ON THE POSITION OF THE SAMPLE MEAN AMONG THE ORDERED SAMPLE ELEMENTS

by

KÁROLY SARKADI, EDIT SCHNELL[1] and ISTVÁN VINCZE

Introduction

Let $\xi_1, \xi_2, \ldots \xi_n$ be independent and identically distributed random variables with continuous distribution function $F(x)$ and with density function $F'(x) = f(x)$. Let us arrange these variables according to their order of magnitude: $\xi_1^* < \xi_2^* < \ldots < \xi_n^*$ and denote their mean value by $\bar{\xi}$, i.e. $\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$.

In the following the probabilities

$$p_k = \mathbf{P}(\xi_{k-1}^* \leqq \bar{\xi} < \xi_k^*), \qquad k = 2, 3, \ldots n \tag{1}$$

are considered.

Although the examination of these probabilities lies close at hand, as far as we know this problem has not yet been treated in the literature[2].

A distribution-free solution cannot be expected. It is obvious that if $F(x)$ is symmetrical (i.e. $F(x) = 1 - F[\mathbf{M}(\xi) - x]$) the equality $p_k = p_{n-k+1}$ must be valid. We shall see in the following that for exponentially distributed variables this symmetry is not fulfilled.

The explicit determination of the probabilities (1) seems to be very complicated in the general case. The formula based on the joint distribution of $(\xi_1^*, \xi_2^*, \ldots \xi_n^*)$ is rather unmanageable; concerning the exact joint distribution of $(\bar{\xi}, \xi_k^*)$ only for normally distributed variables is a recursion formula known and even this is very complicated [4]. In § 1 we shall determine the probabilities (1) explicitly for the exponential case, making use of the additive Markovian property of the ordered sample elements in this case; in § 2 an alternative elementary proof is given. In § 3 for the general case the limiting joint distribution of $\bar{\xi}$ and ξ_k^* is derived under weak conditions and is shown to be a two dimensional Gaussian one. As a consequence of this the limiting Gaussian distribution for the problem will be given in § 4.

§ 1. Exponentially distributed random variables

1. Let the random variables $\xi_1, \xi_2, \ldots \xi_n$ be independent identically and exponentially distributed, i.e.

$$\mathbf{P}(\xi_i < x) = 1 - e^{-\lambda x} \qquad i = 1, 2, \ldots n \tag{1.1}$$

[1] Hungarian Central Statistical Office.

[2] Added in proof: Our paper was already in print as the paper of DAVID [10] appeared which contains some of our results but considers only normal distribution.

We wish to remark that from the identity

$$p_k = \mathbf{P}(\xi^*_{k-1} \leqq \bar{\xi} < \xi^*_k) = \mathbf{P}(\lambda\xi^*_{k-1} \leqq \lambda\bar{\xi} < \lambda\xi^*_k)$$

there follows the independence of the probabilities p_k from the value $\lambda > 0$ and thus it is sufficient to consider only the case $\lambda = 1$.

2. We shall make use of the following well known relation: (see e.g. [1] p. 232) if $\vartheta_1, \vartheta_2, \dots \vartheta_k$ are independent, exponentially distributed random variables with parameters $\lambda_1, \lambda_2, \dots \lambda_k$ then the density function $g_k(t)$ of the variable $\vartheta = \sum_{i=1}^{k} \vartheta_i$ is equal to

(1.2)

$$g_k(t) = (-1)^{k-1}\lambda_1\lambda_2\dots\lambda_k \sum_{i=1}^{k} \frac{e^{-\lambda_i t}}{(\lambda_i-\lambda_1)\dots(\lambda_i-\lambda_{i-1})(\lambda_i-\lambda_{i+1})\dots(\lambda_i-\lambda_k)}\,..$$

If $\lambda_1 = \lambda_2 = \dots = \lambda_k = \lambda$ then we obtain from (1.2) the following well known formula:

$$(1.3)\qquad g_k(t) = \frac{\lambda^k t^{k-1}}{(k-1)!}e^{-\lambda t}.$$

3. We shall apply the following well known theorem: if $\xi_1, \xi_2, \dots, \xi_n$ are independent, identically and exponentially distributed random variables with the parameter $\lambda = 1$ and $\xi^*_1 < \xi^*_2 < \dots < \xi^*_n$ ($\xi^*_0 = 0$) the set of the same variables rearranged in increasing order of magnitude, then the increments $\delta_k = \xi^*_k - \xi^*_{k-1}$ are independent and exponentially distributed with distribution functions

$$F(x) = \mathbf{P}(\delta_k < x) = 1 - e^{-(n-k+1)x}, \qquad x \geqq 0,\ k = 1, 2, \dots n$$

(see for example [1] p. 585).

4. Let us turn now to the determination of probabilities:

$$P_k = p_1 + p_2 + \dots + p_k = \mathbf{P}(\bar{\xi} < \xi^*_k), \qquad k = 1, 2, \dots n$$

where $\xi_1, \xi_2, \dots \xi_n$ are independent exponentially distributed random variables with parameters λ_i $(i = 1, 2, \dots n)$.

