Parametric Inference

Snehal M. Shekatkar

Department of Scientific Computing, Modeling, and Simulation,
Savitribai Phule Pune University, Pune, India 411007

So far we have been talking about *nonparametric inference* in which we have been given sample $X_1, X_2, \cdots, X_n$ that is assumed to have come from an unknown distribution F. Since F is completely unknown, estimation of only those quantities which can be defined for any distribution (i.e. those which are 'distribution-free') makes sense. As we have already seen, the *mean* and *variance* of the distribution are examples of such quantities. But now assume that the 'shape' (i.e. mathematical form) of F is not unknown but the values of the parameters which control the shape are unknown. For example, we may somehow know that the sample $X_1, X_2, \cdots, X_n$ has come from a Binomial distribution but the value of the success probability p is unknown. Such quantity to be estimated is a parameter of the *known* distribution, and since it is defined only for that particular distribution, it is not 'distribution-free'. Such problem of estimating one or more parameters of a given distribution from a sample is known as **parametric inference**.

Suppose that the known distribution F has several parameters collectively denoted as θ. You may think of theta as a vector of length k where k is the total number of parameters whose values are unknown. That is, the distribution F may depend on more than k parameters, but only k of them are uknown and need to be estimated from the sample. Such unknown parameters are called as **parameters of interest**, while the ones whose values are either already known or whose estimation is not of our interest are called **nuisance parameters**. For example, suppose that we know that the underlying distribution is normal $\mathcal{N}(\mu, \sigma^2)$ where both μ and σ are unknown and need to be estimated. Then both these are parameters of interest and then $k = 2$. On the other hand, if we either already know σ or are not interested in its value, then σ is a nuisance parameters and then $k = 1$.

1 Method of Moments (MoM)

The first method in the parametric inference that we are going to look at is called the 'method of moments'. The basic idea of MoM is to equate the first k moments of the distribution F with the k sample moments. This gives us k equations in k unknowns, and whose solution gives us estimates of the k unknown parameters. Formally, let $\alpha_j(\theta) = \mathbb{E}(X^j)$. Then the MoM estimator of θ, denoted by $\widehat{\theta}_n$ is defined as:

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \alpha_2(\widehat{\theta}_n) &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ &\vdots \\ \alpha_k(\widehat{\theta}_n) &= \frac{1}{n}\sum_{i=1}^{n} X_i^k \end{aligned} \tag{1}$$

Recall that $\widehat{\theta}_n$ denotes k separate parameters. These k equations can then be solved simultaneously to get estimators for k parameters of interest. The following example will demonstrate the working of this method.

Example 1

Let $X_1, X_2, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where both μ and σ are unknown. Find the method of moments estimators for these parameters.

Solution: In this case $k = 2$ and $\theta = (\mu, \sigma)$

$$\alpha_1(\theta) = \mathbb{E}(X_1) = \mu$$
$$\alpha_2(\theta) = \mathbb{E}(X_1^2) = \mathbb{V}(X_1^2) + (\mathbb{E}(X_1))^2 = \sigma^2 + \mu^2$$

Replacing the parameters in these equations by the corresponding estimators, and then equating with the corresponding sample moments, we get:

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\widehat{\mu}_n + \widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Using first of these equations into the second, we get:

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i - \widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} X_i$$

Which gives (see plug-in estimators notes for the actual derivation):

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Hence the method of moments estimators in this case are:

$$\boxed{\begin{aligned} \widehat{\mu}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \widehat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \end{aligned}} \tag{2}$$

As you can see, the MoM estimator for σ^2 is identical with the plug-in estimator. However, there is a conceptual difference between the two. The plug-in estimator corresponds to the statistical functional *variance* that is defined for *any* distribution, and so is a nonparametric estimator. On the other hand, σ in the Gaussian distribution is a particular parameter that appears in the mathematical formula for it. You should also note that this estimator is biased. This tends to be a general feature of MoM estimators. However, the bias goes to zero as $n \to \infty$. In fact we can state the following two general properties for MoM estimators:

1. MoM estimators are consistent: $\widehat{\theta}_n \xrightarrow{P} \theta$

2. MoM estimators are asymptotically normal: $\frac{\widehat{\theta}_{n,j} - \theta_j}{\sigma_j} \rightsquigarrow \mathcal{N}(0,1)$
   (Here θ_j represents j^{th} component of the parameter vector θ, and σ_j denotes the standard deviation of the sampling distribution of θ_j)

To find the associated confidence intervals, one can use the bootstrap method.

