# A General Overview of Parametric Estimation and Inference Techniques.

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 $\Theta$ of $k$-dimensional Euclidean space such that the class of dsitributions $\mathcal{P}$ of the underlying population can be written as $\{P_\theta : \theta \in \Theta\}$. One key assumption made at this stage is that of identifiability; namely that the map $\theta \mapsto P_\theta$ is one-one. Thus knowing $\theta$ is equivalent to knowing the underlying distribution.

In many cases important functionals of the underlying distribution $P_\theta$ can be written as nice functions of $\theta$. We shall study classes of parametric families for which this is the case. Also, whenever we talk of a parametric family $\{P_\theta\}$ we shall assume that the distributions in the class all admit densities. Thus instead of talking about $P_\theta$, we can equally well, talk about $\{f(x, \theta) : \theta \in \Theta\}$.

Thus our set-up will be as follows: We observe $X_1, X_2, \ldots X_n$, drawn at random and with replacement from a population with underlying density $f(x, \theta)$, where $\theta$ varies in $\Theta$. For the sake of mathematical nicety, $\Theta$ is usually assumed to be open; however this need not necessarily be the case, and indeed there are situations of practical interest where $\Theta$ can have a boundary. Thus $X_1, X_2, \ldots, X_n$ are i.i.d. $f(x/\theta)$ and we typically seek to estimate $\theta$ or $g(\theta)$, where $g$ is some real valued function of $\theta$. We will estimate $g(\theta)$ by a function $T_n$ of $(X_1, X_2, \ldots, X_n)$, our observed random vector. In what follows, for the sake of compactness, we shall use $T_n$ to denote the value of the estimator itself, instead of (the more formal) $T_n(X_1, X_2, \ldots, X_n)$. Much of statistical inference has to do with coming up with smart/good estimators of $g(\theta)$; of course "smart/good" needs to be made technically precise and once certain criteria have been set and estimators $T_n$'s proposed,

a lot of work goes into studying how far the $T_n$'s meet the established standards.

There are several questions, basically, which raise themselves at this point. Let's try to list some of them.

- How do we construct rich families of parametric models that we can use for data that arise in real life ? In particular what are desirable properties of such families ?
- Having given ourselves a parametric model, how do we construct meaningful estimators $T_n$ of $g(\theta)$ ?
- How do we judge the goodness of an estimator $T_n$ of $g(\theta)$ for fixed sample size and also as the sample size becomes large ? How do we compare two estimators ?

Let's briefly look at these questions. We start off with the third. Since $T_n$ is a function of the $X_i$'s it is itself a random variable. Now, a criterion that naturally suggests itself, when trying to gauge $T_n$ as an estimate of $g(\theta)$, 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, $T_n$ converges in probability to $g(\theta)$ (under the probability distribution $P_\theta$). In other words, for any $\epsilon > 0$,

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

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 doing for any particular sample size, or the rate at which the above probability is going to 0. However there are several well-known probability inequalities that enable us to put upper bounds on probabilities of the above type; the most well-known is the Markov inequality and more particularly (a special case) Chebyshev's inequality. We will deal with these later.

For a fixed sample size $n$, how do we measure the performance of an estimator $T_n(x)$? We have seen that $\mid T_n - g(\theta) \mid$ is itself random and therefore cannot even be computed as a function of $\theta$ 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 $\theta$ but no longer random It is called the mean absloute error and can be written compactly (using acronym) as:

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

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

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

Of course, this is meaningful, only if the above quantity is finite for all $\theta$. Good estimators are those for which the M.S.E. is generally not too high, whatever be the value of $\theta$. There is a standard decomposition of the M.S.E. as follows:

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

where $b(T_n, \theta) = E_\theta(T_n) - g(\theta)$ is the bias of $T_n$ as an estimator of $g(\theta)$. 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 $\theta$, $T_n$ 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 and is quite fortunately so; otherwise statistical inference would hardly be something worth pursuing. To see this, fix any value of $\theta$ and call it $\theta_0$. Now, define the estimator, $S_n(X_1, X_2, \ldots, X_n) \equiv g(\theta_0)$. Then $M.S.E.(S_n, \theta_0) = 0$. Thus, any uniformly best estimate $T_n$ would need to satisfy $M.S.E.(T_n, \theta_0) = 0$, and since $\theta_0$ is arbitrary, this would imply,

