Ann Inst Stat Math (2013) 65:571–587
DOI 10.1007/s10463-012-0382-9

Coupon collector's problems with statistical applications to rankings

Sigeo Aki · Katuomi Hirano

Received: 4 November 2011 / Revised: 28 August 2012 / Published online: 15 November 2012

Abstract Some new exact distributions on coupon collector's waiting time problems are given based on a generalized Pólya urn sampling. In particular, usual Pólya urn sampling generates an exchangeable random sequence. In this case, an alternative derivation of the distribution is also obtained from de Finetti's theorem. In coupon collector's waiting time problems with m kinds of coupons, the observed order of m kinds of coupons corresponds to a permutation of m letters uniquely. Using the property of coupon collector's problems, a statistical model on the permutation group of m letters is proposed for analyzing ranked data. In the model, as the parameters mean the proportion of the m kinds of coupons, the observed ranking can be intuitively understood. Some examples of statistical inference are also given.

Keywords Generalized Pólya urn · Dirichlet distribution · Exchangeability · Likelihood ratio test · Permutation

1 Introduction

Coupon collector's waiting time problems are the following. There are m different kinds of coupons which come with a product. A coupon is obtained randomly from each purchase of the product. The problem is how many products should we purchase

This research was partially supported by the Kansai University Grant-in-Aid for progress of research in graduate course and by Grant-in-Aid for Scientific Research (C) of the JSPS (Grant Number 22540159).

S. Aki (✉)
Department of Mathematics, Kansai University, 3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680, Japan
e-mail: aki@kansai-u.ac.jp

K. Hirano
Department of Mathematics, Josai University, 1-1 Keyakidai, Sakado-shi, Saitama-ken 350-0295, Japan

until all kinds of coupons are collected. Mahmoud (2008) explained the problem using urn models. When each coupon can be obtained independently with an identical probability, the probability generating function (pgf) of the waiting time distribution is well known (see, e.g., Exercise 38 in Chapter 8 of Graham et al. 1989). The exact probability function of the waiting time is given in Charalambides (2005) and generalizations of the problem have been studied (see, e.g., Kobza et al. 2007; Inoue and Aki 2008).

Let $X_1, X_2, \dots$ be $\{1, 2, \dots, m\}$-valued random variables. Suppose that X_i means the type of the coupon at the i-th trial. For $k = 1, 2, \dots, m$, let

$$\tau_k = \inf\{n : |\{X_1, \dots, X_n\}| = k\} \tag{1}$$

and

$$Y_k = X_{\tau_k}. \tag{2}$$

Then τ_m is the usual coupon collector's waiting time and note that $\tau_1 = 1$. Let $T_1 = \tau_1$ and for $k = 2, \dots, m$, let $T_k = \tau_k - \tau_{k-1}$. Here, we derive the joint exact distribution of $(T_1, \dots, T_m)$ under the assumption that $X_1, X_2, \dots$ are generated by a generalized Pólya urn scheme. The generalized Pólya urn sampling includes the cases of the usual Pólya urn sampling and the i.i.d. sampling. We obtain the exact distribution based on the method of conditional probability generating functions. Further, by considering that the usual Pólya urn sampling generates infinitely exchangeable random variables, we give an alternative proof of the case of Pólya urn sampling using de Finetti's theorem. For $k = 1, 2, \dots, m$, let

$$W_k = \inf\{n : X_n = k\},$$

i.e., W_k is the waiting time for the first coupon of type k. Let $\{W_{(1)}, W_{(2)}, \dots, W_{(m)}\}$ be the set of order statistics for $(W_1, W_2, \dots, W_m)$. Then, it is easy to see that $\tau_k = W_{(k)}$ for each $k = 1, 2, \dots, m$. Even if $X_1, X_2, \dots$ are independent and identically distributed discrete random variables, $(W_1, W_2, \dots, W_m)$ are not independent because they do not have ties. Therefore, the random vector $(Y_1, \dots, Y_m)$ is a random permutation of $\{1, 2, \dots, m\}$. By means of coupon collector's waiting time problems, we can construct a natural and meaningful distribution on the symmetric group $\mathfrak{S}_m$, where $\mathfrak{S}_m$ is the set of all permutations of $\{1, \dots, m\}$. We use the notation $\sigma = \begin{pmatrix} 1 & 2 & \cdots & m \\ \sigma_1 & \sigma_2 & \cdots & \sigma_m \end{pmatrix}$ for $\sigma \in \mathfrak{S}_m$. A permutation of $\{1, \dots, m\}$ can be regarded as a ranking of items $\{1, 2, \dots, m\}$. Holland's model is a well-known parametric model to analyze ranked data (see Diaconis 1988). The parametric model is an exponential family with a $(m-1) \times (m-1)$ matrix parameter. In the model, the parameters are theoretically given, and it may be difficult to understand practical meanings of the parameters. In this study, we propose some parametric models on $\mathfrak{S}_m$ through coupon collector's waiting time problems. In our model, the probability with which we observe each type of coupons is parameterized and hence the observed ranking can be explained by the proportion of each type of coupons.

In Sect. 2, the joint *pgf* of $(T_1, \dots, T_m)$ is given under the assumption that $X_1, X_2, \dots$ are generated by a generalized Pólya urn sampling. Further, some special cases of the distribution are studied. In particular, the case of the usual Pólya urn sampling is considered based on the exchangeability of the sequence $\{X_n\}_{n=1}^{\infty}$. Section 3 presents statistical inference on coupon collector's problems. We show that the maximum likelihood estimation of the proportion of each type of coupons can be performed based on observations of $(T_1, \dots, T_m)$ and $(Y_1, \dots, Y_m)$. We also investigate the feasibility of a likelihood ratio test for equality of proportions of two types of coupons among m types of coupons by simulation.

2 Coupon collector's problems based on a generalized Pólya urn scheme

We consider the coupon collector's waiting time problem in the sequence of a generalized Pólya urn sampling. An urn contains N_i balls labeled "i" for $i = 1, 2, \dots, m$. We set $N = N_1 + N_2 + \cdots + N_m$. A ball is drawn at random. If it is labeled "i", then it is returned immediately together with additional $\alpha(\geq 0)$ balls labeled "i" and with additional $\beta(\geq 0)$ balls labeled "j" for every $j(\neq i)$. Suppose that we repeat drawing balls in the above manner until all kinds of balls are drawn. Let X_n be the number labeled on the ball at the n-th drawing in the sampling scheme. Then, $(Y_1, Y_2, \dots, Y_m)$ and $(T_1, T_2, \dots, T_m)$ are defined by Eq. (2) and the statement just after the formula. $\phi(\boldsymbol{t}) = E[t_1^{T_1} t_2^{T_2} \cdots t_m^{T_m}]$ denotes the joint *pgf* of $(T_1, T_2, \dots, T_m)$, where $\boldsymbol{t} = (t_1, \dots, t_m)$. By definition, we obtain the exact *pgf* of τ_m by setting $t_1 = \cdots = t_m = t$.

