KONINKL. NEDERL. AKADEMIE VAN WETENSCHAPPEN – AMSTERDAM
Reprinted from Proceedings, Series A, 60, No. 5 and Indag. Math., 19, No. 5, 1957

MATHEMATICS

THE ASYMPTOTIC DISTRIBUTION FOR LARGE m OF TERPSTRA'S STATISTIC FOR THE PROBLEM OF m RANKINGS [1]).

BY

PH. VAN ELTEREN

(Communicated by Prof. D. VAN DANTZIG at the meeting of May 25, 1957)

1. *Introduction*

Consider m random vectors $\boldsymbol{x}^{(\alpha)}(\alpha = 1, 2, \ldots, m)$ with n components $x_1^{(\alpha)}, x_2^{(\alpha)}, \ldots x_n^{(\alpha)}$ being the results of measurements on n objects. M. FRIEDMAN (1937) constructed a distributionfree test for the hypothesis H_0 that such vectors are stochastically independent and that for each α the components of $\boldsymbol{x}^{(\alpha)}$ are independent and have the same continuous distribution. In an appendix to Friedman's paper S. WILKS showed that the distribution under H_0 of the test statistic tends to a χ^2-distribution with $n-1$ degrees of freedom if $m \to \infty$. In the papers of M. G. KENDALL and B. BABINGTON SMITH (1939) and M. G. KENDALL (1945) another approximation to the distribution of Friedman's statistic was treated and the case that the distributions of the components are not continuous was attacked by the introduction of a correction for ties. A. BENARD and PH. VAN ELTEREN (1953) gave a generalization of Friedman's test and of Wilks' proof for the case that an arbitrary number of observations for each $x_i^{(\alpha)}$ is available. T. J. TERPSTRA (1955) introduced another statistic for the last mentioned case and derived its "asymptotic distribution" [2]) for $n \to \infty$.

The alternatives for these tests are not precisely formulated by the authors but as their statistics can be considered as means of rank correlation measures they will often lead to rejection of H_0 when the vectors $\boldsymbol{x}^{(\alpha)}$ are positively correlated pair by pair.

The main object of this paper is to derive the asymptotic distribution of Terpstra's statistic under hypothesis H_0 for bounded n and $m \to \infty$. The number of observations of each $x_i^{(\alpha)}$ is here restricted to 1 again

[1]) Report SP 58 of the Statistical Department of the Mathematical Centre, Amsterdam.

[2]) Let $\mathbf{t}_n$ be a random variable depending on a parameter n and let there be functions a_n and b_n of n such that the distribution of $a_n^{-1}(\mathbf{t}_n - b_n)$ tends for $n \to \infty$ to the distribution of a random variable $\boldsymbol{u}$ not depending on n. Then it is said that the asymptotic distribution of $\mathbf{t}_n$ is the distribution of $a_n \boldsymbol{u} + b_n$.

except for section 7 where the case is treated that some observations are missing.

2. *Definitions and abbreviations*

Let $\boldsymbol{T}_{\alpha,\beta}$ be Kendall's rank correlation statistic (cf. M. G. KENDALL (1948), chapter 1 and 2) for the vectors $\boldsymbol{x}^{(\alpha)}$ and $\boldsymbol{x}^{(\beta)}$ given by

$$\boldsymbol{T}_{\alpha,\beta} \overset{\text{def}}{=} \sum_{i<j} \mathbf{z}_{i,j}^{(\alpha)} \mathbf{z}_{i,j}^{(\beta)} \text{ [3]} = \tfrac{1}{2} \sum_{i,j} \mathbf{z}_{i,j}^{(\alpha)} \mathbf{z}_{i,j}^{(\beta)} \tag{2.1}$$

where

$$\left\{ \begin{array}{l} \mathbf{z}_{i,j}^{(\alpha)} \overset{\text{def}}{=} \operatorname{sgn} (\boldsymbol{x}_i^{(\alpha)} - \boldsymbol{x}_j^{(\alpha)}) \\ \text{(cf. D. VAN DANTZIG and J. HEMELRIJK (1954)).} \end{array} \right. \tag{2.2}$$

Then Terpstra's statistic $\boldsymbol{T}$ is defined by

$$\left\{ \begin{array}{l} \boldsymbol{T} \overset{\text{def}}{=} \sum_{\alpha<\beta} \boldsymbol{T}_{\alpha,\beta} = \tfrac{1}{2} \sum_{\alpha,\beta} \boldsymbol{T}_{\alpha,\beta} = \tfrac{1}{4} \sum_{\alpha,\beta} \sum_{i,j} \mathbf{z}_{i,j}^{(\alpha)} \mathbf{z}_{i,j}^{(\beta)} = \\ \quad = \tfrac{1}{4} \sum_{i,j} (\sum_{\alpha} \mathbf{z}_{i,j}^{(\alpha)})^2 - \tfrac{1}{4} \sum_{i,j} \sum_{\alpha} (\mathbf{z}_{i,j}^{(\alpha)})^2. \end{array} \right. \tag{2.3}$$

Now suppose that in an experiment the components of vector $\boldsymbol{x}^{(\alpha)}$ assumed $g^{(\alpha)}$ different values $u_1^{(\alpha)} < u_2^{(\alpha)} < \ldots < u_{g_\alpha}^{(\alpha)}$ $(\alpha = 1, 2, \ldots, m)$. Let the value $u_h^{(\alpha)}$ be assumed by $t_h^{(\alpha)}$ components of $\boldsymbol{x}^{(\alpha)}$ $(h = 1, 2, \ldots, g_\alpha)$. Consider for each α the conditional simultaneous distribution of the components of $\boldsymbol{x}^{(\alpha)}$ under the condition that $\boldsymbol{x}_1^{(\alpha)}, \boldsymbol{x}_2^{(\alpha)}, \ldots, \boldsymbol{x}_n^{(\alpha)}$ is a permutation of $t_1^{(\alpha)}$ values $u_1^{(\alpha)}$, $t_2^{(\alpha)}$ values $u_2^{(\alpha)}, \ldots, t_{g_\alpha}^{(\alpha)}$ values $u_{g_\alpha}^{(\alpha)}$. Hypothesis H_0 implies H_0' stating that for each α all possible permutations of this kind have the same probability and that the vectors $\boldsymbol{x}^{(\alpha)}$ $(\alpha = 1, 2, \ldots, m)$ are independent. Hypothesis H_0' is the basis of the computations in the sections 2–6; thus the numbers $g^{(\alpha)}$ and $t_h^{(\alpha)}$ are assumed to be known constants. It follows that the results to be obtained are also valid if the components of $\boldsymbol{x}^{(\alpha)}$ are ranks allotted to n objects in order of some qualitative property.