$$M.S.E.(T_n, \theta) = 0 \ \ \forall \ \ \theta \, ,$$

and this is impossible except in very trivial (pathological) cases. In particular, if we take two different values $\theta_1$ and $\theta_2$, the set $A_1$ on which $T_n = g(\theta_1)$ has probability 1 under $P_{\theta_1}$ and the set $A_2$ on which $T_n = g(\theta_2)$ and which lies in $A_1^C$ (and hence has probability 0 under $P_{\theta_1}$)has probability 1 under $P_{\theta_2}$. Thus, the probability measures $P_{\theta_1}$ and $P_{\theta_2}$ are mutually singular (to be explained) for any two different values of the parameter $\theta_1$ and $\theta_2$ such that $g(\theta_1) \neq g(\theta_2)$. Hardly, any meaningful statistical model belongs to this category.

A way out of this difficulty is to restrict ourselves to a class of procedures so that we do not need to deal with stupid estimates like $S_n$ as defined above, and look among this restricted class of estimates for a best estimate. Thus we look at a class of estimators of $g(\theta)$, subject to certain

pre-specified constraints. One such constraint is that of unbiasedness. Formally, an estimator $T_n$ is said to be **unbiased** for $g(\theta)$ if

$$E_\theta(T_n) = g(\theta) \ \ \forall \theta \, ,$$

or equivalently

$$b_\theta(T_n) = 0 \ \ \forall \theta \, ,$$

As mentioned before, ideally, we want estimates with low bias and high precision (low variance). Since bias represents systematic error, a reasonable approach is to control for bias before trying to control precision. We shall see that the notion of unbiasedness has some very appealing features and leads to a fundamental inequality in statistical inference called the Cramer-Rao inequality which imposes a lower bound (information bound) on the variance of an unbiased estimator in a regular parametric model. In particular, among unbiased estimators, it is possible in many cases (once again under suitable regularity conditions on the model) to produce one that uniformly beats (has lower M.S.E. for all $\theta$ or equivalently the lowest variance) any other unbiased estimator. Such an estimator is necessarily unique and is called the UMVUE (uniformly minimum variance unbiased estimator). However, attractive as the notion of unbiasedness is, it in no way guarantees the admissibility of an estimator, even in meaningful statistical models. We shall later on, see examples where a biased estimator beats an unbiased estimator very comprehensively.

So much for our third question. We shall have more to talk about it later. The next question on the agenda is of course, how to construct meaningful estimators, unbiased or otherwise. There are several different ways of doing this, depending on what sort of crtieria we want these estimators to meet. In this course, we shall primarily deal with two different estimation techniques, these being (a) Method of Moments (b) Maximum Likelihood Estimation. In linear regression we come across a related method, called the method of least squares, which is also extensively applied in a large variety of models. Method of Moments estimates involve expressing the parameter (or a function of the parameter) as a function of the population moments and then approximating the population moments by sample moments to get plug-in estimates of the parameter. Maximum Likelihood Estimation involves maximizing the joint density of the observations as a function of the parameter $\theta$ for a fixed observed sample $X_1, X_2, \ldots, X_n$, and proposing that $\theta$ that maximizes the joint likelihood as an estimate; in other words choose that value of $\theta$ for which the observed data are the most likely. Of all estimation techniques in frequentist inference, MLE's are probably the most extensively studied. In a host of nicely behaved parametric models, MLE's enjoy many attrctive properties. Firstly, there is a lot of intuitive appeal in the whole idea of choosing that value of the parameter under which the observed data is the most plausible. Secondly, not only are MLE's consistent, they are also **asymptotically unbiased** and **asymptotically efficient**. Basically, it can be shown, that under proper regularity conditions on the statistical model,

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \rightarrow_d N(0, I^{-1}(\theta)) \, ,$$

where $\hat{\theta}_n$ is the MLE of $\theta$ based on i.i.d. observations $X_1, X_2, \ldots, X_n$ from $f(x, \theta)$. Here $I^{-1}(\theta)$ is the lower bound (information bound) for an unbiased estimator of $\theta$ based on one realization $X$ from $f(x, \theta)$. Now, what the above display implies is that, the distribution of $\hat{\theta}_n$ for $n$ *sufficiently*

