Distributions with Fixed Marginals and Related Topics
IMS Lecture Notes – Monograph Series Vol. 28, 1996

Order Statistics of Variables with Given Marginal Distributions

By Tomasz Rychlik*
Polish Academy of Sciences

We review some results on order statistics based on random variables with given one-dimensional distributions. We present bounds on the distribution function of each order statistic and conditions on the marginals for attaining the bounds. For identically distributed samples, we show sharp bounds for the expectation and variance of arbitrary function of a given order statistic and for the expectation of an arbitrary L-estimate.

1. Introduction. Order statistics and functions of order statistics have numerous applications in statistical inference (see, e.g., Balakrishnan and Cohen (1991), David (1981)). The theory has been developed mostly for the standard model of independent identically distributed random variables, for which order statistics have simple distribution functions and the limiting distributions with the rates of convergence for various sequences of order statistics and parent distributions have been explicitly described (see Balkema and de Haan (1978), Reiss (1989)). Asymptotic representations for order statistics, especially extreme ones, were also intensively studied under various relaxations of the independence assumption (see, e.g., Leadbetter et al. (1983)). Another direction of research is devoted to formulas and recurrence relations for distributions and moments of order statistics based on independent non-identically distributed or even arbitrarily distributed random variables (cf. Balakrishnan (1992) and Balakrishnan et al. (1992), respectively).

The aim of this paper is to summarize some results on order statistics based on possibly dependent random variables $X_1, \ldots, X_n$, with given one-dimensional distribution functions $F_1, \ldots, F_n$, respectively. We will write $X_{m:n}$ for the mth smallest order statistic and $F_{m:n}$ for its distribution function. In Section 2 we determine bounds on the distribution functions of each order

* The author was partially supported by KBN Grant 2P301-10-4.

AMS 1991 Subject Classification: Primary 62E10; Secondary 62G30

Key words and phrases: Distributions with given marginals, dependent identically distributed random variables, order statistic, L-estimate, sharp inequalities.

statistic and present conditions in terms of marginal distributions for the existence of stochastically extreme order statistics, whose distributions attain the bound, on the whole real axis. In Section 3 we first characterize all possible distributions of each order statistic based on dependent identically distributed random variables with a given marginal. Then we use the characterization to obtain sharp bounds for the expectation and variance of an arbitrary measurable function of the order statistic. In Section 4 we establish accurate bounds for the expectation of arbitrary L-estimates arising from an arbitrarily dependent identically distributed sample. We also indicate some applications to deterministic inequalities for L-estimates and to robust statistics.

It is worth pointing out that the restrictive assumption of identical marginals in Sections 3 and 4 is essential and it cannot be trivially removed, because standard transformations of arbitrary joint distributions onto copulas generally do not preserve the order of variables. Observe finally that some lower bounds follow immediately from upper ones and conversely by a simple change of variables $Y_i = -X_i$, $i = 1, \ldots, n$, implying $Y_{m:n} = -X_{n+1-m:n}$. Consequently, we concentrate below merely on deriving upper bounds except in Section 3, where the trivial transformation does not provide the opposite bounds.

2. Bounds on Distribution Functions of Order Statistics. Two well known pairs of bivariate random variables attaining the Fréchet bounds (cf. Fréchet (1951)) also provide the stochastically largest and smallest maximum and minimum of variables with marginals F_1, F_2:

$$\max\{F_1, F_2\} \leq F_{1:2} \leq \min\{F_1 + F_2, 1\},$$
$$\max\{F_1 + F_2 - 1, 0\} \leq F_{2:2} \leq \min\{F_1, F_2\}.$$

These results can be generalized to the extremes of $n \geq 2$ random variables. The distribution function of the maximum attains the upper bound $\min\{F_1, \ldots, F_n\}$ (and that of the minimum attains the lower bound $\max\{F_1, \ldots, F_n\}$) if $X_i = F_i^{-1}(U)$, $i = 1, \ldots, n$, for a random variable U uniformly distributed on $[0, 1]$. Construction of the stochastically largest maximum, with the distribution function attaining the lower bound $\max\{\sum_{i=1}^n F_i - n + 1, 0\}$, is less obvious. It was first presented by Mallows (1969) for $[0, 1]$-uniformly distributed variables and by Lai and Robbins (1976) for general marginals. Tchen (1980) proved the existence of an infinite sequence of random variables with given marginals such that every partial maximum is stochastically largest. Theorem 1 provides an upper bound for the distribution function of an arbitrary order statistic and necessary and sufficient conditions for the existence of a random sample with given marginals such that the bound is attained.

