東海大学文理融合学部紀要　第２巻（2024年）73頁～76頁
Journal of the School of Humanities and Science, Tokai University, Vol.2. (2024) pp.73-76

Note on Joint Probability Density Function of Dependent Ordered Statistics and its Application

by
Tsunehisa IMADA*

(Received: October 30, 2023, Accepted: November 14, 2023)

Abstract

Assuming that there are plural dependent statistics and their joint probability density function is known, we arrange them in order of a size of value. In this study, we derive the explicit formulae of the probability density function of two statistics among the ordered statistics. Then, we derive the probability density function of the range of ordered statistics and discuss its applications.

Key Words : All-pairwise comparison, Critical value, Multiple comparison with a control, Range of ordered statistics

1. Introduction

Assume there are plural statistics $S_1, S_2,\cdots, S_K$. Arranging them in order of a size of value, assume

$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(K)}.$$

$S_{(1)}, S_{(2)}, \cdots, S_{(K)}$ are called the ordered statistics. When $S_1, S_2,\cdots, S_K$ are independent and are distributed according to a same probability distribution, the explicit formula of the joint probability density function of $S_{(l)}$ and $S_{(m)}$ for $1 \le l < m \le K$ is known. However, when $S_1, S_2,\cdots, S_K$ are dependent, the explicit formula of that is not found. Imada (2022) derived that for dependent $S_1, S_2,\cdots, S_K$ in a certain case.

In this study, under the assumption that $S_1, S_2,\cdots, S_K$ are dependent and their joint probability density function $f(s_1, s_2,\cdots, s_K)$ is known, we derive the explicit formula of the joint probability density function of $S_{(l)}$ and $S_{(m)}$ for $1 \le l < m \le K$. Then, we derive the probability density function of $R_{(l),(m)} = S_{(m)} - S_{(l)}$, which is the range of the ordered statistics

$$S_{(l)} \le S_{(l+1)} \le \cdots \le S_{(m)}.$$

Furthermore, we discuss its applications.

2. Joint probability density function of $S_{(l)}$ and $S_{(m)}$ for $1 \le l < m \le K$

We determine the joint probability density function $h_{(l),(m)}(s_{(l)}, s_{(m)})$ of $S_{(l)}$ and $S_{(m)}$ for $1 \le l < m \le K$. Since

$$h_{(l),(m)}(s_{(l)}, s_{(m)}) = \frac{\partial^2}{\partial s_{(l)} \partial s_{(m)}} P(S_{(l)} \le s_{(l)}, S_{(m)} \le s_{(m)}),$$

we derive the explicit formula of

$$\frac{\partial^2}{\partial s_{(l)} \partial s_{(m)}} P(S_{(l)} \le s_{(l)}, S_{(m)} \le s_{(m)}).$$

Assume $ds_{(l)} \ne 0, ds_{(m)} \ne 0$. Letting

$$G_{(l),(m)}(s_{(l)}, s_{(m)}) = P(S_{(l)} \le s_{(l)} + ds_{(l)}, S_{(m)} \le s_{(m)}) - P(S_{(l)} \le s_{(l)}, S_{(m)} \le s_{(m)}),$$

we obtain

$$h_{(l),(m)}(s_{(l)}, s_{(m)}) = \lim_{ds_{(l)} \to 0, ds_{(m)} \to 0} \times \frac{G_{(l),(m)}(s_{(l)}, s_{(m)} + ds_{(m)}) - G_{(l),(m)}(s_{(l)}, s_{(m)})}{ds_{(l)} ds_{(m)}}. \quad (2.1)$$

Assume $ds_{(l)} > 0, ds_{(m)} > 0$. Then

$$\begin{aligned} &G_{(l),(m)}(s_{(l)}, s_{(m)} + ds_{(m)}) - G_{(l),(m)}(s_{(l)}, s_{(m)}) \\ &= P(s_{(l)} \le S_{(l)} \le s_{(l)} + ds_{(l)}, \\ &\quad s_{(m)} \le S_{(m)} \le s_{(m)} + ds_{(m)}) \\ &= P(S_{(1)} \le s_{(l)}, \cdots, S_{(l-1)} \le s_{(l)}, \\ &\quad s_{(l)} \le S_{(l)} \le s_{(l)} + ds_{(l)}, s_{(l)} + ds_{(l)} \le S_{(l+1)} \le s_{(m)}, \\ &\quad \cdots, s_{(l)} + ds_{(l)} \le S_{(m-1)} \le s_{(m)}, \\ &\quad s_{(m)} \le S_{(m)} \le s_{(m)} + ds_{(m)}, s_{(m)} + ds_{(m)} \le S_{(m+1)}, \\ &\quad \cdots, s_{(m)} + ds_{(m)} \le S_{(K)}) \end{aligned}$$

