

Bias, Variance and MSE of Estimators

Guy Lebanon

February 14, 2006

We assume that we have iid samples X_1, \dots, X_n that follow some (possibly unknown) distribution. The task of statistics is to estimate properties of the unknown distribution. Estimating the pdf or cdf of the unknown distribution is often too difficult and so the common situation is that we are trying to estimate a parameter θ of the unknown distribution. This parameter may be the mean $E(X_i)$, the variance $\text{Var}(X_i)$ or some other quantity such as $P(X_i > 0)$.

The unknown parameter is a fixed real number $\theta \in \mathbb{R}$. To estimate it, we use an estimator which is a function of our observations $\hat{\theta}(X_1, \dots, X_n)$. We will follow the standard practice and omit (in the notation only) the dependency of the estimator on the samples, i.e. we write $\hat{\theta}$. However, it is crucial to remember that $\hat{\theta}$ is a random variable since it is a function of n random variables.

A desirable property of an estimator is that it is correct on average. That is, if there are repeated samplings of n samples X_1, \dots, X_n , the estimator $\hat{\theta}(X_1, \dots, X_n)$ will have, on average, the correct value. Such estimators are called unbiased.

Definition 1. *The bias of an estimator $\hat{\theta}$ is $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$. If it is 0, the estimator is said to be unbiased.*

There is, however, more important performance characterizations for an estimator than just being unbiased. The mean squared error is perhaps the most important of them. It captures the error that the estimator makes. However, since the estimator is a RV, we need to average over its distribution thus capturing the average performance if there are many repeated samplings of X_1, \dots, X_n .

Definition 2. *The mean squared error (MSE) of an estimator is $E((\hat{\theta} - \theta)^2)$.*

Theorem 1. *The mean squared error of an estimator equals $\text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})$.*

Proof.

$$\begin{aligned} E((\hat{\theta} - \theta)^2) &= E(((\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta))^2) = E\{(\hat{\theta} - E(\hat{\theta}))^2 + (E(\hat{\theta}) - \theta)^2 + (\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)\} \\ &= \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) + E((\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta)) = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) + E(\hat{\theta}E(\hat{\theta}) - (E(\hat{\theta}))^2 - \theta\hat{\theta} + E(\hat{\theta})\theta) \\ &= \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) + (E(\hat{\theta}))^2 - (E(\hat{\theta}))^2 - \theta E(\hat{\theta}) + \theta E(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) \end{aligned}$$

□

Since the MSE decomposes into a sum of the bias and variance of the estimator, both quantities are important and need to be as small as possible to achieve good estimation performance. It is common to trade-off some increase in bias for a larger decrease in the variance and vice-versa.

Two important special cases of estimators are $\hat{\theta} = \bar{X}$ which estimates the expectation $E(X_i) = \mu$ and $\hat{\theta} = S^2$ which estimates the variance $\text{Var}(X_i) = \sigma^2$. We show below that both are unbiased and therefore their MSE is simply their variance. Note that we do not make any normality assumptions!

Theorem 2. *\bar{X} is an unbiased estimator of μ and S^2 is an unbiased estimator of the variance σ^2 .*

Proof.

$$E(\bar{X}) = E\left(n^{-1} \sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)/n = nE(X_i)/n = \mu.$$

To prove that S^2 is unbiased, we first need the following results:

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum X_i^2 - 2\bar{X} \sum X_i + n\bar{X}^2 = \sum X_i^2 - 2n\bar{X}\bar{X} + n\bar{X}^2 = \sum X_i^2 - n\bar{X}^2$$

and therefore

$$\mathbb{E} \left(\sum_{i=1}^n (X_i - \bar{X})^2 \right) = \mathbb{E} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = \sum_{i=1}^n \mathbb{E}(X_i^2) - n\mathbb{E}(\bar{X}^2) = n\mathbb{E}(X_1^2) - n\mathbb{E}(\bar{X}^2).$$

Substituting the expectations $\mathbb{E}(X^2) = \text{Var}(X) + (\mathbb{E}(X))^2 = \sigma^2 + \mu^2$ and $\mathbb{E}(\bar{X}^2) = \text{Var}(\bar{X}) + (\mathbb{E}(\bar{X}))^2 = \frac{\sigma^2}{n} + \mu^2$ in the above equation we have

$$\mathbb{E} \left(\sum_{i=1}^n (X_i - \bar{X})^2 \right) = n(\sigma^2 + \mu^2) - n \left(\frac{\sigma^2}{n} + \mu^2 \right) = (n-1)\sigma^2$$

which proves that $\mathbb{E}(S^2) = E(\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)) = \sigma^2$. □

It follows therefore that the MSE of \bar{X} as an estimator of μ is $\text{Var}(\bar{X}) = \sigma^2/n$. This MSE decreases to 0 as $n \rightarrow \infty$ which is a nice result: as we get more samples, the MSE of our estimator goes to 0. For S^2 , the MSE is $\text{Var}(S^2)$ which may be computed with some tedious algebra (it also decreases to 0 as $n \rightarrow \infty$).

Another performance measure for estimators $\hat{\theta}$ is the probability that the estimator's error is within some acceptable error range

$$P(|\hat{\theta} - \theta| < \epsilon) = P(-\epsilon < \hat{\theta} - \theta < \epsilon) = P(\theta - \epsilon < \hat{\theta} < \theta + \epsilon) = \int_{\theta - \epsilon}^{\theta + \epsilon} f_{\hat{\theta}}(r) dr.$$

However, to evaluate the above quantity, we need (i) the pdf $f_{\hat{\theta}}$ which depends on the pdf of X_i (which is typically unknown) and (ii) the true value θ (also typically unknown). If we don't know the above quantities, and if $\hat{\theta}$ is unbiased, may obtain a bound by applying Chebyshev's inequality

$$P(|\hat{\theta} - \theta| < \epsilon) = 1 - P(|\hat{\theta} - \mathbb{E}(\hat{\theta})| \geq \epsilon) \geq 1 - \frac{\text{Var}(\hat{\theta})}{\epsilon^2}.$$