*large* is well-approximated by a $N(\theta, n^{-1}\, I^{-1}(\theta))$ distribution. Thus, for *large* $n$, the behavior of $\hat{\theta}_n$ is captured by a random variable with mean $\theta$, the quantity we are trying to estimate (asymptotic unbiasedness), and variance $n^{-1}\, I^{-1}(\theta)$, this being precisely the lower bound on the variance of an unbiased estimator based on $n$ i.i.d. observations from the population (asymptotic efficiency). In fact, among the class of "regular" estimators, $\hat{\theta}_n$ has the smallest asymptotic variance. Suppose that $T_n$ is a sequence of "regular" estimators of $\theta$ (we do not spit out the technicalities) with

$$\sqrt{n}\,(T_n - \theta) \rightarrow_d N(0, v(\theta))$$

for some $v(\theta) > 0$. Note that $T_n$ is asymptotically unbiased. It can be shown that $v(\theta) \geq I^{-1}(\theta)$; in other words, the limiting ratio of the asymptotic variance of $T_n$ ( $n^{-1}\, v(\theta)$ ) to that of the asymptotic variance of the MLE ($n^{-1}\, I^{-1}(\theta)$) is at least 1. Thus $T_n$ is asymptotically more spread out about $\theta$ than the MLE $\hat{\theta}_n$, and in this sense more efficient.

Construction of unbiased estimators in parametric statistical models can involve various levels of difficulty; the challenge however is to find the UMVUE, if one exists. Note that UMVUE's need not exist (this will be illustrated later by an example), and in fact, even if they do, can be foolish. To obtain UMVUE's, what one does is obtain any unbiased estimator of the parameter and then condition by a **complete sufficient statistic** for the model. The resulting object is the unique UMVUE, regardless of the initial unbiased statistic that we started out with. Roughly speaking, a **sufficient** statistic is a function (and hence a coarsening) of the data that contains all the information about the parameter $\theta$. A statistic $T$ is said to be **complete** if the only function of $T$ that can unbiasedly estimate 0 for all possible values of the parameter is the 0 function itself. Thus no non-trivial function of a complete statistic can unbiasedly estimate 0 under all possible dsitributions in the model. A **complete sufficient statistic** is a statistic that is both complete and sufficient. Why conditioning by a complete sufficient statistic works is the content of two very crucial theorems in statistical inference, called the **Rao-Blackwell** theorem and the **Lehmann-Scheffe** theorem. Basically, the Rao-Blackwell theorem provides a way of reducing the variance of an unbiased estimator by conditioning on a sufficient statistic. Conditioning preserves unbiasedness and sufficiency guarantees that the resulting object is still a bona-fide statistic in that it does not depend on the parameter. Once the statistic we condition upon also has the virtue of completeness, the uniqueness follows from the inherent idea of uniqueness associated with the notion of completeness. These issues will be emphasized in greater detail, later.

There still remain a few words to be said about the classes of parametric models used in statistics. Obviously, the ideas above could not have been used very fruitfully unless there was a sufficiently rich and well-behaved class of models to apply them to. Fortunately, the exponential family of distributions provides us with such a class. They have been studied very extensively and indeed much of the core of parametric statistics deals with models belonging to this class or closely allied classes. Many of the well-known statistical distributions like the Binomial, Poisson, Negative Binomial, Normal, Gamma, Beta etc. belong to this class. Exponential families are mathematically very tractable in that one can immediately read off sufficient statistics for the parameter and check quite easily for completeness. Thus UMVUE's are easy to construct.

Furthermore, maximum likelihood estimates for exponential families can also be obtained quite neatly, as can information bounds. Finally exponential families also have the desirable properties that (a) the joint distribution of $n$ i.i.d. random variables, each coming from a (fixed) distribution in the exponential family, also belongs to an exponential family; (b) the dimension of the natural sufficient statistic for the parameter does not change with sample size; in other words, the natural sufficient statistic for the parameter based on the joint distribution of $n$ i.i.d. observations has the same dimension as that for 1 observation.

In what has been described above we have introduced some key concepts and entities that figure largely in statistical inference. As is easy to see, these concepts are all quite inter-twined and it requires understanding each of these individually as well as how they stand in relation to one another to obtain a frim grasp of the key ideas involved. Sufficiency is the art of condensing the observed data without losing any information on the parameter. In conjunction with the notion of completeness, it leads to methods of constructing best unbiased estimates. Best unbiased estimates can be compared to a benchmark, which is the information bound. Maximum likelihood estimates, though not necessarily unbiased, are in general asymptotically so, and attain the information bound, beating other asymptotically unbiased estimators. Exponential families which provide a plethora of interesting distributions are used extensively in statistical modelling, not only because of underlying scientific reasons, but also because of the fact that the inferential concepts developed can be applied elegantly to these families with neat and satisfactory results.