Principles of Parametric Inference

Moulinath Banerjee

University of Michigan

September 11, 2012

The object of statistical inference is to glean information about an underlying population based on a sample collected from it. The actual population is assumed to be described by some probability distribution. Statistical inference is concerned with learning about the distribution or at least some characteristics of the distribution that are of scientific interest.

In parametric statistical inference, which we will be primarily concerned with in this course, the underlying distribution of the population is taken to be parametrized by a Euclidean parameter. In other words, there exists a subset Θ of k-dimensional Euclidean space such that the class of distributions $\mathcal{P}$ of the underlying population can be written as $\{P_\theta : \theta \in \Theta\}$. You can think of the θ's as labels for the class of distributions under consideration. More precisely this will be our set-up: Our data $X_1, X_2, \ldots X_n$ are i.i.d. observations from the distribution P_θ where $\theta \in \Theta$, the parameter space. We assume identifiability of the parameter, i.e. $\theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2}$. In general, we will also assume that X_1 has a density $f(x, \theta)$ (this can either be a probability mass function or an ordinary probability density function). Here x is a typical value assumed by the random variable. Thus, $f(x, \theta)$ for a discrete random variable X_1 just gives us the probability that X_1 assumes the value x when the underlying parameter is indeed θ. For a continuous random variable, $f(x, \theta)$ gives us the density function of the random variable X_1 at the point x when θ is the underlying parameter. Thus $f(x, \theta)\, dx$ where dx is a very small number is approximately the probability that X_1 lives in the interval $[x, x + dx]$ under parameter value θ.

We will be interested in estimating θ, or more generally, a function of θ, say $g(\theta)$.

Let us consider a few examples that will enable us to understand these notions better.

(1) Let $X_1, X_2, \ldots, X_n$ be the outcomes of n independent flips of the same coin. Here, we code $X_i = 1$ if the i'th toss produces H and 0 otherwise. The parameter of interest is θ,

the probability of H turning up in a single toss. This can be any number between 0 and 1. The X_i's are i.i.d. and the common distribution P_θ is the Bernoulli(θ) distribution which has probability mass function:

$$f(x,\theta) = \theta^x \, (1-\theta)^{1-x} \quad , x \in \{0,1\} \, .$$

Check that this is indeed a valid expression for the p.m.f. Here the parameter space, i.e. the set of all possible values for θ is the closed interval $[0,1]$.

(2) Let $X_1, X_2, \ldots, X_n$ denote the failure times of n different bulbs. We can think of the X_i's as independent and identically distributed random variables from an exponential distribution with an unknown parameter θ which we want to estimate. If $F(x,\theta)$ denotes the distribution function of X_1 under parameter value θ, then

$$F(x,\theta) = P_{\theta \, \text{is true parameter}}(X_1 \leq x) = 1 - e^{-\theta \, x} \, .$$

The common density function is given by,

$$f(x,\theta) = \theta \, e^{-x \, \theta} \, .$$

Here the parameter space for θ is $(0,\infty)$.

Note that θ is very naturally related to the mean of the distribution. We have $E_\theta(X_1) = 1/\theta$. The expression $E_\theta(X_1)$ should be read as the *expected value of* X_1 *when the true parameter is* θ. In general, whenever I write an expression with θ as a subscript, interpret that as *under the scenario that the true underlying parameter is* θ.

(3) Let $X_1, X_2, \ldots, X_n$ be the number of customers that arrive at n different identical counters in unit time. Then the X_i's can be thought of as i.i.d. random variables with a (common) Poisson distribution with mean θ. Once again θ which is also the parameter that completely specifices the Poisson distribution varies in the set $(0,\infty)$ which therefore is the parameter space Θ. The probability mass function is:

$$f(x,\theta) = \frac{e^{-\theta} \, \theta^x}{x \, !} \, .$$

(4) Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from a Normal distribution with mean μ and variance σ^2. The mean and the variance completely specify the normal distribution. We can take the parameter $\theta = (\mu, \sigma^2)$. Thus, we have a two dimensional parameter and Θ, the set of all possible values of θ is the set in $\mathbb{R}^2$ given by $(-\infty,\infty) \times (0,\infty)$. The density function $f(x,\theta)$ is then given by:

$$f(x,\theta) = \frac{1}{\sqrt{2\,\pi}\,\sigma} \exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right] \, .$$

Note that each different value of θ gives you a different normal curve. if you fix μ, the first component of θ and vary σ^2 you get a family of normal (density) curves all centered at μ but with varying spread. A smaller value of σ^2 corresponds to a curve that is more peaked about μ and also more tightly packed around it. If you fix σ^2 and vary μ you get a family of curves that are all translates of a fixed curve (say, the one centered at 0).