Theorem 1 *The joint pgf of $(T_1, \dots, T_m)$ is written as:*

$$\phi(\boldsymbol{t}) = \sum_{\sigma \in \mathfrak{S}_m} \sum_{s_2=1}^{\infty} \cdots \sum_{s_m=1}^{\infty} \frac{G(s_2, \dots, s_m) t_1 t_2^{s_2} \cdots t_m^{s_m}}{(N)_{(1+s_2+s_3+\cdots+s_m)\uparrow(\alpha+(m-1)\beta)}}, \tag{3}$$

where

$$\begin{aligned}
&G(s_2, \dots, s_m) \\
&\quad = N_{\sigma_1}(N_{\sigma_1} + \alpha)_{(s_2-1)\uparrow\alpha}(N_{\sigma_2} + s_2\beta)(N_{\sigma_1} + N_{\sigma_2} \\
&\qquad +(1 + s_2)(\alpha + \beta))_{(s_3-1)\uparrow(\alpha+\beta)}(N_{\sigma_3} + (s_2 + s_3)\beta) \cdots \\
&\qquad \times(N_{\sigma_1} + \cdots + N_{\sigma_{m-1}} + (1 + s_2 + \cdots + s_{m-1})(\alpha + (m-2)\beta))_{(s_m-1)\uparrow(\alpha+(m-2)\beta)} \\
&\qquad \times(N_{\sigma_m} + (s_2 + s_3 + \cdots + s_m)\beta) \\
&\quad = N_{\sigma_1} \prod_{i=2}^{m} \left[\left(\sum_{j=1}^{i-1} N_{\sigma_j} + \left(\sum_{j=1}^{i-1} s_j \right) (\alpha + (i-1)\beta) \right)_{(s_i-1)\uparrow(\alpha+(i-2)\beta)} \right] \\
&\qquad \times \prod_{i=2}^{m} \left(N_{\sigma_i} + \beta \sum_{j=2}^{i} s_j \right).
\end{aligned}$$

Here, for every positive integer k, we define

$$(n)_{k\uparrow\alpha} = \prod_{i=1}^{k}(n + (i-1)\alpha).$$

When $\alpha = 1$, *we sometimes omit* α *like* $(n)_{k\uparrow}$. *In the above formula, we always set* $s_1 = 1$.

Proof Conditioning on Y_1, we obtain

$$\begin{aligned}\phi(\boldsymbol{t}) &= E[E[t_1^{T_1}\cdots t_m^{T_m}|Y_1]]\\ &= \sum_{\sigma_1=1}^{m} P(Y_1=\sigma_1)E[t_1^{T_1}\cdots t_m^{T_m}|Y_1=\sigma_1]\\ &= \sum_{\sigma_1=1}^{m} P(Y_1=\sigma_1)t_1E[t_2^{T_2}\cdots t_m^{T_m}|Y_1=\sigma_1]\\ &= \sum_{\sigma_1=1}^{m} \frac{N_{\sigma_1}}{N}t_1E[t_2^{T_2}\cdots t_m^{T_m}|Y_1=\sigma_1].\end{aligned}$$

Further, conditioning on Y_2 and T_2, we have

$$\begin{aligned}&E[t_2^{T_2}\cdots t_m^{T_m}|Y_1=\sigma_1]\\ &\quad= \sum_{\sigma_2\neq\sigma_1}\sum_{s_2=1}^{\infty} P(Y_2=\sigma_2, T_2=s_2|Y_1=\sigma_1)\\ &\qquad\times t_2^{s_2}E[t_3^{T_3}\cdots t_m^{T_m}|Y_1=\sigma_1, Y_2=\sigma_2, T_2=s_2].\end{aligned}$$

Noting that

$$\begin{aligned}&P(Y_2=\sigma_2, T_2=s_2|Y_1=\sigma_1)\\ &\quad= P(Y_2=\sigma_2|Y_1=\sigma_1, T_2=s_2)P(T_2=s_2|Y_1=\sigma_1)\\ &\quad= \frac{N_{\sigma_2}+s_2\beta}{N-N_{\sigma_1}+s_2(m-1)\beta}\cdot\frac{(N_{\sigma_1}+\alpha)_{(s_2-1)\uparrow\alpha}(N-N_{\sigma_1}+s_2(m-1)\beta)}{(N+\alpha+(m-1)\beta)_{s_2\uparrow(\alpha+(m-1)\beta)}}\\ &\quad= \frac{(N_{\sigma_1}+\alpha)_{(s_2-1)\uparrow\alpha}(N_{\sigma_2}+s_2\beta)}{(N+\alpha+(m-1)\beta)_{s_2\uparrow(\alpha+(m-1)\beta)}},\end{aligned}$$

we see that

$$\begin{aligned}\phi(\boldsymbol{t}) = \sum_{\sigma_1=1}^{m}\sum_{\substack{\sigma_2=1\\ \sigma_2\neq\sigma_1}}^{m}\sum_{s_2=1}^{\infty}\frac{(N_{\sigma_1})_{s_2\uparrow\alpha}(N_{\sigma_2}+s_2\beta)}{(N)_{(1+s_2)\uparrow(\alpha+(m-1)\beta)}}t_1t_2^{s_2}\\ \times E[t_3^{T_3}\cdots t_m^{T_m}|Y_1=\sigma_1, (T_1=1), Y_2=\sigma_2, T_2=s_2].\end{aligned}$$

By repeating the conditioning like above, we obtain the desired result. □

Setting $\beta = 0$ in Theorem 1, we obtain the joint pgf of $(T_1, \dots, T_m)$ based on the usual Pólya urn sampling.

Theorem 2 *If $\beta = 0$, that is, $X_1, X_2, \dots$ are generated by the Pólya urn sampling, the joint pgf of $(T_1, \dots, T_m)$ can be written as:*

$$\phi(\boldsymbol{t}) = \sum_{\sigma \in \mathfrak{S}_m} \sum_{s_2=1}^{\infty} \cdots \sum_{s_m=1}^{\infty} \frac{G(s_2, \dots, s_m) t_1 t_2^{s_2} \cdots t_m^{s_m}}{(N)_{(1+s_2+s_3+\cdots+s_m)\uparrow\alpha}}, \tag{4}$$

where

$$\begin{aligned} &G(s_2, \dots, s_m) \\ &\quad = N_{\sigma_1}(N_{\sigma_1} + \alpha)_{(s_2-1)\uparrow\alpha} N_{\sigma_2}(N_{\sigma_1} + N_{\sigma_2} + (s_2+1)\alpha)_{(s_3-1)\uparrow\alpha} N_{\sigma_3} \\ &\qquad \cdots (N_{\sigma_1} + \cdots + N_{\sigma_{m-1}} + (1 + s_2 + s_3 + \cdots + s_{m-1})\alpha)_{(s_m-1)\uparrow\alpha} N_{\sigma_m} \\ &\quad = N_{\sigma_1} \prod_{i=2}^{m} \left[\left(\sum_{j=1}^{i-1} N_{\sigma_j} + \left(\sum_{j=1}^{i-1} s_j \right) \alpha \right)_{(s_i-1)\uparrow\alpha} N_{\sigma_i} \right], \end{aligned}$$

where $s_1 = 1$.

After stating the corollaries below, we shall give an alternative Proof of Theorem 2 as an application of de Finetti's theorem. By setting $N_1 = N_2 = \cdots = N_m = k$ in Theorem 2, we have the following corollary.

Corollary 1 *If $N_1 = N_2 = \cdots = N_m = k$, the joint pgf of $(T_1, \dots, T_m)$ can be written as:*

$$\phi(\boldsymbol{t}) = m! \sum_{s_2=1}^{\infty} \cdots \sum_{s_m=1}^{\infty} \frac{k^m \prod_{i=2}^{m} \left\{ ((i-1)k + (\sum_{j=1}^{i-1} s_j)\alpha)_{(s_i-1)\uparrow\alpha} \right\}}{(mk)_{(1+s_2+\cdots+s_m)\uparrow\alpha}} t_1 t_2^{s_2} \cdots t_m^{s_m},$$

where $s_1 = 1$.

