A THEOREM ON RANK ORDERS FOR TWO CENSORED SAMPLES[1]

By Carl-Erik Särndal

University of Lund

0. Summary. Let m and n be the sizes, respectively, of two independent random samples both of which may be censored in an arbitrary manner so that $h (1 \leqq h \leqq m)$ and $k (1 \leqq k \leqq n)$ observations, respectively, remain. If arranged in ascending order, the remaining observations can appear in $\binom{h+k}{h}$ possible rank orders. In this paper we prove a theorem which is useful in obtaining the probability associated with any one of these rank orders, provided the two samples are drawn from populations with identical distribution functions.

1. Introduction. Let $F(x)$ and $G(y)$ be continuous cumulative distribution functions. Denote by $x_{(1)} < x_{(2)} < \cdots < x_{(m)}$, the ordered observations in a sample of m from $F(x)$, and let $y_{(1)} < y_{(2)} < \cdots < y_{(n)}$ be an independent ordered sample of size n from $G(y)$. Select any $h \leqq m$ out of the x's and any $k \leqq n$ out of the y's, say

$$x_{(m_1)} < x_{(m_2)} < \cdots < x_{(m_h)}; \qquad y_{(n_1)} < y_{(n_2)} < \cdots < y_{(n_k)},$$

where the indices $m_1, \cdots, m_h$; $n_1, \cdots, n_k$ are fixed integers. Define

$$m_0 = n_0 = 0, \qquad m_{h+1} = m + 1, \qquad n_{k+1} = n + 1,$$

$$x_{(m_0)} = y_{(n_0)} = -\infty, \qquad x_{(m_{h+1})} = y_{(n_{k+1})} = \infty.$$

The number of x's in the interval $(x_{(m_{i-1})}, x_{(m_i)})$ is $a_i = m_i - m_{i-1} - 1$ $(i = 1, \cdots, h + 1)$, and the number of y's in the interval $(y_{(n_{i-1})}, y_{(n_i)})$ is $b_i = n_i - n_{i-1} - 1$ $(i = 1, \cdots, k + 1)$. By definition the a's and the b's are fixed non-negative integers, and $\sum_{i=1}^{h+1} a_i = m - h$, $\sum_{i=1}^{k+1} b_i = n - k$.

Assume $F(x) = G(y)$, let $K = h + k$, and consider the combined sample of the selected x's and y's. Denote the elements of the common ordering by

$$z_{(N_1)} < z_{(N_2)} < \cdots < z_{(N_K)}, \tag{1.1}$$

and call it a z sequence. Defining $z_{(N_0)} = -\infty$, $z_{(N_{K+1})} = \infty$, we have $K + 1$ mutually exclusive intervals $(z_{(N_{\nu-1})}, z_{(N_\nu)})$ $(\nu = 1, \cdots, K + 1)$. (The probability of ties is zero.) Denote the number of x's (y's) in the interval $(z_{(N_{\nu-1})}, z_{(N_\nu)})$ by $\alpha_\nu(\beta_\nu)$. The α's and β's will be referred to as *cell frequencies*. Some of the α's and β's are fixed integers, others are discrete random variables.

Clearly, $\alpha_\nu(\beta_\nu)$ is a fixed interger and identically the same as one of of the a's (b's) if either of the conditions,

(i) $\nu = 1$ and $z_{(N_1)}$ is an $x(y)$,

(ii) $\nu = K + 1$ and $z_{(N_K)}$ is an $x(y)$,

Received 16 June 1964.

[1] Sponsored by the United States Army Research Office (Durham).

(iii) both $z_{(N_{\nu-1})}$ and $z_{(N_\nu)}$ $(\nu = 2, \cdots, K)$ are x's (y's),
holds. We have

$$\sum_{\nu=1}^{K+1} \alpha_\nu = \sum_{i=1}^{h+1} a_i = m - h, \qquad \sum_{\nu=1}^{K+1} \beta_\nu = \sum_{i=1}^{k+1} b_i = n - k,$$

and if none of the three conditions stated above holds, then the α's(β's) are discrete random variables subject to linear restrictions implying that two or more of them add up to a fixed integer $a_i(b_i)$. These points are clarified by an example.

