Statistics for Finance

1. LECTURE 3: ESTIMATION AND LIKELIHOOD.

One of the central themes in mathematical statistics is the theme of parameter estimation. This relates to the fitting of probability laws to data. Many families of probability laws depend on a small number of parameters. For example, the normal distributions are determined by the mean $\mu$ and the standard deviation $\sigma$. Even though, one may make a reasonable assumption on the type of the distribution, e.g. normal, one usually does not know the parameters of the distribution, e.g. mean and standard deviation, and one needs to determine these from the available data.

The philosophical foundation of our approach is that sample data, say, $X_1, X_2, \ldots, X_n$, of a sample of size $n$, are thought of as a (subset of an infinite) collection of independent, identically distributed (i.i.d.) random variables, following the probability distribution in question.

A bit of explanation is required at this point. We are used of sample data having the form of real numbers. When for example we measure the heights of a sample of 5 students in Warwick, we may record heights 178, 189, 170, 160, 164. So what does $X_1, X_2, X_3, X_4, X_5$ stand for? The answer is that, although, we may end up with concrete real numbers, a priori these numbers are unknown and could be anything. That is why we treat them as random and name them $X_1, X_2, \ldots, X_5$.

1.1. Sample Mean and Variance. The method of moments.

We have already introduced the sample mean and variance, but let us view the relation of these quantities to the parameters of the underlying distribution.

Let us remind that the sample mean is defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and the sample variance as

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

**Definition 1.** An estimator $\hat{\theta}$ of a parameter $\theta$ of a distribution is called unbiased estimator if

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

A few words of explanation. The estimator $\hat{\theta}$ will be a function of the measurements $(X_1, \ldots, X_n)$ on the sample, i.e. $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$. As we discussed before
the measurements \((X_1, \ldots, X_n)\), are considered as i.i.d random variable having the underlying distribution. If \(f(x; \theta)\) denotes the pdf of the underlying distribution, with parameter \(\theta\), then the expectation in the above definition should be interpreted as

\[
E[\hat{\theta}] = E[\hat{\theta}(X_1, \ldots, X_n)] = \int \cdots \int \hat{\theta}(x_1, \ldots, x_n) f(x_1; \theta) \cdots f(x_n; \theta) \, dx_1 \cdots dx_n
\]

and the definition of unbiased estimator corresponds to the fact that the above integral should be equal to the parameter \(\theta\) of the underlying distribution.

**Definition 2.** Let \(\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)\) an estimator of a parameter \(\theta\) based on a sample \((X_1, \ldots, X_n)\) of size \(n\). Then \(\hat{\theta}_n\) is called **consistent** if \(\hat{\theta}_n\) converges to \(\theta\) in probability, that is

\[
P(|\hat{\theta}_n - \theta| \geq \epsilon) \to 0, \quad \text{as } n \to \infty
\]

Here, again, as in the previous definition, the meaning of the probability \(P\) is identified with the underlying distribution with parameter \(\theta\).

**Proposition 1.** The sample mean and variance are consistent and unbiased estimators of the mean and variance of the underlying distribution.

**Proof.** It is easy to compute that

\[
E \left[ \frac{X_1 + \cdots + X_n}{n} \right] = \mu
\]

and

\[
E \left[ \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right] = \frac{n}{n-1} E \left[ (X_1 - \bar{X})^2 \right]
\]

\[
= \frac{n}{n-1} \left( E[X_1^2] - 2E[X_1 \bar{X}] + E[\bar{X}^2] \right)
\]

\[
= \frac{n}{n-1} \left( E[X_1^2] - \frac{2}{n} E[X_1^2] - \frac{2(n-1)}{n} E[X_1 X_2] + E[\bar{X}^2] \right)
\]

and now expanding the \(E[\bar{X}^2]\) as

\[
E[\bar{X}^2] = \frac{1}{n^2} \left( nE[X_1^2] + n(n-1)E[X_1 X_2] \right)
\]

and also using the independence, e.g. \(E[X_1 X_2] = E[X_1]E[X_2] = \mu^2\) we get that the above equals to

\[
E[X_1^2] - \mu^2 = \sigma^2.
\]

We, therefore, obtain that the sample mean and sample variance are unbiased estimators.
The fact that the sample mean is a consistent estimator follows immediately from the weak Law of Large Number (assuming of course that the variance $\sigma^2$ is finite).

The fact that the sample variance is also a consistent estimator follows easily. First, we have by an easy computation that

$$s^2 = \frac{n}{n-1} \left( \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}^2 \right).$$

The result now follows from the Law of Large Numbers since $X_i$’s and hence $X_i^2$’s are independent and therefore

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \to E[X_i^2]$$

and

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \to E[X_1]^2.$$

The above considerations introduce us to the **Method of Moments**. Let us recall that the $k$th moment of a distribution is defined as

$$\mu_k = \int x^k f(x) dx$$

If $X_1, X_2, \ldots$ are sample data drawn from a given distribution then the $k$th sample moment is defined as

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

and by the Law of Large Numbers (under the appropriate condition) we have that $\hat{\mu}_k$ approximates $\mu_k$, as the sample size gets larger.

