ORDER STATISTICS OF A SAMPLE AND OF AN EXTENDED SAMPLE FROM DISCRETE DISTRIBUTIONS

VISHNU DAYAL JHA and A. P. KHURANA
University of Indore, Indore

(Received : February, 1982)

SUMMARY

Let $X_{i:n}$ be the ith order statistic given by

$$X_{1:n} \leqslant X_{2:n} \leqslant \ldots \leqslant X_{n:n}$$

when n independent observations X_i, $i = 1, 2, \ldots, n$ are arranged in the ascending order of magnitudes. Some of the results obtained by Siddiqui [2] are extended for discrete distributions. Applications relating to testing the hypothesis and testing the outlier are also stated using the joint distribution of $\{X_{k:n}, X_{k+s:n+m}\}$.

Introduction

Let $X_{i:n}$ be the ith order statistic given by

$$X_{1:n} \leqslant X_{2:n} \leqslant \ldots \leqslant X_{n:n}$$

when n independent observations X_i, $i = 1, 2, \ldots, n$ are arranged in the ascending order of magnitudes. The joint distribution of $\{X_{k:n}, X_{k+s:n+m}\}$ for any k, s, $1 \leqslant k \leqslant n$, $1 \leqslant k + s \leqslant n + m$ has been studied by Siddiqui [2], assuming absolute continuity of the distribution function under the following hypotheses:

$H_0 : X_i$, $i = 1, 2, \ldots, n + m$, are identically distributed with a common distribution function $F(x)$. $H_1 : X_i$, $i = 1, 2, \ldots, n$, are identically distributed with a common distribution function $F(x)$ and X_i, $i = n + 1, n + 2, \ldots, n + m$, are identically distributed with a common distribution function $G(x)$.

Some of the results obtained by Siddiqui [2] are extended for discrete

distributions. Applications relating to testing the hypothesis H_0 and testing the outlier are also stated in section 4 using the joint distribution of

$$\{X_{k:n},\ X_{k+s:n+m}\}$$

Let X_i be integral valued random variables taking the values 0, 1, . . . , with respective probabilities $f(0), f(1), \ldots$, for $i = 1, 2, \ldots, n$ and with respective probabilities $g(0), g(1), \ldots$, for $i = n+1, n+2, \ldots, n+m$. The distribution functions are given by

$$F(x) = \sum_{i \leqslant x} f(i) \quad \text{and} \quad G(x) = \sum_{i \leqslant x} g(i).$$

2. Probability that $X_{k:n} = X_{k+s:n+m}$ under H_1

The probability measures under H_0 and H_1 are denoted by P_0 and P_1 respectively.

Now

$$P_1(X_{k+s:n+m} = X_{k:n}) = \sum_{x=0}^{\infty} P_1(X_{k+s:n+m} = x \mid X_{k:n} = x)\, P_1(X_{k:n} = x).$$

Consider the event $(X_{k+s:n+m} = x \mid X_{k:n} = x)$. This event can occur if and only if, out of m observations from $G(x)$, $s - j$ observations are less than x, $j + i$ observations are equal to x and $m - s - i$ observations are greater than x with respective probabilities $G(x-1)$, $g(x)$ and $1 - G(x)$; $(j = 0, 1, \ldots, s;\ i = 0, 1, \ldots, m - s)$.

Hence

$$P_1(X_{k+s:n+m} = x \mid X_{k:n} = x) = \sum_{i=0}^{m-s} \sum_{j=0}^{s} C(i, j, s, m)\, G^{s-j}(x-1) \cdot g^{j+i}(x)\, (1 - G(x))^{m-s-i} \qquad (2.1)$$

here $G(x-1) = 0$ for $x = 0$ and

$$C(i, j, s, m) = \frac{m!}{(s-j)!\,(j+i)!\,(m-s-i)!}$$

(2.1) can also be written as

$$P_1(X_{k+s:n+m} = x \mid X_{k:n} = x)$$

$$= \binom{m}{s} \sum_{i=0}^{m-s} \sum_{j=0}^{s} \binom{s}{j} \binom{m-s}{i} \frac{\Gamma(j+1)\,\Gamma(i+1)}{\Gamma(i+j+1)}$$

$$\cdot\, G^{s-j}(x-1)\, g^{j+i}(x)\, (1-G(x))^{m-s-i}$$

$$= \binom{m}{s} \sum_{i=0}^{m-s} \sum_{j=0}^{s} \binom{s}{j} \binom{m-s}{i} i\, G^{s-j}(x-1)$$

$$\cdot\, (1-G(x))^{m-s-i} \int_0^1 (wg(x))^j\, (g(x)(1-w))^{i-1}\, g(x)dw \quad (2.2)$$