THEOREM 1. (Rychlik (1995)). (a) *If random variables* $X_1, \dots, X_n$ *have distribution functions* $F_1, \dots, F_n$, respectively, then

$$F_{m:n} \leq \min\left\{\frac{1}{m}\sum_{i=1}^{n} F_i, 1\right\}. \tag{1}$$

(b) *Let* $x^\star$ *stand for the upper end-point of the distribution function* $F^\star = \min\{\frac{1}{m}\sum_{i=1}^{n} F_i, 1\}$, *and* $P_1(x^\star), \dots, P_n(x^\star)$, *and* $P^\star(x^\star)$ *for the jumps of* $F_1, \dots, F_n$, *and* $F^\star$ *at* $x^\star$, *respectively. Then there exist random variables* $X_1, \dots, X_n$, *and* $X_{m:n}$ *on a common probability space, with distribution functions* $F_1, \dots, F_n$, *and* $F^\star$, *respectively, iff*

$$\frac{dF_j}{d(\sum F_i)} \leq \frac{1}{m} \quad on \ (-\infty, x^\star) \quad \left(\sum F_i\right) - a.s., \tag{2}$$

$$\sum_{j=1}^{n} \min\{P_j(x^\star),\ P^\star(x^\star)\} \geq mP^\star(x^\star). \tag{3}$$

IDEA OF THE PROOF. (a) We apply the identity

$$mF_{m:n}(x) = \sum_{i=1}^{n} F_i(x) - \sum_{i=1}^{m-1} P(X_{i:n} \leq x < X_{m:n}) - \sum_{i=m+1}^{n} F_{i:n}(x), \tag{4}$$

which follows directly from

$$\sum_{i=1}^{n} F_{i:n} = \sum_{i=1}^{n} F_i.$$

(b) The proof is constructive. A closer analysis of (4) shows that $F_{m:n} = F^\star$ is possible iff

$$P(X_{1:n} = X_{m:n} \leq x^\star \leq X_{m+1:n}) = 1.$$

This property implies (2) and (3), and is used in the construction. First, we take an $F^\star$-distributed random variable $X^\star \leq x^\star$ and put $X_i = X^\star$ for exactly m variables among $X_1, \dots, X_n$. The choice of indices is random with the distribution depending on the value of $X^\star$. Assumptions (2) and (3) enable us to make m coordinates equal to $X^\star$ and to preserve marginal distributions. The remaining X_i are distributed on $[x^\star, +\infty)$ according to the respective conditional F_i. ∎

COROLLARY 1. (cf. Lai and Robbins (1976)). *For any distribution functions* $F_1, \dots, F_n$, *there exist random variables* $X_1, \dots, X_n$ *with marginals* $F_1, \dots,$

F_n, *respectively, such that*

$$F_{1:n} = \min\left\{\sum_{i=1}^{n} F_i,\ 1\right\}.$$

Corollary 2. (Caraux and Gascuel (1992), Rychlik (1992)). *For every distribution function F and $m \le n$, there exist identically F-distributed random variables $X_1, \ldots, X_n$ such that*

$$F_{m:n} = \min\left\{\frac{n}{m}F,\ 1\right\}.$$

There are marginals for which stochastically smallest order statistics exist and the bounds are tighter than (1) (e.g., a deterministic sample). There are also marginals such that stochastically extreme order statistics do not exist. For instance, take X_1, X_2 and X_3 with the following marginal distributions:

$$\begin{aligned}
&P(X_1 = 1) = P(X_1 = 3) = \frac{1}{2},\\
&P(X_2 = 2) = 1,\\
&P(X_3 = 1) = P(X_3 = 3) = \frac{1}{2}
\end{aligned}$$

(see Rychlik (1995)). Then the joint distribution depends on a single parameter $p = P(X_1 = X_3 = 1) \le \frac{1}{2}$ and $X_{2:3} = 1, 2$ and 3 with probabilities $p, 1-2p$ and p, respectively.

From now on we assume that the X_i are identically distributed.

3. Bounds on Expectations and Variances of Functions of Order Statistics. This section contains a summary of Rychlik (1994).