Consider now, the problem of estimating $g(\theta)$ where g is some function of θ. In many cases $g(\theta) = \theta$ itself; for example, we could be interested in estimating θ, the probability of H in Example 1 above. Generally $g(\theta)$ will describe some important aspect of the distribution P_θ. In Example 1, $g(\theta) = \theta$ describes the probability of the coin landing heads; in Example 3, $g(\theta) = 1/\theta$ is the expected value of the lifetime of a bulb. Our estimate of $g(\theta)$ will be some function of our observed data $X = (X_1, X_2, \ldots, X_n)$. We will generically denote an estimate of $g(\theta)$ by $T_n(X_1, X_2, \ldots, X_n)$ and will write T_n for brevity. Thus T_n is some function of the observed data and is therefore a random variable itself.

Let's quickly look at an example. In Example 1, a natural estimate of θ, as we have discussed before, is $\overline{X}_n$, the mean of the X_i's. This is simply the sample proportion of Heads in n tosses of the coin. Thus

$$T_n(X_1, X_2, \ldots, X_n) = \frac{X_1 + X_2 + \ldots + X_n}{n}\,.$$

By the WLLN $\overline{X}_n$ converges in probability to θ and is therefore a reasonable estimator, at least in this sense. Of course, this is not the only estimator of θ that one can propose (but this is indeed the best estimator in more ways than one). One could also propose the proportion of heads in the first m tosses of the coin as an estimator, m being the floor of $n/2$. This will also converge in probability to θ as $n \to \infty$, but its variance will always be larger than that of $\overline{X}_n$.

In general there will be several different estimators of $g(\theta)$ which may all seem reasonable from different perspectives – the question then becomes one of finding the most optimal one. This requires an objective measure of the performance of the estimator. If T_n estimates $g(\theta)$ a criterion that naturally suggests itself is the distance of T_n from $g(\theta)$. Good estimators are those for which $\mid T_n - g(\theta) \mid$ is generally small. Since T_n is a random variable no deterministic statement can be made about the absolute deviation; however what we can expect of a good estimator is a high chance of remaining close to $g(\theta)$. Also as n, the sample size, increases we get hold of more information and hence expect to be able to do a better job of estimating $g(\theta)$. These notions when coupled together give rise to the consistency requirement for a sequence of estimators T_n; as n increases, T_n ought to converge in probability to $g(\theta)$ (under the probability distribution P_θ). In other words, for any $\epsilon > 0$,

$$P_\theta\left(\mid T_n - g(\theta) \mid > \epsilon\right) \to_P 0\,.$$

The above is clearly a large sample property; what it says is that with probability increasing to 1 (as the sample size grows), T_n estimates $g(\theta)$ to any pre-determined level of accuracy. However, the consistency condition alone, does not tell us anything about how well we are performing for any particular sample size, or the rate at which the above probability is going to 0.

For a fixed sample size n, how do we measure the performance of an estimator T_n? We have seen that $\mid T_n - g(\theta) \mid$ is itself random and therefore cannot even be computed as a function of θ before the experiment is carried out. A way out of this difficulty is to obtain an average measure of the error, or in other words, average out $\mid T_n - g(\theta) \mid$ over all possible realizations of T_n. The resulting quantity is then still a function of θ but no longer random It is called the mean absloute error and can be written compactly (using acronym) as:

$$M.A.D. = E_\theta \mid T_n - g(\theta) \mid .$$

However, it is more common to avoid absolute deviations and work with the square of the deviation, integrated out as before over the distribution of T_n. This is called the mean-squared error (M.S.E.) and is

$$M.S.E.(T_n, g(\theta)) = E_\theta \ (T_n - g(\theta))^2 .$$

Of course, this is meaningful, only if the above quantity is finite for all θ. Good estimators are those for which the M.S.E. is generally not too high, whatever be the value of θ. There is a standard decomposition of the M.S.E. that helps us understand its components. We have,

$$\begin{aligned} M.S.E,(T_n, g(\theta)) &= E_\theta \ (T_n - g(\theta))^2 \\ &= E_\theta \ (T_n - E_\theta(T_n) + E_\theta(T_n) - g(\theta))^2 \\ &= E_\theta \ (T_n - E_\theta(T_n))^2 + (E_\theta(T_n) - g(\theta))^2 + 2\, E_\theta[(T_n - E_\theta(T_n))\,(E_\theta(T_n) - g(\theta))] \\ &= \mathrm{Var}_\theta \,(T_n) + b(T_n, \theta)^2 \,, \end{aligned}$$

where $b(T_n, g(\theta)) = E_\theta(T_n) - g(\theta)$ is the bias of T_n as an estimator of $g(\theta)$ (the cross product term in the above display vanishes since $E_\theta(T_n) - g(\theta)$ is a constant and $E_\theta(T_n - E_\theta(T_n)) = 0$). It measures, on an average, by how much T_n overestimates or underestimates $g(\theta)$. If we think of the expectation $E_\theta(T_n)$ as the center of the distribution of T_n, then the bias measures by how much the center deviates from the target. The variance of T_n, of course, measures how closely T_n is clustered around its center. Ideally one would like to minimize both simultaneously, but unfortunately this is rarely possible. Two estimators T_n and S_n can be compared on the basis of their mean squared errors. Under parameter value θ, T_n dominates S_n as an estimator if $M.S.E.(T_n, \theta) \leq M.S.E.(S_n, \theta)$. We say that S_n is inadmissible in the presence of T_n if

