ELEMENTARY PROOFS OF SOME BASIC FACTS CONCERNING ORDER STATISTICS

By

G. HAJÓS (Budapest), member of the Academy, and A. RÉNYI (Budapest), corresponding member of the Academy

Let $\xi_1, \ldots, \xi_n$ denote a sample of size n from a population with the distribution function $F(x)$. By other words, $\xi_1, \ldots, \xi_n$ are mutually independent random variables with the common distribution function $F(x)$. Let $\xi_1^*, \ldots, \xi_n^*$ be the same set of variables, rearranged in increasing order of magnitude, i. e.

$$\xi_k^* = R_k(\xi_1, \ldots, \xi_n),$$

where $R_k(x_1, \ldots, x_n)$ denotes the k'th term of the sequence obtained by rearranging the numbers $x_1, \ldots, x_n$ in increasing order of magnitude.

The present paper deals with the *order statistic* ξ_k^*. Some basic facts will be proved by simple methods. We aim expressively to avoid the calculus and at reduction of any calculation to possibly minimal extent. As consequence, our results may be easily checked by calculation in various different ways, which we are not intended to mention.

Our results seem us mostly to be known, though we did not find some of them explicitly in the literature. We endeavoured to give an elementary and systematic treatment of our subject. Accordingly, our paper may be of methodical interest. As to the literature we refer to the bibliography compiled by S. S. WILKS [3] and by the second named author [4].

1. In order to obtain distribution-free results, i. e. results independent of the distribution function $F(x)$, we introduce $\eta_k = F(\xi_k)$ $(k = 1, \ldots, n)$. If we suppose that $y = F(x)$ is strictly increasing and continuous, the same holds for the inverse function $x = F^{-1}(y)$ and we have[1]

$$\mathbf{P}(\eta_k < x) = \mathbf{P}(\xi_k < F^{-1}(x)) = F(F^{-1}(x)) = x \qquad (0 \leqq x \leqq 1),$$

what shows that the variables $\eta_1, \ldots, \eta_n$ are *uniformly distributed* in the interval (0, 1). Putting $\eta_k^* = F(\xi_k^*)$ $(k = 1, \ldots, n)$ we have

$$\eta_k^* = F(\xi_k^*) = F(R_k(\xi_1, \ldots, \xi_n)) = R_k(F(\xi_1), \ldots, F(\xi_n)) = R_k(\eta_1, \ldots, \eta_n).$$

Consequently $\eta_1^*, \ldots \eta_n^*$ are order statistics of a sample of size n from a population of uniform distribution in (0, 1).

[1] $\mathbf{P}(A)$ denotes the probability of the event A.

Accordingly, we confine ourselves in what succeeds to the research of the order statistics η_k^*. Our results may be interpreted as facts concerning the original order statistics ξ_k^*.

2. The variables η_k^* are not independent, the relation $\eta_j^* \leqq \eta_k^*$ (in case $j < k$) contradicts the independency.

The joint density function of the variables $\eta_1^*, \ldots, \eta_n^*$ *is*

$$f(x_1, \ldots, x_n) = n! \qquad (0 \leqq x_1 \leqq \cdots \leqq x_n \leqq 1). \tag{1}$$

As a matter of fact, if E denotes any measurable subset ot the n-dimensional simplex defined by the inequalities $0 \leqq x_1 \leqq \cdots \leqq x_n \leqq 1$, we have

$$\mathbf{P}((\eta_1^*, \ldots, \eta_n^*) \in E) = \sum \mathbf{P}((\eta_{i_1}, \ldots, \eta_{i_n}) \in E),$$

where the summation is extended over all permutations $i_1, \ldots, i_n$ of the indices $1, \ldots, n$ and the density function of $(\eta_{i_1}, \ldots, \eta_{i_n})$ is equal to 1 at any point $(x_1, \ldots, x_n)$ of the cube $0 \leqq x_k \leqq 1$ $(k = 1, \ldots, n)$.

Considering now the case that $\eta_k^* = c_k, \ldots, \eta_n^* = c_n$ $(2 \leqq k \leqq n)$ are fixed, we state that $\eta_1^*, \ldots, \eta_{k-1}^*$ *are order statistics of a sample of size* $k-1$ *from a population of uniform distribution in the interval* $(0, c_k)$. In fact, this is true if $\eta_1^*, \ldots, \eta_{k-1}^*$ are furnished by any given $k-1$ variables out of $\eta_1, \ldots, \eta_n$, since these $k-1$ variables are uniformly distributed, even within the cube $0 \leqq x_i \leqq c_k$ $(i = 1, \ldots, k-1)$. Thus, by (1), *the joint density function of the variables* $\eta_1^*, \ldots, \eta_{k-1}^*$, *under condition* $\eta_k^* = c_k, \ldots, \eta_n^* = c_n$, *is*

$$f(x_1, \ldots, x_{k-1} | c_k, \ldots, c_n) = \frac{(k-1)!}{c_k^{k-1}} \qquad (0 \leqq x_1 \leqq \cdots \leqq x_{k-1} \leqq c_k). \tag{2}$$