EXAMPLE. Let $h = 3$, $k = 4$, and consider the z sequence $x_{(m_1)} < y_{(n_1)} < x_{(m_2)} < y_{(n_2)} < y_{(n_3)} < y_{(n_4)} < x_{(m_3)}$. For brevity, let us in the future denote such a sequence simply by $x\ y\ x\ y\ y\ y\ x$. The sequence can be more elaborately described in the following way:

$$\begin{array}{cccccccc} \underline{\alpha_1} & \underline{\alpha_2\ \ \alpha_3} & & \underline{\alpha_4\ \ \alpha_5\ \ \alpha_6\ \ \alpha_7} & & \underline{\alpha_8} \\ a_1 & a_2 & & a_3 & & a_4 \end{array}$$

$$\begin{array}{cccccc} \underline{\beta_1\ \ \beta_2} & \underline{\beta_3\ \ \beta_4} & \underline{\beta_5} & \underline{\beta_6} & \underline{\beta_7\ \ \beta_8} \\ b_1 & b_2 & b_3 & b_4 & b_5 \end{array}.$$

The sum of cell frequencies underscored by the same line equals the a or b written just below the line. For example, β_1 and β_2 are random cell frequencies, one of them being redundant since their sum equals the fixed integer b_1. The number of lines used in underscoring the α's(β's) is $h + 1 = 4$, $(k + 1 = 5)$.

For the future we make it a rule to consider as redundant the last $\alpha(\beta)$ in a sequence of α's (β's) that add up to a fixed integer $a_i(b_i)$. Then, with the example considered above we can associate a vector composed of seven non-redundant random cell frequencies, namely, $(\beta_1, \alpha_2, \beta_3, \alpha_4, \alpha_5, \alpha_6, \beta_7)$. Noting that the z sequence is $x\ y\ x\ y\ y\ y\ x$, the corresponding vector of random cell frequencies is obtained by entering a $\beta(\alpha)$ in the place of an $x(y)$, and furnishing each α and β with the proper order index. This holds, of course, for any z sequence and for any values of h and k. Thus, the vector will contain h β's and k α's.

The main result of this paper is a theorem that gives the probability associated with any vector of random cell frequencies. As a byproduct the probability of any of the $\binom{K}{h}$ different z sequences can be computed.

2. Further notation. It will prove useful to introduce an alternative way of denoting the cell frequencies α_ν and β_ν $(\nu = 1, \cdots, K + 1)$. To this end, let r_i $(i = 1, \cdots, h + 1)$ be positive integers such that $\sum_{i=1}^{h+1} r_i = K + 1$. Assume that the sequences $\alpha_1, \alpha_2, \cdots, \alpha_{K+1}$ and $a_{11}, \cdots, a_{1r_1}, a_{21}, \cdots, a_{2r_2}, a_{31}, \cdots, a_{h+1,r_{h+1}}$ are identically the same, element by element, from left to right. Assume also that $\sum_{j=1}^{r_i} a_{ij} = a_i$ $(i = 1, \cdots, h + 1)$.

Similarly, let s_i $(i = 1, \cdots, k + 1)$ be positive integers such that $\sum_{i=1}^{k+1} s_i = K + 1$. Assume that the sequences $\beta_1, \beta_2, \cdots, \beta_{K+1}$ and $b_{11}, \cdots, b_{1s_1}, b_{21}, \cdots, b_{2s_2}, b_{31}, \cdots, b_{k+1,s_{k+1}}$ are identically the same, element by element, from left to right. Assume also that $\sum_{j=1}^{s_i} b_{ij} = b_i$ $(i = 1, \cdots, k + 1)$.

Referring to the example of the previous section, $r_i(s_i)$ is the number of α's(β's) underscored by the ith line, the lines being counted from left to right.

Obviously, $r_i \geqq 1 (i = 1, \cdots, h+1)$, $s_i \geqq 1 (i = 1, \cdots, k+1)$. At least two among the $h + k + 2$ integers r_i, s_i equal unity (two that equal unity are obtained from the extreme cells). If $r_i(s_i)$ equals unity, then $a_{i1} = a_i$ ($b_{i1} = b_i$), i.e., the cell frequency $a_{i1}(b_{i1})$ is a fixed integer (one of conditions (i), (ii), and (iii) of Section 1 is fulfilled). On the other hand, if $r_i(s_i)$ exceeds unity, then

$$\sum_{j=1}^{r_i} a_{ij} = a_i \ (\sum_{j=1}^{s_i} b_{ij} = b_i), \tag{2.1}$$

where $a_i(b_i)$ is fixed, which means that $a_{i1}, \cdots, a_{i,r_i-1}(b_{i1}, \cdots, b_{i,s_i-1})$ are non-redundant discrete random variables to be entered as components of the vector of random cell frequencies.