With our notation the following relations hold:

$$\xi^*_k = \delta_1 + \delta_2 + \dots + \delta_k$$

$$\bar{\xi} = \frac{1}{n}\sum_{k=1}^{n}\xi_k = \frac{1}{n}\sum_{k=1}^{n}\xi^*_k = \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{k}\delta_i\right) =$$

$$= \frac{1}{n}\sum_{k=1}^{n}(n-k+1)\,\delta_k = \sum_{k=1}^{n}\left(1 - \frac{k-1}{n}\right)\delta_k.$$

Thus

$$
\begin{aligned}
P_k = \mathbf{P}(\bar{\xi} < \xi_k^*) &= \mathbf{P}\left(\sum_{i=1}^{n}\left(1-\frac{i-1}{n}\right)\delta_i < \sum_{i=1}^{k}\delta_i\right) = \\
&= \mathbf{P}\left(\sum_{i=k+1}^{n}\left(1-\frac{i-1}{n}\right)\delta_i < \sum_{j=2}^{k}\frac{j-1}{n}\delta_j\right).
\end{aligned}
\tag{1.4}
$$

Let us introduce the following new variables:

$$\eta_i = \left(1-\frac{i-1}{n}\right)\delta_i, \quad \vartheta_j = j\delta_{j+1}$$

for the distribution function of which

$$\mathbf{P}(\eta_i < x) = 1 - e^{-nx} \text{ resp. } \mathbf{P}(\vartheta_j < y) = 1 - e^{-\left(\frac{n}{j}-1\right)y}$$

is valid.

In consequence of the independence of the variables η_i, the density function $\gamma_{n-k}(x)$ of $\eta = \eta_{k+1} + \eta_{k+2} + \ldots + \eta_n$ is the following:

$$\gamma_{n-k}(x) = \frac{n^{n-k}}{(n-k-1)!}x^{n-k-1}e^{-nx},$$

further — because of (1.2) the density function $g_k(x)$ of the variable $\vartheta = \frac{1}{n}\sum_{i=1}^{k-1}\vartheta_i$ equals

$$g_k(x) = n\binom{n-1}{k-1}e^{nx}\sum_{r=1}^{k-1}(-1)^{r-1}\binom{k-1}{r}\left(-\frac{r}{n}\right)^{k-2}e^{-\frac{n^2}{r}x}.$$

Consequently our formula (1.4) may be written — because of the idependence of η and ϑ — as follows:

$$
\begin{aligned}
P_k = \mathbf{P}(\eta < \vartheta) &= \int_0^\infty\left(\int_x^\infty g_k(u)\,du\right)\gamma_{n-k}(x)\,dx = \\
&= \frac{n^{n-k+1}}{(n-k-1)!}\binom{n-1}{k-1}\sum_{r=1}^{k-1}(-1)^{r-1}\binom{k-1}{r}\times \\
&\times\left(-\frac{r}{n}\right)^{k-2}\int_0^\infty\left(\int_x^\infty e^{-\left(\frac{n^2}{r}-n\right)u}\,du\right)x^{n-k-1}e^{-nx}\,dx.
\end{aligned}
$$

Carrying out the integration and using a slight modification — we obtain

$$P_k = \frac{(-1)^k}{n^{n-1}} \binom{n}{k} \sum_{r=1}^{k-1} (-1)^{r-1} r^{n-1} \frac{k-r}{n-r} \binom{k}{r} \qquad k = 2, 3, \ldots n . \tag{1.5}$$

Thus we have determined the probabilities mentioned in (1.4) for the exponential case.

Obviously P_n must be equal to 1 ($\bar{\xi} < \xi_n^*$ being a certain event). If in (1.5) k is replaced by n we obtain

$$P_n = \frac{(-1)^n}{n^{n-1}} \sum_{r=1}^{n-1} (-1)^{r-1} r^{n-1} \binom{n}{r} = \frac{1}{n^{n-1}} \sum_{r=1}^{n-1} (-1)^{n-1+r} \binom{n}{r} r^{n-1} .$$

N. H. Abel has proved the identity (see [2])

$$(b-n) \sum_{r=0}^{n-1} \binom{n}{r} (a+r)^r (b-r)^{n-r-1} = (a+b)^n ,$$

which holds for every real value of a, b and for every integer $n \geqq 0$.

Let us apply Abel's identity for $a = b = 0$ and take into consideration that for $r = 0$ the left side of the equality equals 0, accordingly:

$$-n \sum_{r=1}^{n-1} \binom{n}{r} r^{n-1} (-1)^{n-r-1} + n^n = 0 ;$$

from this

$$\sum_{r=1}^{n-1} (-1)^{n-r-1} \binom{n}{r} r^{n-1} = n^{n-1} .$$

Obviously $(-1)^{n-r-1} = (-1)^{n-r+1}$ and thus it is verified that our formula gives $P_n = 1$.

5. For the probabilities $p_k = P_k - P_{k-1}$ from (1.5) the following formulae are obtained:

$$p_k = \frac{(-1)^k}{n^{n-1}} \binom{n}{k-1} \sum_{r=1}^{k-1} (-1)^{r-1} r^{n-1} \binom{k-1}{r} \tag{1.6}$$

or

$$p_k = \frac{\binom{n}{k-1} (k-1)!}{n^{n-1}} \mathfrak{S}_{n-1}^{k-1} , \tag{1.7}$$

where the $\mathfrak{S}_r^k$ denote the so called Stirling numbers of the second kind (see e.g. [3] p. 168—181; tabulated in [5]).

The above distribution is a special case of the occupancy distribution. It is known that putting randomly r objects into n cells, the probability that k cells will be occupied and $n - k$ will be empty is

$$\frac{\binom{n}{k} k!}{n^r} \mathfrak{S}_r^k$$

(see e.g. [3] p. 178). Thus (1.7) gives the probability that putting $n-1$ objects into n cells, $k-1$ cells will be occupied and $n-k+1$ will be empty. Confidence limits for the occupancy distribution are tabulated in [9].