By setting $\alpha = 0$ in (4), we have the following corollary.

Corollary 2 *Suppose that one ball is sampled at random from the urn and it is returned immediately. If the trials are repeated, that is, the sampling is i.i.d., then the joint pgf of $(T_1, \dots, T_m)$ is given by*

$$\begin{aligned} \phi(\boldsymbol{t}) = \sum_{\sigma \in \mathfrak{S}_m} & \frac{N_{\sigma_1} t_1}{N} \cdot \frac{N_{\sigma_2} t_2}{N - N_{\sigma_1} t_2} \cdot \frac{N_{\sigma_3} t_3}{N - (N_{\sigma_1} + N_{\sigma_2}) t_3} \cdots \\ & \times \frac{N_{\sigma_m} t_m}{N - (N_{\sigma_1} + N_{\sigma_2} + \cdots + N_{\sigma_{m-1}}) t_m}. \end{aligned} \tag{5}$$

Further, if $N_1 = N_2 = \cdots = N_m$ *hold, then we have*

$$\phi(\boldsymbol{t}) = m!\frac{t_1}{m}\prod_{k=1}^{m-1}\frac{t_{k+1}}{m-kt_{k+1}}.$$

Proof of Corollary 2 By setting $\alpha = 0$ in (4), we have

$$\begin{aligned}\phi(\boldsymbol{t}) &= \sum_{\sigma\in\mathfrak{S}_m}\sum_{s_2=1}^{\infty}\cdots\sum_{s_m=1}^{\infty} t_1 t_2^{s_2}\cdots t_m^{s_m}\\ &\quad\times\frac{N_{\sigma_1}N_{\sigma_1}^{s_2-1}N_{\sigma_2}(N_{\sigma_1}+N_{\sigma_2})^{s_3-1}N_{\sigma_3}\cdots(N_{\sigma_1}+\cdots+N_{\sigma_{m-1}})^{s_m-1}N_{\sigma_m}}{N^{1+s_2+\cdots+s_m}}\\ &= \sum_{\sigma\in\mathfrak{S}_m}\frac{N_{\sigma_1}t_1}{N}\sum_{s_2=1}^{\infty}\left(\frac{N_{\sigma_1}}{N}t_2\right)^{s_2-1}\frac{N_{\sigma_2}}{N}t_2\sum_{s_3=1}^{\infty}\left(\frac{N_{\sigma_1}+N_{\sigma_2}}{N}t_3\right)^{s_3-1}\frac{N_{\sigma_3}}{N}t_3\\ &\quad\cdots\sum_{s_m=1}^{\infty}\left(\frac{N_{\sigma_1}+\cdots+N_{\sigma_{m-1}}}{N}t_m\right)^{s_m-1}\frac{N_{\sigma_m}}{N}t_m\\ &= \sum_{\sigma\in\mathfrak{S}_m}\frac{N_{\sigma_1}t_1}{N}\cdot\frac{\frac{N_{\sigma_2}}{N}t_2}{1-\frac{N_{\sigma_1}}{N}t_2}\cdot\frac{\frac{N_{\sigma_3}}{N}t_3}{1-\frac{N_{\sigma_1}+N_{\sigma_2}}{N}t_3}\cdots\frac{\frac{N_{\sigma_m}}{N}t_m}{1-\frac{N_{\sigma_1}+\cdots+N_{\sigma_{m-1}}}{N}t_m}.\end{aligned}$$

Therefore, Corollary 2 holds. □

As Theorem 1 has been proved using the method of conditional probability generating functions, we have obtained Theorem 2 as a special case of Theorem 1. Theorem 2 derives the joint *pgf* of $(T_1, \ldots, T_m)$ based on the usual Pólya-Eggenberger urn scheme. It is well known that the usual Pólya-Eggenberger urn scheme generates an infinitely exchangeable sequence (see Mahmoud 2008; Johnson and Kotz 1977). From de Finetti's theorem, we see that an infinitely exchangeable sequence is a mixture of i.i.d. sequences with a directing random measure called the de Finetti measure. This means that the distributional results based on infinitely exchangeable sequences can be derived from the corresponding results based on i.i.d. sequences if the de Finetti measure is given. Fortunately, the de Finetti measure corresponding to the usual Pólya-Eggenberger urn scheme is known to be a Dirichlet distribution. By giving an alternative Proof of Theorem 2, let us regard Theorem 2 as an extension of the i.i.d. case.

We need some properties of Dirichlet distributions for giving an alternative Proof of Theorem 2. We denote the $(m-1)$-dimensional simplex by

$$\boldsymbol{S}_{m-1} = \{(x_1, \ldots, x_{m-1}) : x_j \geq 0, x_1 + x_2 + \cdots + x_{m-1} \leq 1\}.$$

Definition 1 The distribution on $\boldsymbol{S}_{m-1}$ with density

$$f(x_1, \ldots, x_{m-1}) = \frac{\Gamma(\nu_1+\cdots+\nu_m)}{\Gamma(\nu_1)\cdots\Gamma(\nu_m)}x_1^{\nu_1-1}\cdots x_{m-1}^{\nu_{m-1}-1}(1-x_1-\cdots-x_{m-1})^{\nu_m-1}$$

is called a Dirichlet distribution of parameter $(\nu_1, \ldots, \nu_m)$ and denoted by $D(\nu_1, \ldots, \nu_{m-1}; \nu_m)$.

Lemma 1 *If* $(X_1, \ldots, X_k)$ *follows* $D(\nu_1, \ldots, \nu_k; \nu_{k+1})$, *it holds that*

$$E[X_1^{r_1} \cdots X_k^{r_k} X_{k+1}] = \frac{(\nu_1)_{r_1\uparrow} \cdots (\nu_k)_{r_k\uparrow} \nu_{k+1}}{(\nu_1 + \cdots + \nu_{k+1})_{(r_1+\cdots+r_k+1)\uparrow}},$$

where $X_{k+1} = 1 - X_1 - \cdots - X_k$.

Proof If $(X_1, \ldots, X_k)$ follows $D(\nu_1, \ldots, \nu_k; \nu_{k+1})$, it holds that

$$\begin{aligned} E[X_1^{r_1} \cdots X_k^{r_k}] &= \frac{\Gamma(\nu_1 + r_1) \cdots \Gamma(\nu_k + r_k)\Gamma(\nu_1 + \cdots + \nu_{k+1})}{\Gamma(\nu_1) \cdots \Gamma(\nu_k)\Gamma(\nu_1 + \cdots + \nu_{k+1} + r_1 + \cdots + r_k)} \\ &= \frac{(\nu_1)_{r_1\uparrow} \cdots (\nu_k)_{r_k\uparrow}}{(\nu_1 + \cdots + \nu_{k+1})_{(r_1+\cdots+r_k)\uparrow}}. \end{aligned}$$

Therefore, the result follows by the next calculations.

$$\begin{aligned} &E[X_1^{r_1} \cdots X_k^{r_k} X_{k+1}] \\ &\quad = E[X_1^{r_1} \cdots X_k^{r_k}(1 - (X_1 + \cdots + X_k))] \\ &\quad = E[X_1^{r_1} X_2^{r_2} \cdots X_k^{r_k}] - E[X_1^{r_1+1} X_2^{r_2} \cdots X_k^{r_k}] - E[X_1^{r_1} X_2^{r_2+1} \cdots X_k^{r_k}] \\ &\qquad - \cdots - E[X_1^{r_1} X_2^{r_2} \cdots X_k^{r_k+1}] \\ &\quad = \frac{(\nu_1)_{r_1\uparrow} \cdots (\nu_k)_{r_k\uparrow} \nu_{k+1}}{(\nu_1 + \cdots + \nu_{k+1})_{(r_1+\cdots+r_k+1)\uparrow}}. \end{aligned}$$

□

The next lemma is a direct extension of the multinomial theorem.

