A NOTE ON THE ESTIMATION OF QUANTILES USING RECORD BREAKING

Carlos N. Bouza[1], Universidad de La Habana, Cuba
Lakshma C. Singh[2], College of Business and Advanced Computation, India

ABSTRACT
A new proof of the Strong Consistency of a non parametric estimator of the quantiles is provided. The estimate uses record breaking data. The proposition uses a smaller number of hypothesis than other similar. The application to some well known problems as the determination of Optimal-burn-in-time, pollution and the newsboy vendor problem are discussed.

Key words: Kernel function, Wasserstein metric, weak convergence.

RESUMEN
Se brinda una nueva demostración de la Consistencia Fuerte de un estimador de los quantiles. El nuevo estimador utiliza mediciones de rupturas. La proposición utiliza menos hipótesis que otras versiones. Se discute su aplicación a algunos conocidos problemas como los de la determinación del tiempo óptimo de Burn-in, la polución y el problema del vendedor de periódicos.

MSC: 62G07.

1. INTRODUCTION

In different experiments the quantile Q(P) is used for measuring uncertainty. It measures the evidence that a null hypothesis, as H_0: $\theta \geq \theta_0$, is true. Taking: $\theta^*[s]$ as a test statistic $P(\theta^*[s] \geq Q(P)) = P = 1 - \alpha$. $1 - P = \alpha$ is the probability of committing a Type I error. Classic models deal with test statistics with a standard normal distribution. In many real life problems it is not possible to assume certain realistic distribution of $\theta^*[s]$.

We will consider the problem of rejecting H_0: $\theta \geq \theta_0$ or of determining a one side confidence interval $[Q(P), \infty[$ when the involved distribution function (DF) is unknown. This problems arises in Quality Control when a new good is developed and the lifetime X is studied for establishing its survival function $F^*(t) = 1 - F(t)$. The manufacturer wants to fix Q(P) such that $Prob\{X > Q(P)\} \leq 1 - P = 1 - \alpha$.

The written warranty for the good will state the compromise of providing labour and parts to fix or to replace the defective product if $X < Q(P)$. The consumer feels safe. The manufacturer fixes P and expects to give service to $n^* = n(1 - P)$ goods in a lot of size n.

A similar situation is present when the environment is studied. For example, Q(P) must be fixed for establishing if the quality of the air or the water is acceptable.

This class of problems can be characterised by experiments where the test implies a costly measurement process. Sometimes they are destructive. That is the case when the goods are stressed up to destroying it or the survival of mice under certain environmental conditions is observed.

Generally observing a sample and calculating the sample P-the quantile makes the estimation of Q(P). A less costly procedure is to apply the test and measure successive minima. That is to stress the first sampled unit up to obtain $X_1 = X_{(1)}$. If it does not survive at X_1 then $X_2 = X_{(2)}$ otherwise $X_i = X_{(2)}$, where i is the first unit with $X_i < X_{(1)}$. The procedure is repeated with the entire sample and $X_{(1)}, X_{(2)},\ldots X_{(r1)}$ minima are observed with $X_{(i)} < X_{(j)}$ if i< j. Take K(i) as the number of sampled units evaluated after observing. $X_{(i)}\ldots$. Therefore we obtain a sequence of record values

$$\{(X_{(1)}, K_{(1)}),\ldots,(X_{(r)},K_{(r)})\}$$

E-mail: [1]bouza@matcom.uh.cu
[2]laks_singh6585@rediff.com

and

$$\sum_{i=1}^{r} K(i) = n - 1$$

Chandler (1952) studied the stochastic behaviour of this sequence. This study has been followed by Pwass (1964) who derived the distribution of the frequency of the record highs. Glick (1978) obtained that the number of records in a sample of size n has a distribution which converges to a N(ln(n), ln(n)). Samaniego-Whitkaker (1988) studied the estimation of the distribution function F using a set of m independent record samples. Gulati- Padgett (994) analysed the estimation of quantiles. In this paper we reanalyse this problem and a new proof for the pointwise consistency of the estimation based on kernel function is developed.

