# Lecture 16. Point Estimation, Sample Analogue Principal

11/14/2011

**Point Estimation**

In a typical statistical problem, we have a random variable/vector $X$ of interest but its pdf $f_X(x)$ or pmf $p_X(x)$ is unknown: either completely unknown, or "known up to a finite dimensional parameter". By the latter, we mean the following: we know (or are willing to assume) $f_X(x) = f(x, \theta)$ or $p_X(x) = p(x, \theta)$ for some known functions $f$ and $p$ and some unknown $\theta \in \Theta \subset R^{d_\theta}$. Typically, we can obtain a random sample from $f_X$ or $p_X$: $\{X_1, ..., X_n\}$. We would like to learn about $f_X$ and $p_X$ from the sample.

If $f_X$ and $p_X$ are completely unknown, it is possible to use the sample to estimate them as a whole and the estimation is classified as nonparametric estimation. The topic is covered in Econ 710.

Often we may not be interested in the whole pdf or pmf. Instead, we are only interested in some finite-dimensional parameters about the pdf or pmf. For example, we may only be interested in the mean of $X$: $\mu = E[X]$, or the variance $\sigma^2 = Var(X)$, or the median $\xi_{0.5} = median(X)$. In general, denote the parameter of interest $\theta$.

If $f_X$ or $p_X$ are known up to a finite dimensional parameter: $f_X(x) = f(x, \theta)$ or $p_X(x) = p(x, \theta)$ , it is often of interest to learn about $\theta$.

In both of the cases discussed above, an estimator $\hat{\theta}_n$ of $\theta$ is a function of the random sample $\{X_1, ..., X_n\}$: $\hat{\theta}_n = \vartheta(X_1, ..., X_n)$. We would like the estimator to be close to the true value. The most common measure of "closeness" of the estimator to the true value is the mean-squared error (MSE):

$$MSE(\hat{\theta}_n, \theta) = E_\theta[(\hat{\theta}_n - \theta)^2], \tag{1}$$

where $E_\theta$ signifies the fact that the expecation is taken under $\theta$. One thing we know for sure is that there is no estimator that minimizes $MSE(\hat{\theta}_n, \theta)$ for all $\theta$. The reason for that is the

following: let $\hat{\theta}_n^{13} = 13$, then $MSE(\hat{\theta}_n^{13}, \theta) = 0$ for $\theta = 13$. An estimator, $\hat{\theta}_n^*$, that minimizes $MSE(\hat{\theta}_n, \theta)$ for all $\theta$ has to minimizes $MSE(\hat{\theta}_n, \theta)$ at $\theta = 13$. Thus, $MSE(\hat{\theta}_n^*, 13)$ has to be zero. By the same argument, $MSE(\hat{\theta}_n^*, c)$ has to be zero for any $c \in R$. This is impossible. Therefore there is no estimator that minimizes $MSE(\hat{\theta}_n, \theta)$ for all $\theta$.

To get around this issue, we rule out the silly estimators like $\hat{\theta}_n^{13}$ by restricting our space of estimators. Then among the restricted space of estimators, there may exists an estimator that minimizes $MSE(\hat{\theta}_n, \theta)$. **Unbiasedness** and **consistency** are two commonly used restrictions.

Let $\Theta$ be the parameter space, i.e., $\Theta$ is the set that we believe $\theta$ lies in, typically a subset of $R^k$, where $k$ is the dimension of $\theta$.

An estimator $\hat{\theta}_n$ is unbiased iff $E_\theta[\hat{\theta}_n] = \theta$ for all $\theta \in \Theta$.

An unbiased estimator is on average right regardless of $\theta$.

An estimator $\hat{\theta}_n$ is consistent iff $\lim_{n\to\infty} \Pr_\theta(|\hat{\theta}_n - \theta| > \varepsilon) = 0$ for all $\theta \in \Theta$ and all $\varepsilon > 0$.

A consistent estimator is right regardless of $\theta$ if given an infinite sample.

**Definition 1** (Uniformly Minimum Variance Unbiased (UMVU) estimator)**.** An estimator $\hat{\theta}_n^*$ of a parameter $\theta \in \Theta \subset R$ if

(a) $E_\theta[\hat{\theta}_n] = \theta$ for all $\theta \in \Theta$ and

(b) $Var_\theta(\hat{\theta}_n^*) \leq Var_\theta(\hat{\theta}_n)$ for any other unbiased estimator $\hat{\theta}_n$ of $\theta$ for all $\theta \in \Theta$.

In certain problems UMVU estimators can be found and this is discussed in the next lecture.