Lemma 2 *Let n be a positive integer and let* β *be a real number. Then the following equation holds.*

$$(x_1 + x_2 + \cdots + x_m)_{n\uparrow\beta} = \sum \binom{n}{k_1, \ldots, k_m} (x_1)_{k_1\uparrow\beta}(x_2)_{k_2\uparrow\beta} \cdots (x_m)_{k_m\uparrow\beta}, \tag{6}$$

where the summation is extended for nonnegative integers $k_1, \ldots, k_m$ *satisfying* $k_1 + k_2 + \cdots + k_m = n$. *In particular, the next equation holds for an extension of binomial theorem.*

$$(x + y)_{n\uparrow\beta} = \sum_{k=0}^{n} \binom{n}{k} (x)_{k\uparrow\beta}(y)_{(n-k)\uparrow\beta}. \tag{7}$$

Remark 1 By setting $\beta = 1$, $\beta = -1$ and $\beta = 0$ in Eq. (7), we have Nörlund's formula, Vandermonde's formula and Newton's formula, respectively. Lemma 2 may be proved easily. The special case of the formula (6) is given in the exercises of Chapter 3 of Charalambides (2002).

We give here an alternative Proof of Theorem 2 using de Finetti's theorem.

Proof (An alternative proof of Theorem 2) It is well known that the infinite sequence $X_1, X_2, \ldots$ is exchangeable when the Pólya-Eggenberger urn model is used for sampling. Then from de Finetti's theorem, the joint distribution of the sequence $\{X_n\}, n \geq 1$ is obtained by randomizing the parameter of a multinomial. This randomization is expressed by a random vector $\boldsymbol{P} = (P_1, \ldots, P_m)$ which takes values $\boldsymbol{p} = (p_1, \ldots, p_{m-1}) \in \boldsymbol{S_{m-1}}$ and $p_m = 1 - p_1 - \cdots - p_{m-1}$. In the Pólya-Eggenberger urn model, the random vector $\boldsymbol{P}$ follows the Dirichlet distribution $D(\frac{N_1}{\alpha}, \ldots, \frac{N_{m-1}}{\alpha}; \frac{N_m}{\alpha})$, see Johnson and Kotz (1977). Therefore, we can write

$$
\begin{aligned}
\phi(\boldsymbol{t}) &= \int_{\boldsymbol{S}_{m-1}} E[t_1^{T_1} \cdots t_m^{T_m} | \boldsymbol{P} = \boldsymbol{p}] f(\boldsymbol{p}) d\boldsymbol{p} \\
&= \int_{\boldsymbol{S}_{m-1}} \sum_{\sigma \in \mathfrak{S}_m} p_{\sigma_1} t_1 \cdot \frac{p_{\sigma_2} t_2}{1 - p_{\sigma_1} t_2} \cdot \frac{p_{\sigma_3} t_3}{1 - (p_{\sigma_1} + p_{\sigma_2}) t_3} \cdots \\
&\quad \times \frac{p_{\sigma_m} t_m}{1 - (p_{\sigma_1} + p_{\sigma_2} + \cdots + p_{\sigma_{m-1}}) t_m} f(\boldsymbol{p}) d\boldsymbol{p} \\
&= \sum_{\sigma \in \mathfrak{S}_m} \sum_{s_2=1}^{\infty} \cdots \sum_{s_m=1}^{\infty} \int_{\boldsymbol{S}_{m-1}} p_{\sigma_1} (p_{\sigma_1})^{s_2 - 1} p_{\sigma_2} \cdots \\
&\quad \times (p_{\sigma_1} + \cdots + p_{\sigma_{m-1}})^{s_m - 1} p_{\sigma_m} f(\boldsymbol{p}) d\boldsymbol{p} \; t_1 t_2^{s_2} \cdots t_m^{s_m} \\
&= \sum_{\sigma \in \mathfrak{S}_m} \sum_{s_2=1}^{\infty} \cdots \sum_{s_m=1}^{\infty} E[P_{\sigma_1} P_{\sigma_1}^{s_2 - 1} P_{\sigma_2} (P_{\sigma_1} + P_{\sigma_2})^{s_3 - 1} P_{\sigma_3} \cdots \\
&\quad \times (P_{\sigma_1} + \cdots + P_{\sigma_{m-1}})^{s_m - 1} P_{\sigma_m}] \; t_1 t_2^{s_2} \cdots t_m^{s_m},
\end{aligned}
$$

where

$$
\begin{aligned}
&E[P_{\sigma_1} P_{\sigma_1}^{s_2 - 1} P_{\sigma_2} (P_{\sigma_1} + P_{\sigma_2})^{s_3 - 1} P_{\sigma_3} \cdots (P_{\sigma_1} + \cdots + P_{\sigma_{m-1}})^{s_m - 1} P_{\sigma_m}] \\
&= E\left[P_{\sigma(1)}^{s_2} P_{\sigma(2)} \left(\sum_{j_{31} + j_{32} = s_3 - 1} \binom{s_3 - 1}{j_{31}} P_{\sigma_1}^{j_{31}} P_{\sigma_2}^{j_{32}} \right) P_{\sigma_3} \right. \\
&\quad \left. \times \cdots \times \left(\sum_{j_{m1} + \cdots + j_{m,m-1} = s_m - 1} \binom{s_m - 1}{j_{m1}, \ldots, j_{m,m-1}} P_{\sigma_1}^{j_{m1}} \cdots P_{\sigma_{m-1}}^{j_{m,m-1}} \right) P_{\sigma_m} \right] \\
&= \sum_{j_{31} + j_{32} = s_3 - 1} \cdots \sum_{j_{m1} + j_{m,m-1} = s_m - 1} \binom{s_3 - 1}{j_{31}} \cdots \binom{s_m - 1}{j_{m1}, \ldots, j_{m,m-1}} \\
&\quad \times E[P_{\sigma_1}^{1 + (s_2 - 1) + j_{31} + \cdots + j_{m1}} P_{\sigma_2}^{1 + j_{32} + \cdots + j_{m1}} \cdots P_{\sigma_{m-1}}^{1 + j_{m,m-1}} P_{\sigma_m}].
\end{aligned}
$$

