The Annals of Statistics
1976, Vol. 4, No. 2, 406-408

SOME PROPERTIES OF THE EMPIRICAL DISTRIBUTION FUNCTION IN THE NON-i.i.d. CASE[1]

By M. C. A. van Zuylen

Mathematisch Centrum, Amsterdam

For $N = 1, 2, \cdots$ let $X_{1N}, X_{2N}, \cdots, X_{NN}$ be independent rv's having continuous df's $F_{1N}, F_{2N}, \cdots, F_{NN}$. For the set $X_{1N}, X_{2N}, \cdots, X_{NN}$, let us denote by $X_{1:N} \leqq X_{2:N} \leqq \cdots \leqq X_{N:N}$ 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 $N = 1, 2, \cdots$,

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

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

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

1. Notation and results. Let X_{nN}, F_{nN}, etc. be as in the abstract. Define $\bar{F}_N^{-1}$ by $\bar{F}_N^{-1}(s) = \inf\{x : \bar{F}_N(x) \geqq s\}$. Because $\bar{F}_N$ is assumed 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 $F_{1N}, F_{2N}, \cdots, F_{NN}$, *such that for* $N = 1, 2, \cdots$,

(a) $P(1 - \beta^{-1}(1 - \bar{F}_N(x)) \leqq \mathbb{F}_N(x) \leqq \beta^{-1}\bar{F}_N(x)$, *for* $x \in (-\infty, \infty)) \geqq 1 - \varepsilon$,
(b) $P(1 - \beta(1 - \bar{F}_N(x)) \geqq \mathbb{F}_N(x) \geqq \beta\bar{F}_N(x)$, *for* $x \in [X_{1:N}, \infty)) \geqq 1 - \varepsilon$.

The theorem is useful for proving asymptotic normality of rank statistics and of linear combinations of functions of order 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; statements appear in Shorack [2]. Note that our theorem fills the gap in extending the proof of Theorem 3 of Shorack [3], so that his Theorem 3 can now be claimed to hold without the assumption $b_1 = b_2 = 0$ (that is, for unbounded J).

Our basic tool is a result of Hoeffding [1]. Suppose that Z_j, $1 \leqq j \leqq N$, are independent Bernoulli (p_j) rv's and suppose

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

Received February 1975; revised June 1975.

[1] Report SW 34/75 of the Department of Mathematical Statistics, Mathematisch Centrum, Amsterdam.

AMS 1970 *subject classifications.* Primary 60G17; Secondary 62G30.

Key words and phrases. Empirical distribution function, order statistics.

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

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

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

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

$$\begin{aligned} &P(\textstyle\sum_{j=1}^{N} Z_j \geqq n) \\ &\qquad \leqq P(|\textstyle\sum_{j=1}^{N} Z_j - N\bar{p}| \geqq n - N\bar{p}) \\ &\qquad \leqq (n - N\bar{p})^{-4}\mathscr{E}(\textstyle\sum_{j=1}^{N} Z_j - N\bar{p})^4 \\ &\qquad \leqq (n - N\bar{p})^{-4} \textstyle\sum_{k=0}^{N} (k - N\bar{p})^4\binom{N}{k}\bar{p}^k(1-\bar{p})^{N-k} \\ &\qquad = (n - N\bar{p})^{-4}\{(\bar{p}(1-\bar{p}))^2(3N^2 - 6N) + N\bar{p}(1-\bar{p})\} \\ &\qquad \leqq (n - N\bar{p})^{-4} \min((3N^2\bar{p}^2 + N\bar{p}), (3N^2(1-\bar{p})^2 + N(1-\bar{p}))) . \end{aligned} \tag{1.1}$$

2. Proof of the theorem. (a) For all $N \geqq 1$ and all $0 < \beta < 1$ we have

$$\begin{aligned} &P(\mathbb{F}_N(x) \leqq \beta^{-1}\bar{F}_N(x), \text{ for } x \in (-\infty, \infty)) \\ &\qquad = P(\mathbb{F}_N(X_{n:N}) \leqq \beta^{-1}\bar{F}_N(X_{n:N}), \text{ for } n = 1, 2, \cdots, N) \\ &\qquad \geqq 1 - \textstyle\sum_{n=1}^{N} P(\bar{F}_N(X_{n:N}) < \beta n N^{-1}) = 1 - \sum_{n=1}^{N} P(\sum_{j=1}^{N} Z_j \geqq n) , \end{aligned} \tag{2.1}$$

where Z_j, $1 \leqq j \leqq N$, are independent Bernoulli (p_j) rv's with

$$p_j = F_{jN}(\bar{F}_N^{-1}(\beta n N^{-1})) .$$

From (2.1) 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 $n = 1, 2, \cdots, N$,

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

so that

$$\textstyle\sum_{n=1}^{N} P(\bar{F}_N(X_{n:N}) < \beta n N^{-1}) \leqq M_1\beta(1-\beta)^{-4} \to 0 \qquad \text{as} \quad \beta \to 0 , \tag{2.2}$$

where M_1 is a finite constant, independent of N and of the df's $F_{1N}, F_{2N}, \cdots, F_{NN}$. Assertion (a) in the theorem now follows from (2.2), (2.1) and the symmetric result that comes from replacing X_{nN} by $-X_{nN}$.

(b) For all $N \geqq 1$ and all $0 < \beta < 1$ we have (using $(n-1)/\beta N > 1$ for $n > [\beta N] + 1$)

$$\begin{aligned} &P(\mathbb{F}_N(x) \geqq \beta\bar{F}_N(x), \text{ for } x \in [X_{1:N}, \infty)) \\ &\qquad \geqq P(\textstyle\bigcap_{n=2}^{N} [\mathbb{F}_N(X_{n:N}-) \geqq \beta\bar{F}_N(X_{n:N})]) \\ &\qquad \geqq 1 - \textstyle\sum_{n=2}^{N} P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}) \\ &\qquad = 1 - \textstyle\sum_{n=2}^{[\beta N]+1} P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}) , \\ &\qquad = 1 - \textstyle\sum_{n=2}^{[\beta N]+1} P(\sum_{j=1}^{N} Z_j \geqq N - n + 1) , \end{aligned} \tag{2.3}$$

where Z_j, $1 \leqq j \leqq N$, are independent Bernoulli (p_j) rv's with

$$p_j = 1 - F_{jN}(\bar{F}_N^{-1}(\beta^{-1}(n-1)N^{-1})) .$$

From (2.3) and (1.1) with now $\bar{p} = 1 - \beta^{-1}(n-1)N^{-1}$ it is immediate again that for $n = 2, 3, \cdots, [\beta N] + 1$,

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

so that

$$\sum_{n=2}^{[\beta N]+1} P(\bar{F}_N(X_{n:N}) > \beta^{-1}(n-1)N^{-1}) \leqq M_2\beta^2(1-\beta)^{-4} \rightarrow 0 \quad \text{as} \quad \beta \rightarrow 0, \tag{2.4}$$

where M_2 is a finite constant, independent of N and of the df's $F_{1N}, F_{2N}, \cdots, F_{NN}$. Assertion (b) in the theorem now follows from (2.4), (2.3) and symmetry.

Acknowledgment. I wish to thank Frits Ruymgaart for suggesting the problem and Professor W. R. van Zwet for his careful reading of the manuscript and his valuable comments.

REFERENCES

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

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

[3] Shorack, G. R. (1973). Convergence of reduced empirical and quantile processes with application to functions of order statistics in the non-i.i.d. case. *Ann. Statist.* **1** 146-152.

Afd. Statistiek
Mathematisch Centrum
2e Boerhaavestraat 49
Amsterdam—0, Holland