The idea behind the Method of Moments is the following: Assume that we want to estimate a parameter $\theta$ of the distribution. Then we try to express this parameter in terms of moments of the distribution.

**Example 1.** Consider the Poisson distribution with parameter $\lambda$, i.e.

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

It is easy to check (check it !) that $\lambda = E[X]$. Therefore, the parameter $\lambda$ can be estimated by the sample mean of a large sample.

**Example 2.** Consider a normal distribution $N(\mu, \sigma^2)$. Of course, we know that $\mu$ is the first moment and that $\sigma^2 = \text{Var}(X) = E[X^2] - E[X]^2 = \mu_2 - \mu_1^2$. So estimating the first two moments, gives us an estimation of the parameters of the normal distribution.
1.2. **Maximum Likelihood.**

Maximum likelihood is another important method of estimation. Many well-known estimators, such as the sample mean and the least squares estimation in regression are maximum likelihood estimators. Maximum likelihood estimation tends to give more efficient estimates than other methods. Parameters used in ARIMA time series models are usually estimated by maximum likelihood.

Let us start describing the method. Suppose that we have a distribution, with a parameter \( \theta = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k \), that we wish to estimate. Let, also, \( X = (X_1, \ldots, X_n) \) a set of sample data. Viewed as a collection of i.i.d variables, the sample data will have a probability density function

\[
 f(X_1, \ldots, X_n; \theta) = \prod_{i=1}^{n} f(X_i; \theta).
\]

This function is viewed as a function of the parameter \( \theta \), we will denote it by \( L(\theta) \) and call it the **likelihood function**. The product structure is due to the assumption of independence.

The **maximum likelihood estimator (MLE)** is the value of the parameter \( \theta \), that maximises the likelihood function, given the observed sample data, \( (X_1, \ldots, X_n) \).

It is often mathematically more tractable to maximise a sum of functions, than a product of function. Therefore, instead of trying to maximise the likelihood function we prefer to maximise the **log-likelihood function**

\[
\log L(\theta) = \sum_{i=1}^{n} \log f(X_i; \theta).
\]

**Example 3.** Suppose that the underlying distribution is a normal \( N(\mu, \sigma^2) \) and we want to estimate the mean \( \mu \) and variance \( \sigma^2 \) from sample data \( (X_1, \ldots, X_n) \), using the maximum likelihood estimator.

First, we start with the log-likelihood function, which in this case is

\[
\log L(\mu, \sigma) = -n \log \sigma - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.
\]

To maximise the log-likelihood function we differentiate with respect to \( \mu, \sigma \) and obtain

\[
\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)
\]

\[
\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \sigma^{-3} \sum_{i=1}^{n} (X_i - \mu)^2
\]
the partials need to be equal to zero and therefore solving the first equation we get that

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i := \bar{X}.$$ 

Setting the second partial equal to zero and substituting $\mu = \hat{\mu}$ we obtain the maximum likelihood estimator for the standard deviation as

$$\hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Remark: Notice that the MLE is biased since

$$E[\hat{\sigma}^2_{ML}] = \frac{n - 1}{n} \sigma^2$$

Example 4. Suppose we want to estimate the parameters of a Gamma($\alpha, \theta$) distribution

$$f(x; \alpha, \theta) = \frac{\theta^{-\alpha} x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha)}$$

The maximum likelihood equations are

$$0 = -n \log \theta + \sum_{i=1}^{n} \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$0 = n\alpha \theta - \sum_{i=1}^{n} X_i$$

Solving these equations in terms of the parameters we get

$$\hat{\theta} = \frac{\bar{X}}{\hat{\alpha}}$$

$$0 = n \log \hat{\alpha} - n \log \bar{X} + \sum_{i=1}^{n} \log X_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}.$$ 

Notice that the second equation is a nonlinear equation which cannot be solved explicitly! In order to solve it we need to resort to numerical iteration scheme. To start the iterative numerical procedure we may use the initial value obtained from the method of moments.

Proposition 2. Under appropriate smoothness conditions on the pdf $f$, the maximum likelihood estimator is consistent.
Proof. We will only give an outline of the proof, which, nevertheless, presents the ideas. We begin by observing that by the Law of Large Numbers, as \( n \) tends to infinity, we have that

\[
\frac{1}{n} L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f(X_i; \theta) \rightarrow E \log f(X; \theta) = \int \log f(x; \theta) f(x; \theta_0) \, dx
\]

In the above \( \theta_0 \) is meant to be the real value of the parameter \( \theta \) of the distribution.