The following abbreviations will be used (cf. TERPSTRA (1955))

$$G_2^{(\alpha)} \overset{\text{def}}{=} \left\{ \begin{array}{ll} 1 - \{n(n-1)\}^{-1} \sum_h t_h^{(\alpha)} (t_h^{(\alpha)} - 1) & (n \geqq 2) \\ 0 & (n < 2), \end{array} \right. \tag{2.4}$$

$$G_3^{(\alpha)} \overset{\text{def}}{=} \left\{ \begin{array}{ll} 1 - \{n(n-1)(n-2)\}^{-1} \sum_h t_h^{(\alpha)} (t_h^{(\alpha)} - 1) (t_h^{(\alpha)} - 2) & (n \geqq 3) \\ 0 & (n < 3). \end{array} \right. \tag{2.5}$$

---

[3]) In this paper except for section 5 α and β are supposed to run through the values $1, 2, \ldots, m$; i, j, k and l through $1, 2, \ldots, n$ and h through $1, 2, \ldots, g_\alpha$ with the restrictions mentioned below the summation symbols Σ. The random character of a variable is denoted by printing its symbol in bold type; an arbitrary value assumed by a random variable is often denoted by the same symbol, printed in italics.

If further

$$G_2 \overset{\text{def}}{=} m^{-1} \sum_{\alpha} G_2^{(\alpha} \tag{2.6}$$

$$G_3 \overset{\text{def}}{=} m^{-1} \sum_{\alpha} G_3^{(\alpha)} \tag{2.7}$$

and

$$\mathbf{z}_{i,j} \overset{\text{def}}{=} m^{-\frac{1}{2}} \sum_{\alpha} \mathbf{z}_{i,j}^{(\alpha)} \tag{2.8}$$

(2.3) can be written as

$$\boldsymbol{T} \overset{\text{def}}{=} \tfrac{1}{4} m \sum_{i,j} \mathbf{z}_{i,j}^2 - \tfrac{1}{4} mn\,(n-1)\, G_2. \tag{2.9}$$

As G_2 is a constant under H_0' the distribution of $\boldsymbol{T}$ is determined by that of $\sum_{i,j} \mathbf{z}_{i,j}^2$, the sum of squares of the statistics $\mathbf{z}_{i,j}$ of sign tests applied to the differences $\mathbf{x}_i^{(\alpha)} - \mathbf{x}_j^{(\alpha)}$ for $\alpha = 1, 2, \ldots, m$.

For the derivation of the asymptotic distribution of $\sum_{i,j} \mathbf{z}_{i,j}^2$ the following simple relation is used

$$\tfrac{1}{2} \sum \mathbf{z}_{i,j}^2 = n^{-1} \left(\sum_{i} \mathbf{z}_i^2 + \tfrac{1}{6} \sum_{j,k,l} \mathbf{z}_{j,k,l}^2 \right) \tag{2.10}$$

where

$$\mathbf{z}_i \overset{\text{def}}{=} \sum_{j} \mathbf{z}_{i,j} \tag{2.11}$$

and

$$\mathbf{z}_{j,k,l} \overset{\text{def}}{=} \mathbf{z}_{j,k} + \mathbf{z}_{k,l} + \mathbf{z}_{l,j}. \tag{2.12}$$

The term $\sum_{i} \mathbf{z}_i^2$ in (2.10) is a linear transform of Friedman's statistic.

3. *Properties of the simultaneous distribution of the variables* $\mathbf{z}_i$ *and* $\mathbf{z}_{j,k,}$

The means and the covariance matrix of the variables $\mathbf{z}_{i,j}$ under hypothesis H_0' are given by (cf. TERPSTRA (1955), section 3.2, lemma 1)

$$E\,\mathbf{z}_{i,j} = m^{-\frac{1}{2}} \sum_{\alpha} E\,\mathbf{z}_{i,j}^{(\alpha)} = 0, \tag{3.1}$$

$$E\,\mathbf{z}_{i,j}^2 = m^{-1} \sum_{\alpha} E\,(\mathbf{z}_{i,j}^{(\alpha)})^2 = m^{-1} \sum_{\alpha} G_2^{(\alpha)} = G_2, \quad (i \neq j), \tag{3.2}$$

$$E\,\mathbf{z}_{i,j}\,\mathbf{z}_{i,l} = m^{-1} \sum_{\alpha} E\,\mathbf{z}_{i,j}^{(\alpha)}\,\mathbf{z}_{i,l}^{(\alpha)} = \tfrac{1}{3}\, m^{-1} \sum_{\alpha} G_3^{(\alpha)} = \tfrac{1}{3}\, G_3, \quad (i,j,k,l) \neq, ^{4}) \tag{3.3}$$

$$E\,\mathbf{z}_{i,j}\,\mathbf{z}_{k,l} = 0, \quad (i,j,k,l) \neq \ ^{4}) \tag{3.4}$$

and as $\mathbf{z}_{i,j} = -\mathbf{z}_{j,i}$

$$E\,\mathbf{z}_{i,j}\,\mathbf{z}_{l,i} = E\,\mathbf{z}_{j,i}\,\mathbf{z}_{i,l} = -\tfrac{1}{3} G_3 \text{ etc.}$$

---

[4]) $(a_1, a_2, \ldots, a_n) \neq$ means $a_i \neq a_j$ for each pair (i, j) where i and j belong to the set $\{1, 2, \ldots, n\}$.

