stichting
mathematisch
centrum

MC

DEPARTMENT OF MATHEMATICAL STATISTICS SW 34/76 NOVEMBER
(AFDELING MATHEMATISCHE STATISTIEK)

M.C.A. VAN ZUYLEN

SOME PROPERTIES OF THE EMPIRICAL DISTRIBUTION FUNCTION IN THE NON-I.I.D. CASE

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS(MOS) subject classification scheme (1970): 60G17, 62G30

First printing January 1975

Second edition November 1976

SOME PROPERTIES OF THE EMPIRICAL DIŞTRIBUTION FUNCTION IN THE NON-I.I.D. CASE

by

M.C.A. van Zuylen

ABSTRACT

For $N = 1,2,\ldots$ let $X_{1N},X_{2N},\ldots,X_{NN}$ be independent rv's having continuous df's $F_{1N},F_{2N},\ldots,F_{NN}$. For the set $X_{1N},X_{2N},\ldots,X_{NN}$, let us denote by $X_{1:N}\leq X_{2:N}\leq\ldots\leq X_{N:N}$ the order statistics, by $\mathbb{F}_N$ the empirical df and by $\bar{F}_N$ the averaged df, i.e. $\bar{F}_N(x) = N^{-1}\sum_{n=1}^{N} F_{nN}(x)$ for $x \in (-\infty,\infty)$. It is shown that for each $\varepsilon > 0$ there exists a $0 < \beta(=\beta_\varepsilon) < 1$, independent of N, such that for $N = 1,2,\ldots,$

(a) $P(\mathbb{F}_N(x) \leq \beta^{-1}\bar{F}_N(x),\text{ for } x \in (-\infty,\infty)) \geq 1-\varepsilon,$

(b) $P(\mathbb{F}_N(x) \geq \beta\bar{F}_N(x),\text{ for } x \in [X_{1:N},\infty)) \geq 1-\varepsilon.$

Moreover, these assertions hold uniformly in all continuous df's $F_{1N},F_{2N},\ldots$ $\ldots,F_{NN}$.

The theorem can be used to prove asymptotic normality of rank statistics in the case where the sample elements are allowed to have different df's.

KEY WORDS & PHRASES: *Empirical distribution function, order statistics.*

1. NOTATION AND RESULTS

For $N = 1,2,\ldots$ let $X_{1N},X_{2N},\ldots,X_{NN}$ be independent random variables (rv's) having continuous distribution functions (df's) $F_{1N},F_{2N},\ldots,F_{NN}$, where all these rv's are supposed to be defined on a single probability space (Ω,A,P). Let us define for $N = 1,2,\ldots$ the empirical df $\mathbb{F}_N$ of $X_{1N},X_{2N},\ldots,X_{NN}$ by taking $N\mathbb{F}_N(x)$ to be the number of elements in the set $\{X_{nN} : X_{nN} \leq x,\ n = 1,2,\ldots,N\}$. The order statistics of $X_{1N},X_{2N},\ldots,X_{NN}$ are denoted by $X_{1:N}\leq X_{2:N}\leq\ldots\leq X_{N:N}$ and let the averaged df $\bar{F}_N$ be defined as $\bar{F}_N(x) = N^{-1}\sum_{n=1}^{N} F_{nN}(x)$ for $x \in (-\infty,\infty)$.

Define $\bar{F}_N^{-1}$ by $\bar{F}_N^{-1}(s) = \inf\{x\colon \bar{F}_N(x) \geq s\}$. Because $\bar{F}_N$ is supposed to be continuous we have $\bar{F}_N(\bar{F}_N^{-1}(s)) = s$ for $s \in (0,1)$.

THEOREM. *For each* $\varepsilon > 0$ *there exists a* $0 < \beta(=\beta_\varepsilon) < 1$, *independent of* N *and of the continuous df's* $F_{1N},F_{2N},\ldots,F_{NN}$, *such that for* $N = 1,2,\ldots$,

(a) $$P(\mathbb{F}_N(x) \leq \beta^{-1}\bar{F}_N(x),\ \textit{for}\ x \in (-\infty,\infty)) \geq 1-\varepsilon,$$

(b) $$P(\mathbb{F}_N(x) \geq \beta\bar{F}_N(x),\ \textit{for}\ x \in [X_{1:N},\infty)) \geq 1-\varepsilon.$$

Replacing X_{nN} by $-X_{nN}$ for $n = 1,2,\ldots,N$, $N = 1,2,\ldots$ the following corollary is immediate.

