Unbiased Estimation of the Standard Deviation for Non-Normal Populations

David E. Giles

Department of Economics, University of Victoria, CANADA

February, 2021

Abstract

The bias of the sample standard deviation as an estimator of the population standard deviation, for a simple random sample of size N from a Normal population, is well documented. Exact and approximate bias corrections appear in the literature, but there has been less discussion of the downward bias of this estimator for non-Normal populations. The appropriate bias correction depends on the kurtosis of the population distribution. We derive and illustrate an approximation for this bias, to $O(N^{-1})$, for several distributions.

Keywords: Standard deviation, unbiased estimation, bias approximation

Mathematics Subject Classifications: 62E15, 62F10

Author contact: 58 Rock Lake Court, L'Amable, ON, CANADA, K0L 2L0

+1-613-332-6833 ; dgiles@uvic.ca

1. INTRODUCTION

Let X follow a distribution, F, with integer moments that are finite, at least up to fourth order. Denote the population central moments by $\mu_j = E\left[(X-\mu_1')^j\right]$, j = 1, 2, 3, …. ; where $\mu_1' = E(X)$ and $Var.(X) = \mu_2 = \sigma^2$, say; and the kurtosis coefficient is $\kappa = (\mu_4/\mu_2^2)$.

Based on a simple random sample of size N, the sample variance is $s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$, where $\bar{x} = \frac{1}{n}\sum_{i=1}^{N} x_i$. For any F with finite first and second moments, $E(s^2) = \sigma^2$ and $E(\bar{x}) = \mu_1'$. In the special case where F is Normal, the sampling distributions of both s^2 and s itself are well known. For example, for the latter see Holtzman (1950). In particular, the bias of s as an estimator of σ, and various approximations to this bias, have been examined in detail *for the Normal case* – *e.g*., see Bolch (1968), Brugger (1969), Cureton (1968), D'Agostino (1970), Gurland and Tripathi (1971), Markowitz (1968) and Stuart (1969),

However, if F is non-Normal, then although s^2 is still an unbiased estimator of σ^2, s is a downward-biased estimator of σ in finite samples, by Jensen's inequality. The *magnitude* of this bias has not been established, in general, and we explore this problem here.

2. MAIN RESULT

Under standard regularity conditions, both $\left(\bar{x}-\mu_1'\right)$ and $\left(s^2-\sigma^2\right)$ are $O_p(N^{-1/2})$; and note that we can write $s = \sigma[1+(s^2-\sigma^2)/\sigma^2]^{1/2}$. So, by the generalized binomial theorem (or using the Maclaurin expansion), we have:

$$s = \sigma\left[1+\frac{1}{2\sigma^2}(s^2-\sigma^2)-\frac{1}{8\sigma^4}(s^2-\sigma^2)^2+\frac{1}{16\sigma^6}(s^2-\sigma^2)^3-\frac{5}{128\sigma^8}(s^2-\sigma^2)^4+\cdots\right]. \qquad (1)$$

Convergence of the infinite series in (1) requires that $|(s^2-\sigma^2)/\sigma^2| < 1$, and this condition will be satisfied for large N as s^2 is a consistent estimator of σ^2. However, convergence is not required for the approximation that follows.

Retaining terms in the expected value of (1) up to $O(N^{-1})$, we have

$$E(s) = \sigma\left[1+\frac{1}{2\sigma^2}E(s^2-\sigma^2)-\frac{1}{8\sigma^4}E[(s^2-\sigma^2)]^2\right]+O\left(N^{-3/2}\right) \ . \qquad (2)$$

Now, $E(s^2-\sigma^2) = 0$, and from Angelova (2012, eq. (19)),

$$E[(s^2-\sigma^2)]^2 = \left[\frac{(\mu_4-\mu_2^2)}{N}+\frac{2\mu_2^2}{N(N-1)}\right]. \qquad (3)$$