Now by (2.11), (2.12) and (3.1)

$$E\,\mathbf{z}_i = E\,\mathbf{z}_{j,k,l} = 0. \tag{3.5}$$

Further by (3.2), (3.3) and (3.4)

$$E\,\mathbf{z}_i\,\mathbf{z}_{j,k,l} = 0 \tag{3.6}$$

for each i, j, k and l. Formula (3.6) is trivial if $i \neq j$, $i \neq k$ and $i \neq l$. And if $i = j$

$$E\,\mathbf{z}_i\,\mathbf{z}_{i,k,l} = E \sum_j \mathbf{z}_{i,j}\,(\mathbf{z}_{i,k} + \mathbf{z}_{k,l} + \mathbf{z}_{l,i}) =$$
$$= \sum_j E\,\mathbf{z}_{i,j}\,\mathbf{z}_{i,k} + \sum_j E\,\mathbf{z}_{i,j}\,\mathbf{z}_{l,i} + E\,\mathbf{z}_{i,l}\,\mathbf{z}_{k,l} + E\,\mathbf{z}_{i,k}\,\mathbf{z}_{k,l} = 0.$$

The covariance matrix of the variables $\mathbf{z}_i$ is given by

$$E\,\mathbf{z}_i\,\mathbf{z}_j = (n\,\delta_{i,j} - 1)\,(G_{2,3} + \tfrac{1}{3}\,n G_3) \tag{3.7}$$

where

$$\delta_{i,j} \overset{\text{def}}{=} \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \tag{3.8}$$

and

$$G_{2,3} \overset{\text{def}}{=} G_2 - \tfrac{2}{3}\,G_3. \tag{3.9}$$

Formula (3.7) is easily derived from (3.2), (3.3) and (3.4) or as a special case from formula (2.3.1) in BENARD and VAN ELTEREN (1953).

Let $\boldsymbol{u}$ be the vector $(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{n-1})$ and $\mathscr{A}$ the matrix with elements

$$a_{i,j} \overset{\text{def}}{=} n^{-1}(1 + \delta_{i,j}) \quad (i = 1, 2, \ldots, n-1;\ j = 1, 2, \ldots, n-1). \tag{3.10}$$

Then by $\sum_{i=1}^{n} \mathbf{z}_i = 0$

$$n^{-1} \sum_{i=1}^{n} \mathbf{z}_i^2 = \boldsymbol{u}'\,\mathscr{A}\,\boldsymbol{u} \quad [5]) \tag{3.11}$$

and the covariance matrix of the components of the vector $\boldsymbol{u}$ is

$$E\,\boldsymbol{u}\boldsymbol{u}' = (G_{2,3} + \tfrac{1}{3}\,n G_3)\,\mathscr{A}^{-1} \quad \text{(cf. (3.7))}. \tag{3.12}$$

The variables $\mathbf{z}_{j,k,l}$ have a covariance matrix given by

$$E\,\mathbf{z}_{j,k,l}^2 = 3\,E\,\mathbf{z}_{j,k}^2 - 6\,E\,\mathbf{z}_{j,k}\,\mathbf{z}_{j,l} = 3 G_{2,3}\ , \quad (j,k,l) \neq \tag{3.13}$$

and

$$E\,\mathbf{z}_{i,k,l}\,\mathbf{z}_{j,k,l} = E\,\mathbf{z}_{k,l}^2 - 2\,E\,\mathbf{z}_{i,k}\,\mathbf{z}_{j,k} + 2\,E\,\mathbf{z}_{i,j}\,\mathbf{z}_{k,l} = G_{2,3}, \ (i,j,k,l) \neq . \tag{3.14}$$

Covariances of two variables $\mathbf{z}_{j,k,l}$ with not more than one suffix the same are zero.

---

[5]) Matrix notation; $\boldsymbol{u}$ denotes the column vector $\boldsymbol{u}'$ the corresponding row vector. Matrices are denoted by script-letters $\mathscr{A}$, $\mathscr{B}$ etc.

Thus if i, j, k and l are $\leqq n-1$

$$E\,\mathbf{z}_{i,j,n}\,\mathbf{z}_{k,l,n}=(\delta_{i,k}-\delta_{i,l}-\delta_{j,k}+\delta_{j,l}+\delta_{i,k}\,\delta_{j,l}-\delta_{i,l}\,\delta_{j,k})\,G_{2,3}. \tag{3.15}$$

Let $\mathbf{v}$ be the vector with $\frac{1}{2}(n-1)\,(n-2)$ components $\mathbf{z}_{j,k,n}$ $i\leqq j<k\leqq n-1$ and $\mathscr{B}$ the matrix with $\frac{1}{2}(n-1)\,(n-2)$ rows and columns and with elements

$$\begin{cases} b_{(i,j),(k,l)} \overset{\text{def}}{=} n^{-1}(-\delta_{i,k}+\delta_{i,l}+\delta_{j,k}-\delta_{j,l})+\delta_{i,k}\,\delta_{j,l}-\delta_{i,l}\,\delta_{j,k}, \\ \qquad\qquad (1\leqq i<j\leqq n-1),\ (1\leqq k<l\leqq n-1). \end{cases} \tag{3.16}$$

Then by

$$\mathbf{z}_{j,k,l}=\mathbf{z}_{j,k,n}+\mathbf{z}_{j,n,l}+\mathbf{z}_{n,k,l} \tag{3.17}$$

is found that

$$\frac{1}{6n}\sum_{j,k,l}\mathbf{z}^2_{j,k,l}=\mathbf{v}'\,\mathscr{B}\,\mathbf{v}. \tag{3.18}$$

Further the covariance matrix of the components of the vector $\mathbf{v}$ is (cf. (3.15))

$$E\,\mathbf{v}\,\mathbf{v}'=G_{2,3}\,\mathscr{B}^{-1}. \tag{3.19}$$

Remark. It is seen from (3.19) that $E\mathbf{v}\mathbf{v}'$ is a zero matrix if $G_{2,3}=0$ or if (cf. (3.9), (2.6) and (2.7))

$$G^{(\alpha)}_{2,3}\overset{\text{def}}{=}G^{(\alpha)}_2-\tfrac{2}{3}\,G^{(\alpha)}_3=0 \qquad \text{for each } \alpha.$$

Now

$$G^{(\alpha)}_{2,3}=\frac{n^3+\sum_h\{2\,(t^{(\alpha)}_h)^3-3n\,(t^{(\alpha)}_h)^2\}}{3n(n-1)\,(n-2)}, \tag{3.20}$$

thus if a tie of size t is divided into two ties of sizes r and s respectively $(r+s=t)$, the value of $G^{(\alpha)}_{2,3}$ is increased by

$$\frac{2(r^3+s^3-t^3)-3n(r^2+s^2-t^2)}{3n(n-1)(n-2)}=\frac{2rs(n-t)}{n(n-1)(n-2)}. \tag{3.21}$$

It follows from (3.20) that $G^{(\alpha)}_{2,3}=0$ if $g_\alpha=1$ and from (3.21) that this also holds if $g_\alpha=2$. In all other cases $G^{(\alpha)}_{2,3}$ will be positive. Consequently $E\mathbf{v}\mathbf{v}'$ will be a zero-matrix if and only if $g_\alpha\leqq 2$ for each α. Then by (3.18) $\sum_{j,k,l}\mathbf{z}^2_{j,k,l}$ is a constant equal to 0 and thus by (2.10) Friedman's and Terpstra's tests are equivalent. It is easily seen that if $g_\alpha=1$ for each α the statistics of both tests have the constant value 0.