COROLLARY. *For each* $\varepsilon > 0$ *there exists a* $0 < \beta(=\beta_\varepsilon) < 1$, *independent of* N *and of the continuous df's* $F_{1N},F_{2N},\ldots,F_{NN}$, *such that for* $N = 1,2,\ldots$,

(a) $$P(\mathbb{F}_N(x) \geq 1 - \beta^{-1}(1-\bar{F}_N(x)),\ \textit{for}\ x \in (-\infty,\infty)) \geq 1-\varepsilon,$$

(b) $$P(\mathbb{F}_N(x) \leq 1 - \beta(1-\bar{F}_N(x)),\ \textit{for}\ x \in (-\infty,X_{N:N})) \geq 1-\varepsilon.$$

The theorem and the corollary are useful for proving asymptotic normality of rank statistics in the case where the sample elements are allowed to have different df's.

For the i.i.d. case these results are well-known. They are given by SHORACK in [3] and proved in [2]. The present theorem will be proved in an entirely different manner. The basic tool is a result of HOEFFDING [1] which is given in the lemma below.

Suppose that $Z_1, Z_2, \ldots, Z_N$ are independent rv's, each assuming the values 0 and 1 only, with

$$\Pr(Z_j = 1) = p_j \qquad \text{for } j = 1,2,\ldots,N$$

and

$$0 < N^{-1} \sum_{j=1}^{N} p_j = \bar{p} < 1.$$

LEMMA (HOEFFDING). *If* f *is a strictly convex function defined on* $(-\infty,\infty)$ *then*

$$E(f(\sum_{j=1}^{N} Z_j)) \le \sum_{k=0}^{N} f(k)\binom{N}{k}\bar{p}^k(1-\bar{p})^{N-k},$$

where equality holds if and only if $p_1=p_2=\ldots=p_N = \bar{p}$.

In particular this lemma together with Markov's equalities implies that for $n > N\bar{p}$,

$$\text{(1.1)} \qquad \Pr(\sum_{j=1}^{N} Z_j \ge n) \le \Pr(|\sum_{j=1}^{N} Z_j - N\bar{p}| \ge n - N\bar{p})$$

$$\le (n-N\bar{p})^{-4}\, E(\sum_{j=1}^{N} Z_j - N\bar{p})^4$$

$$\le (n-N\bar{p})^{-4} \sum_{k=0}^{N} (k-N\bar{p})^4\binom{N}{k}\bar{p}^k(1-\bar{p})^{N-k}$$

$$= (n-N\bar{p})^{-4}\{(\bar{p}(1-\bar{p}))^2(3N^2-6N) + N\bar{p}(1-\bar{p})\}$$

$$\le (n-N\bar{p})^{-4} \min((3N^2\bar{p}^2+N\bar{p}),\ (3N^2(1-\bar{p})^2 + N(1-\bar{p}))).$$

2. PROOF OF THE THEOREM

(a): For all $N \ge 1$ and all $0 < \beta < 1$ we have

$$(2.1)\qquad P(\mathbb{F}_N(x) \le \beta^{-1}\bar{F}_N(x),\ \text{for } x \in (-\infty,\infty))$$

$$= P(\mathbb{F}_N(X_{n:N}) \le \beta^{-1}\bar{F}_N(X_{n:N}),\ \text{for } n = 1,2,\ldots,N)$$

$$\ge 1 - \sum_{n=1}^{N} P(\bar{F}_N(X_{n:N}) < \beta n N^{-1}).$$

Here we have used the Bonferroni inequality and the fact that $\mathbb{F}_N(X_{n:N}) = nN^{-1}$ with probability one for $n = 1,2,\ldots,N$. Now for $n = 1,2,\ldots,N$,

$$(2.2)\qquad P(\bar{F}_N(X_{n:N}) < \beta n N^{-1}) = P(X_{n:N} < \bar{F}_N^{-1}(\beta n N^{-1})) = \Pr(\sum_{j=1}^{N} Z_j \ge n),$$

where $Z_1, Z_2, \ldots, Z_N$ are independent rv's each assuming the values 0 and 1 only, with

$$\Pr(Z_j = 1) = p_j = F_{jN}(\bar{F}_N^{-1}(\beta n N^{-1})) \qquad \text{for } j = 1,2,\ldots,N.$$

From (2.2) and (1.1) with $\bar{p} = N^{-1}\sum_{j=1}^{N} F_{jN}(\bar{F}_N^{-1}(\beta n N^{-1})) = \beta n N^{-1}$ it is now immediate that for $n = 1,2,\ldots,N$,

$$P(\bar{F}_N(X_{n:N}) < \beta n N^{-1}) \le \beta(1-\beta)^{-4}(3\beta n^{-2} + n^{-3}) \le 4\beta(1-\beta)^{-4}n^{-2},$$

so that

$$(2.3)\qquad \sum_{n=1}^{N} P(\bar{F}_N(X_{n:N}) < \beta n N^{-1}) \le M_1\beta(1-\beta)^{-4} \to 0 \quad \text{as } \beta \to 0,$$

where M_1 is a finite constant, independent of N and of the df's $F_{1N}, F_{2N}, \ldots, F_{NN}$. Assertion (a) in the theorem now follows from (2.3) and (2.1).