Using Lemma 1, we obtain

$$
\begin{aligned}
&E[P_{\sigma_1} P_{\sigma_1}^{s_2-1} P_{\sigma_2} (P_{\sigma_1} + P_{\sigma_2})^{s_3-1} P_{\sigma_3} \cdots (P_{\sigma_1} + \cdots + P_{\sigma_{m-1}})^{s_m-1} P_{\sigma_m}] \\
&\quad = \sum_{j_{31}+j_{32}=s_3-1} \cdots \sum_{j_{m1}+\cdots+j_{m,m-1}=s_m-1} \binom{s_3-1}{j_{31}} \cdots \binom{s_m-1}{j_{m1}, \ldots, j_{m,m-1}} \\
&\qquad \times \left(\frac{N_{\sigma_1}}{\alpha}\right)_{1+(s_2-1)+\cdots+j_{m1}\uparrow} \left(\frac{N_{\sigma_2}}{\alpha}\right)_{(1+j_{32}+\cdots+j_{m2})\uparrow} \cdots \\
&\qquad \times \left(\frac{N_{\sigma_{m-1}}}{\alpha}\right)_{(1+j_{m,m-1})\uparrow} \left(\frac{N_{\sigma_m}}{\alpha}\right) \left\{\left(\frac{N}{\alpha}\right)_{(1+s_2+\cdots+s_m)\uparrow}\right\}^{-1} \\
&\quad = \sum_{j_{31}+j_{32}=s_3-1} \cdots \sum_{j_{m1}+\cdots+j_{m,m-1}=s_m-1} \binom{s_3-1}{j_{31}} \cdots \binom{s_m-1}{j_{m1}, \cdots, j_{m,m-1}} \\
&\qquad \times \left(N_{\sigma_1}\right)_{1+(s_2-1)+\cdots+j_{m1}\uparrow\alpha} \left(N_{\sigma_2}\right)_{(1+j_{32}+\cdots+j_{m2})\uparrow\alpha} \cdots \\
&\qquad \times \left(N_{\sigma_{m-1}}\right)_{(1+j_{m,m-1})\uparrow\alpha} \left(N_{\sigma_m}\right) \left\{(N)_{(1+s_2+\cdots+s_m)\uparrow\alpha}\right\}^{-1} \\
&\quad = \frac{\left(N_{\sigma_1}\right)_{s_2\uparrow\alpha} N_{\sigma_2}}{(N)_{(1+s_2+\cdots+s_m)\uparrow\alpha}} \\
&\qquad \times \left(\sum_{j_{31}+j_{32}=s_3-1} \binom{s_3-1}{j_{31}} (N_{\sigma_1} + s_2\alpha)_{j_{31}\uparrow\alpha} (N_{\sigma_2} + \alpha)_{j_{32}\uparrow\alpha} \right) \cdots \\
&\qquad \times \left(\sum_{j_{m1}+\cdots+j_{m,m-1}=s_m-1} \binom{s_m-1}{j_{m1}, \ldots, j_{m,m-1}} (N_{\sigma_1} + (s_2 + j_{31} + \cdots + j_{m1})\alpha)_{j_{m1}\uparrow\alpha} \right. \\
&\qquad \left. \cdots (N_{\sigma_{m-1}} + \alpha)_{j_{m,m-1}\uparrow\alpha} \right) N_{\sigma_m}.
\end{aligned}
$$

Using Lemma 2 for each summation, we complete the proof. □

3 Statistical inference on coupon collector's problems

3.1 Parametric estimation

In this subsection, we study two statistical problems for estimating parameters concerning coupon collector's waiting time problems.

First, we study a very simple problem. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $P(X_i = j) = p_j$. For estimating the probability $(p_1, \ldots, p_m)$ based on observations of $(Y_1, \ldots, Y_m)$ and $(T_1, \ldots, T_m)$, we parameterize $(p_1, \ldots, p_m)$ with a positive parameter θ as $p_i = \frac{\theta^{i-1}}{1+\theta+\theta^2+\cdots+\theta^{m-1}}$, $i = 1, 2, \ldots, m$. Since the joint probability function of $(Y_1, \ldots, Y_m)$ and $(T_1, \ldots, T_m)$ and the marginal probability function of $(Y_1, \ldots, Y_m)$ can be easily obtained by replacing $\frac{N_j}{N}$ with p_j in Eq. (5), we can estimate the parameter θ by the method of maximum likelihood. To be precise, the pgf of the waiting time can be written as:

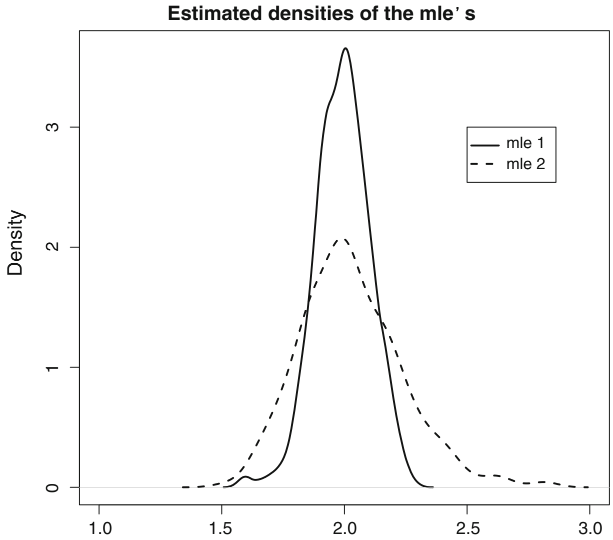

Fig. 1 Estimated densities of the distribution of $\hat{\theta}_1$ and $\hat{\theta}_2$ based on 500 estimates, respectively

$$\phi(\boldsymbol{t}) = \sum_{\sigma \in \mathfrak{S}_m} p_{\sigma_1} t_1 \cdot \frac{p_{\sigma_2} t_2}{1 - p_{\sigma_1} t_2} \cdot \frac{p_{\sigma_3} t_3}{1 - (p_{\sigma_1} + p_{\sigma_2}) t_3} \cdots \frac{p_{\sigma_m} t_m}{1 - (p_{\sigma_1} + p_{\sigma_2} + \cdots + p_{\sigma_{m-1}}) t_m},$$

where each p_i is parametrized by θ as above. Then, for calculating the probability $P((Y_1, \ldots, Y_m) = \sigma, (T_1, \ldots, T_m) = (k_1, \ldots, k_m))$, we can expand

$$p_{\sigma_1} t_1 \cdot \frac{p_{\sigma_2} t_2}{1 - p_{\sigma_1} t_2} \cdot \frac{p_{\sigma_3} t_3}{1 - (p_{\sigma_1} + p_{\sigma_2}) t_3} \cdots \frac{p_{\sigma_m} t_m}{1 - (p_{\sigma_1} + p_{\sigma_2} + \cdots + p_{\sigma_{m-1}}) t_m}$$

in the Taylor series around $\boldsymbol{t} = \boldsymbol{0}$ and pick out the coefficient of $t_1^{k_1} \ldots t_m^{k_m}$.

Example 1 We illustrate the feasibility of estimating the parameter θ using simulated data for $m = 4$ and $\theta = 2$. We set sample size $n = 50$, and we repeat the estimation 500 times. We have estimated the parameter using two kinds of maximum likelihood estimators. One is the mle $\hat{\theta}_1$ based on the observations of $(Y_1, \ldots, Y_4)$ and $(T_1, \ldots, T_4)$. The other is the mle $\hat{\theta}_2$ based on the observations of only $(Y_1, \ldots, Y_4)$. Of course, $\hat{\theta}_1$ is better than $\hat{\theta}_2$, since the former uses more information than the latter. In order to compare the mle's, we have estimated the densities of the mle's based on the 500 mle's $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively. Figure 1 displays the estimated density of the mle's $\hat{\theta}_1$ and $\hat{\theta}_2$. The mean and variance of the maximum likelihood estimates $\hat{\theta}_1$ are 1.99281 and 0.01203. The mean and variance of the maximum likelihood estimates $\hat{\theta}_2$ are 2.026877 and 0.044261. As shown in Fig. 2, the estimators are not highly dependant though they use the same observations of $(Y_1, \ldots, Y_4)$. The value of the correlation coefficient between the 500 pairs of the estimates is 0.5330742. From the simulation study, we see that the additional use of observations of the waiting time fairly improves the mle for θ.