6. From formula (1.5) we obtain for every n:

$$P_2 = \frac{1}{n^{n-2}} \qquad P_3 = \frac{1}{n^{n-2}}\,[2^{n-2}(n-1)-(n-2)]$$

$$P_4 = \frac{1}{n^{n-2}}\,\frac{1}{2}\,[3^{n-2}(n-1)(n-2) - 2^{n-1}(n-1)(n-3) + (n-2)(n-3)]\,.$$

7. From above formulae in the case $n=3$ we obtain

$$p_2 = \frac{1}{3}, \qquad p_3 = \frac{2}{3} \qquad (p_1 = 0)$$

and in the case $n=4$

$$p_2 = \frac{1}{16}, \qquad p_3 = \frac{9}{16}, \qquad p_4 = \frac{6}{16} \qquad (p_1 = 0)\,.$$

§ 2. An alternative proof

1. We now turn to an other proof of (1.7) which is based on combinatorial models.

It will be shown that the determination of the probabilities $p_k = \mathbf{P}(\xi^*_{k-1} \leqq \bar{\xi} < \xi^*_k)$ can be reduced in case of exponential parent distribution (Model A) to the above mentioned occupancy problem (Model B).

2. First of all we will show that Model A is equivalent with the following Model C: let us divide the interval [0,1) by $n-1$ mutually independent variates distributed uniformly in the interval [0,1). We consider the probability of exactly $k-1$ random intervals having lengths $\leqq \frac{1}{n}$. In the following it is shown that this probability equals

$$p_k = \mathbf{P}(\xi^*_{k-1} \leqq \bar{\xi} < \xi^*_k)\,.$$

For this purpose it suffices to verify that the joint distribution of the quantities $\xi_i / \sum\limits_{j=1}^{n} \xi_j$ $(i = 1, 2, \ldots n)$ is identical with that of the random intervals obtained by above procedure. These two random vectors have the same set of possible values: each component must be nonnegative and the sum of all components equals 1. The latter vector has a uniform distribution within this set. We show that the same holds for the former vector variate as well. It is clear from the form of the density function of the vector $\{\xi_i\}$ that the conditional distribution of $\xi_i / \sum\limits_{j=1}^{n} \xi_j$ with respect to $\sum\limits_{j=1}^{n} \xi_j$ is uniform within the set of the possible values. As this set (and thus the distribution) does not depend on the actual value of $\sum\limits_{j=1}^{n} \xi_j$ the unconditional distribution of $\xi_i / \sum\limits_{j=1}^{n} \xi_j$ agrees with the conditional one. This proves the equivalence of Models A and C.

3. Evidently Model B can be formulated in the following way: The interval $[0,1)$ is divided into n intervals of length $1/n$. We are interested in the probability that exactly $k-1$ of these intervals will contain at least one of $n-1$ random points, independently and uniformly distributed in the whole interval.

It must be shown only that Model C is equivalent with Model B. For this purpose we give a transformation, mapping Model B into Model C.

The transformation between the values $0 \leqq \eta_1 \leqq \eta_2 \leqq \ldots \leqq \eta_{n-1} < 1$ and $0 \leqq \eta_1' \leqq \eta_2' \leqq \ldots \leqq \eta_{n-1}' < 1$ given below will have the following properties:

a) it will be 1 : 1,

b) if $\{\eta_1, \eta_2, \ldots, \eta_{n-1}\}$ is distributed according to the order statistics of a sample of size $n-1$ from a $[0,1)$ uniform distribution then $\{\eta_1', \eta_2', \ldots, \eta_{n-1}'\}$ follows the same law,

c) if the interval $\left[\dfrac{i-1}{n}, \dfrac{i}{n}\right)$ contains at least one of the values $\eta_1, \eta_2, \ldots, \eta_n$, then $\eta_1' - \eta_{i-1}' \leqq 1/n$; otherwise $\eta_i' - \eta_{i-1}' > 1/n$ $(i = 1, 2, \ldots, n-1;$ $\eta_0' = 0,\ \eta_n' = 1)$.

Evidently above conditions assure the equivalence of Models B and C.

The transformation is as follows:

Let be

$$\zeta_i = n\,\eta_i - [n\,\eta_i] \qquad (i = 1, 2, \ldots, n-1) \tag{2.1}$$

where $[x]$ denotes the greatest integer $\leqq x$. Let be further

$$\beta_1 = 0$$

$$\beta_i = \begin{cases} \beta_{i-1} & \text{if } \ \zeta_i \geqq \zeta_{i-1} \\ \beta_{i-1} + 1 & \text{if } \ \zeta_i < \zeta_{i-1} \end{cases} \qquad (i = 2, 3, \ldots, n-1) \tag{2.2}$$

and

$$\alpha_i = [n\,\eta_i] - \beta_i \qquad (i = 1, 2, \ldots, n-1)\,. \tag{2.3}$$

Evidently the sequences $\{\alpha_i\}$ and $\{\beta_i\}$ are monotonically increasing,

$$0 \leqq \beta_i + \zeta_i - \beta_{i-1} - \zeta_{i-1} < 1 \qquad (i = 2, 3, \ldots, n-1) \tag{2.4}$$

and

$$\eta_i = \frac{\alpha_i + \beta_i + \zeta_i}{n} \qquad (i = 1, 2, \ldots, n-1)\,. \tag{2.5}$$

Let us denote by $\gamma_1 < \gamma_2 < \ldots < \gamma_s$ the indices of those β_i $(i=2, \ldots, n-1)$ for which

$$\beta_{\gamma_j} = \beta_{\gamma_j - 1} \tag{2.6}$$

and by $\delta_1 < \delta_2 < \ldots \delta_{n-k}$ the indices of those γ_i $(i = 1, 2, \ldots, s)$ for which

$$\alpha_{\gamma_{\delta_j}} = \alpha_{\gamma_{\delta_j} - 1}\,. \tag{2.7}$$

Here the symbol k is used according to its former interpretation. It follows namely that $\alpha_j + \beta_j = \alpha_{j-1} + \beta_{j-1}$ if and only if $j = \gamma_{\delta_i}$ for some i. This means that $\alpha_j + \beta_j$ takes on $k - 1$ different values, in other words, $k - 1$ is the number of "occupied" intervals, i.e. the intervals $\left[\frac{i-1}{n}, \frac{i}{n}\right)$ $(1 \leqq i \leqq n)$ each of which contain at least one of the points η_j.