4. *The asymptotic distribution of* $\sum_{i,j}\mathbf{z}^2_{i,j}$

Let there be $\nu_{1,m}$ vectors $\mathbf{x}^{(\alpha)}$ whose components assume under hypothesis H_0' at least two different values (thus $\nu_{1,m}$ vectors with $g_\alpha\geqq 2$) and $\nu_{2,m}$ vectors with $g_\alpha\geqq 3$. Then the following theorem will be proved.

Theorem I. If for large m, $v_{2,m}^{-1}=O(m^{-1})$ then $G_{2,3}^{-1}$ is bounded and the distribution under H_0' of

$$\mathbf{Z} \overset{\text{def}}{=} \tfrac{1}{2} G_{2,3}^{-1} \sum_{i,j} \mathbf{z}_{i,j}^2 = G_{2,3}^{-1} (2m^{-1} \mathbf{T} + \tfrac{1}{2} n (n-1) G_2) \tag{4.1}$$

tends for $m \to \infty$ to the distribution of a random variable $\mathbf{X}$ defined by

$$\mathbf{X} \overset{\text{def}}{=} (1 + \tfrac{1}{3} n G_3 G_{2,3}^{-1}) \mathbf{X}_1 + \mathbf{X}_2 \tag{4.2}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are mutually independent random variables possessing χ^2-distributions with $n-1$ and $\frac{1}{2}(n-1)(n-2)$ degrees of freedom respectively.

Proof. As $G_{2,3}^{(\alpha)}$ has a positive lower bound for each α with $g_\alpha \geqq 2$ (see remark at the end of section 3) and as $G_{2,3} = m^{-1} \sum_\alpha G_{2,3}^{(\alpha)}$ (cf. (3.9), (2.6) and (2.7)) $G_{2,3}^{-1} = O(m v_{2,m}^{-1}) = O(1)$ if $v_{2,m}^{-1} = O(m^{-1})$.

Consider the vectors $\mathbf{u} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_{n-1})$ and $\mathbf{v} = (\mathbf{z}_{1,2,n}, \mathbf{z}_{1,3,n}, \ldots, \mathbf{z}_{n-2,n-1,n})$ defined in section 3 and square matrices $\mathscr{C}$ with $n-1$ and $\mathscr{D}$ with $\frac{1}{2}(n-1)(n-2)$ rows and columns fulfilling the matrix equations

$$E(\mathscr{C}\mathbf{u})(\mathscr{C}\mathbf{u})' = \mathscr{C}(E\mathbf{u}\mathbf{u}')\mathscr{C}' = \mathscr{I}_{n-1} \tag{4.3}$$

and

$$E(\mathscr{D}\mathbf{v})(\mathscr{D}\mathbf{v})' = \mathscr{D}(E\mathbf{v}\mathbf{v}')\mathscr{D}' = \mathscr{I}_{\frac{1}{2}(n-1)(n-2)} \tag{4.4}$$

where $\mathscr{I}_\nu$ denotes the unit matrix with ν rows and columns.

The existence of matrices like $\mathscr{C}$ and $\mathscr{D}$ is for instance proved in C. R. Rao (1952), p. 18.

By (3.12) it follows from (4.3) that

$$\mathscr{C}'\mathscr{C} = (G_{2,3} + \tfrac{1}{3} n G_3)^{-1} \mathscr{A} \tag{4.5}$$

and by (3.19) from (4.4) that

$$\mathscr{D}'\mathscr{D} = G_{2,3}^{-1} \mathscr{B}. \tag{4.6}$$

Let $\mathbf{w}$ denote the combined vector $(\mathscr{C}\mathbf{u}, \mathscr{D}\mathbf{v})$ with $(n-1) + \frac{1}{2}(n-1) \times (n-2) = \frac{1}{2} n(n-1)$ components. Under hypothesis H_0' this vector has the following properties:

a. It is the sum of m stochastically independent random vectors (cf. the definitions of $\mathbf{u}$ and $\mathbf{v}$, (2.8), (2.11) and (2.12)),
b. Its components have zero means (cf. (3.5)),
c. These components have bounded ranges and thus fulfill the well-known Liapounoff-conditions for the validity of the central limit theorem,
d. Their covariance matrix is for each m the unit matrix $\mathscr{I}_{\frac{1}{2}n(n-1)}$ (cf. (3.6), (4.3) and (4.4)).

Hence for $\mathbf{w}$ the central limit theorem for independent random vectors applies (cf. J. V. Uspensky (1937), p. 318) and thus the simultaneous

distribution of its components tends for $m \to \infty$ to the multinormal distribution with covariance matrix $\mathscr{I}_{\frac{1}{2}n(n-1)}$. This is the simultaneous distribution of $n-1$ independent random normal variables with zero means and unit variances.

