

CONCERNING THE SAMPLE STANDARD DEVIATION

Stanley L. Sclove

University of Illinois at Chicago

Let X_1, X_2, \dots, X_n denote the sample. The sample variance V is equal to $\text{SSD}/(n-1)$, where SSD is the sum of squared deviations from the mean,

$$\text{SSD} = \sum_{i=1}^n (X_i - \bar{X})^2.$$

This is the *definitional* formula for SSD. It is useful to have *computational* formulas.

Computational formulas for SSD. Note that, expanding the square and distributing the summation,

$$\begin{aligned} \text{SSD} &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=1}^n (X_i^2 - 2\bar{X}X_i + \bar{X}^2) \\ &= \sum_{i=1}^n X_i^2 - \sum_{i=1}^n 2\bar{X}X_i + \sum_{i=1}^n \bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - 2\bar{X}\sum_{i=1}^n X_i + n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - 2\bar{X}(n\bar{X}) + n(\bar{X}^2) \\ &= \sum_{i=1}^n X_i^2 - 2n(\bar{X}^2) + n(\bar{X}^2) \\ &= \sum_{i=1}^n X_i^2 - n(\bar{X}^2). \end{aligned}$$

The last expression is a computational formula for SSD. Replacing \bar{X} by $\sum_{i=1}^n X_i/n$ and simplifying gives another equivalent expression convenient for computation,

$$\text{SSD} = \sum_{i=1}^n X_i^2 - \frac{(\sum_{i=1}^n X_i)^2}{n}.$$

Unbiasedness of the sample variance. Now suppose that the sample X_1, X_2, \dots, X_n has been sampled from a process or from a finite population with replacement (as opposed to being sampled without replacement – the result to be obtained about unbiasedness requires modification for this case). Then it is relatively easy to show that the sample variance V (that is, S^2 , where S is the sample standard deviation) is unbiased for the variance σ^2 . That is, $E[V] = \sigma^2$. And, this is true regardless of the nature of the parent population (provided only that it has a finite variance). To see this, recall that $V = \text{SSD}/(n-1)$. So, $E[V] = E[\text{SSD}/(n-1)] = E[\text{SSD}]/(n-1)$ and to prove V is unbiased it suffices to prove that $E[\text{SSD}] = (n-1)\sigma^2$. To do this, we have

$$\begin{aligned}
E[\sum_{i=1}^n (X_i - \bar{X})^2] &= E[\sum_{i=1}^n X_i^2 - n\bar{X}^2] \\
&= n\{E[X_1^2] - E[\bar{X}^2]\} \\
&= n\{(\sigma^2 + \mu^2) - (\sigma^2/n + \mu^2)\} \\
&= (n - 1) \sigma^2.
\end{aligned}$$

The standard deviation. The standard deviation is usually of more interest than the variance, and estimating the standard deviation is not so simple. Obviously, S , the square root of the sample variance, can be used. But S is biased for σ . To see this, recall that for any random variable Y , $\text{Var}[Y] \geq 0$. Since $\text{Var}[Y] = E[Y^2] - (E[Y])^2$, $E[Y^2] - (E[Y])^2 \geq 0$; that is, $E[Y^2] \geq (E[Y])^2$. Now, take Y to be S . Then we have $\sigma^2 = E[S^2] \geq (E[S])^2$. This is the same as $(E[S])^2 \leq \sigma^2$, or, taking square roots, $E[S] \leq \sigma$. That is, S is *downward biased* as an estimator for σ ; it is, on the average, too small. This bias is not small if n is small, as is the case for example in making control charts in SPC. It can be a good idea to make a correction to S if n is small. The correction is found by computing the mathematical expectation of S , $E[S]$. Unlike the value of $E[V]$, the value of $E[S]$ depends upon the parent population. Here it will be derived for a Normal parent populations. That is, the correction obtained will apply to samples from a Normal distribution.

It can be shown (for a Normal parent population with mean μ and standard deviation σ) that $E[S]$ is a constant times σ , where that constant, denoted in the quality control literature by c_4 , depends upon n but is for every n less than 1:

$$E[S] = c_4\sigma.$$

That is, the correction to S takes the form S/c_4 . Since c_4 is less than 1, the resulting estimator S/c_4 will be greater than S . In fact it will be shown that

$$c_4 = [2/(n - 1)]^{1/2}\Gamma(n/2)/\Gamma[(n - 1)/2].$$

Here if m is an integer, $\Gamma(m) = (m-1)!$ and $\Gamma(m+1/2) = \frac{1 \times 3 \times 5 \times 7 \times \dots \times (2m-1)}{2^m} \Gamma(1/2)$, where $\Gamma(1/2) = \sqrt{\pi}$, or about 1.77245. For example, for $n = 6$,

$$\begin{aligned}
c_4 &= [2/(6 - 1)]^{1/2}\Gamma(6/2)/\Gamma[(6 - 1)/2] \\
&= (0.4)^{1/2}\Gamma(3)/\Gamma(2 + 1/2).
\end{aligned}$$

We have $\Gamma(3) = 2! = 2$ and $\Gamma(2 + 1/2) = (1 \times 3/2^2)\sqrt{\pi} = (3/4)\sqrt{\pi}$, or about $0.75(1.77245) = 1.32934$. Thus c_4 is about $(0.4)^{1/2}(2)/1.3293 = (0.632455)(2)/1.32934 = 0.9515$.