The MLE will now try to find the \( \hat{\theta} \) that maximises \( L(\theta)/n \). By the above convergence, we have that this should then be approximately the value of \( \theta \) that maximises \( E \log f(X; \theta) \). To maximise this differentiate with respect to \( \theta \) to get that

\[
\frac{\partial}{\partial \theta} \int \log(f(x; \theta)) f(x; \theta_0) \, dx = \int \frac{\partial f(x; \theta)/\partial \theta}{f(x; \theta)} f(x; \theta_0) \, dx.
\]

Setting \( \theta = \theta_0 \) in the above we get that is is equal to

\[
\int \frac{\partial}{\partial \theta} f(x; \theta_0) \, dx = \frac{\partial}{\partial \theta} \int f(x; \theta_0) \, dx = 0.
\]

Therefore \( \theta_0 \) maximises the \( E[\log f(x; \theta)] \) and therefore the maximiser of the log-likelihood function will approach, as \( n \) grows, to the value \( \theta_0 \). \( \square \)

1.3. Comparisons.

We introduced two methods of estimation: the method of moments and the maximum likelihood estimation. We need some way to compare the two methods. Which one is more likely to give better results? There are several measures of the efficiency of the estimator. One of the most commonly used is the mean square error (MSE). This is defined as follows. Suppose, that we want to estimate a parameter \( \theta \), and we use an estimator \( \hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) \). Then the mean square error is defined as

\[
E \left[ (\hat{\theta} - \theta)^2 \right].
\]

Therefore, one seeks, estimators that minimise the MSE. Notice that it holds

\[
E \left[ (\hat{\theta} - \theta)^2 \right] = (E[\hat{\theta}] - \theta)^2 + \text{Var}(\hat{\theta}).
\]

If the estimator \( \hat{\theta} \) is unbiased, then the MSE equals the \( \text{Var}(\hat{\theta}) \). So having an unbiased estimator may reduce the MSE. However this is not necessary and one should be willing to accept a (small) bias, as long as the MSE becomes smaller.

The sample mean is an unbiased estimator. Moreover it is immediate (why?) that

\[
\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}
\]

where \( \sigma^2 \) is the variance of the distribution. Therefore, the MSE of the sample mean is \( \sigma^2/n \).
In the case of a maximum likelihood estimator of a parameter $\theta$ we have the following theorem

**Theorem 1.** Under smoothness conditions on $f$, the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to standard normal. Here

\[
I(\theta) = E \left[ \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \right] = -E \left[ \frac{\partial^2}{\partial \theta^2} \log f(X; \theta) \right]
\]

We will skip the proof of this important theorem. The reader is referred to the book of Rice.

This theorem tells us that the maximum likelihood estimator is approximately unbiased and that the mean square error is approximately $1/nI(\theta_0)$.

A way to compare the efficiency of two estimators, say $\hat{\theta}$ and $\tilde{\theta}$ we introduce the efficiency of $\hat{\theta}$ in terms of $\tilde{\theta}$ as

\[
\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{Var}(\tilde{\theta})}{\text{Var}(\hat{\theta})}
\]

Notice that the above definition makes sense as a comparison measure between estimators that are unbiased or that have the same bias.

### 1.4. Exercises.

1. Consider the Pareto distribution with pdf

   \[ f(x) = \frac{ac^a}{x^{a+1}}, \quad x > c. \]

   Compute the maximum likelihood estimator for the parameters $a, c$.

2. Consider the Gamma distribution Gamma($\alpha, \theta$). Write the equations for the maximum likelihood estimators for the parameters $\alpha, \theta$. Can you solve them? If you cannot solve them directly, how would you proceed to solve them?

3. Compute the mean of a Poisson distribution with parameter $\lambda$.

4. Consider a Gamma distribution Gamma($\alpha, \theta$). Use the method of moments to estimate the parameters $\alpha, \theta$ of the Gamma distribution.

5. Consider the distribution

   \[ f(x; \alpha) = \frac{1 + \alpha x}{2}, \quad -1 < x < 1. \]

   The parameter $\alpha$ lies in between $\pm 1$.

   A. Use the method of moments to estimate the parameter $\alpha$. 


B. Use the maximum likelihood method to estimate $\alpha$. If you cannot solve the equations explain why is this and describe what would you do in order to find the MLE.

C. Compare the efficiency between the two estimators.

6. Consider the problem of estimating the variance of a normal distribution, with unknown mean from a sample $X_1, X_2, \ldots, X_n$, of i.i.d normal random variables. In answering the following question use the fact (see Rice, Section 6.3) that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

and that the mean and the variance of a chi-square random variable with $r$ degrees of freedom is $r$ and $2r$, respectively.

A. Find the MLE and the moment-method estimators of the variance. Which one is unbiased?

B. Which one of the two has smaller MSE?

C. For what values of $\rho$ does the estimator $\rho \sum_{i=1}^{n} (X_i - \bar{X})^2$ have the minimal MSE?