Consequently the simultaneous distribution of the variables [6])

$$\begin{cases} \boldsymbol{Z}_1 \stackrel{\text{def}}{=} (\mathscr{C}\boldsymbol{u})'\,(\mathscr{C}\boldsymbol{u}) = \boldsymbol{u}'\mathscr{C}'\mathscr{C}\boldsymbol{u} = (G_{2,3} + \tfrac{1}{3}nG_3)^{-1}\,\boldsymbol{u}'\mathscr{A}\boldsymbol{u} = \\ \quad = (G_{2,3} + \tfrac{1}{3}nG_3)^{-1}\,n^{-1}\sum_i \boldsymbol{z}_i^2 \qquad \text{(cf. (4.5) and (3.11))} \end{cases} \tag{4.7}$$

and

$$\begin{cases} \boldsymbol{Z}_2 \stackrel{\text{def}}{=} (\mathscr{D}\boldsymbol{v})'\,\mathscr{D}\boldsymbol{v} = \boldsymbol{v}'\mathscr{D}'\mathscr{D}\boldsymbol{v} = G_{2,3}^{-1}\,\boldsymbol{v}'\mathscr{B}\boldsymbol{v} = (6nG_{2,3})^{-1}\sum_{j,k,l} \boldsymbol{z}_{j,k,l}^2 \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{(cf. (4.6) and (3.18))} \end{cases} \tag{4.8}$$

tends for $m \to \infty$ to the simultaneous distribution of two independent variables $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ specified above. The distribution of $\boldsymbol{Z}_1$ (resp. $\boldsymbol{Z}_2$) tends to the distribution of $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$).

As by (2.10), (4.1), (4.7) and (4.8)

$$\boldsymbol{Z} = G_{2,3}^{-1}\,n^{-1}\,(\sum_i \boldsymbol{z}_i^2 + \tfrac{1}{6}\sum_{j,k,l} \boldsymbol{z}_{j,k,l}^2) = (1 + \tfrac{1}{3}nG_3G_{2,3}^{-1})\,\boldsymbol{Z}_1 + \boldsymbol{Z}_2$$

the distribution of $\boldsymbol{Z}$ tends to the distribution of $\boldsymbol{X}$ defined by (4.2) and theorem I has been proved.

Remark. If for large m $v_{2,m} = o(m)$ but $v_{1,m}^{-1} = O(m^{-1})$ then $G_{2,3} = o(1)$ $\sum_{j,k,l} \boldsymbol{z}_{j,k,l}^2 = o(1)$ and G_3^{-1} is bounded. Thus for $m \to \infty$, $\frac{3}{2}(nG_3)^{-1}\sum_{i,j} \boldsymbol{z}_{i,j}^2$ tends to $\boldsymbol{Z}_1$ and its distribution to a χ^2-distribution with $n-1$ degrees of freedom.

If also $v_{1,m} = o(m)$, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are both $o(1)$ and for $m \to \infty$, $P[\sum_{i,j} \boldsymbol{z}_{i,j}^2 \leqq \varepsilon] \to 1$ for each positive value of ε.

5. *Computation of the distribution function of* $\boldsymbol{X}$

It has been shown in section 4 that the asymptotic distribution function of $\boldsymbol{T}$ is known if the distribution function $F(x)$ of $\boldsymbol{X}$ (cf. (4.2)), has been evaluated. The following methods for the computation of $F(x)$ will be considered. The condition of theorem I is assumed to be valid.

a. Expansion in terms of χ^2-distribution functions

According to its definition $\boldsymbol{X}$ can be considered as a quadratic form in normal variables. Most of the expansions for the distributions of such forms mentioned in the literature converge slowly if they are applied for $\boldsymbol{X}$. The expansion in positive terms due to H. Robbins and E. J. G. Pitman (1949) may be applicable for small values of n. If

$$c \stackrel{\text{def}}{=} 1 + \tfrac{1}{3}nG_3G_{2,3}^{-1} \tag{5.1}$$

[6]) $\boldsymbol{Z}_1$ is Friedman's statistic. The derivation of its asymptotic distribution by Benard and Van Elteren (1953) is essentially the same as that given here.

and $F_\nu(x)$ denotes the distribution function of the χ^2-distribution with ν degrees of freedom, the Robbins–Pitman-expansion gives

$$F(x) = \sum_{j=0}^{\infty} K_j \, F_{\frac{1}{4}n(n-1)+2j}(x) \tag{5.2}$$

with

$$K_j \overset{\text{def}}{=} \frac{(n-3+2)(n-3+4)\ldots(n-3+2j)}{2 \cdot 4 \cdot \ldots \cdot 2j} \cdot \left(\frac{c-1}{c}\right)^j c^{-\frac{1}{4}(n-1)}. \tag{5.3}$$

For *odd* values of n a finite expansion can be given. Let $f(x)$ be the density function of $\boldsymbol{X}$ and $f_\nu(x)$ of a χ^2-distribution with ν degrees of freedom. Then

$$f(x) = c^{-1} \int_0^x f_{\frac{1}{4}(n-1)(n-2)}(x-y) \, f_{n-1}(y/c) \, dy.$$

If $\frac{1}{2}(n-1)$ is denoted by k and $\frac{1}{4}(n-1)(n-2)$ by l, then

$$\begin{cases} f(x) = 2^{-(k+l)} c^{-1} \{\Gamma(k)\,\Gamma(l)\}^{-1} \int_0^x e^{-(x-y)/2} e^{-y/2c} (x-y)^{l-1} (y/c)^{k-1} \, dy \\ \qquad = 2^{-(k+l)} c^{-k} b^{-(k+l-1)} \{\Gamma(k)\,\Gamma(l)\}^{-1} e^{-x/2c} \int_0^{bx} e^{-t/2} t^{l-1} (bx-t)^{k-1} \, dt \end{cases} \tag{5.4}$$

where $b = 1 - c^{-1}$, $t = (x-y)(1-c^{-1})$.

For odd values of n, k is an integer and thus

$$\int_0^{bx} e^{-t/2} t^{l-1} (bx-t)^{k-1} \, dt = \sum_{j=0}^{k-1} (-1)^j \binom{k-1}{j} (bx)^{k-1-j} \int_0^{bx} e^{-t/2} t^{l+j-1} \, dt.$$

Substitution in (5.4) gives

$$f(x) = (2c)^{-k} b^{-l} \{\Gamma(l)\}^{-1} e^{-x/2c} \sum_{j=0}^{k-1} \left(-\frac{2}{b}\right)^j \frac{\Gamma(l+j)}{\Gamma(j+1)\,\Gamma(k-j)} x^{k-1-j} \, F_{2(l+j)}(bx).$$

