Handout 14: Unbiasedness and MSE

We begin making the transition from the ideas and tools of probability theory to the theory and methods of statistical inference. For simplicity, suppose that we have the data $X_1, X_2, \ldots, X_n$ in hand, where the X's are assumed to be independent random variables with a common distribution F_θ indexed by the parameter θ whose value is unspecified. In practice, we would often be willing to assume that the distribution F_θ has a density or probability mass function of some known form. For example, physical measurements might be assumed to be random draws from an $N(\mu, \sigma^2)$ distribution, counts might be assumed to be $Binomial(n, p)$ or $Poisson(\lambda)$ variables, and the failure times of engineered systems in a life-testing experiment might be assumed to be $\Gamma(\alpha, \beta)$ variables. Note that such assumptions do not specify the exact distribution that applies to the experiment of interest, but rather, specifies only its "type."

Suppose that a particular parametric model F_θ has been assumed to apply to the available data X_1, X_2,..., X_n. It is virtually always the case in practice that the exact value of the parameter θ is not known. A naïve but nonetheless useful way to think of *statistical estimation* is to equate it with the process of guessing. The goal of statistical estimation is to make an *educated guess* about the value of the unknown parameter θ . What makes one's guess "educated" is the fact that the estimate of θ is informed by the data. A *point estimator* of θ is a fully specified function of the data which, when the data is revealed, yields a numerical guess at the value of θ . The estimator will typically be represented by the symbol $\widehat{\theta}$, pronounced "theta hat," whose dependence on the data is reflected in the equation:

$$\widehat{\theta} = \widehat{\theta}(X_1, X_2, \ldots, X_n)$$

Let us suppose, for example, that the experiment of interest involves n independent tosses of a bent coin, where n is a known integer. The parameter p, the probability of heads in a given coin toss, is treated as an unknown constant. It seems quite reasonable to assume that the experimental data we will observe is well described as a sequence of iid Bernoulli trials. Most people would base their estimate of p on the variable X, the number of successes in n iid Bernoulli trials; of course, $X \sim B(n,\ p)$. The sample proportion of successes, $\widehat{p} = X/n$ is a natural estimator of p and is, in fact, the estimator of p that most people would use.

What makes a "good" estimator? Ideally, we would hope that it tends to be close to the true parameter it is estimating. The following defintion formalizes this notion.

Definition 1 (Mean Squared Error (MSE)) *The mean squared error of an estimator $\widehat{\theta}$ of the parameter θ is defined as:*

$$MSE(\widehat{\theta}) = \mathbb{E}\left(\widehat{\theta} - \theta\right)^2$$

The MSE has a useful decomposition that we will often use in the analyses of various estimators.

Theorem 1 (Decomposition of MSE) *The mean squared error of $\widehat{\theta}$ can be decomposed as follows:*

$$MSE(\widehat{\theta}) = Var(\widehat{\theta}) + Bias(\widehat{\theta})^2$$

Where:

$$Bias(\widehat{\theta}) = \mathbb{E}(\widehat{\theta} - \theta)$$

Proof. This result can be established with the following:

$$\begin{aligned}
\widehat{\theta} - \theta &= (\widehat{\theta} - \mathbb{E}\widehat{\theta}) + (\mathbb{E}\widehat{\theta} - \theta) \\
(\widehat{\theta} - \theta)^2 &= (\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 + (\mathbb{E}\widehat{\theta} - \theta)^2 + 2 \cdot (\widehat{\theta} - \mathbb{E}\widehat{\theta}) \cdot (\mathbb{E}\widehat{\theta} - \theta)
\end{aligned}$$

Taking expectations on both sides, notice that the cross-term drops out:

$$\begin{aligned}
\mathbb{E}(\widehat{\theta} - \theta)^2 &= \mathbb{E}(\widehat{\theta} - \mathbb{E}\widehat{\theta})^2 + \mathbb{E}(\mathbb{E}\widehat{\theta} - \theta)^2 \\
MSE(\widehat{\theta}) &= Var(\widehat{\theta}) + Bias(\widehat{\theta})^2
\end{aligned}$$

Giving the desired result ■.

Notice that the MSE in general is a function of the value of actual value of θ. The decomposition leads to a definition of bias:

Definition 2 (Unbiasedness) *An estimator $\widehat{\theta}$ is unbiased for θ if $Bias(\widehat{\theta}) = 0$.*

We close with a result that establishes the unbiasness of estimators for the mean and standard deviation of an arbitrary estimator.

Theorem 2 *If $X_1, \dots X_n \overset{iid}{\sim} F$ for some distribution F with finite mean μ and variance σ^2, then*

$$\bar{X} = \frac{\sum_i X_i}{n}$$
$$s^2 = \frac{1}{(n-1)} \sum_i (X_i - \bar{X})^2$$

Are unbiased estimators of μ and σ^2, respectively.

Proof. The first result is a simple application of the linearity of the expectation operator. The second comes from a similar derivation to the decomposition of the MSE:

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

Dividing both sides by $(n-1)$ finishes the proof ■.