This yields the approximation,

$$E(s) \simeq \sigma\left[1 - \frac{1}{8}\left[\frac{\kappa-1}{N} + \frac{2}{N(N-1)}\right]\right] = (\sigma/C_N^*)\ , \tag{4}$$

where

$$C_N^* = [8N(N-1)]/[8N(N-1) - (N-1)(\kappa-3) - 2N]\ . \tag{5}$$

So, $\hat{\sigma} = C_N^* s$ is an unbiased estimator of σ, to $O(N^{-1})$. For a Normal population, the corresponding scale factor for $\hat{\sigma}$ to be *exactly* unbiased for s is known to be

$$C_N = \Gamma[(N-1)/2]\sqrt{(N-1)/2}\,/\Gamma[N/2]. \tag{6}$$

Using (4), and the fact that $E(s^2) = \sigma^2$, we see immediately that $var(s) \simeq \sigma^2\left(C_N^{*2} - 1\right)/C_N^{*2}$ and $var(\hat{\sigma}) \simeq \sigma^2\left(C_N^{*2} - 1\right)$, each to $O(N^{-1})$.

3. DISCUSSION

Some early tabulations for C_N by various authors are discussed by Jarrett (1968). Also, see Holtzman (1950), Bolch (1968), and Gurland and Tripathi (1971). Table 1 compares values of C_N^* with C_N, and with two approximations to C_N suggested by Gurland and Tripathi, for the Normal case. Values of C_N^*, for three other common population distributions, and various values of N, also appear in Table 1. An extended table can be downloaded as an Excel spreadsheet from https://github.com/DaveGiles1949/My-Documents.

The accuracy of C_N^* relative to the exact C_N is apparent in Table 1 – even for sample sizes as small as N = 15. This lends credence to the accuracy of the C_N^* values for the other distributions, which show that this bias adjustment factor increases with the degree of kurtosis, but decreases (to 1) rapidly as N increases.

In practice, the form of the population distribution, and hence the value of κ, may be unknown. In this case an estimate of κ – such as the fourth standardized central sample moment, b_2 – can be used. Johnson and Lowe (1979) show that $b_2 \leq N$, so the corresponding estimate of C_N^* satisfies $(\frac{16}{13}) \leq \widehat{C_N^*} < (\frac{8}{7})$ for $N \geq 2$. In particular, $\widehat{C_N^*} > 1$, as required, but the order of magnitude of our main unbiasedness result is then only approximate.

TABLE 1

VALUES OF C_N AND C_N^* FOR VARIOUS POPULATIONS[1]