Theorem 2. *There exist identically distributed random variables $X_1, \ldots, X_n$ such that X_1 and $X_{m:n}$ have distribution functions F and $F_{m:n}$, respectively, iff*

$$\max\left\{0,\ \frac{nF - m + 1}{n - m + 1}\right\} \le F_{m:n} \le \min\left\{\frac{n}{m}F,\ 1\right\} \tag{5}$$

and

$$\frac{dF_{m:n}}{dF} \le n \qquad F - a.s. \tag{6}$$

Idea of the Proof. Inequalities (5) are special cases of (1) and the analogous lower bound. That (6) is also necessary follows from

$$F_{m:n}(x) - F_{m:n}(y) \le \sum_{i=1}^{n} P(y < X_i = X_{i:n} \le x) \le n[F(x) - F(y)].$$

Assuming (5) and (6), we construct $X_1, \ldots, X_n$ as follows. Define two distribution functions

$$G_1 = \max\left\{0, \frac{nF - F_{m:n} - m + 1}{m - n}\right\},$$
$$G_2 = \min\left\{\frac{nF - F_{m:n}}{m - 1}, 1\right\}.$$

Then $G_1 \le F_{m:n} \le G_2$. Put $X_{i:n} = G_2^{-1}(U)$, $F_{m:n}^{-1}(U)$ or $G_1^{-1}(U)$ for i less than, equal to, or greater than m, respectively, where U is a random variable uniformly distributed on $[0,1]$. Finally, set X_i by a random ordering of $X_{i:n}$, $i = 1, \ldots, n$, so that all orderings are equally probable. ∎

As a consequence of Theorem 2, we can replace the problems of finding extremes of the expectation and variance of an arbitrary function h of a given order statistic by finding extremes of the following functionals

$$\int_{-\infty}^{+\infty} h(x)\, dF_{m:n}(x), \tag{7}$$

$$\inf_{c \in R} \int_{-\infty}^{+\infty} (h(x) - c)^2\, dF_{m:n}(x), \tag{8}$$

respectively, with respect to $F_{m:n} \in \mathcal{F}_{m:n}$, which denotes the family of distribution functions determined by (5) and (6). These latter problems can be further simplified once we prove that the extremes are attainable on the set of extreme points of $\mathcal{F}_{m:n}$, which have a particularly simple form for continuous marginal distribution functions F.

THEOREM 3. *The extreme values of* (7) *and* (8) *over $\mathcal{F}_{m:n}$ are attained at extreme points of $\mathcal{F}_{m:n}$.*

We now describe the extreme points. For a given $F_{m:n} \in \mathcal{F}_{m:n}$, let

$$A = \left\{x \in R:\ \max\left\{0, \frac{nF(x) - m + 1}{n - m + 1}\right\} < F_{m:n}(x) < \min\left\{\frac{n}{m}F, 1\right\}\right\},$$

and define A^- similarly, using the left-continuous versions of the functions. Then $A \cap A^-$ is an open set, i.e., a possibly empty and at most countable union of disjoint open intervals, $A \cap A^- = \bigcup_{j \in J}(a_j, b_j)$, $J \subset N$, and $A \triangle A^- \subset \{a_j, b_j : j \in J\}$.

THEOREM 4. *A distribution function $F_{m:n}$ is an extreme point of $\mathcal{F}_{m:n}$ iff the measure of every set*

$$[a_j, b_j] \cap (A \cup A^-) \cap \left\{0 < \frac{dF_{m:n}}{dF} < n\right\}, \qquad j \in J,$$

with respect to F either equals 0 or is concentrated at a single point.

COROLLARY 3. *If F is continuous, then $F_{m:n}$ is an extreme point of $\mathcal{F}_{m:n}$ iff*

$$\frac{dF_{m:n}}{dF} = 0 \text{ or } n \quad \text{on } A \quad F - a.s.$$

IDEA OF THE PROOF OF THEOREMS 3 AND 4. In the proof of Theorem 3 we use the fact that the set of probability densities $\left\{\frac{dF_{m:n}}{dF} : F_{m:n} \in \mathcal{F}_{m:n}\right\}$ is compact in the weak* topology of the class of F-essentially bounded functions and so are the subsets on which convex functionals are maximal. The extreme points of the subsets satisfy the statement. For the supremum of the variance we also need a version of the minimax theorem. In the proof of Theorem 4 we merely apply the definitions of $\mathcal{F}_{m:n}$ and extreme points. ∎

Usually, given a fixed function h, we are in position to guess immediately the shapes of the extreme densities that yield the extremes of (7) and (8). The precise determination of solutions may pose only numerical problems. Some applications of Theorems 3 and 4 are given in Rychlik (1994).

