

Estimating the Variance of a Sample

Greg Reynolds

15th February 2008

Introduction

Suppose that we have a random variable X that has unknown mean μ and unknown variance σ^2 . Suppose also that we have n samples (instances) of X , denoted x_1, x_2, \dots, x_n , each of which is independent of the others. We seek to find unbiased estimates of the mean and variance of X . The definition of an *unbiased estimate* of a statistic θ (the estimate denoted $\hat{\theta}$) is:

$$\mathbb{E}\{\hat{\theta}\} = \theta \tag{1}$$

For the following proof we also need to define:

$$\mathbb{E}\{X\} = \mu \tag{2}$$

$$\mathbb{E}\{(X - \mu)^2\} = \sigma^2 \tag{3}$$

Mean Estimation

The well known estimate of the mean, \bar{x} is:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \tag{4}$$

to show that this is an unbiased estimate, we find the expectation of \bar{x} :

$$\begin{aligned} \mathbb{E}\{\bar{x}\} &= \mathbb{E}\left\{\frac{1}{n} \sum_{i=1}^n x_i\right\} \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\{x_i\} \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned} \tag{5}$$

where we have used the fact that the expected value of any instance x_i of X is conceptually the same as the expected value of X .

Variance Estimation

An estimate, s^2 of the variance is often quoted as:

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (6)$$

equally, it is also often quoted as:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (7)$$

The proof below shows that the second estimate is in fact the “correct” estimate, although for large n , the difference is negligible. The target is to show that $\mathbf{E}\{s^2\} = \sigma^2$.

$$\begin{aligned} \mathbf{E}\{s^2\} &= \mathbf{E}\left\{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2\right\} \\ &= \frac{1}{n-1} \sum_{i=1}^n \mathbf{E}\{(x_i - \bar{x})^2\} \\ &= \frac{1}{n-1} \sum_{i=1}^n \mathbf{E}\{((x_i - \mu) - (\bar{x} - \mu))^2\} \end{aligned} \quad (8)$$

Expanding the expectation gives,

$$\begin{aligned} \mathbf{E}\{((x_i - \mu) - (\bar{x} - \mu))^2\} &= \mathbf{E}\{(x_i - \mu)^2 - 2(x_i - \mu)(\bar{x} - \mu) + (\bar{x} - \mu)^2\} \\ &= \mathbf{E}\{(x_i - \mu)^2\} - 2\mathbf{E}\{(x_i - \mu)(\bar{x} - \mu)\} + \mathbf{E}\{(\bar{x} - \mu)^2\} \end{aligned} \quad (9)$$

we can now make the following observations:

$$\mathbf{E}\{(x_i - \mu)^2\} = \sigma^2 \quad (10)$$

using (4) we can write:

$$\begin{aligned} (x_i - \mu)(\bar{x} - \mu) &= (x_i - \mu) \left[\frac{1}{n} \sum_{j=1}^n (x_j - \mu) \right] \\ &= \frac{1}{n} \sum_{j=1}^n (x_i - \mu)(x_j - \mu) \end{aligned} \quad (11)$$

observing that:

$$\mathbf{E}\{(x_i - \mu)(x_j - \mu)\} = 0 \quad \forall i \neq j \quad (12)$$

because each sample of X is independent of all the others. All this means that the second term in the expansion becomes:

$$-2\mathbf{E}\{(x_i - \mu)(\bar{x} - \mu)\} = -\frac{2}{n}\sigma^2 \quad (13)$$

Finally, the final term must be addressed, again using (4):

$$\begin{aligned} \mathbf{E}\{(\bar{x} - \mu)^2\} &= \mathbf{E}\left\{ \left[\frac{1}{n} \sum_{j=1}^n (x_j - \mu) \right] \left[\frac{1}{n} \sum_{k=1}^n (x_k - \mu) \right] \right\} \\ &= \frac{1}{n^2} \mathbf{E}\left\{ \sum_{j=1}^n \sum_{k=1}^n (x_j - \mu)(x_k - \mu) \right\} \\ &= \frac{1}{n^2} n\sigma^2 \\ &= \frac{1}{n}\sigma^2 \end{aligned} \quad (14)$$

So we can finally substitute all these terms back in:

$$\begin{aligned} E\{s^2\} &= \frac{1}{n-1} \sum_{i=1}^n \left[\sigma^2 - 2\frac{\sigma^2}{n} - \frac{\sigma^2}{n} \right] \\ &= \frac{1}{n-1} \sum_{i=1}^n \left(\sigma^2 - \frac{\sigma^2}{n} \right) \\ &= \frac{1}{n-1} (n\sigma^2 - \sigma^2) = \frac{1}{n-1} (n-1)\sigma^2 \\ &= \sigma^2 \end{aligned} \tag{15}$$

Thus this estimate of the variance is unbiased.