We denote these $k - 1$ values of $\alpha_j + \beta_j$ by

$$\bar{\varepsilon}_1 < \bar{\varepsilon}_2 < \ldots \bar{\varepsilon}_{k-1} \tag{2.8}$$

whereas the remainder members of the sequence $0, 1, \ldots, n - 1$ are denoted by

$$\varepsilon_1 < \varepsilon_2 < \ldots < \varepsilon_{n-k+1} \cdot$$

Let α'_i be defined by the relations

$$\begin{aligned} \alpha'_{\varepsilon_i+1} &= \delta_i && (i = 1, 2, \ldots, n - k + 1) \\ \alpha'_{\bar{\varepsilon}_i+1} &= \alpha'_{\bar{\varepsilon}_i} && (i = 1, 2, \ldots, k - 1) \end{aligned} \tag{2.9}$$

where $\alpha'_0 = 0$, $\delta_{n-k+i} = s + 1$. (Note that $\delta_{n-k} \leqq s$ by definition).

Let us define now the transformed values

$$\eta'_i = \frac{\alpha'_i + \beta_i + \zeta_i}{n} . \tag{2.10}$$

It follows from (2.9) that $\alpha'_{n-1} \leqq s + 1$ and from (2.2) and (2.6) that

$$\beta_{n-1} = n - 2 - s . \tag{2.11}$$

This and (2.4) assure the fulfilment of the condition $0 \leqq \eta'_1 \leqq \eta'_2 \leqq \leqq \ldots \leqq \eta'_{n-1} < 1$.

We now go over to the proof that the transformation (2.10) is one by one, i.e. the quantities $\{\eta_i\}$ can be also uniquely determined from the quantities $\{\eta'_i\}$. In fact, if $0 \leqq \eta'_1 \leqq \eta'_2 \leqq \ldots \leqq \eta'_{n-1} < 1$ are given, the sequence $\{\eta_i\}$ can be uniquely determined in the following way: the sequences $\{ {}'_1 \}$, $\{\beta_i\}$, $\{\zeta_i\}$, can be determined from (2.1) — (2.3), putting η'_i and α'_i instead of η_i and α_i, respectively. Then the sequence $\{\gamma_i\}$ will follow from (2.6).

Putting $\alpha'_0 = 0$, $\alpha'_n = s + 1$ ($\alpha'_n \geqq \alpha'_{n-1}$ is assured by (2.11)) and, in accordance with (2.9), denoting by

$$\delta_0 = 0 < \delta_1 < \delta_2 < \ldots < \delta_{n-k} < \delta_{n-k+1} = s + 1$$

the values taken on by $\alpha'_0, \alpha'_1, \ldots, \alpha'_n$, the sequences $\{\varepsilon_i\}$ and $\{\bar{\varepsilon}_i\}$ will be defined by (2.9).

Let be $\varphi_1 < \varphi_2 < \ldots < \varphi_{n-1}$ the complementary set of the sequence $\gamma_{\delta_1}, \gamma_{\delta_2}, \ldots, \gamma_{\delta_{n-k}}$ within the sequence $1, 2, \ldots, n - 1$ and let us define in accordance with (2.7)

$$\begin{aligned} \alpha_{\varphi_i} &= \bar{\varepsilon}_i - \beta_{\varphi_i} && (i = 1, 2, \ldots, k - 1), \\ \alpha_{\gamma_{\delta_i}} &= \alpha_{\gamma_{\delta_i} - 1} && (i = 1, 2, \ldots, n - k) \end{aligned}$$

from which by (2.5) we obtain the sequence $0 \leqq \eta_1 \leqq \eta_2 \leqq \ldots \leqq \eta_{n-1} < 1$

Thus it is proved that the transformation (2.10) is one by one and that the vectors $\{\eta_i\}$ and $\{\eta'_i\}$ have the same set of possible values.

We now have to prove that the transformation preserves the measure. The space of the vector $\{\eta_i\}$ can be divided into a finite number of subsets for each of which the vectors $\{\alpha_i\}$ and $\{\beta_i\}$ are constant; evidently these subsets are measurable. Our transformation means a simple translation for such a subset thus it is measure-preserving. As under the circumstances of the condition b) the distribution of $\{\eta_i\}$ is uniform over the set of its possible values and $\{\eta_i'\}$ has the same set of possible values, the distributions of $\{\eta_i\}$ and $\{\eta_i'\}$ are identical.

Finally we see from (2.8) that the interval $\left[\frac{\varepsilon_i}{n}, \frac{\varepsilon_i+1}{n}\right)$ does not contain any of the values $\eta_1, \eta_2, \ldots, \eta_{n-1}$ and the interval $\left[\frac{\bar{\varepsilon}_j}{n}, \frac{\bar{\varepsilon}_j+1}{n}\right)$ contains at least one of them; on the other hand, (2.9), (2.10) and (2.4) assure that

$$\eta'_{\varepsilon_i+1} - \eta'_{\varepsilon_i} > 1/n, \qquad \eta'_{\bar{\varepsilon}_j+1} - \eta'_{\bar{\varepsilon}_j} \leqq 1/n$$

$$(i = 1, 2, \ldots, n-k+1; \;\; j = 1, 2, \ldots, k-1).$$

Thus the transformation has the property c) as well.