4. Bounds on Expectations of L-Estimates. Linear combinations are the most popular functions of order statistics. Many examples and applications are presented by Balakrishnan and Cohen (1991). Asymptotic expansions for the case of independent identically distributed random variables are given in Helmers (1982). Dropping the assumpion of independence, we have

THEOREM 5. (Rychlik (1993a)). *If $X_1, \ldots, X_n$ have a common distribution function F, then*

$$\mathrm{E} \sum_{i=1}^{n} c_i X_{i:n} \le \int_0^1 F^{-1}(x)\, dC(x), \tag{8}$$

where C is the greatest convex function such that

$$C(0) = 0 \quad \text{and} \quad C\left(\frac{j}{n}\right) \le \sum_{i=1}^{j} c_i, \qquad j = 1, \ldots, n.$$

Inequality (9) is best-possible.

IDEA OF THE PROOF. First, we show that $F_{1:n}, \ldots, F_{n:n}$ are the distribution functions of $X_{1:n}, \ldots, X_{n:n}$, respectively, iff conditions

$$\sum_{i=1}^{n} F_{i:n} = nF \text{ and } F_{i:n} \ge F_{i+1:n}, \qquad i = 1, \ldots, n-1,$$

hold, which are evidently necessary. The sufficiency is proved by the construction:

$$X_j = F_{i:n}^{-1}(U) \text{ with probability } \frac{1}{n}, \qquad i,j = 1,\dots,n, \tag{10}$$

where U is uniformly distributed on $[0,1]$. It follows that

$$\begin{aligned} \mathrm{E}\sum_{i=1}^{n} c_i X_{i:n} &= \int_{-\infty}^{+\infty} x\, d\left(\sum_{i=1}^{n} c_i F_{i:n}(x)\right) \\ &= \int_0^1 F^{-1}(x)\, d\left(\sum_{i=1}^{n} c_i G_i(x)\right), \end{aligned}$$

where

$$\sum_{i=1}^{n} G_i(x) = nx \quad \text{and} \quad 1 \geq G_i(x) \geq G_{i+1}(x) \geq 0, \tag{11}$$

$i = 1,\dots,n-1$, $x \in [0,1]$. We replace the maximization of the expectation by the pointwise minimizations of $\sum_{i=1}^{n} c_i G_i(x)$, $x \in [0,1]$, with constraints (11). These are linear programming problems, whose solutions $G_i^\star(x)$ satisfy

$$\sum_{i=1}^{n} c_i G_i^\star(x) = C(x),$$

and so (9) holds. Moreover, since $x \mapsto G_i^\star(x)$, $i = 1,\dots,n$, are continuous distribution functions on $[0,1]$, we can let $F_{i:n} = G_i^\star \circ F$ in (10) in order to construct random variables attaining the equality in (9). ∎

Calculating the right-hand side of (9) is a simple matter, because C is a piecewise linear function and the integral can thus be split into a linear combination of Lebesgue integrals.

Theorem 5 has an application to the robust analysis of L-estimates against dependence of observations. It was shown in Rychlik (1993b) that the sample mean is the most bias-robust L-estimate of location, whereas, for instance, the sample median is very sensitive and, for some models, it is the most sensitive one in a reasonable class of L-estimates. This contrasts sharply with the classical theory of robust estimation (see Huber (1981)), where only the marginal distributions are violated, and with the common opinion of practitioners.

Another application of Theorem 5, stemming from ideas of Arnold (1980), (1985), is the determination of bounds on the expectation of L-estimates under various moment conditions. To illustrate the method, we assume that X_i are identically distributed, with expectation μ and variance σ^2. Then, by (9) and the Schwarz inequality, we obtain

$$\begin{aligned}
\mathrm{E}\sum_{i=1}^{n} c_i(X_{i:n}-\mu) &\leq \int_0^1 C'(x)[F^{-1}(x)-\mu]\,dx \\
&= \int_0^1 [C'(x)-c][F^{-1}(x)-\mu]\,dx \\
&\leq \left(\int_0^1 [C'(x)-c]^2\,dx\right)^{\frac{1}{2}} \left(\int_0^1 [F^{-1}(x)-\mu]^2\right)^{\frac{1}{2}} \\
&= K\sigma,
\end{aligned}$$

where C' denotes the derivative of C, and $c = \sum_{i=1}^{n} c_i$, and $K^2 = \int_0^1 [C'(x) - c]^2\,dx$. Here c is chosen so to minimize the integral. Note that equality holds iff $F^{-1} - \mu$ and $C' - c$ are proportional, which, together with the variance condition, imply

$$P\left(X_j = x_j = \mu + \frac{\sigma}{K}[C'(\frac{j}{n}) - c]\right) = \frac{1}{n}, \qquad j = 1, \ldots, n. \tag{12}$$