* Professor , Department of Human Information Engineering, Tokai University

$$= \sum_{(q_1,\cdots,q_{l-1}),q_l,(q_{l+1},\cdots,q_{m-1}),q_m,(q_{m+1},\cdots,q_K)} \times P(S_{q_1} \le s_{(l)},\cdots,S_{q_{l-1}} \le s_{(l)}, s_{(l)} \le S_{q_l} \le s_{(l)} + ds_{(l)}, s_{(l)} + ds_{(l)} \le S_{q_{l+1}} \le s_{(m)}, \cdots, s_{(l)} + ds_{(l)} \le S_{q_{m-1}} \le s_{(m)}, s_{(m)} \le S_{q_m} \le s_{(m)} + ds_{(m)}, s_{(m)} + ds_{(m)} \le S_{q_{m+1}}, \cdots, s_{(m)} + ds_{(m)} \le S_{q_K})$$

$$= \sum_{(q_1,\cdots,q_{l-1}),q_l,(q_{l+1},\cdots,q_{m-1}),q_m,(q_{m+1},\cdots,q_K)} \int_{-\infty}^{s_{(l)}} \cdots \int_{-\infty}^{s_{(l)}} \int_{s_{(l)}}^{s_{(l)}+ds_{(l)}} \int_{s_{(l)}+ds_{(l)}}^{s_{(m)}} \cdots \int_{s_{(l)}+ds_{(l)}}^{s_{(m)}} \int_{s_{(m)}}^{s_{(m)}+ds_{(m)}} \int_{s_{(m)}+ds_{(m)}}^{\infty} \cdots \int_{s_{(m)}+ds_{(m)}}^{\infty} f(s_1, s_2,\cdots, s_K) ds_{q_1} \cdots ds_{q_{l-1}} ds_{q_l} ds_{q_{l+1}} \cdots ds_{q_{m-1}} ds_{q_m} ds_{q_{m+1}} \cdots ds_{q_K}.$$

Here, the summation is taken for all sorts of decompositions $(q_1,\cdots,q_{l-1}), q_l, (q_{l+1},\cdots,q_{m-1}), q_m, (q_{m+1},\cdots,K)$ of $(1,2,\cdots,K)$. Therefore, we obtain

$$\lim_{ds_{(l)}\to+0, ds_{(m)}\to+0} \times \frac{G_{(l),(m)}(s_{(l)}, s_{(m)} + ds_{(m)}) - G_{(l),(m)}(s_{(l)}, s_{(m)})}{ds_{(l)} ds_{(m)}} = \sum_{(q_1,\cdots,q_{l-1}),q_l,(q_{l+1},\cdots,q_{m-1}),q_m,(q_{m+1},\cdots,K)} \int_{-\infty}^{s_{(l)}} \cdots \int_{-\infty}^{s_{(l)}} \int_{s_{(l)}}^{s_{(m)}} \cdots \int_{s_{(l)}}^{s_{(m)}} \int_{s_{(m)}}^{\infty} \cdots \int_{s_{(m)}}^{\infty} f(s_1, \cdots, s_{(l)},\cdots,s_{(m)},\cdots, s_K) ds_{q_1} \cdots ds_{q_{l-1}} ds_{q_{l+1}} \cdots ds_{q_{m-1}} ds_{q_{m+1}} \cdots ds_{q_K}.$$

Here $f(s_1, \cdots, s_{(l)},\cdots,s_{(m)},\cdots, s_K)$ is obtained by substituting $s_{(l)}$ and $s_{(m)}$ into q_lth component and q_mth component of $f(s_1, s_2,\cdots, s_K)$. The above derivation is similar when

$$dx_{(l)} > 0, dx_{(m)} < 0,$$
$$dx_{(l)} < 0, dx_{(m)} > 0$$

or

$$dx_{(l)} < 0, dx_{(m)} < 0.$$

Therefore, we obtain

$$h_{(l),(m)}(s_{(l)}, s_{(m)}) = \sum_{(q_1,\cdots,q_{l-1}),q_l,(q_{l+1},\cdots,q_{m-1}),q_m,(q_{m+1},\cdots,K)} \int_{-\infty}^{s_{(l)}} \cdots \int_{-\infty}^{s_{(l)}} \int_{s_{(l)}}^{s_{(m)}} \cdots \int_{s_{(l)}}^{s_{(m)}} \int_{s_{(m)}}^{\infty} \cdots \int_{s_{(m)}}^{\infty} f(s_1, \cdots, s_{(l)},\cdots,s_{(m)},\cdots, s_K) ds_{q_1} \cdots ds_{q_{l-1}} ds_{q_{l+1}} \cdots ds_{q_{m-1}} ds_{q_{m+1}} \cdots ds_{q_K}.$$