Thus we have proved the equivalency of Models C and B, i.e. that of A and B too.

§ 3. The asymptotic joint distribution of the mean and the k-th ordered sample element

It is supposed in this and in the first part of the following § that the common density function of the independent random variables $\xi_1, \xi_2; \ldots, \xi_n$ is continuous and positive in intervals, which contain the expected value $\mathbf{M}(\xi_i)$ and the quantiles considered in the following as inner points. Let be for the sake of simplicity

$$\mathbf{M}(\xi) = 0, \quad \mathbf{M}(\xi^2) = \mathbf{D}^2(\xi) = 1.$$

We shall introduce in the following some new notations for quantities depending on n but without indicating this circumstance. Let us have now a sequence of integers $k = k(n)$ for which we assume that

$$q = q_n = \frac{k(n)}{n} \to \bar{q}, \qquad 0 < \bar{q} < 1,$$

as $n \to \infty$ or in short $\lim\limits_{n \to \infty} q = \bar{q}$.

We shall use further the following notations:

$$t = t(q_n) = F^{-1}(q_n), \qquad \bar{t} = F^{-1}(\bar{q}),$$

$$\sigma^2 = \sigma^2(q_n) = \frac{q(1-q)}{f^2(t)}, \quad \bar{\sigma} = \sigma(\bar{q}),$$

$$m = m(q_n) = \frac{1}{q_n} \int\limits_{-\infty}^{t} u f(u)\, du, \quad \bar{m} = m(\bar{q})$$

and

$$\eta_k^* = \frac{\xi_k^* - t}{\sigma}\sqrt{n},$$

(here naturally $k = k(n)$, $t = t(q_n)$ and $\sigma = \sigma(q_n)$).

It will be shown *that in the case of* $f(\bar{q}) \neq 0$ *the joint limiting distribution of* $\bar{\xi}$ *and* ξ_k^* *is normal, more precisely, that*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\bar{\xi} + z\,\eta_k^*\,\dfrac{\sigma}{\sqrt{n}}}{\sqrt{1 - 2\dfrac{qm}{f(q)}z + \sigma^2 z^2}}\sqrt{n} < y,\ \eta_k^* < w\right) =$$

$$(3.1) \qquad = \frac{1}{2\pi(1-\varrho^2)} \int_{-\infty}^{w}\int_{-\infty}^{y} \exp\left(-\frac{u^2 + v^2 - 2\,\varrho\,uv}{2\,(1-\varrho^2)}\right) du\,dv\,,$$

where

$$\varrho = \frac{\bar{\sigma}\left(z - \dfrac{f(\bar{t})\,\overline{m}}{1-\bar{q}}\right)}{\sqrt{1 - 2\dfrac{\bar{q}\,\overline{m}}{f(\bar{t})}z + \sigma^2 z^2}}\,.$$

Proof. We begin with investigating the conditional distribution of $\bar{\xi}$ with respect to ξ_k^*. Evidently under the condition $\xi_k^* = t'$ the conditional distribution of $\xi_1^* + \xi_2^* + \ldots + \xi_{k-1}^*$ is identical to that of the sum of $k-1$ independent variates with the common distribution function

$$\mathbf{P}(\zeta_i < x) = \frac{1}{F(t')}\int_{-\infty}^{x} f(u)\,du\,, \qquad (-\infty < x \leqq t';\, i = 1, 2, \ldots, k-1)$$

and, similarly $\xi_{k+1}^* + \ldots + \xi_n^*$ is conditionally distributed as the sum of $n-k$ independent variates with the common distribution function

$$\mathbf{P}(\zeta_i < x) = \frac{1}{1-F(t')}\int_{t'}^{x} f(u)\,du \qquad (t' \leqq x \leqq \infty,\, i = k+1, \ldots, n)\,.$$

The variables ζ_i have the expectations and variances

$$\mathbf{M}(\zeta_i) = \frac{1}{F(t')}\int_{-\infty}^{t'} u f(u)\,du\,,$$

$$\mathbf{D}^2(\zeta_i) = \frac{1}{F(t')} \int\limits_{-\infty}^{t'} u^2 f(u)\, du - \left(\frac{1}{F(t')} \int\limits_{-\infty}^{t'} uf(u)\, du \right)^2,$$

$$i = 1, 2, \ldots k-1,$$

$$\mathbf{M}(\zeta_i) = -\frac{1}{1-F(t')} \int\limits_{-\infty}^{t'} uf(u)\, du\,,$$

$$\mathbf{D}^2(\zeta_i) = \frac{1}{1-F(t')} \left(1 - \int\limits_{-\infty}^{t'} u^2 f(u)\, du \right) - \left(\frac{1}{1-F(t')} \int\limits_{-\infty}^{t'} uf(u)\, du \right)^2,$$

$$i = k+1, k+2, \ldots n\,.$$

Thus it follows that the conditional expectation and variance of $\bar{\xi}$ is — with the notation $q' = \dfrac{k-1}{n}$ —

$$\mathbf{M}(\xi \mid \xi_k^* = t') = \frac{n-1}{n} \int\limits_{-\infty}^{t'} uf(u)\, du \frac{q' - F(t')}{F(t')\bigl(1 - F(t')\bigr)} + \frac{t'}{n},$$

$$\mathbf{D}^2(\bar{\xi} \mid \xi_k^* = t') = \frac{n-1}{n^2} \left[\frac{1-q'}{1-F(t')} + \int\limits_{-\infty}^{t'} u^2 f(u)\, du\, \frac{q' - F(t')}{F(t')\bigl(1-F(t')\bigr)} - \right.$$

$$\left. - \left(\int\limits_{-\infty}^{t'} uf(u)\, du \right)^2 \frac{q'\bigl(1 - F(t')\bigr)^2 + (1-q')\, F(t')^2}{F(t')^2\bigl(1 - F(t')\bigr)^2} \right].$$