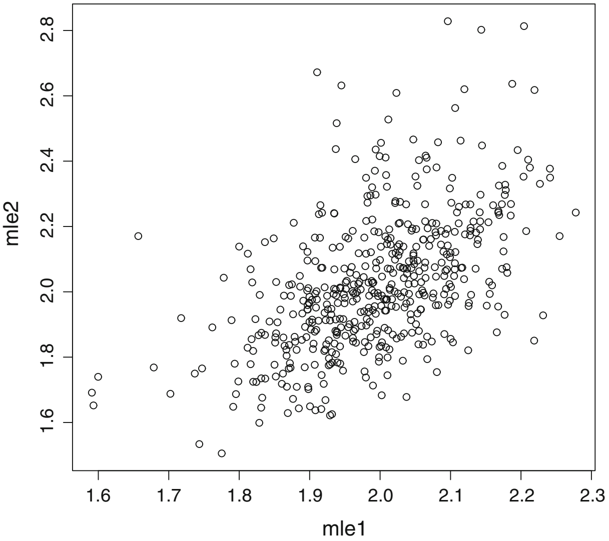

Fig. 2 Scatterplot for 500 estimates $\hat{\theta}_2$ against the corresponding $\hat{\theta}_1$ based on the simulated data of sample size 50

Next, we give a numerical example of maximum likelihood estimation of the parameters α and β by applying Theorem 1 in Sect. 2. Under the generalized Pólya urn sampling given in Sect. 2, we can estimate the number of additional balls α and β based on observations of the waiting times for drawing all the kinds of balls.

Example 2 We assume that $m = 3$, $N_1 = N_2 = N_3 = 10$ are known and that the parameters α and β are unknown positive real numbers. The following data are simulated by setting $m = 3$, $\alpha = 2$ and $\beta = 1$.

```
4, 9, 3, 11, 3, 13, 4, 3, 9, 5, 3, 5, 7, 5, 3, 15, 5, 4, 4, 10, 10, 3, 4, 3, 4,
4, 4, 4, 8, 5, 5, 10, 11, 3, 9, 4, 4, 4, 7, 6, 3, 5, 6, 4, 13, 5, 5, 7, 5, 3.
```

Using Theorem 1, we can calculate the log-likelihood function $\ell(\alpha, \beta)$ based on the data. The graph of $\ell(\alpha, \beta)$ is given in Fig. 3. Maximizing the log-likelihood function $\ell(\alpha, \beta)$ with respect to α and β, we obtain the *mle*'s $\hat{\alpha} = 1.73949$ and $\hat{\beta} = 0.76000$.

3.2 Statistical inference on the symmetric group

In this subsection, we introduce a parametric model on the symmetric group by means of the coupon collector's waiting time problem. Holland's model is well known to analyze ranked data. Let ρ be the $m-1$ dimensional irreducible representation of $\mathfrak{S}_m$. Then, the probability function of $\pi \in \mathfrak{S}_m$ is given as:

$$P_\theta(\pi) = c(\theta)e^{\mathrm{Tr}[\theta\rho(\pi)]}, \quad \text{for } \theta \in \mathrm{Mat}(m-1),$$

where $c(\theta)^{-1} = \sum_\pi e^{\mathrm{Tr}(\theta\rho(\pi))}$, $\mathrm{Mat}(m-1)$ is the set of matrices of size $(m-1) \times (m-1)$, and $Tr[A]$ means the trace of A (see Diaconis 1988). The model

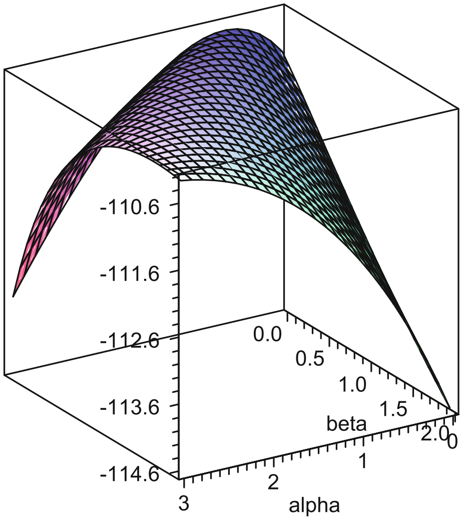

Fig. 3 The graph of the log-likelihood function $\ell(\alpha, \beta)$

is an exponential family introduced theoretically based on an irreducible representation of the symmetric group $\mathfrak{S}_m$. However, it may not be easy to understand the practical meaning of the matrix parameter θ. On the other hand, in many cases of treating practical ranking data, it is supposed that an independent random variable for each item is observed and a permutation is obtained from the ranks of the values of random variables (see, e.g., Hall and Miller 2010 and the references therein). Then, the distribution on the symmetric group $\mathfrak{S}_m$ depends on the random variables which determine rankings, and it may be difficult to study statistical inference on the symmetric group $\mathfrak{S}_m$ generally. Further, though the distributions of the random variables are assumed to be continuous theoretically, ties may occur in practical data.

Here, we give a parametric model on the symmetric group $\mathfrak{S}_m$ based on the coupon collector's waiting time problem. Assuming that $X_1, X_2, \ldots$ are $\{1, \ldots, m\}$-valued independent identically distributed random variables with $P(X_i = j) = \theta_j$, we have from (5),

$$\phi(\boldsymbol{t}) = \sum_{\sigma \in \mathfrak{S}_m} \theta_{\sigma_1} t_1 \cdot \frac{\theta_{\sigma_2} t_2}{1 - \theta_{\sigma_1} t_2} \cdot \frac{\theta_{\sigma_3} t_3}{1 - (\theta_{\sigma_1} + \theta_{\sigma_2}) t_3} \cdots \frac{\theta_{\sigma_m} t_m}{1 - (\theta_{\sigma_1} + \theta_{\sigma_2} + \cdots + \theta_{\sigma_{m-1}}) t_m}.$$

Therefore, $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ is a $\mathfrak{S}_m$-valued random variable (random permutation) and the probability $P_\theta(\sigma) = P(\boldsymbol{Y} = \sigma)$ for $\sigma \in \mathfrak{S}_m$is given as

$$P_{\boldsymbol{\theta}}(\sigma) = \frac{\theta_1 \theta_2 \cdots \theta_m}{(1 - \theta_{\sigma_1})(1 - \theta_{\sigma_1} - \theta_{\sigma_2}) \cdots (1 - \theta_{\sigma_1} - \theta_{\sigma_2} - \cdots - \theta_{\sigma_{m-1}})}, \tag{8}$$

where

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & m \\ \sigma_1 & \sigma_2 & \cdots & \sigma_m \end{pmatrix},$$

and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{m-1}) \in \boldsymbol{S}_{m-1}$ is the parameter of the distribution. Here, we set $\theta_m = 1 - \theta_1 - \theta_2 - \cdots - \theta_{m-1}$.

Based on the coupon collector's waiting time problem, a statistical model on the symmetric group $\mathfrak{S}_m$ is constructed. The meaning of the parameter is clear since it is the vector of probabilities with which the coupons of the corresponding type occur. An observation of the coupon collector's waiting time problem determines a permutation on $\mathfrak{S}_m$ uniquely, because a tie in $\tau_1, \ldots, \tau_m$ does not occur.

