J. Fac. Edu. Saga Univ.
Vol. 33, No. 2 (II) (1985) 185 ~ 190

A Note on the Coupon Collector's Problems

Akihiro NISHI and Kentaro NOMAKUCHI *

1. Introduction

Suppose there are infinite number of coupons consisting of unequal proportions of $k(k \geqq 2)$ types. We randomly select one from this aggregate. Let $p_i (p_i > 0, \sum p_i = 1)$ be the probability that a coupon of the i-th type C_i, say, will be selected at each collection. We denote $p=(p_1,...,p_k)$ and $p^0=(1/k,...,1/k)$ which corresponds to the equiprobable case.

We continue our collections until each C_i has been obtained a_i times, where $a_1,...,a_k$ are the prescribed nonnegative integers. Let $\tau=\tau_{a_1 \cdots a_k}$ be the minimal number of collections needed for the purpose stated above. When $a_1=...=a_k=a>0$, $\tau_{a...a}$ will simply be denoted by τ_a. We will consider τ in case of $k=2$ in full generality. Newman and Shepp[5], under $p^0=(1/k,...,1/k)$ and $a_1=...=a_k=a>0$, give an asymptotic evaluation of the mean of τ_a, as $k \to \infty$. Let $Y_{i,n}$ be the number of coupons of the i-th type after n collections. It is clear that $Y=(Y_{1,n},...,Y_{k,n})$ posseses the multinomial distribution $M_k(n,p)$

$$P[Y=(y_1,...,y_k)]=\frac{n!}{y_1!...y_k!}p_1^{y_1}\cdots p_k^{y_k}.$$

Let $a+M$ be defined by, for some $i(1 \leqq i \leqq k)$ $Y_{i,a+M}=a$ and $Y_{j,a+M}<a(\forall j \neq i)$, namely, at the $a+M$ th collection one of k-types is obtained a times for the first time. Evidently M is a nonnegative integer valued random variable which is at most $(k-1)(a-1)$. When $k=2$, Bradley[2] calls $L \equiv a-M$ *excess*. In case $a_1=a_2=a>0$, it will be easily seen that τ_a can be expressed as

$$\tau_a=a+M+N_L=2a-L+N_L, \tag{1-1}$$

where the conditional distribution of N_L under $L=l$ and the type i(i.e. $Y_{i,a+M}=a$, $Y_{3-i,a+M}<a$) is negative binomial

$$P[N_L=l+t|L=l, \textit{ type } i]=\binom{l+t-1}{t}p_i^t p_{3-i}^t, \quad 1 \leqq l \leqq a, \ t=0,1,2,... . \tag{1-2}$$

Note that, under $p^0=(1/2,1/2)$, the conditional distribution of N_L does not depend on i

$$P_{p^0}[N_L=l+t|L=l]=\binom{l+t-1}{t}\left(\frac{1}{2}\right)^{l+t}, \tag{1-3}$$

where $P_{p^0}[A]$ represents the probability of an event A under p^0. Bradley[2], under $p^0=(1/2,1/2)$, obtains the probability function of the *excess* L

$$P_{p^0}[L=l]=\binom{2a-l-1}{a-1}\left(\frac{1}{2}\right)^{2a-l-1}, \quad 1 \leqq l \leqq a,. \tag{1-4}$$

* Department of Mathematics Faculty of Science Kyushu University

The probability function of τ_a, under $p^0=(1/2,1/2)$, is immediately obtained by (1-1), (1-3) and (1-4)

$$\begin{aligned} P_{p^0}[\tau_a=2a+t] &= \sum_{l=1}^{a} P_{p^0}[L=l]P_{p^0}[N_L=l+t|L=l] \\ &= \left(\frac{1}{2}\right)^{2a+t-l} \sum_{l=1}^{a} \binom{2a-l-1}{a-1}\binom{l+t-1}{t} \\ &= \left(\frac{1}{2}\right)^{2a+t-l} \binom{2a+t-1}{a+t} \end{aligned} \tag{1-5}$$

which is identical with (2-3) given in the following section.