2. THE MAIN RESULT

Take a random sample of size n from the continuous distribution F. Its quantile of order P is given by

$$Q(P) = F^{-1}(P) = \inf \{x \mid F(x) \geq P\}.$$

We measure the succesive minima $X_{(i)}$ and K(i), the number of data between it and its succesor or the last unit. The number of records R is a random variable. The estimator of

$$\bar{F}(t) = 1 - F(t) \tag{2.1}$$

proposed by Samaniego-Whitaker (1988) is

$$\bar{F}^*(t) = \prod_{\{i \mid X_{(i)} \leq t\}} \frac{\sum_{j=1}^{r} K(j) - 1}{\sum_{j=1}^{r} K(j)}$$

We will select m independent samples with the same size n sequentially from F. Then we have m sequences $\{(X_{(i)1}, K(1)), (X_{(i)2}, K(2)), \ldots, (X_{(i)ri}, K(r_i))\}$, i = 1,…,m. They may be combined in an ordered sample of size

$$r^* = \sum_{i=1}^{m} r(i)$$

Now we can use the sequence $S_{mn}=\{(X_{(j)}, K(j)) \mid j=1,\ldots,r^*\}$. An estimator of (2.1) based on the combined sequence is given by:

$$\bar{F}_{mn}(t) = \prod_{\{j \mid X_{(j)} \leq t\}} \frac{\sum_{h=j}^{r^*} K(h) - 1}{\sum_{h=j}^{r^*} K(h)}$$

Therefore

$$F_{mn}(t) = F^*(t) = 1 - \bar{F}_{mn}(t)$$

The nonparametric estimator of Q(P) , derived directly from $F_{mn}(t)$ using S_{mn} , is

$$Q_{mn}(P) = \begin{cases} X_{(1)} & \text{if} \quad 0 < P \leq F_{mn}(X_{(1)}) \\ X_{(j)} & \text{if} \quad F_{mn}(X_{(j-1)}) < P \leq F_{mn}(X_{(j)}) \\ \infty & \text{otherwise} \end{cases}$$

The consistency of $Q_{mn}(P)$ is easily derived from the convergence of the empirical measure. In the proof is assumed that F is absolutely continuous with a density $f(t) = \partial F(t)/\partial t$.

Take a suitable kernel $G[(x-t)/h_{mn}]$ such that the bandwidth $h_{mn} > 0$ is such that $h_{mn} \to 0$ and $mh \to 0$ when $m \to \infty$. Gulati-Padgett (1994) proved the following theorem:

Theorem 2.1. (Gulati-Padgett, 1994). Take

$$x_{mn}(P) = \tilde{F}_{mn}^{-1}(P)$$

where

$$\tilde{F}_{mn}^{-1}(P) = \int_0^{\infty} h_{mn}^{-1} G\left(\frac{x-t}{h_{mn}}\right) F_{mn}(t)dt = \int_0^{x} f_{mn}(u)du$$

and

h1) $h_{mn} \to 0$ as $m \to \infty$

G1) $\int |G(y)|\,dy < \infty$, $\text{Sup}\,|G(y)| < \infty$ and $|yG(y)| \to 0$ as $y \to \infty$

G2) $G(y) \geq 0$ and $\int G(y)dy = 1$

then $x_{mn}(P) \to Q(P)$ a.s. as $m \to \infty$ and $\forall P \in]0,1[$

We will derive the Strong Consistency of $x_{mn}(P)$ using a single hypothesis on the kernel.

Proposition 2.2. For any $P \in]0,1[$ if $G(\bullet)$ is a bounded continuous function then $x_{mn}(P) \to Q(P)$ as $m \to \infty$ almost surely.

Proof.