Interchanging the summation and integral signs, we have

$$P_1(X_{k+s:n+m} = x \mid X_{k:n} = x)$$

$$= \binom{m}{s} (m-s) \int_0^1 (wg(x) + G(x-1))^s\, (g(x)\,(1-w)$$

$$+ 1 - G(x))^{m-s-i}\, g(x)dw \quad (2.3)$$

Putting $wg(x) + G(x-1) = v$, we have

$$P_1(X_{k+s:n+m} = x \mid X_{k:n} = x) = (m-s) \binom{m}{s} \int_{G(x-1)}^{G(x)} v^s (1-v)^{m-s-1}\, dv \quad (2.4)$$

Hence $P_1(X_{k+s:n+m} = X_{k:n})$

$$= k(m-s) \binom{m}{s} \binom{n}{k} \sum_{x=0}^{\infty} \int_{G(x-1)}^{G(x)} \int_{F(x-1)}^{F(x)}$$

$$\cdot\, v^s (1-v)^{m-s-1}\, w^{k-1}\, (1-w)^{n-k}\, dv\, dw \quad (2.5)$$

$$= \sum_{x=0}^{\infty} [I_{G(x)}(s+1, m-s) - I_{(Gx-1)}(s+1, m-s)]$$

$$[I_{F(x)}(k, n-k+1) - I_{F(x-1)}(k, n-k+1)] \quad (2.6)$$

where $I_P(a, b)$ is defined by

$$I_P(a, b) = \frac{1}{B(a, b)} \int_0^P t^{a-1}(1-t)^{b-1}\, dt; \quad a, b > 0$$

and tabulated by Pearson (1934).

Under H_0,

$$P_0(X_{k+s:n+m} = X_{k:n}) = \sum_{x=0}^{\infty} [I_{F(x)}(s+1, m-s) - I_{F(x-1)}(s+1, m-s)]\,[I_{F(x)}(k, n-k+1) - I_{F(x-1)}(k, n-k+1)] \qquad (2.7)$$

3. Joint Distribution of $(X_{k:n}, X_{k+s:n+m})$ under Alternative H_1

Let $h(x, y) = P(X_{k:n} = x, X_{k+s:n+m} = y)$.

Case I. $y = x$

Then $h(x, x) = P(X_{k+s:n+m} = x \mid X_{k:n} = x)\, P(X_{k:n} = x)$

Case II. $y > x$.

(i) Among $X_1, X_2, \ldots, X_n$, let $k - a_1 - 1$ observations be $< x$, $a_1 + a_2 + 1$ observations be $= x$, $i - a_2 - b_1$ observations be $> x$ but $< y$, $b_1 + b_2 + 1$ observations be $= y$, $n - k - b_2 - i - 1$ observations be $> y$ and among $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$, $s - i - c_1 - 1$ observations be $< y$, $m - s - c_2 + i + 1$ observations be $> y$ and $c_1 + c_2$ observations be $= y$.

(ii) Among $X_1, X_2, \ldots, X_n$, let $k - a_1 - 1$ observations be $< x$, $a_1 + a_2 + 1$ observations be $= x$, $i - a_2 - b_1$ observations be $> x$ but $< y$, $b_1 + b_2$ observations be $= y$, $n - k - i - b_2$ observations be $> y$ and among $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$, $s - i - c_1 - 1$ observations be $< x$, $c_1 + c_2 + 1$ observations be $= y$, and $m - s - c_2 + i$ observations be $> y$.

These two cases are equally likely except when $b_1 = b_2 = c_1 = c_2 = 0$.

$$\begin{aligned} h(x, y) &= P(X_{k:n} = x, X_{k+s:n+m} = y \mid X_{k+s:n+m} \in F(x)) \\ &\quad + P(X_{k:n} = x, X_{k+s:n+m} = y \mid X_{k+s:n+m} \in G(x)) \\ &= 2P(X_{k:n} = x, X_{k+s:n+m} = y \mid X_{k+s:n+m} \in F(x)) \end{aligned}$$

when b_1, b_2, c_1, c_2 are not all equal to zero.

