Empirical likelihood/distribution with Tied data

Mai Zhou [1]

Maximization of empirical likelihood with tied observations

First, in maximizing the likelihood, the observations are considered fixed. We maximize it by changing the distribution functions (parameters).

Suppose we have independent observations $X_1, \cdots, X_n$ from distribution $F(\cdot)$. Suppose there are tied observations.

Let $t_1 < \cdots < t_m$ be the distinctive values from the X_i's. $(m \le n)$ and $w_1, \cdots, w_m$ be integers representing the number of tied data at t_i:

$$w_i = \#\{X_j = t_i; j = 1, \cdots, n\} \ .$$

But we allow the weights to be fractions for the applications later.

The empirical likelihood based on the weighted observations is

$$EL = \prod_{i=1}^{m} (p_i)^{w_i}$$

where $p_i = P(X = t_i)$. and the log empirical likelihood is

$$\sum w_i \log p_i \ . \qquad (1)$$

Theorem *The maximization of the log empirical likelihood (1) with respect to p_i subject to the constraints:*

$$\sum p_i \le 1 \ , \qquad p_i \ge 0$$

is given by the formula

$$p_i = \frac{w_i}{\sum_{j=1}^{k} w_j}$$

Proof: It is obvious that we can take $\sum p_i = 1$ in finding the miximizer, since those p_i such that the summation is < 1 cannot be the maximizer.

[1]Mai Zhou, Department of Statistics, University of Kentucky, Lexington, KY 40506 USA E-mail: mai@ms.uky.edu

We use Lagrange multiplier method. First form the target function

$$G = \sum_{i=1}^{m} w_i \log p_i + \lambda(\sum_{i=1}^{m} p_i - 1) \ .$$

Taking partial direvative of G wrt p_i and λ and set them equal to zero we get

$$p_i = \frac{w_i}{\lambda} \ .$$

Now use the $\sum p_i = 1$ condition, we have $\lambda = \sum w_j$.

Theorem *The maximization of the log empirical likelihood with respect to p_i subject to the constraints:*

$$p_i \geq 0, \qquad \sum p_i = 1 \ , \qquad \sum g(t_i) p_i = \mu$$

is given by the formula

$$p_i = \frac{w_i}{\sum_j w_j + \lambda(g(t_i) - \mu)}$$

where λ is the solution of the equation

$$\sum_i \frac{w_i (g(t_i) - \mu)}{\sum_j w_j + \lambda(g(t_i) - \mu)} = 0 \ .$$

For μ in the range of the the $g(t_i)$'s there exist a unique solution of the λ and the p_i given above is also positive.

The proof of the above theorem is similar to (Owen 1990). (and is left as homework)

Discussion

We see that the maximum value of the EL achieved when there are tied data is no longer $(1/n)^n$ but a larger value:

$$\prod_{i=1}^{m} (p_i)^{w_i} \ ,$$

with $p_i = w_i / \sum_j w_j$.

However, the maximizer distribution can still be written as

$$\frac{1}{n} \sum I_{[X_i \leq t]} = F_n(t) \ ,$$

the empirical distribution function, even when there are tied X_i's.

Think of each observation carries $1/n$ probability but they pool them together when observations are tied.

References

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75** 237-249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18**, 90-120.

Owen, A. (2001). *Empirical likelihood.* Chapman & Hall, London.