$$M.S.E.(T_n, \theta) \leq M.S.E.(S_n, \theta) \ \ \forall \ \ \theta \,.$$

The use of the term "inadmissible" hardly needs explanation. If, for all possible values of the parameter, we incur less error using T_n instead of S_n as an estimate of $g(\theta)$, then clearly there is no point in considering S_n as an estimator at all. Continuing along this line of thought, is there an estimate that improves all others? In other words, is there an estimator that makes every other estimator inadmissible ? The answer is no, execpt in certain pathological situations.

As we have noted before, it is generally not possible to find a universally best estimator. One way to try to construct optimal estimators is to restrict oneself to a subclass of estimators and try to find the best possible estimator in this subclass. One arrives at subclasses of estimators by constraining them to meet some desirable requirements. One such requirement is that of *unbiasedness*. Below, we provide a formal definition.

Unbiased estimator: An estimator T_n of $g(\theta)$ is said to be unbiased if $E_\theta(T_n) = g(\theta)$ for all possible values of θ; i.e. $b(T_n, g(\theta)) = 0$.

Thus, unbiased estimators, on an average, hit the target value. This seems to be a reasonable constraint to impose on an estimator and indeed produces meaningful estimates in a variety of situations. Note that for an unbiased estimator T_n, the M.S.E under θ is simply the variance of T_n under θ. In a large class of models, it is possible to find an unbiased estimator of $g(\theta)$ that has the smallest possible variance among all possible unbiased estimators. Such an estimate is called an MVUE (minimum variance unbiased estimator). Here is a formal definition.

MVUE: We call S_n an MVUE of $g(\theta)$ if (i) $E_\theta(S_n) = g(\theta)$ for all values of θ and (ii) if T_n is an unbiased estimate of $g(\theta)$, then $\mathrm{Var}_\theta(S_n) \leq \mathrm{Var}_\theta(T_n)$.

Here are a few examples to illustrate some of the various concepts discussed above.

(a) Consider Example (4) above. A natural unbiased estimator of $g_1(\theta) = \mu$ is $\overline{X}_n$, the sample mean. It is also consistent for μ by the WLLN. It can be shown that this is also the MVUE of μ. In other words, *any* other unbiased estimate of μ will have a larger variance than $\overline{X}_n$. Recall that the variance of $\overline{X}_n$ is simply σ^2/n.

Consider now, the estimation of σ^2. Two estimates of this that we have considered in the past are

$$(i)\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \ \text{ and } \ (ii)\ s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\,.$$

Out of these $\hat{\sigma}^2$ is not unbiased for σ^2 but s^2 is, as you will show in a homework

exercise. In fact s^2 is also the MVUE of σ^2.

(b) Let $X_1, X_2, \ldots, X_n$ be i.i.d. from some underlying density function or mass function $f(x, \theta)$. Let $g(\theta) = E_\theta(X_1)$. Then the sample mean $\overline{X}_n$ is *always* an unbiased estimate of $g(\theta)$. Whether it is MVUE or not depends on the underlying structure of the model.

(c) In Example 1 above, $\overline{X}_n$ is the MVUE of θ. Now define $g(\theta) = \theta/(1-\theta)$. This is a quantity of interest because it is precisely the odds in favour of Heads. It can be shown that there is *no unbiased estimator* of $g(\theta)$ in this model. However an intuitively appealing estimate of $g(\theta)$ is $T_n \equiv \overline{X}_n/(1-\overline{X}_n)$. It is *not unbiased* for $g(\theta)$; however it does converge in probability to $g(\theta)$. This example illustrates an important point – unbiased estimators may not always exist. Hence imposing unbiasedness as a constraint may not be meaningful in all situations.

(d) Unbiased estimators are not always better than biased estimators. Remember, it is the MSE that gauges the performance of the estimator and a biased estimator may actually outperform an unbiased one owing to a significantly smaller variance. Consider the situation where $X_1, X_2, \ldots, X_n$ are i.i.d. Uniform(0,θ). Here $\Theta = (0, \infty)$. A natural estimate of θ is the maximum of the X_i's, which we denote by $X_{(n)}$. Another estimate of θ is obtained by observing that $\overline{X}$ is an unbiased estimate of $\theta/2$, the common mean of the X_i's; hence $2\,\overline{X}_n$ is an unbiased estimate of θ. It can be shown that $X_{(n)}$ in the sense of M.S.E outperforms $2\,\overline{X}$ by an order of magnitude. The best unbiased estimator (MVUE) of θ is $(1 + n^{-1})\,X_{(n)}$.