Consider the z sequence (1.1) consisting of h x's and k y's. Denote the observed rank order by $z = (z_1, \cdots, z_K)$, where $z_\nu = 0(1)$ if $z_{(N_\nu)}$ is an $x(y)$. Furthermore, let $\gamma = (\gamma_1, \cdots, \gamma_K)$ be the corresponding vector of observed random cell frequencies, i.e., $\gamma_\nu = \alpha_\nu(\beta_\nu)$ if $z_\nu = 1(0)$. Denote the vector random variables corresponding to z and γ by Z and Γ, respectively. The probability that $\Gamma = \gamma$ is thus to be interpreted as the probability that $Z = z$ and that, in addition, the magnitudes of the individual random cell frequencies are as specified by the components of γ. Thus, $P(Z = z) = \sum P(\Gamma = \gamma)$, where the summation is a multiple one extending over all components of γ which are non-negative integers fulfilling the restraints imposed by (2.1), i.e.,

$$\sum_{j=1}^{r_i-1} a_{ij} \leqq a_i \quad (i = 1, \cdots, h+1; r_i > 1), \tag{2.2}$$

$$\sum_{j=1}^{s_i-1} b_{ij} \leqq b_i \quad (i = 1, \cdots, k+1; s_i > 1). \tag{2.3}$$

3. The theorem. We shall prove the following

THEOREM. *If $F(x) = G(y)$, then the probability associated with any vector of random cell frequencies can be written in the form*

$$P(\Gamma = \gamma) = \prod_{\nu=1}^{K+1} \binom{\alpha_\nu + \beta_\nu}{\alpha_\nu} \Big/ \binom{m+n}{m}.$$

We recall that if $\gamma_\nu = \alpha_\nu$ then β_ν is given, and, conversely, if $\gamma_\nu = \beta_\nu$, then α_ν is given.

PROOF. Let us find the joint probability of the event $\Gamma = \gamma$, $X = x$, $Y = y$, where $x = (x_{(m_1)}, x_{(m_2)}, \cdots, x_{(m_h)})$, $y = (y_{(n_1)}, y_{(n_2)}, \cdots, y_{(n_k)})$, and X, Y denote the corresponding random vectors. We have

$$\begin{aligned} &P(X = x, Y = y, \Gamma = \gamma) \\ &= \frac{m!}{\prod_{i=1}^{h+1} a_i!} \prod_{i=1}^{h+1} \left\{ \frac{a_i!}{\prod_{j=1}^{r_i} a_{ij}!} \prod_{j=1}^{r_i} (F_{ij} - F_{i,j-1})^{a_{ij}} \right\} \prod_{i=1}^{h} f(x_{(m_i)})\, dx_{(m_i)} \\ &\cdot \frac{n!}{\prod_{i=1}^{k+1} b_i!} \prod_{i=1}^{k+1} \left\{ \frac{b_i!}{\prod_{j=1}^{s_i} b_{ij}!} \prod_{j=1}^{s_i} (F_{ij} - F_{i,j-1})^{b_{ij}} \right\} \prod_{i=1}^{k} f(y_{(n_i)})\, dy_{(n_i)}, \end{aligned} \tag{3.1}$$

where the three sequences (with $K + 2$ elements each)

$$F_{10}, F_{11}, \cdots, F_{1r_1} = F_{20}, \cdots, F_{2r_2} = F_{30}, \cdots, F_{h+1,r_{h+1}},$$

$$F_{10}, F_{11}, \cdots, F_{1s_1} = F_{20}, \cdots, F_{2s_2} = F_{30}, \cdots, F_{k+1,s_{k+1}},$$

$$0 = F(z_{(N_0)}), F(z_{(N_1)}), \cdots, F(z_{(N_K)}), F(z_{(N_{K+1})}) = 1$$

are identically the same, element by element, from left to right. Summing over all the relations (2.2) and (2.3), thus eliminating the vector γ, we obtain the marginal probability

$$P(X = x, Y = y) = \frac{m!}{\prod_{i=1}^{h+1} a_i!} \prod_{i=1}^{h+1} [F(x_{(m_i)}) - F(x_{(m_{i-1})})]^{a_i} \prod_{i=1}^{h} f(x_{(m_i)})\, dx_{(m_i)}$$

$$\cdot \frac{n!}{\prod_{i=1}^{k+1} b_i!} \prod_{i=1}^{k+1} [F(y_{(n_i)}) - F(y_{(n_{i-1})})]^{b_i} \prod_{i=1}^{k} f(y_{(n_i)})\, dy_{(n_i)}.$$

i.e., the joint probability of the order statistics under consideration, the two samples being independent.

