Prof. Sergio A. Alvarez
Maloney Hall, room 569
Computer Science Department
Boston College
Chestnut Hill, MA 02467 USA

http://www.cs.bc.edu/∼alvarez/
alvarez@cs.bc.edu
voice: (617) 552-4333
fax: (617) 552-6790

CS244, Randomness and Computation
Spring 2013
Why $\frac{1}{n-1}$ instead of $\frac{1}{n}$? Unbiasing the sample variance

Measuring expected average behavior based on a sample

Suppose that X is a random variable with numerical values, and that $x_1 \cdots x_n$ are n independent samples of X. We can estimate the expected value and variance of X based on this sample. These notes discuss a subtlety that arises when doing this in the "obvious" way.

Sample mean as an estimate of expected value. Consistency of an estimate. As an estimate of the expected value, consider the sample mean $\overline{x}$, which is just the average of the n sample values x_i:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The Law of Large Numbers states that the sample mean $\overline{x}$ approaches the expected value $E(X)$ as the number of samples, n, approaches infinity. This is a good property. It makes the sample mean what is known as a *consistent* estimate of the expected value.

Bias of an estimate. The sample mean is also an *unbiased* estimate of the expected value. The latter refers to the behavior of the estimate for a fixed value of n. To understand what this means, you must view the sample mean as a random variable in its own right. The sample mean $\overline{x}$ is a random variable because it depends on the specific random selection of n samples of X. The statement that $\overline{x}$ is an unbiased estimate of $E(X)$ means that the expected value of the random variable $\overline{x}$, for a given fixed value of n, is precisely $E(X)$. We can see that this is true by computing $E(\overline{x})$ directly, using the fact that the expected value of a linear combination (weighted sum) is the corresponding linear combination of the expected values (same weights):

$$\begin{aligned} E(\overline{x}) &= E\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} E(x_i) \\ &= E(X) \end{aligned}$$

The last equality above holds because each of the expected values in the sum is equal to $E(X)$ (since x_i is just a sample of X). We conclude, therefore, that the sample mean is a consistent, unbiased estimate of the expected value. So far, so good.

Sample mean sample squared deviation (SMSSD) as an estimate of variance. Let's go on to estimate the variance. Mimicking the above discussion of sample mean, the obvious estimate of variance is the sample mean of the squared deviations from the sample mean:

$$\text{SMSSD} = \overline{(x - \overline{x})^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

In other words, we first estimate the expected value by using the sample mean, and then estimate the variance as the average of the squared deviations of the individual samples from the estimated expected value. I'll refer to this estimate of the variance as the sample mean of the sample squared deviation (SMSSD). We'll consider the consistency and bias of this estimate (see the above discussion of sample mean for definitions of these notions).

Bias of the SMSSD estimate of variance. Unbiased correction of the SMSSD. As discussed in connection with the sample mean, bias refers to any difference between the expected value of the estimate for a fixed value of n, on the one hand, and the actual target value, in this case the variance of X, on the other. In the SMSSD variance estimate, one must remember that $\overline{x}$ is just an estimate of the expected value $E(X)$. To bring out the relationship between the two quantities, it is convenient to express the SMSSD as follows:

$$\begin{aligned}
\text{SMSSD} &= \frac{1}{n}\sum_{i=1}^{n}((x_i - E(X)) - (\overline{x} - E(X)))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2 + \frac{1}{n}\sum_{i=1}^{n}(\overline{x} - E(X))^2 - \frac{2}{n}\sum_{i=1}^{n}(x_i - E(X))(\overline{x} - E(X)) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2 + (\overline{x} - E(X))^2 - 2(\overline{x} - E(X))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))^2 - (\overline{x} - E(X))^2
\end{aligned}$$

The first term on the right-hand side is the mean of the sample squared deviations relative to the actual expected value. The second term is a measure of the error of the sample mean as an estimate of the expected value. In order to get the expected value of the SMSSD, we take the expected value of the sum of the latter two terms. Notice that the expected value of each of the n terms in the sum on the right is just the variance of X:

$$\begin{aligned}
E(\text{SMSSD}) &= \frac{1}{n}\sum_{i=1}^{n}E((x_i - E(X))^2) - E((\overline{x} - E(X))^2) \\
&= E((X - E(X))^2) - E((\overline{x} - E(X))^2)
\end{aligned} \tag{1}$$

Again, the first term on the right in Eq. 1 is just the variance of X. It only remains to calculate the second term, at the far right in Eq. 1. Before doing that, note that we can already see that the only way that the expected value of the SMSSD can equal the variance is for the sample mean $\overline{x}$ to be precisely equal to the actual expected value $E(X)$. This is simply not possible unless the variable X isn't really variable at all. Furthermore, it is clear that the expected value of SMSSD will be less than the variance unless X is constant. Hence, the SMSSD is *biased low*. We will now proceed to calculate the magnitude of the bias. Doing so will allow us to define an improved, unbiased estimate of variance.

$$\begin{aligned} E((\overline{x} - E(X))^2) &= E\left(\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - E(X))\right)^2\right) \\ &= \frac{1}{n^2}E\left(\sum_{i=1}^{n}(x_i - E(X))^2 + \sum_{i=1}^{n}\sum_{j\neq i}(x_i - E(X))(x_j - E(X))\right) \\ &= \frac{1}{n^2}\left(\sum_{i=1}^{n}E\left((x_i - E(X))^2\right) + \sum_{i=1}^{n}\sum_{j\neq i}E\left((x_i - E(X))(x_j - E(X))\right)\right) \end{aligned}$$

Each summand in the first term on the right is just the variance of X. Also, since samples x_i and x_j are independent whenever $i \neq j$, each term in the nested sum at the far right is 0. This leaves the following:

$$E((\overline{x} - E(X))^2) = \frac{1}{n}E\left((X - E(X))^2\right)$$

Finally, we can substitute the latter conclusion into the expression for the expected value of the SMSSD in Eq. 1.

$$\begin{aligned} E(\text{SMSSD}) &= E((X - E(X))^2) - E((\overline{x} - E(X))^2) \\ &= \left(1 - \frac{1}{n}\right)E\left((X - E(X))^2\right) \end{aligned}$$

What this means is that the expected value of the SMSSD is *not* the variance of X, but instead is slightly smaller than the variance, by a factor of $1 - \frac{1}{n}$, or, equivalently, $\frac{n-1}{n}$. In order to obtain an unbiased estimate of the variance, we simply divide the SMSSD by this factor. The result is the following corrected SMSSD:

$$\text{corrected SMSSD (unbiased estimate of variance)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2}$$

So, to summarize, the reason for the using $n - 1$ instead of n in the denominator for the sample estimate of variance is that this is the value that produces an unbiased estimate of the variance.

Consistency of the corrected SMSSD estimate of variance. Fortunately, the corrected SMSSD in Eq. 2 is also a consistent estimate of variance, that is, it converges to the actual variance as the sample size n approaches infinity. Without going into too much detail, the reason for this is the Law of Large Numbers. For large values of n, $\overline{x}$ is very close to $E(X)$. Hence, most of the terms in the sum that defines the corrected SMSSD will be very close to $(x_i - E(X))^2$, and division by $n-1$ (or n) will yield a value that is very close to the variance of X.