This version: December 28, 2023

Standard errors of sample statistics

This memo provides some intuition for the minimum number of data points needed to compute a correlation and/or volatility. The first section covers the asymptotics, while the second section verifies the asymptotics and provides some simulation results.

1 Asymptotics

1.1 Asymptotics of normal variance

Let $x \in \mathbb{R}^T$ be a vector of iid normal variables.

$$x_t \sim N\left(0, \sigma^2\right)$$

The zero mean is without loss of generality, as linear transformations will not affect the variance or the sample variance. The sample variance is given by:

$$s^2 = \frac{1}{T}\sum_t x_t^2 - \left[\frac{1}{T}\sum_t x_t\right]^2$$

Note that

$$\begin{aligned} Var\left(s^2\right) =& Var\left(\frac{1}{T}\sum_t x_t^2 - \left[\frac{1}{T}\sum_t x_t\right]^2\right) \\ =& E\left[\left(\frac{1}{T}\sum_t x_t^2 - \left[\frac{1}{T}\sum_t x_t\right]^2\right)^2\right] - E\left[\frac{1}{T}\sum_t x_t^2 - \left[\frac{1}{T}\sum_t x_t\right]^2\right]^2 \end{aligned}$$

Asymptotically, the last quantity is $\left(\sigma^2\right)^2$. Then:

$$\begin{aligned} Var\left(s^2\right) + \left(\sigma^2\right)^2 \approx& E\left[\left(\frac{1}{T}\sum_t x_t^2 - \left[\frac{1}{T}\sum_t x_t\right]^2\right)^2\right] \\ =& E\left[\left(\frac{1}{T}\sum_t x_t^2 - \frac{1}{T^2}\left[\sum_t x_t\right]^2\right)^2\right] \\ =& E\left[\frac{1}{T^2}\left(\sum_t x_t^2\right)^2\right] \end{aligned}$$

The last equation assumes large T. Next, note that

$$\frac{\sigma^2}{T}\sum_t \frac{x_t^2}{\sigma^2} \sim \frac{\sigma^2}{T}\chi_T^2$$

This implies:

$$E\left[\left(\frac{1}{T}\sum_t x_t^2\right)^2\right] = \frac{3\left(\sigma^2\right)^2}{T}$$

Therefore:

$$Var\left(s^2\right) \approx \frac{2\left(\sigma^2\right)^2}{T}$$

Unfortunately, a variance (or standard deviation) of a variance presents little intuition. This is remedied asymptotically by using the delta method to determine the variance of the standard deviation. The below is the functional variant of the delta method, applied as such for consistency with the next section (it also makes a good warmup for the next section).

The influence function is $IF = s^2 - \sigma^2$. Then define $g\left(s^2\right) = \left(s^2\right)^{1/2}$. Applying the traditional Taylor expansion:

$$\begin{aligned} g_T\left(s^2\right) \approx & g\left(\sigma^2\right) + g'\left(\sigma^2\right)\left(s^2 - \sigma^2\right) \\ & s.t. \\ g'\left(\sigma^2\right) = & \frac{1}{2}\left(\sigma^2\right)^{-1/2} \end{aligned}$$

Define

$$\begin{aligned} g_T \approx & \frac{1}{T}\sum_t g_t \\ g_t \equiv & \left(\sigma^2\right)^{1/2} + \frac{1}{2}\left(\sigma^2\right)^{-1/2}\left(x_t^2 - \overline{x_t}^2 - \sigma^2\right) \end{aligned}$$

Calculating the variance:

$$\begin{aligned} Var\left(g_t\right) = & Var\left(\frac{1}{2}\left(\sigma^2\right)^{-1/2}\left(x_t^2 - \overline{x_t}^2\right)\right) \\ = & \frac{Var\left(x_t^2 - \overline{x_t}^2\right)}{4\sigma^2} \\ = & \frac{\sigma^2}{2} \end{aligned}$$

Then by the CLT:

$$\sqrt{T}\left(g_T - \left(\sigma^2\right)^{1/2}\right) \rightsquigarrow N\left(0, \frac{\sigma^2}{2}\right)$$

1.2 Multivariate covariance matrix

While the below was inspired by Zepeda-Tello et al 2022, it deviates in that it harnesses use of the Wishart distribution to represent the sample covariance matrix and establish a converging sequence for the purposes of deploying the central limit theorem.