By the same argument, *if* $\eta_1^* = c_1, \ldots, \eta_k^* = c_k$ $(1 \leqq k \leqq n-1)$ *are fixed,* $\eta_{k+1}^*, \ldots, \eta_n^*$ *are order statistics of a sample of size* $n-k$ *from a population of uniform distribution in the interval* $(c_k, 1)$ *and the joint density function of the variables* $\eta_{k+1}^*, \ldots, \eta_n^*$, *under condition* $\eta_1^* = c_1, \ldots, \eta_k^* = c_k$, *is*

$$f(x_{k+1}, \ldots, x_n | c_1, \ldots, c_k) = \frac{(n-k)!}{(1-c_k)^{n-k}} \qquad (c_k \leqq x_{k+1} \leqq \cdots \leqq x_n \leqq 1). \tag{3}$$

Since (2) and (3) depend only on c_k, our statements hold also under the only condition $\eta_k^* = c_k$, i. e. (2) and (3) give also the values of the functions $f(x_1, \ldots, x_{k+1} | c_k)$ and $f(x_{k+1}, \ldots, x_n | c_k)$, and the same holds under any restriction on the non-occurring variables. By the same argument, under condition $\eta_k^* = c_k$, the sets of variables $(\eta_1^*, \ldots, \eta_{k-1}^*)$ and $(\eta_{k+1}^*, \ldots, \eta_n^*)$ are independent. By other words, order statistics form a *Markov chain.*[2]

3. *The joint density function of the variables* $\eta_{i+1}^*, \ldots, \eta_k^*$ $(1 \leqq i < k \leqq n)$ *is*

$$f_{ik}(x_{i+1}, \ldots, x_k) = \frac{n!}{i!\,(n-k)!}\, x_{i+1}^i (1-x_k)^{n-1} \qquad (0 \leqq x_{i+1} \leqq \cdots \leqq x_k \leqq 1). \tag{4}$$

[2] A. N. KOLMOGOROFF [1] was the first to remark this.

Indeed, the joint density function of the variables $\eta_1^*, \ldots, \eta_i^*, \eta_{k+1}^*, \ldots, \eta_n^*$, under the condition

$$(C_1) \qquad \eta_{i+1}^* = x_{i+1}, \ldots, \eta_k^* = x_k,$$

is clearly given by

$$f(x_1, \ldots, x_i, x_{k+1}, \ldots, x_n | x_{i+1}, \ldots, x_k) = \frac{f(x_1, \ldots, x_n)}{f_{ik}(x_{i+1}, \ldots, x_k)}.$$

On the other hand, since the variables $\eta_1^*, \ldots, \eta_i^*$ are by the Markov chain property, under condition (C_1), independent of the variables $\eta_{k+1}^*, \ldots, \eta_n^*$, we have

$$f(x_1, \ldots, x_i, x_{k+1}, \ldots, x_n | x_{i+1}, \ldots, x_k) =$$
$$= f(x_1, \ldots, x_i | x_{i+1}, \ldots, x_n) \cdot f(x_{k+1}, \ldots, x_n | x_1, \ldots, x_k).$$

Taking into account (1), (2) and (3), comparison of both statements establishes (4).

By (5), taking the special case $i+1 = k$ *the density function of* η_k^* *is*

$$(5) \qquad f_k(x_k) = \frac{n!}{(k-1)!\,(n-k)!} x_k^{k-1}(1-x_k)^{n-k} = \beta_{nk}(x),$$

i. e. *order statistics* η_k^* *have Beta-distribution.*

The joint density function of any $\eta_{k_1}^*, \ldots, \eta_{k_r}^*$ $(1 \leqq k_1 < \cdots < k_r \leqq n)$ *variables is*

$$f_{k_1, \ldots, k_r}(x_1, \ldots, x_r) = C_{k_1, \ldots, k_r} x_1^{k_1-1}(x_2-x_1)^{k_2-k_1-1} \ldots (x_r-x_{r-1})^{k_r-k_{r-1}-1}(1-x_r)^{n-k_r},$$

where

$$C_{k_1, \ldots, k_r} = \frac{n!}{(k_1-1)!\,(k_2-k_1-1)! \ldots (k_r-k_{r-1}-1)!\,(n-k_r)!}$$

and

$$0 \leqq x_1 \leqq \ldots \leqq x_r \leqq 1.$$

The proof is given immediately by the above argument, if we consider instead of (C_1) the condition

$$(C_2) \qquad \eta_{k_1}^* = x_1, \ldots, \eta_{k_r}^* = x_r,$$

divide $(0, 1)$ by $x_1, \ldots, x_r$ into $r+1$ subintervals, and take into account that, by the Markov chain property, the sets of variables η_i^* lying in these subintervals are independent under condition (C_2).

