Tilburg University

The empirical distribution function as a tail estimator

Einmahl, J.H.J.

Published in:
Statistica Neerlandica

Publication date:
1990

[Link to publication in Tilburg University Research Portal](#)

Citation for published version (APA):
Einmahl, J. H. J. (1990). The empirical distribution function as a tail estimator. *Statistica Neerlandica*, 44(2), 79-82.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

The empirical distribution function as a tail estimator

J.H.J. Einmahl

Department of Medical Informatics and Statistics

P.O. Box 616

NL-6200 MD Maastricht

The Netherlands

It is shown that the empirical distribution function performs well in tail estimation. Consistency as well as the relative error are considered. The results hold true without any condition on the theoretical distribution function.

Key Words & Phrases: consistency, relative error.

1. INTRODUCTION

Very recently a number of results are established concerning the estimation of the (right) tail of a distribution function (df) F using a random sample from that distribution. It is clear that for estimation of the tail of F beyond (or near) the greatest observation the empirical df F_n is of no use. Even more so, nothing reasonable can be done in this region without additional assumptions on F . Therefore most of the published results on this subject are established under certain extreme value type conditions. For details the reader is referred to PICKANDS III (1975, section 4), DAVIS and RESNICK (1984, Theorem 3.2) and BEIRLANT and TEUGELS (1986, Theorem 3), where consistency of an estimator is shown and SMITH (1987, sections 8 and 9) and SMITH and WEISSMAN (1987), where the relative error of estimation (the estimator divided by $1 - F$) is considered.

However, if F has to be estimated "near" $F^{-1}(1 - k_n/n)$, where F^{-1} is the inverse of F and $\{k_n\}_{n=1}^{\infty}$ a sequence of numbers with $0 < k_n \leq n$ and $k_n \rightarrow \infty$, the empirical df F_n is a good choice for the estimator. On the other hand, a number of the results in the papers cited above are concerned with this region instead of the one mentioned in the first paragraph. It is the aim of this note to support the just made claim concerning F_n and to compare our results with those in the literature. It will become clear that F_n is wrongly neglected in (intermediate) tail estimation.

Apart from the fact that F_n is a simple estimator, one advantage of our choice is that the results hold true without *any* condition on the df F . Theorems 1 and

2 below improve parts of the aforementioned papers also with respect to the rate of convergence (Theorem 1) and the range over which the supremum is taken (Theorem 2).

2. NOTATION AND RESULTS

In order to be more explicit let us specify our setup and introduce some notation. Let $X_1, X_2, \dots, X_n$ be a random sample of size n from an arbitrary df F on $\mathbb{R}$. Denote the survival function $1-F$ by $\bar{F}$ and the inverse of F by F^{-1} , i.e.

$$F^{-1}(x) = \inf\{t : F(t) \geq x\}, \quad 0 < x < 1.$$

The empirical df F_n is defined by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{(-\infty, t]}(X_i), \quad t \in \mathbb{R};$$

write $\bar{F}_n$ for $1-F_n$. Moreover, throughout $\{k_n\}_{n=1}^\infty$ will denote an arbitrary sequence of positive constants, such that $k_n \leq n$ and $k_n \rightarrow \infty$, $k_n/n \rightarrow 0$ ($n \rightarrow \infty$). The results are presented in a weak approximation setting. The first one is of the consistency type; a somewhat related result can be found in PICKANDS III (1975, Theorem 11).

THEOREM 1. *On a rich enough probability space there exist a sequence $X_1, X_2, \dots$ of i.i.d. random variables with common df F and a sequence $W_1, W_2, \dots$ of standard Wiener processes on $[0, 1]$ such that as $n \rightarrow \infty$*

$$\sup_{t \geq F^{-1}(1-k_n/n)} |(n/k_n^{1/2})(\bar{F}_n(t) - \bar{F}(t)) - W_n((n/k_n)\bar{F}(t))| \rightarrow_P 0.$$

COROLLARY 1. *Let $\{m_n\}_{n=1}^\infty$ be a sequence of positive constants, such that $m_n/k_n^{1/2} \rightarrow \infty$ ($n \rightarrow \infty$), then*

$$\sup_{t \geq F^{-1}(1-k_n/n)} (n/m_n)|\bar{F}_n(t) - \bar{F}(t)| \rightarrow_P 0.$$

The corollary compares with the aforementioned results in DAVIS and RESNICK (1984) and BEIRLANT and TEUGELS (1986); a possible choice is to set m_n equal to k_n . Note that the corollary is not trivial in the sense that both terms tend to zero in probability, since for continuous F $\sup_{t \geq F^{-1}(1-k_n/n)} (n/m_n)\bar{F}(t) = k_n/m_n$, which may tend to ∞ (e.g. by taking $m_n = k_n^{3/4}$).