It will be worthwhile to mention a related situation. Let $\tau_{j:1}$ be the number of collections needed for the attainment of some $j(1\leqq j\leqq k)$ different types has been obtained. Obviously $\tau_{k:1}$ is equal to τ_1. Nath[4], under $p=(p_1,...,p_k)$, calculates the probability $P[\tau_{j:1}\geqq r]$ and obtains the exact expressions of the mean and variance of $\tau_{j:1}$. Baum and Billingsley[1], under $p^0=(1/k,...,1/k)$, consider the asymptotic behaviour of $\tau_{j:1}$, as $k\to\infty$ posing appropriate conditions on j and k-j. Samuel-Cahn [8] generalizes the results of Baum and Billingsley to a situation where there exists some probability of missing collected coupons.

2. Distribution of τ

For real vectors $x=(x_1,...,x_k)$ and $y=(y_1,...,y_k)$, x is said to be majorized by y (denoted by $x<y$) if

$$\sum_{i=1}^{j} x_{[i]}\leqq \sum_{i=1}^{j} y_{[i]},\ j=1...,k-1,\ \text{and}\ \sum_{i=1}^{k} x_{[i]}=\sum_{i=1}^{k} y_{[i]},$$

where $x_{[1]}\geqq\cdots\geqq x_{[k]}(y_{[1]}\geqq\cdots\geqq y_{[k]})$ is the rearrangement of $x_1,...,x_k$ $(y_1,...,y_k)$(Marshall and Olkin[3], p.7).

A real-valued function ϕ defined on a set A of k-dimensional Euclidean space is said to be *Schur-convex(Schur-concave) if*

$$x<y \text{ on } A\Rightarrow \phi(x)\leqq\phi(y)(\phi(x)\geqq\phi(y))$$

([3], p.54). Let $P=\{p=(p_1,...,p_k)|p_i\geqq 0,\ \sum p_i=1\}$. It is obvious that $p^0=(1/k,...,1/k)<p$ and $(1,0,...,0)>p$ for all $p\in P$.

When $k\geqq 3$, it seems difficult to obtain, under $p=(p_1,...,p_k)$, the exact probability function of $\tau_{a_1\cdots a_k}$. It is clear that the events $\{\tau_{a_1\cdots a_k}\leqq x\}$ and $\{Y_{i,x}\geqq a_i,\ 1\leqq\forall i\leqq k\}$ are identical for all nonnegative integers x.

Thus we have

$$P_p[\tau_{a_1\cdots a_k}\leqq x]=P_p[Y_{i,x}\geqq a_i, 1\leqq\forall i\leqq k]. \tag{2-1}$$

When $a_1=\cdots=a_k=a>0$, the inequality (3-4) given by Olkin[6]([3], E.II.e., p.306) is identical with the Schur-concavity of $P_p[\tau_a\leqq x]$, $p\in P$. Thus this can be stated as:

If $p>q$ on P, then the distribution of τ_a under p is stochastically larger than that of τ_a under q. In particular we have

$$p > q \text{ on } P \Longrightarrow E_p(\tau_a) \geqq E_q(\tau_a)$$

where $E_p(X)$ represents the mean of a random variable X. Note that $E_p(\tau_a)$ is a Schur-convex function of p.

The proof of the above statement is immediately obtained if we use a theorem of Rinott [7] (cited in [3], p.304)

Theorem (Rinott): Let $Y=(Y_{1,x},...,Y_{k,x})$ be a random vector having the multinomial distribution $M_k(x,p)$. If ϕ is a Schur-concave function, then $E_p[\phi(Y)]$ is a Schur-concave function of p.

Let $I_{R_a}(y)$ be the indicator function of the set $R_a = \{y=(y_1,...,y_k) \mid y_i \geqq a,\ 1 \leqq \forall i \leqq k\}$. It is easily verified that $I_{R_a}(y)$ is a Schur-concave function. The theorem of Rinott and the equality (2-1) show that $E_p[I_{R_a}(Y)] = P_p[Y_{i,x} \geqq a, 1 \leqq \forall i \leqq k] = P_p(\tau_a \leqq x)$ is a Schur-concave function of p.