*A few words on unbiasedness and consistency:* Unbiasedness and consistency are useful restrictions on estimators that restrict us to a class of estimators in which a "best" estimator can be discussed properly. They are useful properties of an estimator in their own right (not just to help the discussion of best estimator). It is considered a basic requirement for a good estimator to be either "unbiased" or "consistent" or both. "Unbiasedness" is considered a finite sample property in that $E_\theta[\hat{\theta}_n] = \theta$ is required to hold for finite $n$. Unbiasedness has the undesirable property that a nonlinear function of an unbiased estimator is typically not unbiased. For this reason, unbiased estimators are hard to find in most problems, especially nonlinear problems. Consistency, on the other hand, is considered an asymptotic property (or large-sample property) in that $\hat{\theta}_n \to_p \theta$ is a requirement on the sequence $\{\hat{\theta}_n\}_{n=1}^\infty$ not on each $\hat{\theta}_n$ for a finite $n$. Consistent estimators are much easier to find because convergence in probability survives continuous nonlinear transformations (by virtue of the Slustky Theorem). This course, as well as modern econometrics, focuses on consistent estimators for this reason.

It may seem that unbiasedness is stronger than consistency – this is not true. Unbiasedness and consistency do not imply each other. To see why, consider the estimator $\hat{\theta}_n = X_1$ for the mean, $\theta$, of the poputation distribution from which the random sample $\{X_1, ..., X_n\}$ is drawn. Then clearly, $\hat{\theta}_n$ is unbiased but it is not consistent because the probability that $\hat{\theta}_n$ is $\varepsilon$ away from $\theta$ is not zero (as long as $X_1$ is not a constant) and does not change with $n$. There may be ways to write down assumptions to rule out silly estimators like $\hat{\theta}_n = X_1$ and ensure that unbiasedness implies consistency, but these assumptions won't be of much practical use because finding consistency estimators is typically easier than finding unbiased estimators. It incurs no practical disadvantage to study unbiasedness and consistency seperately.

**Sample Analogue Principle**

In this section, we discuss a very useful way of coming up with estimators for a given parameter of interest: the sample analogue principle. It applies generally whether we have any knowledge about the population distribution or not. We illustrate this principle by several examples.

**Example 1** (Covariance)**.** Suppose we are interested in the covariance $\sigma_{XY}$, of two random variables $X$ and $Y$ and we observe a random $n$-sample $\{(X_i, Y_i)\}_{i=1}^n$. A sample analogue estimator is the sample covariance:

$$\tilde{\sigma}_{XY} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n).$$

Assume that $X$ and $Y$ both have finite second moments ($E[X^2], E[Y^2] < \infty$). One can easily show that $\tilde{\sigma}_{XY} \to_p \sigma_{XY}$ using the same arguments for $\tilde{S}_X \to_p \sigma_X^2$ given in a previous lecture. One can also easily show that $E[\tilde{\sigma}_{XY}] = n\sigma_{XY}/(n-1)$, suggesting that an unbiased estimator for $\sigma_{XY}$ is $\hat{\sigma}_{XY} = (n-1)^{-1} n \tilde{\sigma}_{XY} \equiv (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$.
(details left for exercise)

**Example 2** (Projection)**.** Let $Y$ be a random variable and $X$ be a random $k$-vector, both having finite second moments. Also assume that $X$ is not multicolinear, i.e. $E(XX')$ is positive definite. Suppose that we are interested in the projection of $Y$ onto the linear space spanned by $X$ under the $L_2$ norm. That is, we would like to find a linear combination of $X$, $\theta' X$, such that $E[(Y - \theta' X)^2] \leq E[(Y - c' X)^2]$ for all $c \in R^k$. The minimization problem $\min_{c \in R^k} E[(Y - c' X)^2]$ has closed form solution:

$$\theta = [E(XX')]^{-1} [E(XY)].$$

Suppose we observe a random $n$-sample $\{(Y_i, X_i)\}_{i=1}^n$. A sample analogue estimator of $\theta$ is

$$\hat{\theta}_n = [n^{-1}\sum_{i=1}^n X_i X_i']^{-1}[n^{-1}\sum_{i=1}^n (X_i Y_i)].$$

The consistency of $\hat{\theta}_n$ as an estimator of $\theta$ is easy to show. First apply the strong law of large numbers to $n^{-1}\sum_{i=1}^n X_i X_i'$ and $n^{-1}\sum_{i=1}^n (X_i Y_i)$ . The probability limit (in fact, almost sure limit) of them are $E(XX')$ and $E(XY)$, respectively. Then apply the Slutsky theorem on the function $g(A, b) = A^{-1}b$ to conclude. The Slutsky theorm applies because $g(A, b)$ is continuous in $(A, b)$ at $(A, b) = (E(XX'), E(XY))$.

In general $\hat{\theta}_n$ is not an unbiased estimator of $\theta$.

Note: some may have noticed the resemblance of $\hat{\theta}_n$ to the ordinary least square estimator of the regression model $Y = \theta' X + \varepsilon$ and may be confused about the lack of unbiasedness of $\hat{\theta}_n$. The $\theta$ above is merely the projection coefficient of $Y$ onto the space spanned by $X$. It is an aspect of the joint distribution of $Y$ and $X$ and is well defined without a model (i.e. assumptions) on the relationship between $Y$ and $X$. If one is willing to make the assumption that $Y = \theta' X + \varepsilon$ $E(\varepsilon|X) = 0$, then one has a linear regression model and $\hat{\theta}_n$ is an unbiased (and consistent) estimator of the model coefficient.