PROOF. It easily follows from Theorem 1 in MASON and VAN ZWET (1987) (see also MASON (1988, (1.1)-(1.3))) that there exists a probability space such that

$$\sup_{1-k_n/n \leq s \leq 1} |(n/k_n^{1/2})((1-\Gamma_n(s)) - (1-s)) - W_n((n/k_n)(1-s))| \rightarrow_P 0,$$

where Γ_n is the empirical df based on the first n of a sequence $U_1, U_2, \dots$ of independent uniform-(0,1) random variables and $W_1, W_2, \dots$ is a sequence of standard Wiener processes on $[0, 1]$. Now setting $s = F(t)$ and noting that

$X_i := F^{-1}(U_i)$ is distributed according to F , completes the proof. $\square$

Our second and last result considers the relative error $\bar{F}_n/\bar{F}$.

THEOREM 2. *On a rich enough probability space there exist a sequence $X_1, X_2, \dots$ of i.i.d. random variables with common df F and a sequence $W_1, W_2, \dots$ of standard Wiener processes on $[0, 1]$ such that as $n \rightarrow \infty$*

$$\sup_{t < F^{-1}(1-k_n/n)} \left| k_n^{1/2} \left[\frac{\bar{F}_n(t)}{\bar{F}(t)} - 1 \right] - W_n \left[\frac{k_n}{n\bar{F}(t)} \right] \right| \rightarrow_P 0.$$

COROLLARY 2. *Let $\{r_n\}_{n=1}^\infty$ be a sequence of positive constants, such that $r_n/k_n^{1/2} \rightarrow 0 (n \rightarrow \infty)$, then*

$$\sup_{t < F^{-1}(1-k_n/n)} r_n |(\bar{F}_n(t)/\bar{F}(t)) - 1| \rightarrow_P 0.$$

Theorem 2 and Corollary 2 are in the same spirit as the aforementioned results in SMITH (1987) and SMITH and WEISSMAN (1987). A precise comparison is difficult because of the "conditional" setup and the complicated conditions in both papers and the fact that F is estimated there in just one (moving) point instead of on a whole region.

PROOF. It follows immediately from Theorem 1 and its proof in MASON (1985) that there exists a probability space such that

$$\sup_{0 \leq s < 1-k_n/n} \left| k_n^{1/2} \left[\frac{1-\Gamma_n(s)}{1-s} - 1 \right] - W_n \left[\frac{k_n}{n} \frac{s}{1-s} \right] \right| \rightarrow_P 0.$$

(Here the notation is the same as in the proof of Theorem 1.) It easily follows from oscillation modulus type results for W_n that

$$\sup_{0 \leq s < 1-k_n/n} \left| W_n \left[\frac{k_n}{n} \frac{s}{1-s} \right] - W_n \left[\frac{k_n}{n(1-s)} \right] \right| \rightarrow_P 0,$$

since $k_n/(n(1-s)) - k_ns/(n(1-s)) = k_n/n$. Hence

$$\sup_{0 \leq s < 1-k_n/n} \left| k_n^{1/2} \left[\frac{1-\Gamma_n(s)}{1-s} - 1 \right] - W_n \left[\frac{k_n}{n(1-s)} \right] \right| \rightarrow_P 0.$$

Again setting $s = F(t)$ completes the proof. $\square$

REMARK. Using results in EINMAHL and MASON (1988), MASON (1988) and WELLNER (1978) it is possible to obtain under some mild additional conditions on k_n similar results concerning the almost sure (instead of "in probability") behaviour of $\bar{F}_n$ in the tail.

REFERENCES

- BEIRLANT, J. and J.K. TEUGELS (1986), The asymptotic behavior of Hill's estimator, *Theory of Probability and its Applications* 31, 463-469.
- DAVIS, R. and S. RESNICK (1984), Tail estimation motivated by extreme value theory, *The Annals of Statistics* 12, 1467-1487.
- EINMAHL, J.H.J. and D.M. MASON (1988), Laws of the iterated logarithm in the tails for weighted uniform empirical processes, *The Annals of Probability* 16, 126-141.
- MASON, D.M. (1985), The asymptotic distribution of generalized Rényi statistics, *Acta Scientiarum Mathematicarum (Szeged)* 48, 315-323.
- MASON, D.M. (1988), A strong invariance theorem for the tail empirical process, *Annales de l'Institut Henri Poincaré Probabilités et Statistiques* 24, 491-506.
- MASON, D.M. and W.R. VAN ZWET (1987), A refinement of the KMT inequality for the uniform empirical process, *The Annals of Probability* 15, 871-884.
- PICKANDS III, J. (1975), Statistical inference using extreme order statistics, *The Annals of Statistics* 3, 119-131.
- SMITH, R.L. (1987), Estimating tails of probability distributions, *The Annals of Statistics* 15, 1174-1207.
- SMITH, R.L. and I. WEISSMAN (1987), Large deviations of tail estimators based on the Pareto approximation, *Journal of Applied Probability* 24, 619-630.
- WELLNER, J.A. (1978), Limit theorems for the ratio of the empirical distribution function to the true distribution function, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 45, 73-88.

Received September 1989