It is also true that

$$\text{Var}[S] = \sigma^2(1 - c_4^2).$$

Further, the variance of the sample standard deviation can, again for sampling from a Normal parent, be approximated as $\sigma^2/(2n)$. There is no contradiction between these results: For large n , the factor $(1 - c_4^2)$ is very close to $1/(2n)$. It is interesting that $\text{Var}[\bar{X}] = \sigma^2/n$, while $\text{Var}[S] \approx \sigma^2/(2n) = (1/2)\sigma^2/n = (1/2)\text{Var}[\bar{X}]$.

Derivations. How are these facts derived? Let S denote the sample standard deviation and $V = S^2$ the sample variance. Let W be distributed according to the chi-square distribution with $n-1$ d.f. Then S^2 is distributed as $\sigma^2W/(n-1)$ and

$$E[S] = E\{[\sigma^2W/(n-1)]^{1/2}\} = \sigma E[W^{1/2}]/(n-1)^{1/2}.$$

Further, $E[W^{1/2}] = \int w^{1/2}g(w; (n-1)/2, 1/2)dw$, where $g(w; m, b)$ denotes the p.d.f. of the Gamma distribution with parameters k and b ,

$$g(x; k, b) = b^k e^{-bx} x^{k-1} / \Gamma(k).$$

That is,

$$g(w; (n-1)/2, 1/2) = e^{-w/2} w^{(n-1)/2-1} / \{2^{(n-1)/2} \Gamma[(n-1)/2]\}.$$

The factor $w^{1/2}$ will combine with $w^{(n-1)/2-1}$ in the Gamma p.d.f. to give another Gamma p.d.f., except for constants. Now factor out constants, leaving the integral of the other Gamma p.d.f., which is 1. So the constants factored out give the answer for $E[W^{1/2}]$, which in turn will give $E[S]$. The details are

$$\begin{aligned} E[W^{1/2}] &= \int w^{1/2} g(w; (n-1)/2, 1/2) dw \\ &= \int w^{1/2} e^{-w/2} w^{(n-1)/2-1} dw / \{2^{(n-1)/2} \Gamma[(n-1)/2]\} \\ &= \int e^{-w/2} w^{n/2-1} dw / \{2^{(n-1)/2} \Gamma[(n-1)/2]\} \\ &= [\Gamma(n/2) 2^{n/2}] / \{2^{(n-1)/2} \Gamma[(n-1)/2]\} \\ &= 2^{1/2} \Gamma(n/2) / \Gamma[(n-1)/2]. \end{aligned}$$

Thus

$$\begin{aligned}
E[S] &= \sigma E[W^{1/2}]/(n-1)^{1/2} \\
&= \sigma \{2^{1/2}\Gamma(n/2)/\Gamma[(n-1)/2]\}/(n-1)^{1/2} \\
&= \sigma [2/(n-1)]^{1/2}\Gamma(n/2)/\Gamma[(n-1)/2] \\
&= c_4\sigma,
\end{aligned}$$

where

$$c_4 = [2/(n-1)]^{1/2}\Gamma(n/2)/\Gamma[(n-1)/2].$$

Also,

$$\begin{aligned}
Var[S] &= E[S^2] - \{E[S]\}^2 \\
&= E[V] - \{E[S]\}^2 \\
&= \sigma^2 - E[S]^2 \\
&= \sigma^2 - (c_4\sigma)^2 \\
&= \sigma^2 - \sigma^2 c_4^2 \\
&= \sigma^2(1 - c_4^2).
\end{aligned}$$

To get the result $Var[S] \approx \sigma^2/(2n)$, recall that for a differentiable function g of a random variable Y with mean μ_Y and variance σ_Y^2 ,

$$Var[g(Y)] \approx \sigma_Y^2 [g'(\mu_Y)]^2.$$

Take $Y = S^2$ and $g(Y) = g(S^2) = \sqrt{S^2} = S$. The sample X_1, X_2, \dots, X_n is from a distribution with mean μ and variance σ^2 . One has $\mu_Y = E[S^2] = \sigma^2$, $\sigma_Y^2 = Var[S^2]$, $g'(x) = 1/(2\sqrt{x})$, and

$$Var[S] \approx Var[S^2](1/(2\sqrt{\sigma^2}))^2 = Var[S^2]/(4\sigma^2).$$

Now, for a Normal parent distribution,

$$\begin{aligned}
Var[S^2] &= Var[\sigma^2\chi_{n-1}^2/(n-1)] \\
&= [\sigma^2/(n-1)]^2 Var[\chi_{n-1}^2] \\
&= [\sigma^2/(n-1)]^2 [2(n-1)] \\
&= 2\sigma^4/(n-1).
\end{aligned}$$

Thus,

$$Var[S] \approx Var[S^2]/(4\sigma^2) = [2\sigma^4/(n-1)]/(4\sigma^2) = \sigma^2/[2(n-1)] \approx \sigma^2/(2n).$$

In general, for a parent distribution that is not necessarily Normal, one has [Kendall and Stuart, Vol. 1, p. 245, (10.9); Cramér, p. 365]

$$\text{Var}[S^2] \approx (\mu_4 - \mu_2^2 + 4\mu_2\mu^2 - 4\mu\mu_3)/n,$$

where μ_r is the r -th central moment. This will give

$$\begin{aligned}\text{Var}[S] &\approx \text{Var}[S^2]/(4\sigma^2) \\ &= (1/n)(\mu_4 - \mu_2^2 + 4\mu_2\mu^2 - 4\mu\mu_3)/(4\sigma^2),\end{aligned}$$

or $\text{Var}[S] \approx (1/n)(\mu_4 - \sigma^4 + 4\sigma^2\mu^2 - 4\mu\mu_3)/(4\sigma^2)$.

Copyright © 2005 Stanley Louis Sclove

Created 2004: May 9; updated 2005: Aug 29