(b): For all $N \ge 1$ and all $0 < \beta < 1$ we have

$$(2.4)\qquad P(\mathbb{F}_N(x) \ge \beta\bar{F}_N(x),\ \text{for } x \in [X_{1:N},\infty))$$

$$\ge P(\bigcap_{n=2}^{N} [\mathbb{F}_N(X_{n:N}-) \ge \beta\bar{F}_N(X_{n:N})])$$

$$\ge 1 - \sum_{n=2}^{N} P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}) = 1 - \sum_{n=2}^{[\beta N]+1} P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}),$$

where $[\beta N]$ is the greatest integer in βN. The terms with $n > [\beta N] + 1$ may be omitted because $\beta^{-1}(n-1)N^{-1} > 1$ for these terms. Since $n \geq 2$ we are guaranteed that $0 < \beta(n-1)N^{-1} \leq 1$ for every term, for each $N \geq 1$ and $0 < \beta < 1$.

Now for $n = 2,3,\ldots,[\beta N]+1$,

$$(2.5) \qquad P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}) = P(\sum_{j=1}^{N} Z_j \geq N - n + 1),$$

where $Z_1, Z_2, \ldots, Z_N$ are independent rv's each assuming the values 0 and 1 only, where in this case

$$\Pr(Z_j = 1) = p_j = 1 - F_{jN}(\bar{F}_N^{-1}(\beta^{-1}(n-1)N^{-1})) \qquad \text{for } j = 1,2,\ldots,N.$$

From (2.5) and (1.1) with now $\bar{p} = 1 - \beta^{-1}(n-1)N^{-1}$ it is immediate again that for $n = 2,3,\ldots,[\beta N]+1$,

$$P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}) \leq 4\beta^{-2}(1-\beta^{-1})^{-4}(n-1)^{-2},$$

so that

$$(2.6) \qquad \sum_{n=2}^{[\beta N]+1} P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}) \leq M_2\beta^{-2}(1-\beta^{-1})^{-4} \to 0 \quad \text{as } \beta \to 0,$$

where M_2 is a finite constant, independent of N and of the df's $F_{1N}, F_{2N}, \ldots, F_{NN}$. Assertion (b) in the theorem now follows from (2.6) and (2.4).

ACKNOWLEDGEMENT. I wish to thank Dr. F.H. RUYMGAART for suggesting the problem and Prof. Dr. W.R. VAN ZWET for his careful reading of the manuscript and his valuable comments.

REFERENCES

[1] HOEFFDING, W., *On the distribution of the number of successes in independent trials.* Ann. Math. Statist. 27 (1956) pp. 713-721.

[2] SHORACK, G.R., *A uniformly convergent empirical process.* Tech. Report No. 20, Math. Dept., Univ. of Washington, 1970a.

[3] SHORACK, G.R., *Functions of order statistics.* Ann. Math. Statist. 43 (1972) pp. 412-427.

LIST OF SYMBOLS

		LATIN			GREEK	MATHEMATICS
Normal		Italics		Script		
A	a	*A*	*a*	$\mathcal{A}$		$-$: "bar"
B	b	*B*	*b*		β	0 : zero
C	c	*C*	*c*			1 : one
D	d	*D*	*d*			∞ : infinity
E	e	*E*	*e*	$\mathcal{E}$	Σ	$\sum$: summation
F	f	*F*	*f*			$\in$: element of
G	g	*G*	*g*			$\cap$: intersection
H	h	*H*	*h*		Ω	$\mathbb{F}_N$: empirical distribution function
I	i	*I*	*i*			
J	j	*J*	*j*			
K	k	*K*	*k*			
L	l	*L*	*l*			
M	m	*M*	*m*			
N	n	*N*	*n*			
O	o	*O*	*o*			
P	p	*P*	*p*			
Q	q	*Q*	*q*			
R	r	*R*	*r*			
S	s	*S*	*s*			
T	t	*T*	*t*			
U	u	*U*	*u*			
V	v	*V*	*v*			
W	w	*W*	*w*			
X	x	*X*	*x*			
Y	y	*Y*	*y*			
Z	z	*Z*	*z*			