A Note on Discrete Approximations of Continuous Distributions

John Kennan[1]
University of Wisconsin-Madison and NBER

September 2006

Suppose F is a strictly increasing distribution function defined on the real line, and $\hat{F}$ is an approximation of this distribution. Define the distance between F and $\hat{F}$, using the L^p metric (with $p > 0$), as

$$\left\|F - \hat{F}\right\|_p^p = \int_{-\infty}^{\infty} \left|F(x) - \hat{F}(x)\right|^p dF(x) = \int_0^1 \left|t - \hat{F}(F^{-1}(t))\right|^p dt$$

Proposition

A. The best discrete approximation $\hat{F}$ to a given distribution F using the fixed support points $\{x_i\}_{i=1}^n$ is given by

$$\hat{F}(x_i) = \frac{F(x_i) + F(x_{i+1})}{2}$$

for $1 \leq i \leq n-1$, with $F(x_n) = 1$.

B. The best n-point approximation $\hat{F}$ to a given distribution F has equally-weighted support points $\{x_i^*\}$ given by

$$F(x_i^*) = \frac{2i-1}{2n}$$

for $1 \leq i \leq n$.

C. The minimal distance between F and $\hat{F}$ is

$$\left\|F - \hat{F}\right\|_p = \frac{1}{2n}(p+1)^{-\frac{1}{p}}$$

[1]Department of Economics, University of Wisconsin, 1180 Observatory Drive, Madison, WI 53706; jkennan@ssc.wisc.edu.

Proof:

For arbitrary support points $\{x_i\}_{i=1}^{n}$, define $q_i = F(x_i)$, with $q_0 = 0$, and assume (without loss) that these points are ordered so that q_i is increasing in i. Also define $Q_i = \hat{F}(x_i)$, with $Q_0 = 0$ and $Q_n = 1$. Then the best approximation solves

$$\min_{q_i, Q_i} \sum_{i=0}^{n-1} \int_{q_i}^{q_{i+1}} |t - Q_i|^p \, dt$$

The first-order condition for Q_i gives

$$\frac{d}{dQ_i} \int_{q_i}^{q_{i+1}} |t - Q_i|^p \, dt = |Q_i - q_i|^p - |q_{i+1} - Q_i|^p = 0$$

This implies

$$\frac{|q_{i+1} - Q_i|}{|Q_i - q_i|} = 1$$

Thus if the support points are given, $\hat{F}$ is determined by

$$Q_i = \frac{q_i + q_{i+1}}{2}$$

This proves part A.

Note that Q_i is increasing. If the support points can be varied, the first-order condition for q_i gives

$$|q_i - Q_{i-1}|^p - |q_i - Q_i|^p = (q_i - Q_{i-1})^p - (Q_i - q_i)^p = 0$$

This implies

$$q_i = \frac{Q_{i-1} + Q_i}{2}$$

Combining these results,

$$4Q_i = 2q_i + 2q_{i+1} = Q_{i-1} + 2Q_i + Q_{i+1}$$

So

$$Q_{i+1} - Q_i = Q_i - Q_{i-1}$$

for $1 \le i \le n-1$. Since $Q_0 = 0$ and $Q_n = 1$, this implies

$$Q_i = \frac{i}{n}$$

and

$$q_i = \frac{2i-1}{2n}$$

This proves part B.

The approximation error is given by

$$\|F - \hat{F}\|_p^p = \sum_{i=0}^{n-1} \int_{\frac{2i-1}{2n}}^{\frac{2i+1}{2n}} \left| t - \frac{i}{n} \right|^p dt$$

The result is

$$\|F - \hat{F}\|_p = \frac{1}{2n} (p+1)^{-\frac{1}{p}}$$

This proves part C.

Remark: The distance between F and $\hat{F}$ is increasing in p. In the limit, the (sup-norm) distance is

$$\|F - \hat{F}\|_\infty = \frac{1}{2n}$$

Discussion

It is noteworthy that the best approximate distribution does not depend on which L^p norm is used –

minimizing the squared distance between F and $\hat{F}$ gives the same answer as minimizing the absolute value of the distance, or minimizing the maximum distance, for example.

Another interesting feature of the result is that the best discrete approximation is a uniform distribution. One might have thought that the approximation could be improved, in general, by allowing for non-uniform weights. But in fact if such an improvement were possible, it would just indicate that the support points are not in the right place.

In econometric applications, it is common to model unobserved heterogeneity by assuming that there is a finite set of "types", and taking the characteristics of these types and also the type proportions as parameters to be estimated. There may be situations where there is reason to believe that the true type distribution is discrete, in which case this procedure seems reasonable. But there are also situations in which the true type distribution is continuous, and in that case if the aim is to find the best discrete approximation of the type distribution, then the probabilities should not be treated as parameters to be estimated: each type should be assigned equal weight.