Now

$$F(x) = (2c)^{-k} b^{-l} \{\Gamma(l)\}^{-1} \sum_{j=0}^{k-1} \left(-\frac{2}{b}\right)^j \frac{\Gamma(l+j)}{\Gamma(j+1)\,\Gamma(k-j)} I_{k-1-j,\,2(l+j)}(x), \tag{5.5}$$

where $I_{r,s}(x) \overset{\text{def}}{=} \int_0^x e^{-t/2c} t^r F_s(bt) dt$ (r integer, $r \geqq 0$, $s > 0$) and by induction [7])

$$I_{r,s}(x) = 2c \sum_{i=0}^{r} c^i \, r!\,i \left\{ 2^r b^{s/2} \frac{\Gamma(r+s/2-i)}{\Gamma(s/2)} F_{2r+s-2i}(x) - 2^i e^{-x/2c} x^{r-i} F_s(bx) \right\}. \tag{5.6}$$

Substituting (5.6) in (5.5), interchanging the order of summation and putting $h = k-1-i$ the following expression is found

$$\begin{cases} F(x) = \frac{1}{\Gamma(l)} \sum_{h=0}^{k-1} \sum_{j=0}^{h} \frac{(-1)^j}{c^h \Gamma(h-j+1)\,\Gamma(j+1)} \Big\{ \Gamma(h+l) \, F_{2(h+l)}(x) - \\ - \left(\frac{c}{c-1}\right)^{l+j} \Gamma(l+j) \, e^{-x/2c} \left(\frac{x}{2}\right)^{h-j} F_{2(l+j)}(x(1-c^{-1})) \Big\}. \end{cases} \tag{5.7}$$

---

[7]) $r!\,i \overset{\text{def}}{=} \begin{cases} r(r-1)\ldots(r-i+1) & \text{if } i = 1, 2, 3, \ldots \\ 1 & \text{if } i = 0. \end{cases}$

For instance for $n=3$ $(k=1, l=\frac{1}{2})$

$$(5.8)\quad \begin{cases} F(x)=\frac{1}{\Gamma(\frac{1}{2})}\left\{\Gamma(\frac{1}{2})\,F_1(x)-\left(\frac{c}{c-1}\right)^{\frac{1}{2}}\Gamma(\frac{1}{2})\,e^{-x/2c}\,F_1(x(1-c^{-1}))\right\}= \\ \qquad =F_1(x)-\sqrt{\frac{c}{c-1}}\,e^{-x/2c}\,F_1(x(1-c^{-1})) \end{cases}$$

or for large x

$$F(x)\approx 1-\sqrt{\frac{c}{c-1}}\,e^{-x/2c}.$$

b. Numerical convolution

The methods described under *a* are very cumbersome except for small values of n. For moderate values of n (say $n\geqq 6$) it will be easier to approximate $F(x)$ by numerical convolution of the distributions of $c\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ using tables of the χ^2-distributions.

c. Normal approximation

By the asymptotic normality of the χ^2-distributions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ for $n\to\infty$, $(\boldsymbol{X}-E\boldsymbol{X})/\sigma\{\boldsymbol{X}\}$ and thus $(\boldsymbol{Z}-E\boldsymbol{Z})/\sigma\{\boldsymbol{Z}\}$ (cf. section 6) will be approximately normally distributed for large n.

6. *Comparison of the distributions of* $\boldsymbol{Z}$ *and of* $\boldsymbol{X}$

The distribution of $\boldsymbol{X}$ (cf. section 4) will in practice be used as an approximation to the distribution of $\boldsymbol{Z}$ for relatively large values of m. For the distribution of $\boldsymbol{Z}$ mean and variance can be derived by (4.1) from Terpstra's results for $\boldsymbol{T}$ (cf. TERPSTRA (1955) formula's (2.3.1), (2.3.8) and (2.3.12)). They are

$$E\boldsymbol{Z}=\tfrac{1}{2}n(n-1)G_2G_{2,3}^{-1}$$

and

$$\sigma^2\{\boldsymbol{Z}\}=4m^{-2}G_{2,3}^{-2}\,\sigma^2\{\boldsymbol{T}\}=\tfrac{1}{9}\,n\,(n-1)\,G_{2,3}^{-2}\,[2\,(n-2)\,\{G_3^2-m^{-2}\sum_\alpha(G_3^{(\alpha)})^2\}+ \\ +9\,\{G_2^2-m^{-2}\sum_\alpha(G_2^{(\alpha)})^2\}].$$

For $\boldsymbol{X}$ is found (cf. (4.2) and (5.1))

$$E\boldsymbol{X}=cE\boldsymbol{X}_1+E\boldsymbol{X}_2=\tfrac{1}{2}n(n-1)G_2G_{2,3}^{-1}$$

and

$$\sigma^2\{\boldsymbol{X}\}=c^2\sigma^2\{\boldsymbol{X}_1\}+\sigma^2\{\boldsymbol{X}_2\}=\tfrac{1}{9}n(n-1)G_{2,3}^{-2}\{2(n-2)G_3^2+9G_2^2\}.$$

$\boldsymbol{Z}$ and $\boldsymbol{X}$ have the same mean and the variance of $\boldsymbol{Z}$ is always smaller than the variance of $\boldsymbol{X}$. In proportion to $\sigma^2\{\boldsymbol{Z}\}$ the difference is $O(m^{-1})$ for large m.

The author computed the distribution of $\boldsymbol{Z}$ for $n=3, 4, 5, 6$, $g_\alpha=3$ $(\alpha=1, 2, \ldots, m)$ and thus $c=n+1=4$ and compared it to the distribution of $\boldsymbol{X}$ (cf. (5.8)).