Take $f^*(x) = \int_0^{\infty} G[(x-t)/h_{mn}]f(t)dt$ where $f(t) = \partial F(t)/\partial t$ and $\underset{x\geq 0}{\text{Sup}}\,|f_m(x) - f(x)| \leq S$ defining

$$S = \underset{x\geq 0}{\text{Sup}}\left\{\left|\int_0^{\infty} h_{mn}^{-1} G\left(\frac{x-t}{h_{mn}}\right) d(F_{mn}(t) - F(t))\right|\right\} + \underset{x\geq 0}{\text{Sup}}\left\{\left|f^*(x) - f(t)\right|\right\} = \text{Sup}\,U_{mn}(x) + \text{Sup}\,U_m^*(x)$$

From Nadaraya (1965) we have that $U_m^*(x) \to 0$ if $m \to \infty$.

Consider that $F_{mn} = F_1$ and $F = F_2$ are DF generated by probability measures from a class $\wp$ and take

$$D(\nu_1, \nu_2) = \{\eta \in \wp \subset (\Re^+ \times \Re^+) \mid \eta \circ \pi_i = \nu_i\}$$

where π_i is the projection of i. Let us define an appropiate distance d(x,t*) between t* and x, denote by ρ a radius with center x and

$$L_n = \text{Sup}\left\{\left|G\left(\frac{x-t}{h_{mn}}\right) - G\left(\frac{x-t'}{h_{mn}}\right)\right| d(t,t') \,\middle|\, t,t \in \left\{t^* \geq 0 \,\middle|\, d(x,t^*) \leq \rho, \rho > 0\right\}\right\}$$

then

$$U_{mm}(x) = \left|\int_0^{\infty}\int_0^{\infty} h_{mn}^{-1}\left[G\left(\frac{x-t}{h_{mn}}\right) - G\left(\frac{x-t'}{h_{mn}}\right)\right]\eta(d(t,t'))\right| \leq \int_0^{\infty}\int_0^{\infty} h_{mn}^{-1}\left|\left[G\left(\frac{x-t}{h_{mn}}\right) - G\left(\frac{x-t'}{h_{mn}}\right)\right]\right|\eta(d(t,t')) \leq$$

$$\leq \int_0^{\infty} L_h[\text{Max}\{d(x,t), d(x,t')\}]d(t,t')\eta[t,t']$$

Hence we have that

$$U_{mn}(x) = \left[\left[\int_0^{\infty} L_h(d(x,t)^2)\nu_1(t)\right]^{1/2} + \left[\int_0^{\infty} L_h(d(x,t')^2)\nu_{21}(t)\right]^{1/2}\right]\left[\int_0^{\infty}\int_0^{\infty}(d(t,t')^2)\eta(d(t,t'))\right]^{1/2}$$

Proposition 2.1 of Römisch-Scultz (1989) for a bidimensional variable holds in our case because p = q =2.

The Wasserstein metric , see Pflug (1996) for bidimensional variables is

$$W_b(\nu_1,\nu_2) = \inf\left\{\int_{\Re^b \times \Re^b} (d(t,t))^{'b}\eta(d(t,t))\,\middle|\,\eta \in D(\nu_1,\nu_2)\right\}^{1/b}$$

Using the related DF´s we have that

$$U_{mn}(x) \le \text{Max}\,(L_h)W_1\,(\nu_1,\,\nu_2) = \text{Max}(L_h)\int_0^{\infty} |\,F_{mn}(t) - F(t)\,|\,dt$$

Using the consistency of F_{mn} obtained by Samaniego-Whitaker (1988) we complete the proof.□

3. APPLICATIONS

We can illustrate through some classic examples cases where the estimation of Q(P) is needed and the results obtained in this paper provide frame where a cheaper method permits of enhance good results.