3. Explicit formulae of probability density function of $R_{(l),(m)} = S_{(m)} - S_{(l)}$

We derived $h_{(l),(m)}(s_{(l)}, s_{(m)})$ in Section 2. Then, the probability density function of $R_{(l),(m)} = S_{(m)} - S_{(l)}$ is given by

$$g(r_{(l),(m)}) = \int_{-\infty}^{\infty} h_{(l),(m)}(s_{(l)}, r_{(l),(m)} + s_{(l)})\, ds_{(l)}.$$

If $l = 1, m = K$, the probability density function of $R_{(1),(K)} = S_{(K)} - S_{(1)}$ is given by

$$g(r_{(1),(K)}) = \int_{-\infty}^{\infty} h_{(1),(K)}(s_{(1)}, r_{(1),(K)} + s_{(1)})\, ds_{(1)}.$$

4. Applications

4.1. Determination of critical value of all-pairwise multiple comparison for normal means

Assume there are K normal populations $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2), \cdots, N(\mu_K, \sigma^2)$. We consider the all-pairwise multiple comparison for $\mu_1, \mu_2, \cdots, \mu_K$. Specifically, we set up a null hypothesis and its alternative hypothesis as

$$H_{i,j}: \mu_i = \mu_j \text{ vs. } H_{i,j}^A: \mu_i \ne \mu_j \qquad (4.1)$$

for $1 \le i < j \le K$ and consider the simultaneous test for them. Let $X_{k1}, X_{k2},\cdots, X_{kn_k}$ be a sample from $N(\mu_k, \sigma^2)$ for $k = 1,2,\cdots,K$. Let

$$\bar{X}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki} \quad (k = 1,2,\cdots,K), N = \sum_{k=1}^{K} n_k,$$

$$s = \sqrt{\frac{1}{\nu}\sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)^2}$$

where $\nu = N - K$. Tukey (1953) used a statistic

$$S_{i,j} = \frac{|\bar{X}_i - \bar{X}_j|}{\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}\, s}$$

for testing (4.1). Specifying the critical value c for testing hypotheses (4.1), if $S_{i,j} > c$, we reject $H_{i,j}$. Otherwise we retain $H_{i,j}$. We determine the critical value c so that

$$P\left(\max_{1\le i<j\le K} S_{i,j} > c\right) = \alpha$$

for a specified significance level α. Here, $P(\cdot)$ is the probability measure under the assumption that all $H_{i,j}$s are true. If all $H_{i,j}$s are true, we can assume $\mu_1 = \mu_2 = \cdots = \mu_K = 0$ without loss of generality. If $n_1 = n_2 = \cdots = n_K = n$,

$$P\left(\max_{1\le i<j\le K} S_{i,j} > c\right)$$

is formulated. Specifically,

$$P\left(\max_{1\le i<j\le K} S_{i,j} > c\right)$$

$$= 1 - P\left(\max_{1\le i<j\le K} S_{i,j} \le c\right)$$

$$= 1 - \sum_{k=1}^{K} P\left(0 \le \frac{\sqrt{\frac{n}{2}}(\bar{X}_k - \bar{X}_i)}{s} \le c \text{ for } i \ne k\right)$$

$$= 1 - K\int_0^\infty \left[\int_{-\infty}^{\infty} \left\{\Phi(z) - \Phi\left(z - \sqrt{2}cs_0\right)\right\}^{K-1} \phi(z)dz\right] \times g(s_0)ds_0.$$

Here $\Phi(z)$ is the cumulative distribution function of $N(0,1)$, $\phi(z)$ is the probability density function of $N(0,1)$ and $g(s_0)$ is the probability density function of $s_0 = s/\sigma$ given by

$$g(s_0) = \frac{\psi^{\psi/2}}{2^{(\psi-2)/2}\Gamma[\psi/2]} {s_0}^{\psi-1} \exp\left[-\frac{\psi s_0^2}{2}\right].$$

On the other hand, let

$$S_k = \frac{\sqrt{n}\bar{X}_k}{s}$$

for $1 \le i < j \le K$. $(S_1, S_2, \cdots, S_K)'$ is distributed according to K -variate t-distribution with ν degrees of freedom and covariance matrix I_K. Here I_K is K-dimensional unit matrix. Arranging statistics $S_1, S_2, \cdots, S_K$ in order of a size of value, assume

$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(K)}.$$

Letting $R_{(1),(K)} = S_{(K)} - S_{(1)}$, we obtain

$$P\left(\max_{1\le i<j\le K} S_{i,j} > c\right) = P(R_{(1),(K)} > \sqrt{2}c)$$

Letting $g(r_{(1),(K)})$ be the probability density function of $R_{(1),(K)}$, we obtain

$$P\left(R_{(1),(K)} > \sqrt{2}c\right) = \int_{\sqrt{2}c}^{\infty} g(r_{(1),(K)})d\, r_{(1),(K)}.$$

Specifically, the critical value c of all-pairwise multiple comparison for a specified significance level can be obtained by the probability density function of $R_{(1),(K)}$.