**Example 3** (Quantile Estimation)**.** Suppose that we are interested in the $p$ quantile, $\xi_p$, of a continuous random variable $Y$ with continuous and strictly increasing cdf $F_Y(y) : R \to [0, 1]$:

$$\xi_p := \inf\{y : F_Y(y) \geq p\} \equiv F_Y^{-1}(p). \tag{2}$$

We observe a random $n$-sample $\{Y_1, ..., Y_n\}$. Then a sample analogue estimator of $\xi_p$ is the sample quantile:

$$\hat{\xi}_p = \inf\{y : \hat{F}_Y(y) \geq p\},$$

where $\hat{F}_Y(y) = n^{-1}\sum_{i=1}^n 1\{Y_i \leq y\}$ is the empirical distribution of $Y$.

The estimator $\hat{\xi}_p$ is consistent and the proof of consistency is a bit different the consistency proofs we have seen so far. Consider an arbitrary $\varepsilon > 0$. To show consistency, it suffices to show that

$$\Pr(|\hat{\xi}_p - \xi_p| > \varepsilon) \to 0 \text{ as } n \to \infty. \tag{3}$$

The left-hand-side equals $\Pr(\hat{\xi}_p - \xi_p > \varepsilon) + \Pr(\hat{\xi}_p - \xi_p < -\varepsilon)$. Thus, it suffices to show that

each summand converges to zero. First, consider the first summand:

$$\begin{aligned}
\Pr(\hat{\xi}_p - \xi_p > \varepsilon) &= \Pr(\hat{\xi}_p > \xi_p + \varepsilon) \\
&\leq \Pr(\hat{F}_Y(\xi_p + \varepsilon) < p) \\
&= \Pr\left( n^{-1} \sum_{i=1}^{n} 1\{Y_i \leq \xi_p + \varepsilon\} < p \right) \\
&= \Pr\left( F_Y(\xi_p + \varepsilon) - n^{-1} \sum_{i=1}^{n} 1\{Y_i \leq \xi_p + \varepsilon\} > F_Y(\xi_p + \varepsilon) - p \right) \\
&\leq \Pr\left( \left| F_Y(\xi_p + \varepsilon) - n^{-1} \sum_{i=1}^{n} 1\{Y_i \leq \xi_p + \varepsilon\} \right| > F_Y(\xi_p + \varepsilon) - p \right) \\
&\to 0,
\end{aligned}$$

where the first inequality holds by the definition of $\hat{\xi}_p$, and the convergence holds because $F_Y(y)$ is strictly increasing (and thus $F_Y(\xi_p + \varepsilon) - p > 0$) and because of the law of large numbers applied on $n^{-1} \sum_{i=1}^{n} 1\{Y_i \leq \xi_p + \varepsilon\}$.

Second, consider the second summand:

$$\begin{aligned}
\Pr(\hat{\xi}_p - \xi_p < -\varepsilon) &= \Pr(\hat{\xi}_p < \xi_p - \varepsilon) \\
&\leq \Pr(n^{-1} \sum_{i=1}^{n} 1\{Y_i \leq \xi_p - \varepsilon\} \geq p) \\
&= \Pr\left( n^{-1} \sum_{i=1}^{n} 1\{Y_i \leq \xi_p - \varepsilon\} - F_Y(\xi_p - \varepsilon) \geq p - F_Y(\xi_p - \varepsilon) \right) \\
&\leq \Pr\left( \left| n^{-1} \sum_{i=1}^{n} 1\{Y_i \leq \xi_p - \varepsilon\} - F_Y(\xi_p - \varepsilon) \right| \geq p - F_Y(\xi_p - \varepsilon) \right) \\
&\to 0, \qquad (4)
\end{aligned}$$

where the first inequality holds by the definition of $\hat{\xi}_p$ and the convergence holds for the same reason as the convergence in the previous part of the proof.

Notice that the strict monotonicity of $F_Y$ is crucial in the proof. The continuity of $F_Y$ is not used - and thus does not need to be assumed.

**Example 4.** Estimating $\theta$ in $Uniform[0, \theta]$. Suppose that we are interested in the distribution of $Y$ and we know $Y$ is uniformly distributed between $[0, \theta]$ but we don't know $\theta$. We observe a random $n$-sample of $Y$: $\{Y_1, ..., Y_n\}$. There are different ways to apply the sample analogue principle to this example. First, we know that the mean of $Y$ is $\theta/2$. Thus

$\theta = 2E[Y]$ and a sample analogue estimator motivated from this equation is

$$\hat{\theta}_{1n} = 2\bar{Y}_n. \tag{5}$$

Second, we know the median of $Y$ is also $\theta/2$. Thus, $\theta = 2\xi_{0.5}$. A sample analogue estimator motivated from this form is

$$\hat{\theta}_{2n} = 2\hat{\xi}_{0.5}. \tag{6}$$

Third, we know that $\theta$ is the largest possible value that $Y$ can take. A sample analogue from this perspective is

$$\hat{\theta}_{3n} = \max_i Y_i.$$

All three estimators are consistent. The first estimator is also unbiased. But they are not equally good in terms of accuracy. This will become clear in the next lecture.