Remark. Let $H(p) = -\sum_{i=1}^{k} p_i \log p_i$ be the entropy of $p \in P$. It is well known that H(p) is Schur-concave([3], p.71). It will be of some interest to ascertain whether $H(p) \leqq H(q)$ implies $E_p(\tau_a) \geqq E_q(\tau_a)$ or not and/or, more strongly, $H(p) \leqq H(q)$ implies that the distribution of τ_a under p is stochastically larger than that of τ_a under q. Note that, in case of $k=2$, the majorization and the entropy induce a equivalent order relation in P namely, $p > q \Leftrightarrow H(p) \leqq H(q)$.

In the following we will confine our attention to the case $k=2$. We use $\tau = \tau_{a,b}$ and (p,q) instead of $\tau = \tau_{a_1 a_2}$ and (p_1,p_2) respectively, where a and b are positive integers and $p>0$, $q>0$, $p+q=1$.

Lemma 2-1.

$$P(\tau = a+b+t) = \left\{ \binom{a+b+t-1}{a-1} q^t + \binom{a+b+t-1}{b-1} p^t \right\} p^a q^b, \quad t = 0,1,2,\ldots, . \qquad (2\text{-}2)$$

Proof. If $Y_{1,a+b+t-1} = a-1$, then $Y_{2,a+b+t-1} = b+t \geqq b$. If $Y_{2,a+b+t-1} = b-1$, then $Y_{1,a+b+t-1} = a+t \geqq a$. Therefore it is easily seen that $\{\tau = a+b+t\} = \{Y_{1,a+b+t-1} = a-1$ and $Y_{1,a+b+t} = a\} \cup \{Y_{2,a+b+t-1} = b-1$ and $Y_{2,a+b+t} = b\}$ (disjoint union). Since

$$P(Y_{1,a+b+t-1} = a-1 \text{ and } Y_{1,a+b+t} = a) = \binom{a+b+t-1}{a-1} p^{a-1} q^{b+t} p \text{ and } P(Y_{2,a+b+t-1} = b-1$$

and $Y_{2,a+b+t} = b) = \binom{a+b+t-1}{b-1} q^{b-1} p^{a+t} q$, the equation (2-2) is verified. (q.e.d.)

When $b=a$ and $p=q=1/2$, (2-2) becomes

$$P(\tau = 2a+t) = \binom{2a+t-1}{a-1} \left(\frac{1}{2}\right)^{2a+t-1} \qquad (2\text{-}3)$$

that is identical with (1-5).

Lemma 2-2.

$$\sum_{t=a+1}^{\infty}\binom{b+t}{b}p^t q^{b+1}=\sum_{t=0}^{b}\binom{a+t}{a}p^{a+1}q^t \tag{2-4}$$

where a and b are nonnegative integers and $p>0$, $q>0$, $p+q=1$.

Proof. $Y_n \equiv Y=(Y_{1,n}, Y_{2,n})$ can be viewed as a random walk on the plane starting at the origin. According to the Fig.2-1, it is seen that the trajectories that pass through the lattice points on the line A must pass through those on the line B and vice versa. Thus we have

$$\sum_{t=0}^{b}\binom{a+t}{a}p^{a+1}q^t=P(Y_n \textit{ arrives at the line } A \textit{ after some epochs})=$$

$$P(Y_n \textit{ arrives at the line } B \textit{ after some epochs})=\sum_{t=a+1}^{\infty}\binom{b+t}{b}p^t q^{b+1}. \quad \text{(q.e.d.)}$$

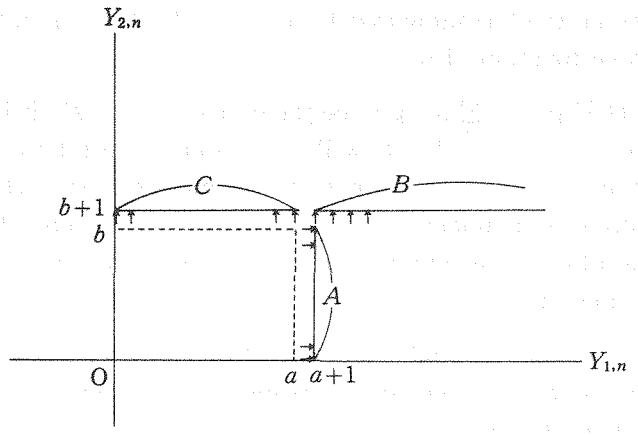

Fig. 2-1

Considering

$$\sum_{t=0}^{a}\binom{b+t}{b}p^t q^{b+1}+\sum_{t=a+1}^{\infty}\binom{b+t}{b}p^t q^{b+1}=1\left(=\sum_{t=0}^{\infty}\binom{b+t}{b}p^t q^{b+1}=q^{b+1}(1-p)^{-b-1}\right) \tag{2-5}$$