Example 1. A product has three phases in its life cycle. At first the failure or hazard rate decreases up to a certain moment t*. The manufacturer wants to determine Q(P) such that P{t* < Q(P)} < 1 - P. A sample is selected from a lot and the selected units are operated under extreme conditions for identifying failures. The use of record data permits to ship the K(i) products sampled which exhibit a performance not worse than the unit with lifetime $X_{(i)}$. The estimation of Q(P) permits to fix the warranty time . See Chou–Tang (1992) and Bouza (2001) for a discussion of the theoretical problem.

Example 2. The concentration of 239.246 Pu in surface oil over an area in the surroundings of a nuclear power station is to be measured. Samples of quadrates to a depth of 10 cm will be collected. The interest is to calculate Q(P) for establishing if the station operates safely. The initial samples are used for estimating the quantile. Changes in the emissions will be analyzed by comparing the posterior monitoring results with the threshold fixed by the estimate, see a detailed description in Gilbert (1987). The collector determines “in situ“ if the selected quadrate has a smaller concentration. Then each sampled quadrate is compared with the previous minimum before deciding to collect it or not.

Example 3. The newsboy vendor problem models a wide variety of economical problems. It is described in the following way:

a. The newsboy has an initial wealth. He buys newspapers at a price c(1) and sells them at price c(2). The unsold papers are returned to the dealer and he obtains c(3) for each unit. When a lot is insufficient he buys additional newspapers at a cost c(4). The demand is a random variable with distribution function F(y). The payoff is:
   Z(y,x) = Z* + c(2)y - c(1)x + c(3)[Max (0, -x -y) - c(4)[Max(0, y - x)] with c(3) < c(1) < c (2)

b. Using the utility function u[Z(x,a)] the optimization problem to be solved is: $\text{Max}_{a\ge 0}E[u(Z(x,a))]$.

It is a Stochastic Linear Programming problem. The optimal decision for a distribution function F is x(F) = Q(P).

If the newsboy is risk neutral: P = c(4)c(1)/[c(4) - c(3)], see Eeckhoudt-Golleer (1995).

If the newsboy is risk averse P < c(4)c(1)/[c(4)-c(3)].

In practice the newsboy does not know F, as he is maximizing a small change is needed for computing Q(P) using the records of the successive maxima.

ACKNOWLEDGMENTS

These results were partially obtained under the support of a TWAS project that sustained a visit of the first author to India.

REFERENCES

BOUZA, C. (2001): “Investigation of Burn-in-time problems with unknown failure time distribution”, **J. of Statistics and Management Sc**. 37, 1-7.

CHOU, K. and K. TAG (1992): ¨”Burn in time and estimation of change point with Weibull exponential mixture distribution”, **Decision Sc**. 23, 973-988.

CHANDLER, K.N. (1952): “The distribution and frequencies of record values”, **J. Royal Stat. Soc. B**. 14, 220-228.

EECKHOUDT, L.C.; C. GOLLER and H. SCHLESINGER (1995): “The risk averse (and prudent) newsboy”, **Manag. Sc**. 41, 786-794.

GLICK, N. (1987): “Breaking records and breaking boards”, **Amer. Math, Monthly**. 85, 2-26.

GILBERT, R.O. (1987): Statistical methods for environmental pollution monitoring. Van Nostrand Reinhold, N. York.

GULATI, S. and W.J. PADGETT (1994): “Non parametric quantile estimation form record breaking data”, **Australian J. Stat**. 36, 211-223.

NADARAYA, E.A. (1965): “On non-parametric estimates of distribution functions and regression curves”, **Theory of Prob. and Appl**. 10, 186-190.

PFLUG, G.Ch. (1996): Optimization of stochastic models. The interface between Simulation and Optimization. Kluwer Acad. Pub. Boston.

RÖMISCH, W. and R. SCHULTZ (1989): “Stability for stochastic programs”, **Preprint** 242, Humboldt University.

SAMANIEGO, F.J. and L.R. WHITAKER (1998): “On estimating population characteristics from record breaking observation II: non-parametric results”, **Naval. Res. Logistics Quart**. 35, 221-236.