Let be $\varphi(x) = (2\pi)^{-\frac{1}{2}} \exp\left(-\dfrac{x^2}{2}\right)$ and $\Phi(x) = \displaystyle\int\limits_{-\infty}^{t'} \varphi(t)\, dt$.

According to the central limit theorem we have

$$\lim_{n\to\infty} \mathbf{P} \left(\frac{\bar{\xi} - \mathbf{M}(\bar{\xi} \mid \xi_k^* = t')}{\mathbf{D}(\bar{\xi} \mid \xi_k^* = t')} < y \mid \xi_k^* = t' \right) = \Phi(y) \tag{3.2}$$

uniformly in y.

It is known furthermore that denoting the distribution function and density function of ξ_k^* by $F_k(u)$ and $f_k(u)$, respectively, we have [8] $\left(t = F^{-1}\left(\dfrac{k}{n}\right) \right)$

$$\lim_{n\to\infty} \frac{\sigma}{\sqrt{n}} f_k\left(t + \frac{\sigma}{\sqrt{n}}\,\omega\right) = \varphi(w) \tag{3.3}$$

$$\lim_{n\to\infty} F_k\left(t + \frac{\sigma}{\sqrt{n}} w\right) = \Phi(w) \tag{3.4}$$

uniformly in w.

For the conditional expectation and variance we have

$$\lim_{n\to\infty} \mathbf{M}(\overline{\xi} \mid \xi_k^* = t + x) = -\, x \int\limits_{-\infty}^{\bar{t}} u f(u)\, du \, \frac{f(\bar{t})}{\bar{q}(1-\bar{q})} + o(x) =$$

$$= -\,\frac{x \overline{m} f(\bar{t})}{1-\bar{q}} + o(x)\,,$$

$$\lim_{n\to\infty} n\, \mathbf{D}^2(\overline{\xi} \mid \xi_k^* = t + x) = 1 - \frac{\overline{m}^2 \bar{q}}{1-\bar{q}} + O(x)\,.$$

Hence

$$\lim_{n\to\infty} \mathbf{M}(\overline{\xi} + z\,\xi_k^* \mid \xi_k^* = t + x) = z\bar{t} + x\left(z - \frac{\overline{m} f(\bar{t})}{1-\bar{q}}\right) + o(x)\,,$$

$$\lim_{n\to\infty} n\, \mathbf{D}^2(\overline{\xi} + z\,\xi_k^* \mid \xi_k^* = t + x) = 1 - \frac{\overline{m}^2 \bar{q}}{1-\bar{q}} + O(x)\,.$$

Thus from (3.2) — ξ_k^* being given — it follows that for each $\varepsilon > 0$, we can find a positive constant δ, such that for sufficiently large n

$$\left| \mathbf{P}\left(\overline{\xi} + z\,\xi_k^* < zt + x\left(z - \frac{mf(t)}{1-q}\right) + \right.\right.$$
$$\left.\left. + \frac{y'}{\sqrt{n}} \sqrt{1 - \frac{m^2 q}{1-q}} \,\right|\, \xi_k^* = t + x\right) - \Phi(y') \Bigg| < \frac{\varepsilon}{4K}\,, \tag{3.5}$$

whenever $|x| < \delta$; here K shall fulfill the condition

$$\Phi(-K) < \frac{\varepsilon}{6}\,. \tag{3.6}$$

Further for sufficiently large n (see (3.4)) we have

$$F_k\left(t - \frac{\sigma}{\sqrt{n}}\, K\right) < \frac{\varepsilon}{2}\,. \tag{3.7}$$

Now the probability in the left hand side of (3.1) equals

$$I_n(q) = \int_{-\infty}^{w} \mathbf{P}\left(\overline{\xi} + z\,\xi_k^* < zt + \frac{y}{\sqrt{n}}\sqrt{1 - 2\frac{qm}{f(t)}z + \sigma^2 z^2}\;\middle|\; \xi_k^* = t + \frac{\sigma}{\sqrt{n}}\tau\right) \times \tag{3.8}$$

$$\times f_k\left(t + \frac{\sigma}{\sqrt{n}}\tau\right)\frac{\sigma}{\sqrt{n}}\,d\tau\,.$$

Let be

$$I(\bar{q}) = \int_{-\infty}^{w} \Phi\left(y\sqrt{\frac{1 - 2\frac{\bar{q}\overline{m}}{f(\bar{t})}z + \bar{\sigma}^2 z^2}{1 - \frac{\overline{m}^2\bar{q}}{1-\bar{q}}}} - \tau\frac{z - \frac{\overline{m}f(\bar{t})}{1-\bar{q}}}{\sqrt{1 - \frac{\overline{m}^2\bar{q}}{1-\bar{q}}}}\right)\varphi(\tau)\,d\tau\,. \tag{3.9}$$

The difference between $I_n(q)$ and $I(\bar{q})$ can be made smaller than ε as shown in the following:

First we can denote the argument of Φ in $I(\bar{q})$ by y', then apply (3.5) putting $x = \frac{\sigma}{\sqrt{n}}\tau$. Dividing the integration interval $(-\infty, w)$ into $(-\infty, -K)$ and $(-K, w)$ (here $-K$ can be chosen smaller than $-|w|$), in consequence of (3.6) and (3.7) the first part of the difference becomes smaller then $\frac{\varepsilon}{2}$. Concerning the second part we refer to (3.3) and (3.5), namely for sufficiently large n we obtain from (3.3) the inequality

$$\left|\frac{\sigma}{\sqrt{n}}f_k\left(t + \frac{\sigma}{\sqrt{n}}\tau\right) - \varphi(\tau)\right| < \frac{\varepsilon}{6K}\,.$$

I.e. for sufficiently large n

$$|\,I_n(q) - I(\bar{q})\,| < \frac{\varepsilon}{3}\,.$$

Let us turn now to the evaluation of $I(\bar{q})$; denoting the coefficients of y and τ by A and B, resp.

$$\int_{-\infty}^{w}\Phi(By - A\,\tau)\,\varphi(\tau)\,d\tau = \frac{1}{2\pi}\int_{-\infty}^{w}\left(\int_{-\infty}^{By-A\tau} e^{-\frac{u^2}{2}}du\right)e^{-\frac{\tau^2}{2}}d\tau =$$

$$= \frac{B}{2\pi}\int_{-\infty}^{w}\int_{-\infty}^{y} e^{-\frac{B^2}{2}\left(u^2 - 2\frac{A}{B}u\tau + \tau^2\right)}du\,d\tau.$$

From this we have the relation (3.1) and the expression of ϱ.

§ 4. The asymptotic normality of the distribution P_k

In this § we shall use of our previous assumptions

$$\mathbf{M}(\xi) = 0\,, \quad \mathbf{M}(\xi^2) = \mathbf{D}^2(\xi) = 1\,.$$

Let us denote by q_0 the quantile corresponding to the expectation $\mathbf{M}(\xi) = 0$, i.e.

$$q_0 = F(\mathbf{M}(\xi)) = F(0)\,.$$

Let further be

$$m_0 = \mathbf{M}(\xi \mid \xi < 0) = \frac{1}{q_0} \int\limits_{-\infty}^{0} u f(u)\, du\,.$$

In order to obtain for the probabilities $P_k = \mathbf{P}(\bar{\xi} < \xi_k^*)$ a reasonable limiting distribution let $q \to q_0$ according to $\dfrac{k}{n} = q = q_0 + \dfrac{x}{\sqrt{n}}$.

In this case we obtain the following relations

$$m = m(q) = m_0 + O\left(\frac{1}{\sqrt{n}}\right),$$

$$t = t(q) = \frac{x}{f(0)\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right),$$

$$f(t) = f(0) + O\left(\frac{1}{\sqrt{n}}\right).$$

Substituting $z = -1$, $\omega = +\infty$ and

$$y = \frac{t}{\sqrt{1 - 2\dfrac{mq}{f(t)} + \sigma^2}} =$$

$$= \frac{x}{\sqrt{q_0(1-q_0) + 2\, q_0\, m_0\, f(0) + f(0)^2}} + o\left(\frac{1}{\sqrt{n}}\right)$$

in (3.1) we obtain for large n the asymptotic relation

$$\mathbf{P}(\bar{\xi} < \xi_k^*) \sim \Phi\left(x \frac{1}{\sqrt{q_0(1-q_0) + 2\, m_0\, q_0\, f(0) + f(0)^2}}\right).$$

Denoting by $\varkappa_n = \varkappa$ the random variable for which the event $(\xi_{\varkappa-1}^* \leqq \bar{\xi} < \xi_\varkappa^*)$ occurs, the relation

$$\mathbf{P}(\varkappa_n < k + 1) = \mathbf{P}(\bar{\xi} < \xi_k^*)$$

holds and we obtain

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\dfrac{\varkappa_n}{n} - q_0}{\dfrac{1}{\sqrt{n}}} < x\right) = \Phi\left(x \frac{1}{\sqrt{q_0(1-q_0) + 2\, m_0\, q_0\, f(0) + f(0)^2}}\right).$$

and hence

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\varkappa_n - nq_0}{\sqrt{q_0(1-q_0) + 2\, m\, q_0\, f(0) + f(0)^2}} < y\sqrt{n}\right) = \Phi(y)\,. \tag{4.1}$$

We show now that this theorem can be extended for the case $f(0) = 0$ under the assumption that $f(x)$ is continuous in a neighbourhood of the origin.

In this case we obtain the simple form

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\varkappa_n - nq_0}{\sqrt{q_0(1-q_0)}} < y\sqrt{n}\right) = \Phi(y)\,.$$

This means that $\varkappa_n$ has the same limiting distribution as the variate $\overline{\varkappa}_n$ defined by the relation[3]

$$\xi^*_{\overline{\varkappa}-1} < \mathbf{M}(\xi) < \xi^*_{\overline{\varkappa}}\,.$$

Proof. By virtue of the central limit theorem we can find a positive constant K such that for sufficiently large n $\mathbf{P}\left(|\bar{\xi}| > \frac{K}{\sqrt{n}}\right) < \frac{\varepsilon}{2}$, but in this case

$$\mathbf{P}\left(\xi^*_k > \frac{K}{\sqrt{n}},\ |\bar{\xi}| > \frac{K}{\sqrt{n}}\right) < \frac{\varepsilon}{2} \quad \text{too.} \tag{4.2}$$