4.2. Critical value of multiple comparison with a control for normal means

We consider the multiple comparison with a control for comparing μ_1 with $\mu_2, \cdots, \mu_K$ simultaneouly. Here, we assume

$$\mu_1 \ge \mu_k$$

for $k = 2,3, \cdots, K$ in advance of the test. We set up a null hypothesis and its alternative hypothesis as

$$H_{1,i}: \mu_1 = \mu_i \text{ vs. } H_{1,i}^A: \mu_1 > \mu_i \qquad (4.2)$$

for $k = 2,3, \cdots, K$ and consider the simultaneous test for them. Dunnett (1955) used a statistic

$$S_{1,i} = \frac{\bar{X}_1 - \bar{X}_i}{\sqrt{\frac{1}{n_1} + \frac{1}{n_i}}s}$$

for testing (4.2). Specifying the critical value c for testing hypotheses (4.2), if $S_{1,i} > c$, we reject $H_{1,i}$. Otherwise we retain $H_{1,i}$. We determine the critical value c so that

$$P\left(\max_{i=2,3,\cdots,K} S_{1,i} > c\right) = \alpha.$$

Here

$$P\left(\max_{i=2,3,\cdots,K} S_{1,i} > c\right)$$

can be formulated. Specifically, letting

$$\lambda_{1,i} = \frac{n_i}{n_1 + n_i},$$

we obtain

$$P\left(\max_{i=2,3,\cdots,K} S_{1,i} > c\right)$$

$$= 1 - P\left(\max_{i=2,3,\cdots,K} S_{1,i} \le c\right)$$

$$= 1 - P\left(\frac{\bar{X}_1 - \bar{X}_i}{\sqrt{\frac{1}{n_1} + \frac{1}{n_i}}s} \le c \text{ for } i = 2,3, \cdots, K\right)$$

$$= 1 - P\left(\sqrt{n_i}\bar{X}_i \le \frac{\sqrt{\lambda_{1,i}}\sqrt{n_1}\bar{X}_1 + cs}{\sqrt{1-\lambda_{1,i}}} \text{ for } i = 2,3, \cdots, K\right)$$

$$= 1 - \sum_{i=2}^{K} \int_0^\infty \int_{-\infty}^{\infty} \Phi\left(\frac{\sqrt{\lambda_{1,i}}z + cs_0}{\sqrt{1-\lambda_{1,i}}}\right) \phi(z)\, g(s_0)dzds_0.$$

On the other hand, $(S_{1,2}, S_{1,3}, \cdots, S_{1,K})$ is distributed according to $K-1$ -variate t-distribution with ν degrees of freedom and covariance matrix $(\sqrt{\lambda_{1,i}\lambda_{1,j}})$. Arranging statistics $S_{1,2}, S_{1,3}, \cdots, S_{1,K}$ in order of a size of value, assume

$$S_{(1)} \le S_{(2)} \le \cdots \le S_{(K-1)}.$$

Letting $R_{(1),(K-1)} = S_{(K-1)} - S_{(1)}$, we obtain

$$P\left(\max_{i=2,3,\cdots,K} S_{1,i} > c\right) = P(R_{(1),(K-1)} > c)$$

Letting $g(r_{(1),(K-1)})$ be the probability density function of $R_{(1),(K-1)}$, we obtain

$$P\left(R_{(1),(K-1)} > c\right) = \int_c^{\infty} g(r_{(1),(K-1)}) d\, r_{(1),(K-1)}.$$

Specifically, the critical value c of multiple comparison with a control for a specified significance level can be obtained by the probability density function of $R_{(1),(K-1)}$.

5. Conclusions

In this study, we discussed the joint probability density function of the ordered statistics constructed from plural dependent statistics. Specifically, we derived the probability density function of two statistics among the ordered statistics. Then, we derived the probability density function of the range of ordered statistics and applied it to determining the critical values of the all-pairwise multiple comparison procedure and the multiple comparison procedure with a control for normal means.

However, there remain some problems to be solved in the future. We should clarify the advantage of determination of the critical values of the multiple comparison procedures by using the probability density function of the range of ordered statistics. Furthermore, we want to apply the probability density function of the range of ordered statistics to constructing more powerful stepwise multiple comparison procedures.

References

[1] Imada, T. (2022). Note on Joint Probability Density Function of Ordered Statistics. *Bulletin of the School of Humanities and Science of Tokai University*, **1**, 112--115.

[2] Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, **50**, 1096-1121.

[3] Tukey, J. W. (1953). The problem of multiple comparisons. Unpublished manuscript, Princeton University.