Suppose x_t is p dimensional such that

$$x_t \sim N(0, \Sigma)$$

Then:

$$\frac{1}{T}\sum_t x_t x_t' - \overline{xx}' \sim \frac{1}{T} W_p(\Sigma, T)$$

where W_p is a Wishart distribution. If all we care about is a single pairwise covariance, then

$$V(\sigma_{12}) = \frac{1}{T_{12}}\left(\sigma_{12}^2 + \sigma_1^2\sigma_2^2\right)$$

where $T \geq 2$ and

$$V\left(\sigma_1^2\right) = \frac{2\left(\sigma_1^2\right)^2}{T}$$

as previously calculated. As the sample mean converges in distribution to a normal with the true variance, we have:

$$\begin{aligned}
\sqrt{T_1}\left(\sigma_{1T}^2 - \sigma_1^2\right) &\rightsquigarrow N\left(0, 2\left(\sigma_1^2\right)^2\right) \\
\sqrt{T_2}\left(\sigma_{2T}^2 - \sigma_2^2\right) &\rightsquigarrow N\left(0, 2\left(\sigma_2^2\right)^2\right) \\
\sqrt{T_{12}}\left(\sigma_{12T} - \sigma_{12}\right) &\rightsquigarrow N\left(0, \sigma_{12}^2 + \sigma_1^2\sigma_2^2\right)
\end{aligned}$$

Unfortunately, we cannot directly use the above variances from the Wishart, since the estimates are correlated. The functional variant of the Delta method provides a path forward. Start with a

Taylor approximation of the test statistic:

$$\begin{aligned} g_T =& \left(\frac{\sigma_{12T}^2}{\sigma_{1T}^2\sigma_{2T}^2}\right)^{1/2} \\ \approx& g + \nabla g' \left[IF\right] \\ s.t.& \\ g \equiv& \left(\frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}\right)^{1/2} \\ \nabla g\left(\mu\right) \equiv& \left(\frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}\right)^{1/2} \times \begin{bmatrix} \frac{1}{\sigma_{12}} \\ \frac{-1}{2\sigma_1^2} \\ \frac{-1}{2\sigma_2^2} \end{bmatrix} \\ \left[IF\right] \equiv& \begin{bmatrix} \sigma_{12T} - \sigma_{12} \\ \sigma_{1T}^2 - \sigma_1^2 \\ \sigma_{2T}^2 - \sigma_2^2 \end{bmatrix} \end{aligned}$$

Note that because by the continuous mapping theorem, each of the plug-in estimators converge in probability to their population moments. This implies the approximation is asymptotically exact.

Given the approximation, we can express g_T as:

$$\begin{aligned} g_T \approx& \frac{1}{T}\sum_t g_t \\ s.t.& \\ g_t \equiv& \rho + \left(\frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}\right)^{1/2} \left[\frac{x_{1t}x_{2t} - \overline{x}_1\overline{x}_2 - \sigma_{12}}{\sigma_{12}} - \frac{x_{1t}^2 - \overline{x}_1^2 - \sigma_1^2}{2\sigma_1^2} - \frac{x_{2t}^2 - \overline{x_2}^2 - \sigma_2^2}{2\sigma_2^2}\right] \end{aligned}$$

Now apply the CLT:

$$\sqrt{T}\left(g_T - \rho\right) \rightsquigarrow N\left(0, V\left(g_t\right)\right)$$

To estimate, simply replace all population moments with their plug-in estimators. The solution reconciles with the previous solution.