Now

$$P(X_{k:n} = x,\ X_{k+s:n+m} = y \mid X_{k+s:n+m} \in F(x))$$

$$= \sum_{i=0}^{n} \sum_{c_2=0}^{m-s+i+1} \sum_{c_1=0}^{s-i-1} \sum_{b_2=0}^{n-k-i-1} \sum_{b_1=0}^{i} \sum_{a_2=0}^{i-b_1} \sum_{a_1=0}^{k-1}$$

$$\cdot \frac{n!\,m!}{(k-a_1-1)!\,(i-a_2-b_1)!\,(n-k-b_2-i-1)!\,(a_1+a_2+1)!\,(b_1+b_2+1)!}$$

$$\cdot \frac{1}{(s-i-c_1-1)!\,(m-s-c_2+i+1)!\,(c_1+c_2)!}\, F_{(x-1)}^{k-a_1-1}$$

$$\cdot [G(y-1)]^{s-i-c_1-1}\,[1-G(y)]^{m-c_2+i+1-s}\,[F(y-1)-F(x)]^{i-a_2-b_1}$$

$$\cdot [f(x)]^{a_1+a_2+1}[1-F(y)]^{n-k-b_1-i-1}[f(y)]^{b_1+b_2+1}[g(y)]^{c_1+c_2},\ i > a_2 + b_1$$

Using the technique applied in section 2, we have, on simplication.

$$2P_1(X_{k:n} = x,\ X_{k+s:n+m} = y \mid X_{k+s:n+m} \in F(x))$$

$$= 2\sum_{i=0}^{n} \frac{n!\,m!}{(m-s+i)!\,i!\,(k-1)!}$$

$$\cdot \frac{1}{(n-k-i+1)!\,(s-i-1)!} \int_{G(y-1)}^{G(y)} \int_{F(y-1)}^{F(y)} \int_{F(x-1)}^{F(x)}$$

$$\cdot u^{k-1}(z-u)^i\,(1-z)^{n-k-1-i} \cdot w^{s-i-1}(1-w)^{n-s+1}\,du\,dz\,dw \tag{3.1}$$

Case III. $y < x$.

(i) Among $X_1, X_2, \ldots, X_n$, let $k-i-b_1-2$ observations be $<y$, b_1+b_2+1 observations be $=y$, $i-a_2-b_2$ observations be $>y$ but $<x$, a_1+a_2+1 obseravations be $=x$ and $n-a_2-k$ observations be $>x$, and among $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$, $s+i-c_1+1$ observations be $<y$, c_1+c_2 observations be $=y$ and $m-s-i-c_2-1$ observations be $>y$.

(ii) Among $X_1, X_2, \ldots, X_n$, let $k-i-b_1-1$ observations be $<y$, b_1+b_2 observations be $=y$, $i-a_1-b_2$ observations be $>y$ but $<x$, a_1+a_2+1 observations be $=x$ and $n-k-a_2$ observations be $>x$, and among $X_{n+1}, X_{n+2}, \ldots, X_{n+m}$, let $s+i+c_1$ observations be $<y$, c_1+c_2+1 observations be $=y$ and $n-s-i-c_2-1$ observations be $>y$.

These two cases are equally likely except when $c_1 = c_2 = b_1 = b_2 = 0$.

Thus

$$h(x,\ y) = 2 \sum_{i=0}^{n} \sum_{c_2=0}^{m-s-i-1} \sum_{c_1=0}^{s+i+1} \sum_{a_2=0}^{n-k} \sum_{a_1=0}^{i} \sum_{b_2=0}^{i-a_1} \sum_{b_1=0}^{k-i-2}$$

$$\cdot \frac{n\,!\,m\,!}{(k-i-b_1-2)\,!\,(i-a_1-b_2)\,!\,(n-a_2-k)\,!\,(s+i-c_1+1)\,!\,(b_1+b_2+1)\,!}$$

$$\cdot \frac{1}{(m-s+i-c_2-1)\,!\,(a+a_2+1)\,!\,(c_1+c_2)\,!}\,[F(y-1)]^{k-i-b_1-2}$$

$$\cdot [F(x-1)-F(y)]^{i-a_1-b_2}\,[1-F(x)]^{n-a_2-k}\,[G(y-1)]^{s+i-c_1+1}$$

$$\cdot [f(y)]^{b_1+b_2+1}\,[1-G(y)]^{m-s+i-c_2-1}\,[f(x)]^{a_1+a_2+1}\,[g(y)]^{c_1+c_2}$$

