Lsn21

Clark

Admin

We can think of a statistical model as a data generating mechanism. How the model generates data depends on both the distribution of the response variable as well as the *parameters* of the statistical model. Oftentimes we write:

For instance, we might consider $Y \sim \text{Binom}(10, .3)$. That is, we assume our random variable, which is some mathematical representation for a real world experiment, can take on the values of $Y \in \{0, 1, 2, \cdots, 10\}$ and the probability that Y takes on each of these values is given by

$$P(Y = y) = p(y|p = .3) = \binom{10}{y} .3^y (1 - .3)^{n-y}$$

So for instance, we might say that the number of Cadets in a squad of ten who returned late for recall formation follow this distribution. However, the goal of statistics is usually not to use statistical models to generate probabilities, but rather to use data to determine the value of a parameter. For instace, we might assume that $Y \sim \text{Binom}(10, p)$ and want to say something about the likely value of p. For instance, if we consider above and we look at one squad of 10 from last night and 2 Cadets came back late, what could we say about the most likely value of p?

While there is a common sense way to estimate p, is this the only way? What if, we estimate p by $\hat{p}_{silly} = \frac{y}{15}$? Would that be better or worse that what we used above?

In oprder to formalize this we need the notions of *bias* and *mean square error*. The Bias of a point estimator is:

Let's consider the *bias* of $\hat{p}_{silly}$ for our experiment above.

Let's consider the bias of $\hat{p}$

Note that this brings us to a very important note in statistics. It *could be* that $\hat{p}_{silly}$ is more correct that $\hat{p}$ that we came up with above. For instance if $p = .133$. In that case, $\hat{p}_{silly}$ would be absolutely correct and $\hat{p}$ would be off. All we are saying with bias (and MSE) is that if we would consistently use this as an esimator, $\hat{p}$ would be correct, on average, and $\hat{p}_{silly}$ would not be.

Mean Square Error is similar to Bias, but it also penalizes variance of an estimator. That is, if we have an esimator that is correct, on average, but has a high variability maybe we wouldn't want to use it (after all we

probably aren't doing our experiment an infinite number of times!) For instance let's consider $Y_i \sim N(\mu, \sigma^2)$. How would we normally estimate σ^2?

Let's first find the bias of S^2.

Next we can find the MSE by noting that the MSE is equal to the bias plus the Variance of our estimator, that is

In order to find the MSE we would also need the Variance of our estimator.

Therefore, the MSE of S^2 is

Let's consider a different estimator, $\hat{\sigma}_{mle} = \frac{(\sum_{i=1}^{n} Y_i - \bar{Y})^2}{n}$. Let's find the MSE of this estimator.

So while sometimes it's obvious, like in the case of $\hat{p}$ how we should estimate our parameter, in other instances it's not. Let's work through 8.8 on page 394. At the boards come up with $E[\hat{\theta}_i]$ for $i = 1, 2, 3, 4, 5$.

During WWII the Allies wanted to know how many Tanks the Germans were producing. Fortunately the Germans sequentially numbered their tanks. Each day the Allies would record the serial number of the tank they observed. Let's come up with a statistical model that can help answer the question, how many tanks do the Germans have. We'll make the simplifying assumption that each tank is equally likely to be observed and once you observe a tank you can still observe it in the future.

Come up with two ways to estimate the number of tanks. Call them $\hat{\theta_1}$ and $\hat{\theta_2}$. What method, if any, is unbiased?

What method has the lowest MSE.