2 Empirics

2.1 Bootstrapping assessment of asymptotic convergence

For the purposes of this analysis, use the following parameters:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim MVN\left(0, S\right)$$
$$S \equiv diag\left(s\right) * R * diag\left(s\right)$$
$$s \equiv \begin{bmatrix} \frac{0.15^2}{12} \\ \frac{0.15^2}{12} \end{bmatrix}$$
$$R \equiv \begin{bmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{bmatrix}$$

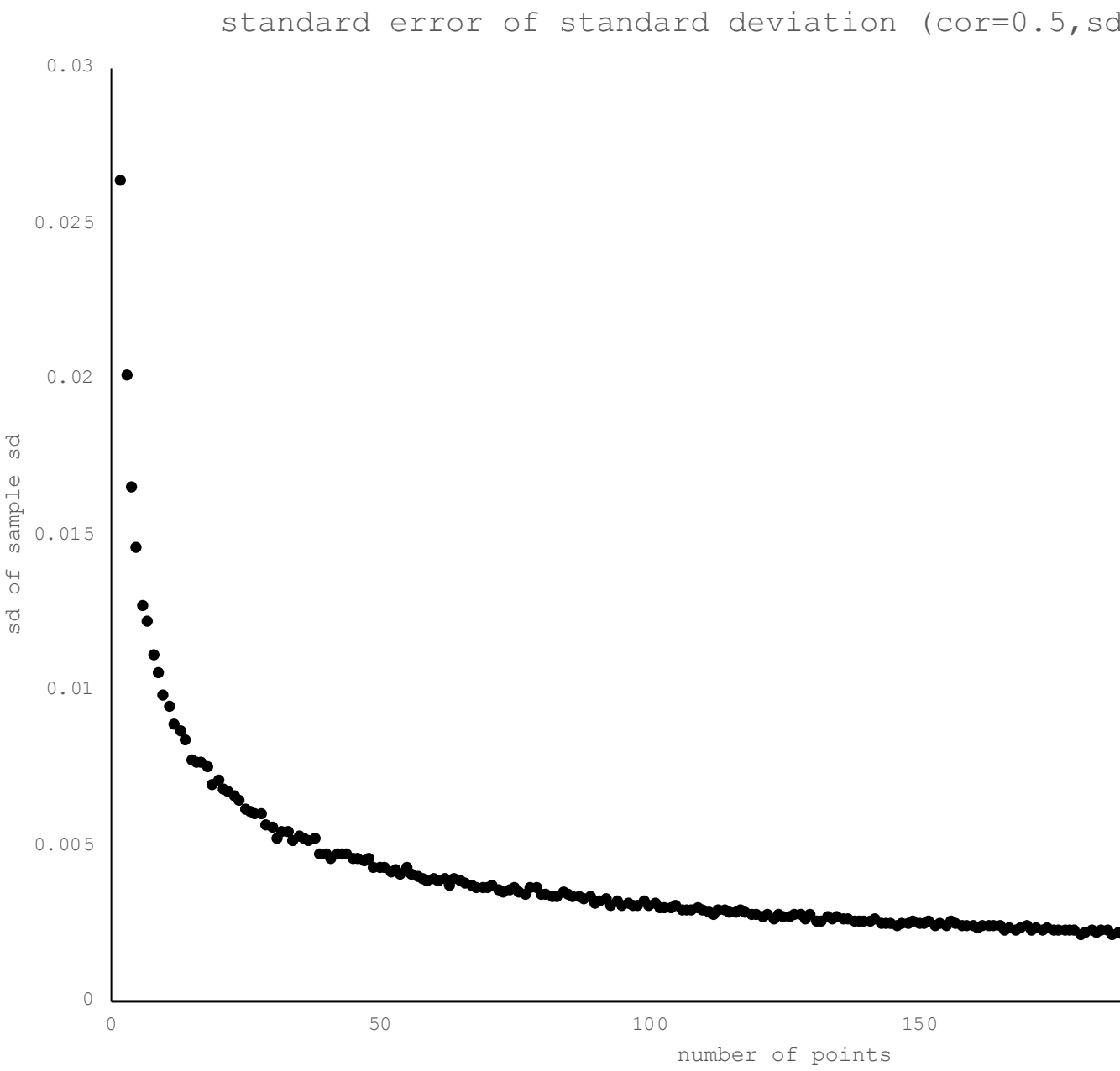
The standard deviation of each series corresponds to approximately 4.3%. Simulate the series 1000 times. For each simulation, compute the asymptotic estimate of the standard error for the sample standard deviation and correlation, and compute the empirical standard deviation of each measure. This leads to the following results:

T	SE(boot σ)	SE(asymp σ)	SE dif (σ)	SE (boot ρ)	SE (asymp ρ)	SE dif (ρ)
3	0.0179	0.0202	(0.0023)	0.6446	0.8455	(0.2009)
4	0.0154	0.0169	(0.0015)	0.5156	0.4827	0.0329
5	0.0138	0.0153	(0.0015)	0.4141	0.4113	0.0028
8	0.0109	0.0114	(0.0005)	0.3111	0.2680	0.0431
10	0.0096	0.0099	(0.0003)	0.2663	0.2376	0.0287
20	0.0068	0.0067	0.0001	0.1743	0.1625	0.0118
50	0.0043	0.0043	0.0000	0.1112	0.1056	0.0057
100	0.0031	0.0031	0.0000	0.0766	0.0745	0.0021
250	0.0019	0.0019	0.0001	0.0477	0.0472	0.0005

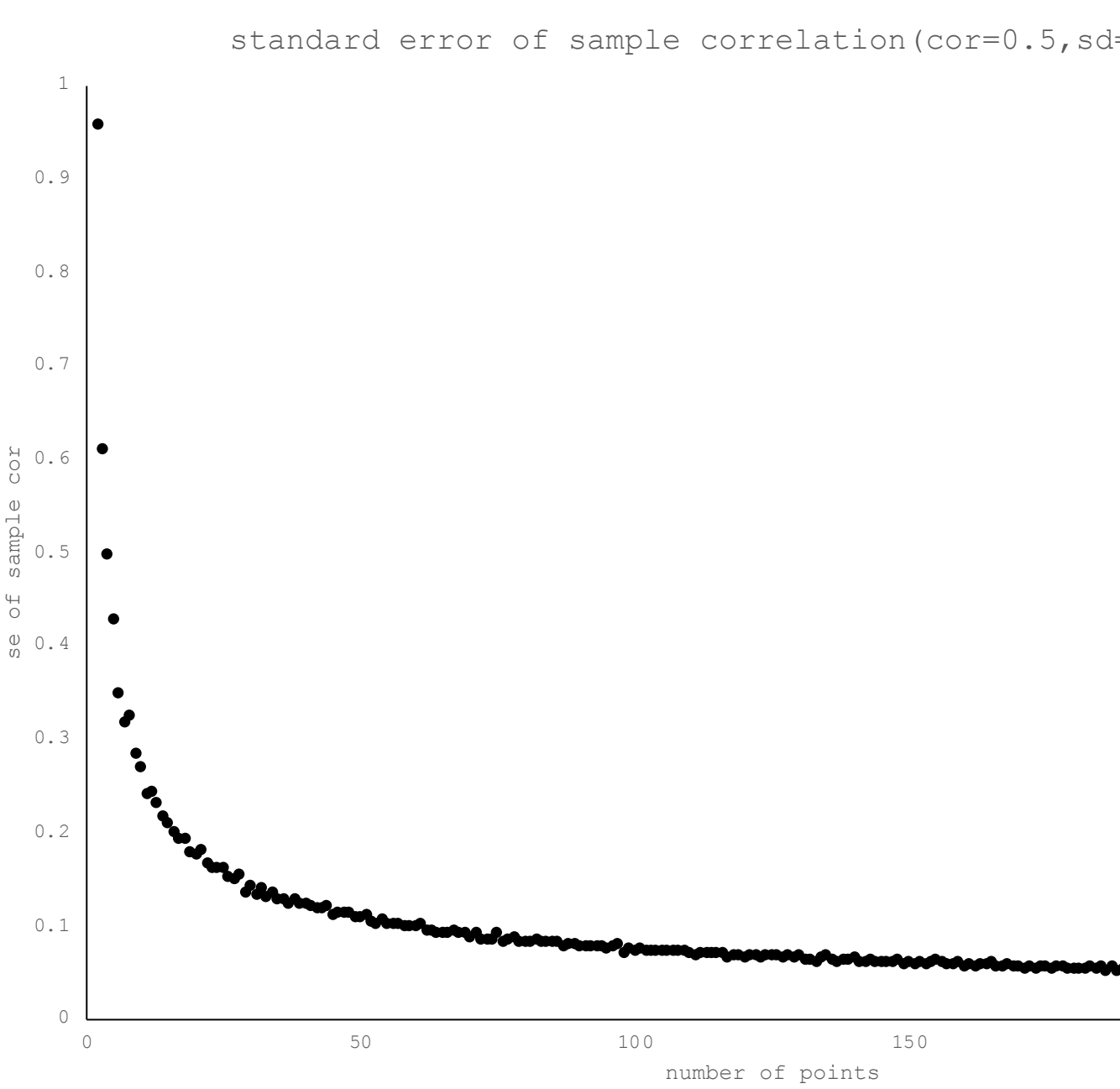
The results verify convergence between the asymptotic and simulation results. . Convergence is faster for standard deviation than for the correlation. The asymptotics seem to require only a few data points to provide effective estimates of the standard errors.