4. We define $\eta_0^* = 0$, $\eta_{n+1}^* = 1$, and introduce the variables

$$\delta_k = \eta_k^* - \eta_{k+1}^* \qquad (k = 1, \ldots, n+1).$$

Since $\delta_1, \ldots, \delta_n$ are obtained from $\eta_1^*, \ldots, \eta_n^*$ by a measure-preserving linear transformation, their joint density functions are equal at corresponding places. By corresponding transformation $y_1 = x_1$, $y_k = x_k - x_{k-1}$ $(k = 2, \ldots, n)$ of (1) *the joint density function of the random variables* $\delta_1, \ldots, \delta_n$ *is*

$$(6) \qquad g(y_1, \ldots, y_n) = n! \qquad (y_1 \geqq 0, \ldots, y_n \geqq 0;\ y_1 + \cdots + y_n \leqq 1).$$

We conclude from (6) that the variables $\delta_1, \ldots, \delta_n$ have the same distribution. As the distribution of η_k is symmetric with respect to the point $\frac{1}{2}$, this symmetry holds also for the joint distribution of $(\eta_1^*, \ldots, \eta_n^*)$. Especially, $\delta_1 = \eta_1^*$ and $\delta_{n+1} = 1 - \eta_n^*$ have equal distributions. We draw the conclusion that the *variables* $\delta_1, \ldots, \delta_{n+1}$ *are equally distributed.*

Their *common density function* is that of η_1^*, given by (5),

$$\beta_{n1}(x) = n(1-x)^{n-1} \qquad (0 \leqq x \leqq 1).$$

Their *common mean value* is, because of $\delta_1 + \cdots + \delta_{n+1} = 1$,

$$\mathbf{M}(\delta_k) = \frac{1}{n+1} \qquad (k = 1, \ldots, n+1). \tag{7}$$

5. The variables $\delta_1, \ldots, \delta_{n+1}$ are not only equally distributed, but are also *equivalent variables,* i. e. their joint distribution remains invariant under any permutation of them. This is established by (6) for permutations of $\delta_1, \ldots, \delta_n$, by the above mentioned symmetry for the permutation $(\delta_1, \ldots, \delta_{n+1}) \to (\delta_{n+1}, \ldots, \delta_1)$, and by successive application of these for any permutation of $\delta_1, \ldots, \delta_{n+1}$.

In consequence, the distribution of the *difference*

$$\eta_{i+k}^* - \eta_i^* = \delta_{i+1} + \cdots + \delta_{i+k} \qquad (0 \leqq i < i+k \leqq n+1)$$

depends only on k, is therefore equal to that of η_k^*, has the density function (5), and, by (7), the mean value $\frac{k}{n+1}$.

Especially, the *range* $\Delta = \eta_n^* - \eta_1^*$ has the density function

$$\beta_{n,\,n-1}(x) = n(n-1)x^{n-2}(1-x) \qquad (0 \leqq x \leqq 1)$$

and the mean value

$$\mathbf{M}(\Delta) = \frac{n-1}{n+1}.$$

6. As previously stated, under condition $\eta_k^* = c_k, \ldots, \eta_n^* = c_n$ the variable η_i^* $(1 \leqq i < k \leqq n)$ is the i'th order statistic of a sample of size $k-1$ from a population of uniform distribution in the interval $(0, c_k)$. Hence, the distribution of the *quotient* $\frac{\eta_i^*}{\eta_k^*}$ does not depend even on c_k, remains therefore unaltered if $\eta_k^*, \ldots, \eta_n^*$ are not fixed, and has, by (6), the density function $\beta_{i+k-1,\,k}(x)$. Thus, both differences and quotients of order statistics η_k^* have Beta-distribution.

The variables $\frac{\eta_k^*}{\eta_{k+1}^*}$ $(k = 1, \ldots, n)$ are *mutually independent.* Indeed, by above statement, the distribution of $\frac{\eta_k^*}{\eta_{k+1}^*}$ does not depend on the values of $\eta_{k+1}^*, \ldots, \eta_n^*$, is therefore independent of $\frac{\eta_{k+1}^*}{\eta_{k+2}^*}, \ldots, \frac{\eta_{n-1}^*}{\eta_n^*}, \frac{\eta_n^*}{\eta_{n+1}^*} = \eta_n^*$.

