Estimators, Mean Square Error, and Consistency

January 20, 2006

1 Statistics and Mean Square Error

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a random sample from a distribution $f(x|\theta)$, with θ unknown. For example, $X_1, \ldots, X_n \sim N(\theta, 1)$. Our goal is to use the information available in the data to make guesses about θ. Ideally, we would like these to be educated guesses that are likely to be close to the true value of θ. If the data is our only available source of information, we must estimate θ by a function of the data, $\delta(\mathbf{x})$. One such function is $\delta(\mathbf{x}) = \bar{x}$, others are $\delta(\mathbf{x}) = median(\mathbf{x})$, $\delta(\mathbf{x}) = max(\mathbf{x})$, or $\delta(\mathbf{x}) = 3x_1/(x_2x_3)$.

Any such function of the data is called a *statistic*. One of the main goals of this course is to figure out how to choose the right statistic to estimate θ. Of course, we need some definition of being "right". Our vague notion of wanting something that is "likely to be close to θ" could be interpreted in many ways, and we have to pick one. One of the most common measures is *Mean Square Error*, or MSE, which is defined as

$$MSE(\theta) = E_\theta[(\delta(\mathbf{X}) - \theta)^2]$$

Estimators $\delta(\mathbf{X})$ that have small MSE are considered good because their expected distance from θ is small (if the squared error is small then the actual distance will be small as well). Note that the MSE is a function of θ, which means some estimators might work well for some values of θ and not for others.

Computing MSE requires the *sampling distribution* of $\delta(\mathbf{x})$, which is where the prerequisite of probability appears in this course. The calculation is

somewhat simplified by noting that MSE can be divided into two parts. Let $\mu_\delta = E_\theta[\delta(\mathbf{X})]$ (note μ_δ is a constant, not a random variable).

$$\begin{aligned} E_\theta[(\delta(\mathbf{X})-\theta)^2] &= E_\theta[(\delta(\mathbf{X})-\mu_\delta+\mu_\delta-\theta)^2] \\ &= E_\theta[(\delta(\mathbf{X})-\mu_\delta)^2+2(\delta(\mathbf{X})-\mu_\delta)(\mu_\delta-\theta)+(\mu_\delta-\theta)^2] \\ &= E_\theta[(\delta(\mathbf{X})-\mu_\delta)^2]+E_\theta[2(\delta(\mathbf{X})-\mu_\delta)(\mu_\delta-\theta)]+E_\theta[(\mu_\delta-\theta)^2] \\ &= V_\theta[\delta(\mathbf{X})]+2(\mu_\delta-\theta)E_\theta[(\delta(\mathbf{X})-\mu_\delta)]+(\mu_\delta-\theta)^2 \\ &= V_\theta[\delta(\mathbf{X})]+(\mu_\delta-\theta)^2 \end{aligned} \quad (1)$$

Thus, the mean square error can be decomposed into a variance term and a bias term. The *bias* is defined as $(\mu_\delta-\theta)$, the distance between the estimator's mean and the parameter θ. An estimator is called *unbiased* if the bias is 0 (which occurs if $E[\delta(\mathbf{X})] = \mu_\delta = \theta$), in which case the MSE is just the variance of the estimator.

For example, suppose $X_1,\ldots,X_n \sim N(\theta,1)$ and $\delta(\mathbf{X}) = \bar{X}$, the sample mean. The distribution of $\bar{X}$ is $N(\theta,1/n)$ ($1/n$ is the variance). In this case $\mu_\delta = E[\delta(\mathbf{X})] = \theta$ and $V_\theta[\delta(\mathbf{X})] = 1/n$, so the MSE is $1/n + 0^2 = 1/n$. Note in this example the MSE does not depend on the parameter θ. The sample mean performs equally well for all values of θ.

Choosing an estimator depends strongly on the likelihood. It turns out $\delta(\mathbf{x}) = \bar{x}$ is one of the best estimators for the normal mean in the previous example. If $X_1,\ldots,X_n \sim Uni(0,\theta)$, $\bar{x}$ doesn't perform nearly as well. To find the MSE, we need the mean and variance of $\bar{x}$. Note that $E[X_i] = \theta/2$ and $V[X_i] = \theta^2/12$. The sample mean therefore has mean $\theta/2$ and variance $\theta^2/(12n)$. The MSE is therefore

$$\frac{\theta^2}{12n} + \left(\frac{\theta}{2}-\theta\right)^2 = \frac{(3n+1)\theta^2}{12n}$$

In this example the MSE depends on θ. In turns out this MSE is much larger than other available estimators. One quick improvement, for example, is to remove the bias. Suppose instead of $\delta(\mathbf{x}) = \bar{x}$ we use $\delta(\mathbf{x}) = 2\bar{x}$. Then $E[\delta(\mathbf{x})] = 2(\theta/2) = \theta$ and $V[\delta(\mathbf{x})] = 4(\theta^2/(12n)) = \theta^2/(3n)$. The MSE is

$$\frac{\theta^2}{3n} + 0 = \frac{4\theta^2}{12n}$$

For $n > 1$, $2\bar{x}$ has a smaller MSE than $\bar{x}$ for all θ.