N	C_N			C_N^*				
	Normal			Normal	Logistic	Laplace	Uniform	Exponential
	Exact	GT(5)(6)	GT(7)	(κ = 3)	(κ = 4.2)	(κ = 6)	(κ = 1.8)	(κ = 9)
2	1.2533	1.2649	1.2500	1.3333	1.4815	1.7778	1.2121	2.6667
3	1.1284	1.1314	1.1250	1.1429	1.2121	1.3333	1.0811	1.6000
4	1.0854	1.0864	1.0833	1.0909	1.1374	1.2152	1.0480	1.3714
5	1.0638	1.0643	1.0625	1.0667	1.1019	1.1594	1.0336	1.2698
6	1.0509	1.0512	1.0500	1.0526	1.0811	1.1268	1.0256	1.2121
7	1.0424	1.0425	1.0417	1.0435	1.0673	1.1053	1.0207	1.1748
8	1.0362	1.0363	1.0357	1.0370	1.0576	1.0900	1.0173	1.1487
9	1.0317	1.0317	1.0313	1.0323	1.0503	1.0787	1.0148	1.1294
10	1.0281	1.0282	1.0278	1.0286	1.0447	1.0698	1.0129	1.1146
11	1.0253	1.0253	1.0250	1.0256	1.0402	1.0628	1.0115	1.1028
12	1.0230	1.0230	1.0227	1.0233	1.0365	1.0571	1.0103	1.0932
13	1.0210	1.0210	1.0208	1.0213	1.0335	1.0523	1.0094	1.0852
14	1.0194	1.0194	1.0192	1.0196	1.0309	1.0482	1.0086	1.0785
15	1.0180	1.0180	1.0179	1.0182	1.0287	1.0448	1.0079	1.0728
16	1.0168	1.0168	1.0167	1.0169	1.0267	1.0418	1.0073	1.0679
17	1.0157	1.0157	1.0156	1.0159	1.0251	1.0392	1.0068	1.0635
18	1.0148	1.0148	1.0147	1.0149	1.0236	1.0368	1.0064	1.0597
19	1.0140	1.0140	1.0139	1.0141	1.0223	1.0348	1.0060	1.0564
20	1.0132	1.0132	1.0132	1.0133	1.0211	1.0330	1.0057	1.0534
21	1.0126	1.0126	1.0125	1.0127	1.0200	1.0313	1.0054	1.0507
22	1.0120	1.0120	1.0119	1.0120	1.0191	1.0298	1.0051	1.0482
23	1.0114	1.0114	1.0114	1.0115	1.0182	1.0285	1.0049	1.0460
24	1.0109	1.0109	1.0109	1.0110	1.0174	1.0272	1.0046	1.0440
25	1.0105	1.0105	1.0104	1.0105	1.0167	1.0261	1.0044	1.0421
26	1.0100	1.0100	1.0100	1.0101	1.0160	1.0250	1.0042	1.0404
27	1.0097	1.0097	1.0096	1.0097	1.0154	1.0241	1.0041	1.0388
28	1.0093	1.0093	1.0093	1.0093	1.0148	1.0232	1.0039	1.0374
29	1.0090	1.0090	1.0089	1.0090	1.0143	1.0223	1.0038	1.0360
30	1.0087	1.0087	1.0086	1.0087	1.0138	1.0216	1.0036	1.0348

1. $C_N = \Gamma[(N-1)/2]\sqrt{(N-1)/2}\,/\Gamma[N/2]$; $C_N^* = [8N(N-1)]/[8N(N-1)-(N-1)(\kappa-3)-2N]$;

GT(5)(6) and GT(7) refer to values imputed from equations (5) and (6), and equation (7) respectively in Gurland and Tripathi (1971).

REFERENCES

Angelova, J. A. (2012), “On Moments of Sample Mean and Variance,” *International Journal of Pure and Applied Mathematics*, 79 (1), 67-85.

Bolch, B. W. (1968), “More on Unbiased Estimation of the Standard Deviation,” *The American Statistician*, 22 (3), 27.

Brugger, R. M. (1969), “A Note on Unbiased Estimation of the Standard Deviation,” *The American Statistician*, 23 (4), 32.

Cureton, E. E. (1968), “Unbiased Estimation of the Standard Deviation,” *The American Statistician*, 22 (1), 22.

D’Agostino, R. B. (1970), “Linear Estimation of the Normal Distribution Standard Deviation,” *The American Statistician*, 24 (3), 14–15.

Gurland, J. and Tripathi, R. C. (1971), “A Simple Approximation for Unbiased Estimation of the Standard Deviation,” *The American Statistician*, 25 (4), 30-32.

Holtzman, W. H. (1950), “The Unbiased Estimate of the Population Variance and Standard Deviation,” *American Journal of Psychology*, 63, 615-617.

Jarrett, R. F. (1968), “A Minor Exercise in History,” *The American Statistician*, 22 (3), 25.

Johenson, M. E. and Lowe, V. W. Jr. (1979), “Bounds on the Sample Skewness and Kurtosis,” *Technometrics*, 21 (3), 377-378.

Markowitz, E. (1968), “Minimum Mean-Square-Error Estimation of the Standard Deviation of the Normal Distribution,” *The American Statistician*, 22 (3), 26.

Stuart, A. (1969), “Reduced Mean-Square-Error Estimation of σ^p in Normal Samples,” *The American Statistician*, 23 (4), 27.