We are, however, mainly interested in the marginal probability for γ which is obtained by eliminating the vectors X, Y through integration. We note that (3.1) can be rewritten in the form

$$P(X = x, Y = y, \Gamma = \gamma) = \frac{m!}{\prod_{i=1}^{h+1} \prod_{j=1}^{r_i} a_{ij}!} \prod_{i=1}^{h+1} \prod_{j=1}^{r_i} (F_{ij} - F_{i,j-1})^{a_{ij}}$$

$$\cdot \frac{n!}{\prod_{i=1}^{k+1} \prod_{j=1}^{s_i} b_{ij}!} \prod_{i=1}^{k+1} \prod_{j=1}^{s_i} (F_{ij} - F_{i,j-1})^{b_{ij}} \cdot \prod_{\nu=1}^{K} f(z_{(N_\nu)})\, dz_{(N_\nu)}$$

$$= \frac{m!}{\prod_{\nu=1}^{K+1} \alpha_\nu!} \frac{n!}{\prod_{\nu=1}^{K+1} \beta_\nu!} \prod_{\nu=1}^{K+1} (F_\nu - F_{\nu-1})^{\alpha_\nu+\beta_\nu} \prod_{\nu=1}^{K} f(z_{(N_\nu)})\, dz_{(N_\nu)},$$

where $F_\nu = F(z_{(N_\nu)})(\nu = 0, 1, \cdots, K+1)$. Integrating the last expression over $-\infty < z_{(N_1)} < z_{(N_2)} < \cdots < z_{(N_K)} < \infty$ we obtain the result stated by the theorem, and the proof is complete.

The theorem can be used to compute the probability of any of the $\binom{K}{h}$ different z sequences (rank orders) obtained from the h x's and the k y's. As pointed out in Section 1, this probability is obtained by summing $P(\Gamma = \gamma)$ over non-negative γ_ν fulfilling those linear restrictions that are pertinent to the sequence under consideration. These summations are easily performed explicitly over either the α_ν or the β_ν by the use of the following formulas.

Let p and q be integers satisfying $1 \leqq p < q \leqq K+1$, and set $\sum_{\nu=p}^{q} \alpha_\nu = c$, $\sum_{\nu=p}^{q} \beta_\nu = d$. Then

$$\sum \prod_{\nu=p}^{q} \binom{\alpha_\nu+\beta_\nu}{\alpha_\nu} = \binom{c+q-p+d}{c}, \tag{3.2}$$

where the summation extends over all non-negative $\alpha_\nu (\nu = p, \cdots, q-1)$ such that $\sum_{\nu=p}^{q-1} \alpha_\nu \leqq c$. Similarly,

$$\sum \prod_{\nu=p}^{q} \binom{\alpha_\nu+\beta_\nu}{\alpha_\nu} = \binom{c+q-p+d}{d}, \tag{3.3}$$

where the summation instead extends over all non-negative β_ν $(\nu = p, \cdots, q-1)$ fulfilling $\sum_{\nu=p}^{q-1} \beta_\nu \leqq d$.

EXAMPLE. Let us compute the probability of the z sequence $x\ y\ x\ y\ y\ x$ considered in the example of Section 1. Letting $z = (0, 1, 0, 1, 1, 1, 0)$ we obtain by summing over the α's using formula (3.2),

$$P(Z = z) = \sum \binom{a_1+\beta_1}{a_1}\binom{a_2+1+\beta_2+\beta_3}{a_2}\binom{a_3+3+\beta_4+\beta_5+\beta_6+\beta_7}{a_3}\binom{a_4+\beta_8}{a_4},$$

where the remaining summations are over non-negative β's such that $\beta_1 + \beta_2 = b_1$, $\beta_3 + \beta_4 = b_2$, $\beta_5 = b_3$, $\beta_6 = b_4$, $\beta_7 + \beta_8 = b_5$.