2 Maximum Likelihood Estimators (MLE)

The maximum likelihood method was introduced by Ronald Fisher, and is one of the most widely methods in inferential statistics. The fundamental principle behind the method is that the most likely values of the parameters of the model are those which maximize the probability of observing the given data. Thus, when the parameters θ are unknown, choose their values which maximize $\mathbb{P}(X_1, X_2, \cdots, X_n \mid \theta)$. This probability of observing data given parameters is known as *Likelihood*, and hence the name.

Consider a simple example: you toss a biased coin 10 times and observe that 9 out of 10 tosses resulted into *Heads*. What is your best guess for the probability p of observing *Heads*? It looks rather unlikely that $p = 0.1$ because in that case we would have probably observed many *Tails*, not *Heads*. In fact with $p = 0.1$, the probability of observing the sample with 9 *Heads* and 1 *Tails* would then be 9×10^{-9} which is too small (prove this!). Although it is not impossible to observe a sample with such small probability, the maximum likelihood simply tells us to maximize this probability. If so, about $p = 1$? This also looks impossible because we have observed one *Tails*, and $p = 0$ tells us that the probability of observing this particular sample is precisely zero! Thus, both choices $p = 0.1$ and $p = 1$ fail to maximize the probability of observed data.

Since we have been assuming that the random variables X_i are independent, we can factorize the likelihood:

$$\mathbb{P}(X_1, X_2, \cdots, X_n \mid \theta) = \mathbb{P}(X_1 \mid \theta) \times \mathbb{P}(X_1 \mid \theta) \cdots \times \mathbb{P}(X_1 \mid \theta) = \prod_{i=1}^{n} \mathbb{P}(X_i \mid \theta)$$

Here $\prod$ denotes the product. When random variables are continuous, we replace the probabilities by the probability densities $p(X_i \mid \theta)$. observe that the product written above is a function of unknown parameters θ, not of the data X_i which is fixed. Thus, formally we defined the *likelihood* or *likelihood function* as:

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} p(X_i \mid \theta) \tag{3}$$

If X_i are discrete, $p(X_i \mid \theta)$ would denote probability, otherwise it is density. Given the data $X_1, X_2, \cdots, X_n$, how can we maximize $\mathcal{L}(\theta)$? Since it is a function of θ, we can directly equate its derivative w.r.t. θ to 0 and get the local extrema. Then second derivative would tell us which of these are maxima and we could pick the maximum among them. Although this is straightforward in theory, in practice the calculations may become very complicated. However, there exists a trick that is almost always used to simply this task. To understand it, let us define **log-likelihood** as the logartihm of the likelihood function:

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p(X_i \mid \theta) \tag{4}$$

Now observe that since logarithm is a monotonic increasing function, the value of θ for which log-likelihood becomes maximum is the same value for which likelihood becomes maximum! Therefore, we can actually maximize the log-likelihood instead of the original likelihood.

But how does this help? First of all, observe that after taking log, we get sum instead of product. When it comes to differentiating a function to maximize it, this itself is a huge simplification. Also, quite often we deal with probability densities involving exponentials (e.g. Gaussian, Gamma etc), and taking the log gets rid of the exponentials. Finally, if you are trying to maximize the likelihood computationally, then you may encounter underflow or overflow on your computer. Since logarithm is a slowly increasing function, this greatly reduces the chances of encoutering values that are too large or too small.

Let us demonstrate the method with an example.

Example 2

Let $X_1, X_2, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ where both μ and σ are unknown and X_i are IID. Find the maximum likelihood estimators for these parameters.

Solution: The likelihood function is:

$$\mathcal{L}_n(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X_i - \mu)^2}{2\sigma^2}\right) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)$$

Thus, the log-likelihood is:

$$\ell_n(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Equating the derivatives of $\ell_n(\mu, \sigma)$ w.r.t. μ and σ, we get the corresponding MLEs $\widehat{\mu}_n$ and $\widehat{\sigma}_n$. First differentiate w.r.t. μ :

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(X_i - \widehat{\mu}_n) = 0$$

Rearranging (with the assumption that $\widehat{\sigma}$), we get:

$$\boxed{\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i} \tag{5}$$

Now differentiating $\ell_n(\mu, \sigma)$ w.r.t. σ and equating with zero, we get:

$$-\frac{n}{\widehat{\sigma}} + \frac{2}{2\widehat{\sigma}^3}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2 = 0 \tag{6}$$

Again rearranging assuming $\widehat{\sigma} \neq 0$, we get:

$$\boxed{\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2}$$

Thus, in this case, ML estimators are same as the MoM estimators although this is not always true.

Maximum Likelihood estimators, just like the MoM estimators, are consistent and asymptotically normal. Even in this case one can use bootstrap to construct confidence intervals. This is called 'parametric bootstrap'. However, for MLEs, it is often possible to find the estimate of the standard error analytically, but we won't go into that discussion here.