Since

$$\lim_{n\to\infty} \frac{F\left(\dfrac{K}{\sqrt{n}}\right) - F\left(-\dfrac{K}{\sqrt{n}}\right)}{\sqrt{\dfrac{q(1-q)}{\sqrt{n}}}} = \lim_{n\to\infty} \frac{2\,Kf\left(\dfrac{K}{\sqrt{n}}\right)}{\sqrt{q(1-q)}} = 0$$

moreover, $F(\xi^*_k)$ is asymptotically normally distributed with parameters q, $\frac{q(1-q)}{n}$, therefore for sufficiently large n

$$\mathbf{P}\left(|\xi^*_k| < \frac{K}{\sqrt{n}}\right) = \mathbf{P}\left(F\left(-\frac{K}{\sqrt{n}}\right) < F(\xi^*_k) < F\left(\frac{K}{\sqrt{n}}\right)\right) < \frac{\varepsilon}{2}\,, \tag{4.3}$$

[3] This remark is due to A. Rényi.

i.e. it follows from (4.2) and (4.3) that

$$\mathbf{P}(|\xi_k^*| < |\bar{\xi}|) < \varepsilon .$$

This proves that

$$\lim_{n\to\infty} \mathbf{P}(|\xi_k^*| < |\bar{\xi}|) = \lim_{n\to\infty} \mathbf{P}(|\xi_k^*| < 0).$$

i.e. $\varkappa_n$ and $\bar{\varkappa}_n$ have the same limiting distribution.

Consider now the special case of (4.1) when ξ is exponentially distributed in the interval $(-1, \infty)$, i.e.

$$F(x) = 1 - e^{-(x+1)} .$$

Then $\mathbf{M}(\xi) = 0$, $q_0 = 1 - e^{-1}$, $m_0 = (e - 1)^{-1}$ and we have

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\varkappa_n - n\dfrac{e-1}{e}}{\dfrac{1}{e}\sqrt{n(e-2)}} < y \right) = \Phi(y) .$$

This agrees with the known limiting form of the occupancy distribution [5, 6].

(Received June 6, 1961)

REFERENCES

[1] Rényi, A.: *Valószínűségszámítás.* Tankönyvkiadó, Budapest, 1954.

[2] Abel, N. H.: *Oeuvres Complètes.* C. Groendahl, Christiania, 1839. Vol. 1., p. 102.

[3] Jordan, C.: *Calculus of finite differences.* Eggenberger, Budapest, 1939.

[4] Pearson, E. S. and Hartley, H. O. (eds.): *Biometrika Tables for Statisticians,* I. Cambridge University Press, 1954.

[5] Schäfer, W.: „Das Mutungsproblem der Besetzungs-Verteilung.“ *Mitteilungsblatt für mathematische Statistik* **6** (1954) 1—38.

[6] Rényi, A.: "Some remarks on the theory of trees." *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **4** (1959) 73—85.

[7] Wilks, S. S.: "Order statistics". *Bulletin of the American Mathematical Society* **54** (1948) 6—50.

[8] Cramér, H.: *Mathematical methods of statistics.* Princeton University Press, 1946.

[9] Nicholson, W. L.: "Occupancy probability distribution critical points". *Biometrika* **48** (1961) 175—181.

[10] David, H. T.: "On sample mean among the moderate order statistics." *Annals of Mathematical Statistics* **33** (1962) 1160—1166.

МЕСТО ВЫБОРОЧНОГО СРЕДНЕГО СРЕДИ ЭЛЕМЕНТОВ ВАРИДЦИОННОГО РЯДА

K. SARKADI, E. SCHNELL и I. VINCZE

Резюме

Авторы в своей статье занимаются следующей проблемой:

Пусть будут $\xi_1, \xi_2, \ldots, \xi_n$ независимые элементы выборки из некоторого совокупности с непрерывным распределением, спрашивается, какая вероятность того, что выборочное среднее

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

попадет между $(k-1)$-ым и k-ым елементом вариационного ряда, иными словами, надо определить вероятности

$$p_k = \mathbf{P}(\xi^*_{k-1} \leqq \bar{\xi} \leqq \xi^*_k) .$$

где через ξ^*_k $(k = 1, 2, \ldots, n)$ обозначается k-тый элемент вариационного ряда. Удалось определить в явном виде вероятности p_k для случая экспоненциального распределения; в § 1 при помощи аддитивного Марковского свойства вариационного ряда следующая формула:

$$p_k = \frac{\binom{n}{k-1}(k-1)\,!}{n^{n-1}} \mathfrak{S}^{k-1}_{n-1} \tag{1.6}$$

где через $\mathfrak{S}^k_n$ обозначается так называемое Стирлинговое число второго рода.

Распределение, характеризуемое формулой (1.6) является частным случаем распределения занятия (см. напр. [5]).

С использованием этого факта авторы в § 2 вторично доказывают формулу (1.6). Сущность этого доказательства следующая: Задается такое преобразование с сохранением меры, которое последовательность $0 \leqq \eta_1 \leqq \leqq \eta_2 \leqq \ldots \leqq \eta_{n-1} < 1$ преобразует в последовательность $0 = \eta'_0 \leqq \eta'_1 \leqq \leqq \eta'_2 \leqq \ldots \leqq \eta'_{n-1} < \eta'_n = 1$ таким образом, что неравнество $\eta'_i - \eta'_{i-1} < < 1/n$ имеет силу тогда и только тогда, если среди чисел $\eta_1, \eta_2, \ldots \eta_{n-1}$ по крайней мере одно попадает в интервал $\left(\frac{i-1}{n}, \frac{i}{n}\right)$.

В § 3—4 доказывается относительно общего случая при некоторых слабых допущениях, что предельным распределением является Гауссовое распределение.