2 Consistency

One desirable property of estimators is *consistency*. If we collect a large number of observations, we hope we have a lot of information about any unknown parameter θ, and thus we hope we can construct an estimator with a very small MSE. We call an estimator *consistent* if

$$\lim_n MSE(\theta) = 0$$

which means that as the number of observations increase the MSE descends to 0. In our first example, we found if $X_1, \ldots, X_n \sim N(\theta, 1)$, then the MSE of $\bar{x}$ is $1/n$. Since $\lim_n(1/n) = 0$, $\bar{x}$ is a consistent estimator of θ.

Remark: To be specific we may call this "MSE-consistant". There are other type of consistancy definitions that, say, look at the probability of the errors. They work better when the estimator do not have a variance.

If $X_1, \ldots, X_n \sim Uni(0, \theta)$, then $\delta(\mathbf{x}) = \bar{x}$ is not a consistent estimator of θ. The MSE is $(3n + 1)\theta^2/(12n)$ and

$$\lim_n \frac{(3n+1)\theta^2}{12n} = \frac{\theta^2}{4} \neq 0$$

so even if we had an extremely large number of observations, $\bar{x}$ would probably not be close to θ. Our adjusted estimator $\delta(\mathbf{x}) = 2\bar{x}$ is consistent, however. We found the MSE to be $\theta^2/3n$, which tends to 0 as n tends to infinity. This doesn't necessarily mean it is the optimal estimator (in fact, there are other consistent estimators with MUCH smaller MSE), but at least with large samples it will get us close to θ.

3 The uniform distribution in more detail

We said there were a number of possible functions we could use for $\delta(\mathbf{x})$. Suppose that $X_1, \ldots, X_n \sim Uni(0, \theta)$. We have already discussed two estimators, $\bar{x}$ and $2\bar{x}$, and found their MSE. There are a variety of others. Instead of the mean $\bar{x}$ we could look at the median of the x values. Analogously to the mean, $2median(x)$ is an improvement. Another estimator is the maximum of the x values. This can be made unbiased by multiplying by $(n + 1)/n$. Finally, there is the possibility of more complicated functions.

Clearly θ must be bigger than $\max(x)$, otherwise $\max(x)$ couldn't be in the sample. If $2\bar{x} < \max(x)$, then $\max(x)$ must be closer to θ than $2\bar{x}$, so we can use the estimator $\max(2\bar{x}, \max(x))$. This results in 6 estimators shown in the table.

We have already derived the MSEs for $\bar{x}$ and $2\bar{x}$. It is also fairly easy to derive the MSEs for $2median(x)$, $\max(x)$, and $(n+1)\max(x)/n$. The MSE for $\max(2\bar{x}, \max(x))$ is more difficult to derive. To demonstrate a little more explicitly what is going on than theoretical calculations allow, we turn to simulations. I generated a dataset $X_1, \ldots, X_{11} \sim Uni(0, \theta = 5)$. For this dataset I computed each of the 6 estimators. Ideally, we would like these estimators to be close to 5, the correct answer. I then tossed away the 11 observations, generated another 11, and computed the 6 estimators for this second set of observation. I then generated a third set, a fourth set, and so on to a total of 100000 sets of 11 observations. The figure shows histograms of the 6 estimators. These histograms approximate the *sampling distributions* of the estimators for $n = 11$. I then approximated the MSE for each estimator. This was done by looking at the values of these estimators for each of the 100000 datasets and using the sample mean and variance as guesses of μ_δ and $V[\delta(\mathbf{x})]$. Note that for $\bar{x}$ and $2\bar{x}$, the estimators whose MSEs are known from the previous section, the values in the table are almost identical to the theoretical values. For $n = 11$, we find the theoretical MSE for $\bar{x}$ is $(3n+1)\theta^2/(12n) = (34)(25)/(132) = 6.44$ and the theoretical MSE for $2\bar{x}$ is $\theta^2/3n = (25/33) = 0.76$.

Estimator	Mean	Bias	Variance	MSE(simulated)
$d1 = \bar{x}$	2.50	-2.50	0.19	6.43
$d2 = 2\bar{x}$	5.00	0.00	0.76	0.76
$d3 = 2median(x)$	4.99	-0.01	1.92	1.92
$d4 = \max(x)$	4.58	-0.42	0.15	0.32
$d5 = (n+1)\max(x)/n$	5.00	0.00	0.18	0.18
$d6 = \max(2\bar{x}, \max(x))$	5.14	0.14	0.52	0.54

Figure 1: Simulated Results for 6 Estimators