2.2 Aggregate error

In contrast to error produced through the use of asymptotics, the error of the sample statistics remains notable even for larger samples. In addition to the values produced in the previous section, the following plots provide further intuition.



The above plot shows the standard error of the sample standard deviation as a function of the sample size. Given the true standard deviation of 4.3%, the sample volatility starts to become reasonable at about 1.1% at around 8 points of data. On a relative basis, this represents a standard error of roughly 25%.



As with standard deviation, most the sample size gains to precision occur within the first 10-20 data points,. However, the standard error of the correlation is always measured with significant error on a relative basis. Future work could enhance naive correlation estimates (say by incorporating estimates from the factor model).

Appendix: alternate calculation for the asymptotic variance.

Zepeda-Tello et al 2022 directly calculate the variance of the expansion as follows:

$$\begin{aligned} V\left(g_T\right) \approx & V\left(\left(\frac{\sigma_{12}^2}{\sigma_1^2 \sigma_2^2}\right)^{1/2}\left[\frac{x_{1t}x_{2t}-\sigma_{12}}{\sigma_{12}}-\frac{x_{1t}^2-\sigma_1^2}{2\sigma_1^2}-\frac{x_{2t}^2-\sigma_2^2}{2\sigma_2^2}\right]\right) \\ = & V\left(\sum_{t\in 1:T}\left(\frac{\sigma_{12}^2}{\sigma_1^2 \sigma_2^2}\right)^{1/2}\left[\frac{x_{1t}x_{2t}}{T\sigma_{12}}-\frac{x_{1t}^2}{2\sigma_1^2 T}-\frac{x_{2t}^2}{2\sigma_2^2 T}\right]\right) \\ = & T \times V\left(\left(\frac{\sigma_{12}^2}{\sigma_1^2 \sigma_2^2}\right)^{1/2}\left[\frac{x_{1t}x_{2t}}{T\sigma_{12}}-\frac{x_{1t}^2}{2\sigma_1^2 T}-\frac{x_{2t}^2}{2\sigma_2^2 T}\right]\right) \end{aligned}$$

The version in the main text provides more intuition via the CLT application.