After simplification we have

$$h(x,\ y) = 2 \sum_{i=0}^{n} \frac{n\,!\,m\,!}{(s+i+1)\,!\,(m-s-i-2)\,!\,i\,!\,(n-k)\,!\,(k-i-2)\,!}$$

$$\cdot \int_{F(y-1)}^{F(y)} \int_{F(x-1)}^{F(x)} \int_{G(y-1)}^{G(y)} t^{k-i-2}\,(z-t)^{i}\ (1-z)^{n-k}\,w^{s+i+1}\,(1-w)^{m-s-i-2}$$

$$dt\,dz\,dw.$$

Let us also consider the case when $b_1 = b_2 = c_1 = c_2 = 0$

$$P(X_{k\,;\,n} = x,\ X_{k+s\,:\,n+m} = y \mid X_{k+s\,:\,n+m} \in F(x))$$

$$= \frac{n\,!\,m\,!}{(n-k)\,!} \sum_{i} \frac{1}{(k-i-2)\,!\,(s+i+1)\,!\,(m-s-i-1)\,!\,i!}$$

$$\cdot [F(y-1)]^{k-i-2}\,[G(y-1)]^{s+i+1}\,[1-G(y)]^{m-s-i-1}$$

$$\cdot \int_{F(x-1)}^{F(x)} (1-w)^{n-k}\,(w-F(y))^{i}\,dw$$

and

$$P(X_{k;n} = x,\ X_{k+s:n+m} = y \mid X_{k+s;n+m} \in G(x))$$

$$= \frac{n\,!\,m\,!}{(n-k)\,!} \sum_{i} \frac{1}{i\,!\,(k-i-1)\,!\,(s+i)\,!\,(m-s-i-1)\,!}\,[G(y-1)]^{s+i}\,g(y)$$

$$\cdot[1 - G(y)]^{m-s-i-1}\,[F(y-1)]^{k-i-1} \int_{F(x-1)}^{F(x)} (w - F(y))^i (1-w)^{n-k}\,dw$$

Hence under H_1

$$P_1(X_{k:n} = x,\ X_{k+s:n+m} = y)$$

$$= \frac{n!\,m!}{(n-m)!} \sum_i \frac{1}{(s+i)!\,(m-s-i-1)!\,i!}\ [G(y-1)]^{s+i}\,[F(y-1)]^{k-i-1}$$

$$\cdot[1 - G(y)]^{m-s-i-1} \left[\frac{G(y-1)}{(s+i+1)} + \frac{g(y)\,F(y-1)}{(k-i-1)}\right]$$

$$\int_{F(x-1)}^{F(x)} (1-w)^{n-k}\,(w - F(y))^i\,dw.$$

4. Applications

(*a*) For testing the hypothesis H_0 that X_i, $i = 1, 2, \ldots, n+m$ are identically distributed with a common distribution function $F(x)$, we can make use of the medians of sample and extended sample. Under H_0, the medians do not differ significantly on the intuition that almost half of the observations of the additional sample will be less than or equal to and the other half of them are greater than the median of the original sample.

Suppose that n is odd, the sample median would be $X_{k:n}$ where $k = n + 1/2$. For m even, the median of the extended sample is $X_{k+s:n+m}$ where $k + s = n + m + 1/2 \Rightarrow s = m/2$.

For given n and m, we can obtain k and s and making use of the result obtained in section 3 under H_0

$$P_0(X_{k+s:n+m} = X_{k:n}) = \sum_{x=0}^{\infty} [I_{F(x)}(s+1, m-s) - I_{F(x-1)}$$

$$(s+1, m-s)] \cdot [I_{F(x)}(k, n-k+1) - I_{F(x)}(k, n-k+1)] \quad (4.1)$$

Using (4.1) the test for testing H_0 would be developed. For given $F(x)$, $I_{F(x)}(s+1, m-s)$ etc. are obtained from the table due to Pearson [1]. For given α, the level of significance we accept H_0 if

$$P_0(X_{k+s:n+m} \neq X_{k:n}) < \alpha$$

or

$$P_0(X_{k+s:n+m} = X_{k:n}) > 1 - \alpha$$

(*b*) Using the largest order statisties of a sample of an extended sample we can develop the test for testing the presence of outlier on the intuition that for $X_{n:n}$ to be an outlier, all the additional observations should be less than $X_{n\,n}$, resulting the largest order statistic of the combined sample unchanged, that is $X_{n:n} = X_{m+n\,m+n}$ with probability one.

Let H_0^* be the hypothesis that $X_{n\,n}$ is not on outlier. The hypothesis H_0^* is to be accepted at level of significance α, if

$P_0(X_{n:n} = X_{m+n:m+n}) < \alpha$ where $P_0(X_{n:n} = X_{m+n\,n+m})$ is obtained by using the table of Incomplete Beta function for $k = n$, $s = m$, and for different values of $F(x)$.

ACKNOWLEDGEMENT

We are very grateful to the referee for his valuable suggestions.

REFERENCES

[1] Pearson, K. (1934) : *Tables of the Incomplete B-Functions*, Cambridge University Press.

[2] Siddiqui, M. M. (1970) : Order statistics of a sample and of an extended sample. In: M. L. Puri (ed.), *Nonparametric Techniques in Statistical Inference*; Cambridge University Press.