TABLE I

Distribution of $\boldsymbol{Z} = 6m^{-1}\boldsymbol{T} + 9$ for $n = 3$ and

x	$m = 3$		$m = 4$		$m = 5$		$m = 6$		$(m = \infty)$
	$N(x)$[8]	$P[\boldsymbol{Z} \geqq x]$	$N(x)$[8]	$P[\boldsymbol{Z} \geqq x]$	$N(x)$[8]	$P[\boldsymbol{Z} \geqq x]$	$N(x)$[8]	$P[\boldsymbol{Z} \geqq x]$	$P[\boldsymbol{X} \geqq x]$
0	—	—	15	1	—	—	310	1	1
1,8	—	—	—	—	370	1	—	—	0,8756
2	—	—	—	—	—	—	1200	0,9601	0,8581
3	17	1	48	0,9306	—	—	—	—	0,7708
4	—	—	—	—	—	—	1680	0,8058	0,6876
6	—	—	60	0,7083	—	—	825	0,5898	0,5413
6,6	—	—	—	—	430	0,7145	—	—	0,5030
8	—	—	—	—	—	—	300	0,4837	0,4234
9	—	—	28	0,4306	—	—	—	—	0,3741
10	—	—	—	—	—	—	1080	0,4451	0,3303
11	12	0,5278	—	—	—	—	—	—	0,2917
11,4	—	—	—	—	240	0,3827	—	—	0,2775
12	—	—	6	0,3009	—	—	900	0,3062	0,2575
15	—	—	24	0,2731	—	—	—	—	0,1770
16	—	—	—	—	—	—	300	0,1905	0,1563
16,2	—	—	—	—	95	0,1975	—	—	0,1524
18	—	—	20	0,1620	—	—	470	0,1519	0,1217
19	6	0,1944	—	—	—	—	—	—	0,1074
20	—	—	—	—	—	—	120	0,0914	0,0948
21	—	—	—	—	100	0,1242	—	—	0,0836
22	—	—	—	—	—	—	120	0,0760	0,0738
24	—	—	6	0,0694	—	—	66	0,0606	0,0575
25,8	—	—	—	—	30	0,0471	—	—	0,0459
26	—	—	—	—	—	—	120	0,0521	0,0448
27	1	0,0278	8	0,0417	—	—	—	—	0,0395
28	—	—	—	—	—	—	180	0,0367	0,0349
30,6	—	—	—	—	20	0,0239	—	—	0,0252
34	—	—	—	—	—	—	42	0,0135	0,0165
35,4	—	—	—	—	10	0,0085	—	—	0,0138
36	—	—	1	0,0046	—	—	20	0,0081	0,0128
38	—	—	—	—	—		30	0,0055	0,0100
44	—	—	—	—	—	—	12	0,0017	0,0047
45	—	—	—	—	1	0,0008	—	—	0,0042
54	—	—	—	—	—	—	1	0,0001	0,0014

[8] $N(x) \overset{\text{def}}{=} 6^{m-1} P[\boldsymbol{Z} = x]$.

In Table I the frequencies $N(x) \overset{\text{def}}{=} 6^{m-1} P[\boldsymbol{Z} = x]$ and the tail probabilities $P[\boldsymbol{Z} \geqq x]$ and $P[\boldsymbol{X} \geqq x] = 1 - F(x)$ are given and in chart I a grafical representation of $P[\boldsymbol{Z} \geqq x]$ and $1 - F(x)$. For the values of x till about $x = 20$, $1 - F(x)$ apparently underestimates the tail probabilities of $\boldsymbol{Z}$, for values larger than about $x = 30$ it overestimates them. The long tail of $1 - F(x)$ to the right may explain the fact that $\boldsymbol{X}$ has a larger variance than $\boldsymbol{Z}$. Near its 5 percent point (round $x = 25$) the distribution of $\boldsymbol{X}$ gives a relatively good approximation to the distribution of $\boldsymbol{Z}$ even for such small values of m as considered here.