4. Applications. As a first application, consider the case $h = k = 1$. We have $a_1 = m_1 - 1$, $a_2 = m - m_1$, $b_1 = n_1 - 1$, $b_2 = n - n_1$. If the z sequence is $x\ y$ (i.e., $x_{(m_1)} < y_{(n_1)}$), then $\alpha_1 = a_1$, $\alpha_2 + \alpha_3 = a_2$, $\beta_1 + \beta_2 = b_1$, $\beta_3 = b_2$, while, if the z sequence is $y\ x$ (i.e., $y_{(n_1)} < x_{(m_1)}$), then $\alpha_1 + \alpha_2 = a_1$, $\alpha_3 = a_2$, $\beta_1 = b_1$, $\beta_2 + \beta_3 = b_2$. Let $\Gamma_1 = (\beta_1, \alpha_2)$, $\Gamma_2 = (\alpha_1, \beta_2)$. According to the theorem,

$$P(\Gamma = \Gamma_1) = P(\Gamma = \Gamma_2) = \prod_{\nu=1}^{3} \binom{\alpha_\nu+\beta_\nu}{\alpha_\nu} \Big/ \binom{m+n}{m}. \tag{4.1}$$

The same thing can be written in a more detailed manner by inserting into (4.1) those α's and β's which are fixed or given by the linear bands. Thus,

$$P(\Gamma = \Gamma_1) = \binom{a_1+\beta_1}{\beta_1}\binom{\alpha_2+b_1-\beta_1}{\alpha_2}\binom{a_2-\alpha_2+b_2}{b_2}\Big/\binom{m+n}{m}, \quad 0 \leqq \beta_1 \leqq b_1, \quad 0 \leqq \alpha_2 \leqq a_2,$$

$$P(\Gamma = \Gamma_2) = \binom{\alpha_1+b_1}{\alpha_1}\binom{a_1-\alpha_1+\beta_2}{\beta_2}\binom{a_2+b_2-\beta_2}{a_2}\Big/\binom{m+n}{m}, \quad 0 \leqq \alpha_1 \leqq a_1, \quad 0 \leqq \beta_2 \leqq b_2.$$

In order to obtain the probability that, for instance, $X_{(m_1)} < Y_{(n_1)}$, let $z = (0, 1)$. Then, by eliminating α_2 through (3.2),

$$P(Z = z) = \sum_{\nu=0}^{b_1} \binom{a_1+\nu}{a_1}\binom{a_2+n-\nu}{a_2}\Big/\binom{m+n}{m}. \tag{4.2}$$

This expression may also be obtained by reasoning that at most $b_1 = (n_1 - 1)$ y's are allowed to be $< x_{(m_1)}$ in order for $x_{(m_1)} < y_{(n_1)}$ to hold. Formula (4.2) was originally derived by Thompson [3], compare also [1], pp. 395–397. By summing over β_1 instead we obtain by using (3.3) the equivalent expression

$$P(Z = z) = \sum_{\nu=0}^{a_2} \binom{m-\nu+b_1}{b_1}\binom{\nu+b_2}{b_2}\Big/\binom{m+n}{m}.$$

As another application we choose the case $h = m$, $k = n$, i.e., all the x's and the y's are selected to start with. Then all cell frequencies $\alpha_\nu = \beta_\nu = 0$ $(\nu = 1, \cdots, K+1)$. For any z sequence, the vector of random cell frequencies is $(0, 0, \cdots, 0)$, and we obtain by the theorem the result that all z sequences are equally probable, each having probability $1/\binom{m+n}{m}$, a well-known result widely used in the theory of rank order statistics.

Possible applications of the theorem lie in the field of designing tests based on rank order statistics in various cases of censoring of one or both samples. The

general form of censoring considered in the theorem makes it possible to deal with a wide range of censoring schemes.

Rank tests in connection with a simple censoring situation were considered in [2]. It was assumed that the experiment is discontinued after the N^* smallest observations of the combined sample of $N = m + n$ have been observed. If N^* contains m^* x's and n^* y's, then the probability of any sequence of the N^* z's is, by our theorem,

$$\binom{m-m^*+n-n^*}{m-m^*} \Big/ \binom{m+n}{m},$$

which confirms the result obtained in [2].

REFERENCES

[1] MOOD, ALEXANDER MCFARLANE (1950). *Introduction to the Theory of Statistics*. McGraw-Hill, New York.

[2] RAO, U. V. R., SAVAGE, I. R., and SOBEL, M., (1960). Contributions to the theory of rank order statistics: The two sample censored case. *Ann. Math. Statist.* **31** 415–426.

[3] THOMPSON, WILLIAM R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **25** 285–294.