Notes: Estimation, Bias and Variance

CS 3130 / ECE 3530: Probability and Statistics for Engineers

March 23, 2023

Parameters of a Distribution. All of the distributions that we have discussed come with a set of parameters that fully describe the equation for the pdf (or pmf). For example, a Normal random variable, $X \sim N(\mu, \sigma^2)$, has the mean μ and variance σ^2 as parameters. An exponential random variable, $X \sim \text{Exp}(\lambda)$, has the rate λ as its only parameter. A Bernoulli random variable, $X \sim \text{Ber}(p)$, has the probability of success, p, as its only parameter. It's important to remember that these parameters are *constants.*

Notation: When we want to make a generic statement about distributions and parameters, it is customary to use the Greek letter θ to denote a parameter.

Estimation of a Distribution's Parameters. We have talked about making assumptions that our data comes from a particular distribution. For example, if we believe that our data is the summation of many small random effects, we can argue that it probably comes from a Normal distribution as a result of the Central Limit Theorem (CLT). Sometimes we just have to make a *modeling decision*, that is, we must decide to model our data with a particular type of distribution. This might be as simple as looking at a histogram of our data and saying "I think that an exponential distribution will fit this."

Once we have decided what type of distribution to use, the next question is: what are the parameters for this distribution? For example, if we have chosen to model our data using a Normal (maybe based on a CLT argument), we still have an infinite number of Normal distributions to choose from because there are an infinite number of μ and σ^2 parameters that will give rise to different Normals! Now, we could make further assumptions about the distribution and pick particular values for μ and σ^2. However, this modeling decision starts to feel too restrictive and is much harder to justify. Fortunately, we can use the data we have collected to *estimate* these parameters. In the case of a Normal, the sample mean $\bar{x}_n$ and variance s_n^2 of our data seem like reasonable choices to estimate the parameters μ and σ^2.

Definition: Let $X_1, X_2, \ldots, X_n$ be iid random variables coming from a distribution with parameter θ. An **estimator** of θ is a statistic $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$. Note: the "hat" notation is to indicate that we are hoping to estimate a particular parameter. For instance, if we are trying to estimate the mean parameter μ of a Normal, we might call our estimator $\hat{\mu}$.

Definition: The estimator $\hat{\theta}$ for a parameter θ is said to be **unbiased** if

$$E[\hat{\theta}] = \theta.$$

The **bias** of $\hat{\theta}$ is how far the estimator is from being unbiased. It is defined by

$$\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

Example: Estimating the mean μ of a Normal. If we choose the sample mean as our estimator, i.e., $\hat{\mu} = \bar{X}_n$, we have already seen that this is an unbiased estimator:

$$E[\bar{X}_n] = E[X_i] = \mu.$$

Example: Estimating the variance σ^2 of a Normal. If we choose the sample variance as our estimator, i.e., $\hat{\sigma}^2 = S_n^2$, it becomes clear why the $(n-1)$ is in the denominator: it is there to make the estimator unbiased. First, remember the formula $\mathrm{Var}(X) = E[X^2] - E[X]^2$. Using this, we can show that

$$E[X_i^2] = \mathrm{Var}(X_i) + E[X]^2 = \sigma^2 + \mu^2, \text{ and}$$

$$E[\bar{X}_n^2] = \mathrm{Var}(\bar{X}_n) + E[\bar{X}_n]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Also, notice that because the X_i are independent, we have $Cov(X_i, X_j) = 0$ if $i \neq j$. Thus, for $i \neq j$,

$$E[X_i X_j] = \mathrm{Cov}(X_i, X_j) + E[X_i]E[X_j] = \mu^2.$$

Using this, we can compute

$$E\left[X_i \bar{X}_n\right] = E\left[X_i \frac{1}{n}\sum_{j=1}^{n} X_j\right] = \frac{1}{n}\sum_{j=1}^{n} E\left[X_i X_j\right] = \frac{\sigma^2}{n} + \mu^2$$

Putting these all together, we get

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

If we had put n in the denominator, we would have gotten

$$E\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = \frac{n-1}{n}\sigma^2.$$

Example: Estimating the proportion parameter p for a Bernoulli distribution. If X_i are iid $\mathrm{Ber}(p)$ random variables, then we know that $E[X_i] = p$. Therefore, the mean statistic also has $E[\bar{X}_n] = p$ and is thus an unbiased estimator of p.

It is possible that two estimates, $\hat{\theta}_1, \hat{\theta}_2$, of a parameter θ are *both* unbiased. How do we decide which is the best to use? Well, we'd want the one that has the least amount of variability, that is, it is more likely to fall close to the true answer. The estimator $\hat{\theta}_1$ is said to be **more efficient** than $\hat{\theta}_2$ if

$$\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2).$$