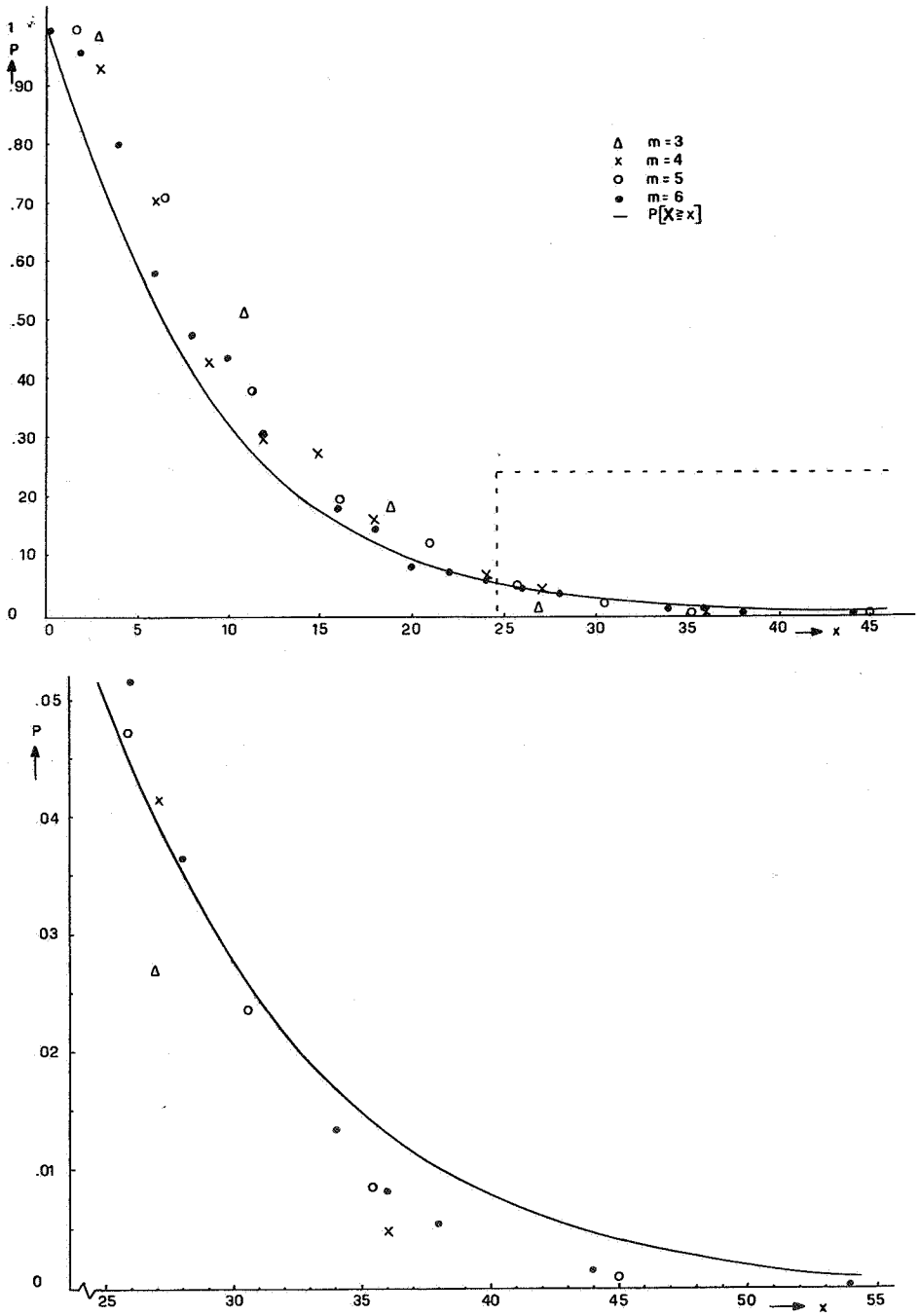

CHART I

Tailprobability function $P[\boldsymbol{Z} \geqq x]$ of $\boldsymbol{Z} = 6m^{-1}\boldsymbol{T} + 9$ for $n = 3$ and $m = 3, 4, 5, 6$ compared to $P[\boldsymbol{X} \geqq x]$ ($\boldsymbol{X}$ defined by (4.2)).
(Second figure is a reproduction of the right hand part of the first (within interrupted line) on enlarged scale.)

7. *Designs with observations omitted at positions chosen at random*

Sometimes the method of m rankings has to be applied on results of experiments, where not all components of each $\boldsymbol{x}^{(\alpha)}$ have been observed. In this section only the case will be treated that the observed components have been chosen at random, i.e. that for each α they are (or can be considered as) a random sample without replacement of $n^{(\alpha)}<n$ components out of the components of $\boldsymbol{x}^{(\alpha)}$. It may be that some observations failed, and that the causes of the failures can be supposed to work at random, or that observations have been omitted deliberately according to a random procedure in order to reduce the size of the experiment.

In these cases the theory treated in the sections 2—5 remains valid if some definitions are appropriately modified. Hypothesis H_0' has to be modified such that all permutations of $t_1^{(\alpha)}$ values $u_1^{(\alpha)}$, $t_2^{(\alpha)}$ values $u_2^{(\alpha)}$, ..., $t_{g_\alpha}^{(\alpha)}$ values $u_{g_\alpha}^{(\alpha)}$ and $n-n^{(\alpha)}$ values $u_0^{(\alpha)}$ are considered, where $u_0^{(\alpha)}$ is an arbitrary value not equal to $u_1^{(\alpha)}$, $u_2^{(\alpha)}$, ... or $u_{g_\alpha}^{(\alpha)}$ allotted to all not observed components. The definitions (2.3), (2.4) and (2.5) are changed as follows:

$$\mathbf{z}_{i,j}^{(\alpha)} \overset{\text{def}}{=} \operatorname{sgn} \left|\boldsymbol{x}_i^{(\alpha)}-u_0^{(\alpha)}\right| \left|\boldsymbol{x}_j^{(\alpha)}-u_0^{(\alpha)}\right| (\boldsymbol{x}_i^{(\alpha)}-\boldsymbol{x}_j^{(\alpha)}), \tag{7.1}$$

$$G_2^{(\alpha)} \overset{\text{def}}{=} \begin{cases} \{n(n-1)\}^{-1}\{n^{(\alpha)}(n^{(\alpha)}-1)-\sum_h t_h^{(\alpha)}(t_h^{(\alpha)}-1)\} & \text{if } n^{(\alpha)} \geqq 2 \\ 0 & \text{if } n^{(\alpha)} < 2 \end{cases} \tag{7.2}$$

and

$$G_3^{(\alpha)} \overset{\text{def}}{=} \begin{cases} \{n(n-1)(n-2)\}^{-1}\{n^{(\alpha)}(n^{(\alpha)}-1)(n^{(\alpha)}-2)- \\ \qquad -\sum_h t_h^{(\alpha)}(t_h^{(\alpha)}-1)(t_h^{(\alpha)}-2)\} & \text{if } n^{(\alpha)} \geqq 3 \\ 0 & \text{if } n^{(\alpha)} < 3. \end{cases} \tag{7.3}$$

The other definitions and abbreviations remain unchanged. Theorem I and its consequences remain valid if $v_{2,m}$ is defined as the number of vectors $\boldsymbol{x}^{(\alpha)}$ whose components assume under H_0' at least three different values, $u_0^{(\alpha)}$ *included*. If $v_{1,m}$ is the number of vectors $\boldsymbol{x}^{(\alpha)}$ whose components assume under H_0' at least two different values $u_0^{(\alpha)}$ *excluded* the remark at the end of section 4 holds.

Finally I want to thank Prof. Dr D. VAN DANTZIG for his helpful suggestions which gave the paper its final form, CONSTANCE VAN EEDEN who read the paper thouroughly and J. TH. RUNNENBURG, A. BENARD and J. FABIUS who suggested many improvements.

Mathematical Centre, Amsterdam

REFERENCES

BENARD, A. and PH. VAN ELTEREN, A generalization of the method of m rankings, Proc. Kon. Ned. Ak. v. Wet. **A 56**, 358–369, Indagationes Mathematicae **15**, 358–369 (1953).

DANTZIG, D. VAN and J. HEMELRIJK, Statistical methods based on a few assumptions, Bull. of the Int. Stat. Inst. **24**, 2, 239–267 (1954).

FRIEDMAN, M., The use of ranks to avoid the assumption of normality implicit in the analysis of variance, Journ. of the Am. Stat. Ass. **32**, 675–699 (1937).

KENDALL, M. G. and B. BABINGTON SMITH, The problem of m rankings, Ann. of Math. Statistics **10**, 275–287 (1939).

KENDALL, M. G., The treatment of ties in ranking problems, Biometrika **33**, 239–251 (1945).

————, Rank correlation methods (2nd edition) Ch. Griffin & Co, London (1955).

RAO, C. R., Advanced statistical methods in Biometric Research, J. Wiley & Sons, New York, Chapman & Hall, London (1952).

ROBBINS, H. and E. J. G. PITMAN, Application of the method of mixtures to quadratic forms in normal variates, Ann. of Math. Statistics **20**, 552–560 (1949).

TERPSTRA, T. J., A generalization of Kendall's rank correlation statistic, Proc. Kon. Ned. Ak. v. Wet. **A 58**, 690–696 and **A 59**, 59–66, Indagationes Mathematicae **17**, 690–696 and **18**, 59–66 (1955).

USPENSKY, J. V., Introduction to mathematical probability, Mc Graw Hill Book Comp., New York and London (1937).

0204