{1}{9}\left(3\sigma_X^2 + 2\sigma_X^2 \sum_{i,j\,:\,i<j}^{3} \rho(i-j)\right) \\
=& \frac{1}{9}\left(3\sigma_X^2 + 2\sigma_X^2(2\rho(1) + \rho(2))\right) \\
=& \frac{\sigma_X^2}{9}\left(3 + 2(2\rho(1) + \rho(2))\right).
\end{aligned}$$

5. Hence show that

$$\mathrm{Var}(\bar{x}) = \frac{\sigma_X^2}{N}\left(1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho(k)\right)$$

where the data are N observations from a stationary process with vari-

ance σ_X^2 and acf $\rho(\cdot)$.

$$
\begin{aligned}
Var(\bar{x}) =& Var\left(\frac{1}{N}\sum_{t=1}^{N} X(t)\right) \\
=& \frac{1}{N^2} Var\left(\sum_{t=1}^{N} X(t)\right) \\
=& \frac{1}{N^2}\left(\sum_{t=1}^{N} Var(X(t)) + 2\sum_{i,j\,:\,i<j} Cov(X(i), X(j))\right) \\
=& \frac{1}{N^2}\left(\sum_{i=1}^{N} \sigma_X^2 + 2\sum_{i,j\,:\,i<j} \sigma_X^2 \rho(i-j)\right) \\
=& \frac{1}{N^2}\left(N\sigma_X^2 + 2\sigma_X^2 \sum_{i,j\,:\,i<j}^{N} \rho(i-j)\right) \\
=& \frac{1}{N^2}\left(N\sigma_X^2 + 2\sigma_X^2((N-1)\rho(1) + (N-2)\rho(2) + \cdots + (1)\rho(N-1)))\right) \\
=& \frac{1}{N^2}\left(N\sigma_X^2 + 2\sigma_X^2 N \sum_{k=1}^{N-1}(1 - \frac{k}{N})\rho(k)\right) \\
=& \frac{\sigma_X^2}{N}\left(1 + 2\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)\rho(k)\right).
\end{aligned}
$$

6. Re-cap what you have done during this activity. What did you learn?
   We have explored properties of the mean of variables with common variance but that are not independent. We have seen that the variance of the mean is not the same as in the i.i.d. case, but is instead scaled by a factor that involves the sum of all cross–correlations between the variables. This, it turns out, may scale either up or down the variance of the mean from the i.i.d. case.
   The learning outcome here relates to an appreciation that estimation with time–dependent data is not going to always be the same as in cases where data are i.i.d.; in particular we explored:

   - *Explain issues regarding estimation of the mean of a time series, being able to identify in particular cases how the variance of the sample mean differs from sampling from uncorrelated data.*