The analogous sharp deterministic inequality

$$\sum_{i=1}^{n} c_i(x_{i:n} - \bar{x}) \leq K\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{\frac{1}{2}} \tag{13}$$

for arbitrary numbers $x_1, \ldots, x_n$ can be deduced by the following reasoning. The urn model with n balls, labelled $x_1, \ldots, x_n$, and exhaustive random sampling without replacement generates n identically distributed random variables with expectation $\bar{x}$ and variance $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Therefore (13) holds and, moreover, it becomes equality if we take the x_j's from (12).

Rychlik (1993c) proved several sharp bounds on the expectation of L-estimates in terms of various parameters of location and dispersion of the marginal distribution as well as their deterministic counterparts. This approach, based on metric projections onto convex sets, enables us to determine tight bounds for monotone sequences and functions and more general inequalities. The results will be published elsewhere.

References

ARNOLD, B. C. (1980). Distribution-free bounds on the mean of the maximum of a dependent sample. *SIAM J. Appl. Math.* **38**, 163–167.

ARNOLD, B. C. (1985). p-Norm bounds on the expectation of the maximum of possibly dependent sample. *J. Multivar. Anal.* **17**, 316–332.

Balakrishnan, N. (1992). Relations between single moments of order statistics from non-identically distributed variables. In: *Order Statistics and Nonparametrics* (Sen, P. K., and Salama, I. A., eds.), North-Holland, Amsterdam, pp. 65–78.

Balakrishnan, N., Bendre, S. M. and Malik, H. J. (1992). General relations and identities for order statistics from non-independent non-identical variables. *Ann. Inst. Statist. Math.* **44**, 177–183.

Balakrishnan, N. and Cohen, A. C. (1991). *Order Statistics and Inference*, Academic Press, London.

Balkema, A. A. and Haan, L. de (1978). Limit distributions for order statistics I, II. *Theory Probab. Appl.* **23**, 77–92, 341–358.

Caraux, G. and Gascuel, O. (1992). Bounds on distribution functions of order statistics for dependent variates. *Statist. Probab. Lett.* **14**, 103–105.

David, H. A. (1981). *Order Statistics*, 2nd ed., J. Wiley, New York.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sc.* **4**, 53–84.

Helmers, R. (1982). Edgeworth Expansions for Linear Combinations of Order Statistics. *Math. Centre Tracts* **105**, Amsterdam.

Huber, P. J. (1981). *Robust Statistics*, J. Wiley, New York.

Lai, T. L. and Robbins, H. (1976). Maximally dependent random variables. *Proc. Nat. Acad. Sci. U.S.A.* **73**, 286–288.

Leadbetter, M. R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York.

Mallows, C. L. (1969). Extrema of expectations of uniform order statistics. *SIAM Rev.* **11**, 410–411.

Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics*, Springer-Verlag, New York.

Rychlik, T. (1992). Stochastically extremal distributions of order statistics for dependent samples. *Statist. Probab. Lett.* **13**, 337–341.

Rychlik, T. (1993a). Bounds for expectation of L-estimates for dependent samples. *Statistics* **24**, 1–7.

Rychlik, T. (1993b). Bias-robustness of L-estimates of location against dependence. *Statistics* **24**, 9–15.

Rychlik T. (1993c). Sharp bounds on L-estimates and their expectations for dependent samples. *Commun. Statist. – Theor. Meth.* **22**, 1053–1068.

RYCHLIK, T. (1994). Distributions and expectations of order statistics for possibly dependent random variables. *J. Mult. Anal.* **48**, 31–42.

RYCHLIK, T.(1995). Bounds for order statistics based on dependent variables with given nonidentical distributions. *Statist. Probab. Lett.* **23**, 351–358

TCHEN, A. (1980). Inequalities for distributions with given marginals. *Ann. Probab.* **8**, 814–827.

INSTITUTE OF MATHEMATICS
POLISH ACADEMY OF SCIENCES
CHOPINA 12
87100 TORUN, POLAND