Let $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ be independent random permutations which follow the distribution (8). Since each $\boldsymbol{Y}_i$ is $\mathfrak{S}_m$-valued, we denote it by $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})$, where $Y_{i1} = \sigma_1, \ldots, Y_{im} = \sigma_m$ if $\boldsymbol{Y}_i = \sigma = \begin{pmatrix} 1 & 2 & \cdots & m \\ \sigma_1 & \sigma_2 & \cdots & \sigma_m \end{pmatrix}$. For $k = 1, 2, \ldots, m-1$, we define the statistics

$$N_{j_1, j_2, \ldots, j_k} = \sum_{i=1}^{n} 1(\{Y_{i1}, Y_{i2}, \ldots, Y_{ik}\} = \{j_1, j_2, \ldots, j_k\}),$$

where $1(A)$ is the indicator function of A and $\{Y_{i1}, Y_{i2}, \ldots, Y_{ik}\} = \{j_1, j_2, \ldots, j_k\}$ means the equality as sets.

Then, we have the next proposition.

Proposition 1 *If $\mathfrak{S}_m$-valued random permutations $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ independently follow the distribution $P_{\boldsymbol{\theta}}$, $\boldsymbol{\theta} \in \boldsymbol{S}_{m-1}$, then $\{N_{j_1, j_2, \cdots, j_k}\}$, ($k = 1, 2, \ldots, m$, $1 \le j_1 < j_2 < \cdots < j_k \le m$) are sufficient statistics for $\boldsymbol{\theta}$.*

Proof The joint probability function of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$ is written as:

$$\begin{aligned} & P(\boldsymbol{Y}_1 = (x_{11}, \ldots, x_{1m}), \ldots, \boldsymbol{Y}_n = (x_{n1}, \ldots, x_{nm})) \\ & \quad = \prod_{i=1}^{n} \frac{\theta_1 \theta_2 \cdots \theta_m}{(1 - \theta_{x_{i1}})(1 - \theta_{x_{i1}} - \theta_{x_{i2}}) \cdots (1 - \theta_{x_{i1}} - \theta_{x_{i2}} \cdots - \theta_{x_{im-1}})} \\ & \quad = (\theta_1 \cdots \theta_m)^n \exp\left(- \sum_{j=1}^{m} N_j \log(1 - \theta_j) - \sum_{j_1 < j_2} N_{j_1, j_2} \log(1 - \theta_{j_1} - \theta_{j_2}) \right. \\ & \qquad \left. - \cdots - \sum_{j_1 < j_2 < \cdots < j_{m-1}} N_{j_1, \ldots, j_{m-1}} \log(1 - \theta_{j_1} - \theta_{j_2} - \cdots - \theta_{j_{m-1}}) \right), \end{aligned}$$

where $\theta_m = 1 - \theta_1 - \cdots - \theta_{m-1}$. Then, the result holds from the factorization theorem. □

Example 3 The following $\mathfrak{S}_4$-valued data are simulated by setting $m = 4$ and $\boldsymbol{\theta} = (0.4, 0.3, 0.2) \in \boldsymbol{S}_3$ in the above model. The sample size is $n = 50$.

```
(2, 3, 4, 1), (3, 4, 1, 2), (3, 1, 2, 4), (1, 3, 2, 4), (4, 3, 1, 2), (3, 1, 2, 4),
(2, 3, 4, 1), (2, 3, 1, 4), (2, 4, 3, 1), (3, 1, 4, 2), (4, 2, 1, 3), (2, 1, 4, 3),
(2, 1, 3, 4), (3, 1, 4, 2), (1, 2, 4, 3), (1, 3, 2, 4), (1, 2, 3, 4), (1, 4, 2, 3),
(1, 2, 3, 4), (1, 3, 2, 4), (2, 1, 3, 4), (1, 3, 4, 2), (2, 1, 3, 4), (1, 2, 3, 4),
(1, 2, 4, 3), (1, 2, 4, 3), (4, 2, 3, 1), (1, 2, 3, 4), (4, 3, 1, 2), (2, 4, 3, 1),
(1, 3, 4, 2), (1, 4, 2, 3), (1, 3, 2, 4), (2, 4, 1, 3), (1, 2, 3, 4), (3, 2, 1, 4),
(2, 3, 1, 4), (1, 2, 3, 4), (1, 3, 2, 4), (2, 3, 1, 4), (1, 2, 3, 4), (2, 3, 1, 4),
(1, 2, 3, 4), (1, 2, 4, 3), (1, 4, 2, 3), (1, 3, 2, 4), (2, 1, 4, 3), (4, 3, 1, 2),
(1, 2, 4, 3), (3, 1, 4, 2)
```

For the above data, the values of the sufficient statistics are given by

$$\begin{aligned}&(n_1, n_2, n_3, n_4, n_{12}, n_{13}, n_{14}, n_{23}, n_{24}, n_{34}, n_{123}, n_{124}, n_{134}, n_{234})\\&\quad = (24, 14, 7, 5, 18, 13, 3, 7, 5, 4, 24, 12, 9, 5).\end{aligned}$$

Then, from Proposition 1, we can write the log-likelihood function of $\boldsymbol{\theta}$ as

$$\begin{aligned}&ll(\theta_1, \theta_2, \theta_3)\\&\quad = -24\log(1-\theta_1) - 14\log(1-\theta_2) - 7\log(1-\theta_3) - 5\log(\theta_1+\theta_2+\theta_3)\\&\qquad -18\log(1-\theta_1-\theta_2) - 13\log(1-\theta_1-\theta_3) - 7\log(1-\theta_2-\theta_3)\\&\qquad -3\log(\theta_2+\theta_3) - 5\log(\theta_1+\theta_3) - 4\log(\theta_2+\theta_1) + 26\log(1-\theta_1-\theta_2-\theta_3)\\&\qquad +45\log(\theta_1) + 41\log(\theta_2) + 38\log(\theta_3).\end{aligned}$$

Maximizing the log-likelihood function with respect to $\boldsymbol{\theta}$, we obtain the *mle* of $\boldsymbol{\theta}$ as $\widehat{\boldsymbol{\theta}} = (0.40454, 0.27290, 0.20484)$.

Next, we study a testing problem of the statistical model. For example, when $m = 4$, we consider testing the null hypothesis $H_0 : \theta_2 = \theta_3$ based on $\mathfrak{S}_4$-valued observations. In the following example, we assess the likelihood ratio test of our model.

Example 4 Setting $m = 4$ and $\Theta_0 = \{\boldsymbol{\theta} \in \boldsymbol{S}_3 : \theta_2 = \theta_3\}$ we consider testing the null hypothesis $H_0 : \boldsymbol{\theta} \in \Theta_0$ versus the alternative hypothesis $H_1 : \boldsymbol{\theta} \in \boldsymbol{S}_3 \setminus \Theta_0$ based on $\mathfrak{S}_4$-valued observations. Since the hypotheses are composite, the likelihood ratio test may be used for the testing problem. Let λ be the likelihood ratio for testing the hypothesis, i.e.,

$$\lambda = \frac{\sup_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{S}_3} L(\boldsymbol{\theta})},$$

where $L(\boldsymbol{\theta})$ is the likelihood function. If sample size is large enough, the likelihood ratio test can be used and the distribution of $-2\log\lambda$ is expected to be approximated by the Chi-squared distribution with 1 degree of freedom ($\chi^2(1)$). In order to assess whether the likelihood ratio test can be useful for a moderate sample size, we have done the following simulation study. We repeated 100000 times to calculate values