Moreover the independent variables

$$\zeta_k^* = \left(\frac{\eta_k^*}{\eta_{k+1}^*}\right)^k \qquad (k \leqq 1, \ldots, n)$$

are uniformly distributed in the interval (0, 1).[3] As a matter of fact, the distribution function of η_n^* is

$$\mathbf{P}(\eta_n^* < x) = \prod_{k=1}^{n} \mathbf{P}(\eta_k < x) = x^n \qquad (0 \leqq x \leqq 1).$$

Correspondingly, since $\frac{\eta_k^*}{\eta_{k+1}^*}$ is the k'th order statistic of a sample of size k from a population of uniform distribution in (0, 1), we have

$$\mathbf{P}\left(\frac{\eta_k^*}{\eta_{k+1}^*} < x\right) = x^k.$$

Whence

$$\mathbf{P}(\zeta_k^* < x) = \mathbf{P}\left(\frac{\eta_k^*}{\eta_{k+1}^*} < \sqrt[k]{x}\right) = x,$$

establishing our statement.

7. We introduce the random variables

$$\vartheta_k = -\ln \zeta_k = -k \ln \frac{\eta_k^*}{\eta_{k+1}^*} \qquad (k = 1, \ldots, n). \tag{8}$$

These are, according to our preceding result, mutually independent and equally distributed, with the distribution function

$$\mathbf{P}(\vartheta_k < x) = \mathbf{P}(\zeta_k > e^{-x}) = 1 - e^{-x} \qquad (x \geqq 0),$$

i. e. have exponential distribution with mean value 1.

From equations (8) we get

$$\ln \eta_k^* = -\sum_{j=k}^{n} \frac{\vartheta_j}{j} \qquad (k = 1, \ldots, n). \tag{9}$$

Consequently, the *logarithms* of the order statistics η_k^* form not only a Markov chain, but also an *additive chain*, i. e. they are consecutive partial sums of a sequence of mutually independent random variables.

We expressed by (9) the order statistics η_k^* as simple functions of independent and equally distributed random variables. Starting from this fact, *limit theorems* on order statistics may be obtained, by means of the central limit theorem, in a simple and straightforward way. This has been shown by the second named author [4].

(Received 4 May 1954)

[3] This was proved by S. MALMQUIST [2]; his proof is rather complicated.

Bibliography

[1] A. N. KOLMOGOROFF, Sulla determinazione empirica di una legge di distribuzione, *Gior. d. Att.*, **4** (1933), pp. 83—91.

[2] S. MALMQUIST, On a property of order statistics from a rectangular distribution, *Skand. Aktuerietidsskrift*, **33** (1950), pp. 214—222.

[3] S. S. WILKS, Order statistics, *Bull. Amer. Math. Soc.*, **54** (1948), pp. 6—50.

[4] A. RÉNYI, On the theory of order statistics, *Acta Math. Acad. Sci. Hung.*, **4** (1953), pp. 191—231.

ЭЛЕМЕНТАРНОЕ ДОКАЗАТЕЛЬСТВО НЕКОТОРЫХ ОСНОВНЫХ ФАКТОВ ТЕОРИЙ ВАРИАЦИОННЫХ РЯДОВ

Г. ХАЙОШ и А. РЕНЬИ (Будапешт)

(Резюме)

В работе, которая имеет преимущественно методический характер, дается систематическое и элементарное изложение некоторых основных фактов теорий вариационных рядов. Пусть $\eta_1, \eta_2, \ldots, \eta_n$ — независимые случайные величины, которые равномерно распределены в интервале $(0, 1)$, пусть $\eta_1^* \leqq \eta_2^* \leqq \ldots \leqq \eta_n^*$ — те же величины, расположенные в возрастающем порядке. Доказывается, между прочим, очень просто, что величины η_k^* $(k = 1, 2, \ldots, n)$ образуют цепь Маркова (теорема А. Н. Колмогорова, см. [1]), более того, что величины $\ln \eta_k^*$ $(k = 1, 2, \ldots, n)$ образуют аддитивную цепь Маркова, так как случайные величины $\left(\frac{\eta_k^*}{\eta_{k+1}^*}\right)^k$ $(k = 1, 2, \ldots, n;\ \eta_{n+1}^* \equiv 1)$ независимы и равномерно распределены в интервале $(0, 1)$ (теорема С. Малмквиста, см. [2]). С помощью этих фактов возможно доказательство многих предельных теорем теорий вариационных рядов на центральную предельную теорему теории вероятностей (см. [4]).