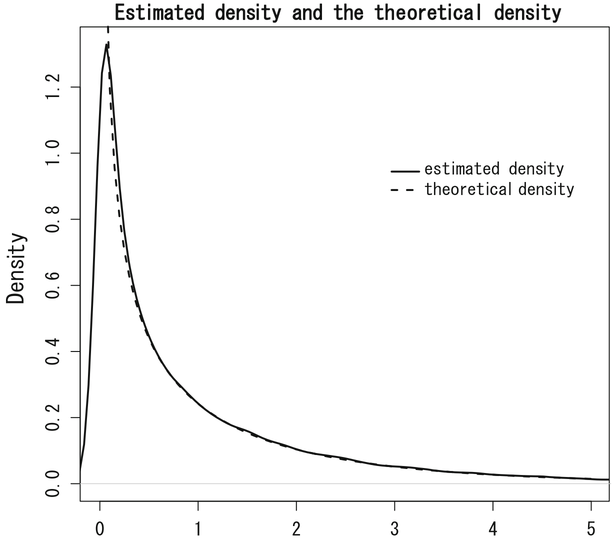

Fig. 4 Estimated density of the 10000 values of the test statistic based on 50 ranked data with parameter $\boldsymbol{\theta}$ = (1/3, 1/4, 1/4) and the density of $\chi^2(1)$

of $-2\log\lambda$ based on simulated 50 realizations of $\mathfrak{S}_4$-valued random variables which follow the distribution with $\boldsymbol{\theta} = (1/3, 1/4, 1/4) \in \Theta_0$. The mean and variance of the 100000 values of $-2\log\lambda$ are 1.016940 and 2.081478, respectively. Further, the number of values which exceed the value 3.841459 with $P(\chi^2(1) > 3.841459) = 0.05$ is 5131 in 100000. Figure 4 compares the estimated density of the 100000 values of the test statistic with the density of $\chi^2(1)$. Two densities look almost the same. Further, Fig. 5 shows the quantile plots against the theoretical distribution [$(\chi^2(1)$ distribution in this case]. Since the line of the sample quantiles of the 100000 values of the test statistic against the $\chi^2(1)$ distribution is almost straight, we may consider that the test statistic follows the $\chi^2(1)$ distribution. To investigate the distribution of the likelihood ratio statistic $-2\log\lambda$ under the alternative hypothesis, we considered two values of the parameter $\boldsymbol{\theta}_i \in \boldsymbol{S}_3 \setminus \Theta_0$ for $i = 1, 2$. Here, $\boldsymbol{\theta}_1 = (1/5, 1/3, 1/4)$ and $\boldsymbol{\theta}_2 = (1/5, 1/6, 1/3)$. We repeated 10000 times to calculate values of $-2\log\lambda$ based on simulated 50 realizations of $\mathfrak{S}_4$-valued random variables which follow the distribution with $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively.

For $\boldsymbol{\theta}_1$, the mean and variance of the 10000 values of $-2\log\lambda$ are 2.387879 and 7.654581, respectively. The number of values which exceed the value 3.841459 with $P(\chi^2(1) > 3.841459) = 0.05$ is 2173 in 10000. For $\boldsymbol{\theta}_2$, the mean and variance of the 10000 values of $-2\log\lambda$ are 8.402736 and 30.64335, respectively. The number of values which exceed the value 3.841459 with $P(\chi^2(1) > 3.841459) = 0.05$ is 7855 in 10000.

Figure 6 shows the estimated densities of the likelihood ratio test statistic under the null hypothesis based on the 100000 values of the above test statistic, under $\boldsymbol{\theta} = \boldsymbol{\theta}_1 \in H_1$ and under $\boldsymbol{\theta} = \boldsymbol{\theta}_2 \in H_1$ based on the 10000 values of the test statistic. Since the

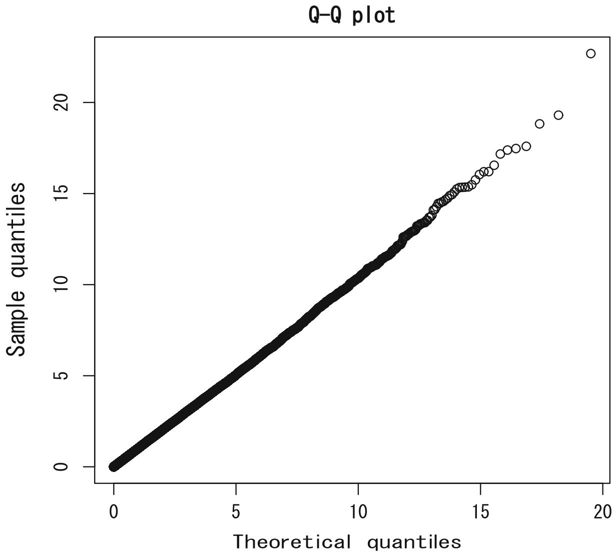

Fig. 5 Quantile plots against $\chi^2(1)$ distribution

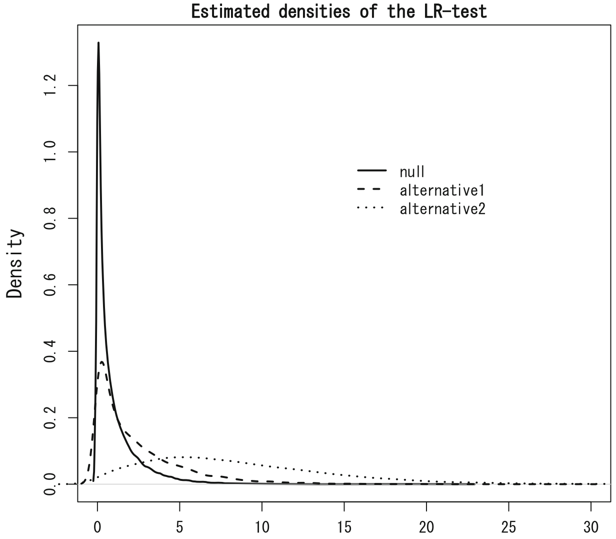

Fig. 6 Estimated densities of the test statistic based on 50 ranked data

distribution of the likelihood ratio test statistic is far from the distribution of $\chi^2(1)$, we can use the likelihood ratio test in our model.

Acknowledgment The authors wish to express their gratitude to anonymous referees for helpful suggestions and valuable remarks.

References

Charalambides, Ch A. (2002). *Enumerative combinatorics*. Boca Raton: Chapman & Hall/CRC.

Charalambides, Ch A. (2005). *Combinatorial methods in discrete distributions*. New York: Wiley.

Diaconis, P. (1988). Group representations in probability and statistics. Lecture notes—Monograph Series 11, IMS.

Graham, R. L., Knuth, D. E., Patashnik, O. (1989). *Concrete mathematics*. Massachusetts: Addison-Wesley Publishing Company.

Hall, P., Miller, H. (2010). Modeling the variability of rankings. *Annals of Statistics*, *38*, 2562–2677.

Inoue, K., Aki, S. (2008). Method for studying generalized birthday and coupon collection problems. *Communications in Statistics—Simulation and Computation*, *37*, 844–862.

Johnson, N. L., Kotz, S. (1977). *Urn models and their applications*. New York: Wiley.

Kobza, J. E., Jacobson, S. H., Vaughan, D. E. (2007). A survey of the coupon collector's problem with random sample sizes. *Methodology and Computing in Applied Probability*, *9*, 573–584.

Mahmoud, H. M. (2008). *Pólya urn models*. Boca Raton: CRC Press.