and (2-4), we have

$$\sum_{t=0}^{a}\binom{b+t}{b}p^t q^{b+1}+\sum_{t=0}^{b}\binom{a+t}{a}p^{a+1}q^t=1, \tag{2-6}$$

where the left hand side of (2-5) represents the probability of the event that Y_n arrives at the line B or C after some epochs. When $b=a$ and $p=q=1/2$, (2-4) and (2-6) yield

$$\sum_{t=a+1}^{\infty}\binom{a+t}{a}\left(\frac{1}{2}\right)^{a+t}=\sum_{t=0}^{a}\binom{a+t}{a}\left(\frac{1}{2}\right)^{a+t}=1, \tag{2-7}$$

the latter half of which is found in Bradley[2].

Theorem 2-1

$$E(\tau)=(a+b)\binom{a+b}{a}p^a q^b+\frac{a}{p}+\left(\frac{b}{q}-\frac{a}{p}\right)\sum_{t=0}^{b}\binom{a+t}{a}p^{a+1}q^t. \tag{2-8}$$

$$\mathrm{E}[\tau(\tau+1)]=[a(a+1)\left\{\binom{a+b+1}{a+1}+\binom{a+b+2}{a+1}p\right\}+b(b+1)\left\{\binom{a+b+1}{b+1}\right.$$
$$\left.+\binom{a+b+2}{b+1}q\right\}]p^aq^b+\frac{a(a+1)}{p^2}+\left\{\frac{b(b+1)}{q^2}-\frac{a(a+1)}{p^2}\right\}\sum_{t=0}^{b+1}\binom{a+1+t}{a+1}p^{a+2}q^t. \quad (2\text{-}9)$$

Proof. According to (2-2),

$$\mathrm{E}(\tau)=\sum_{t=0}^{\infty}(a+b+t)\mathrm{P}(\tau=a+b+t)=\sum_{t=0}^{\infty}\left\{a\binom{a+b+t}{a}q^t+b\binom{a+b+t}{b}p^t\right\}p^aq^b$$
$$=a\sum_{t=b}^{\infty}\binom{a+t}{a}p^aq^t+b\sum_{t=a}^{\infty}\binom{b+t}{b}p^tq^b=(a+b)\binom{a+b}{a}p^aq^b+\frac{a}{p}\sum_{t=b+1}^{\infty}\binom{a+t}{a}p^{a+1}q^t$$
$$+\frac{b}{q}\sum_{t=a+1}^{\infty}\binom{b+t}{b}p^tq^{b+1}.$$

Using (2-4) and (2-5), we obtain (2-8). The proof of (2-9) is similarly obtained. (q.e.d.)

In case $a:b=\mathrm{p}:\mathrm{q}$, (2-8) reduces

$$\mathrm{E}(\tau)=(a+b)\left\{1+\frac{a^ab^b}{(a+b)^{a+b}}\binom{a+b}{a}\right\}\sim(a+b)\left\{1+\frac{1}{\sqrt{2\pi}}\sqrt{\frac{1}{a}+\frac{1}{b}}\right\}. \quad (2\text{-}10)$$

When $b=a$ and $\mathrm{p}=\mathrm{q}=1/2$, after some calculations, we have

Corollary.

$$\mathrm{E}(\tau)=2a\left\{1+\binom{2a}{a}\left(\frac{1}{2}\right)^{2a}\right\}\sim 2a+\frac{2}{\sqrt{\pi}}a^{1/2}-\frac{1}{4\sqrt{\pi}}a^{-1/2}+O(a^{-3/2}). \quad (2\text{-}11)$$

$$V_{ar}(\tau)=2a-4a^2\frac{\binom{2a}{a}}{2^{2a}}+(2a+2)(2a+1)\frac{\binom{2a}{a-1}}{2^{2a}}-\left\{2a\frac{\binom{2a}{a}}{2^{2a}}\right\}^2$$
$$\sim(2-\frac{4}{\pi})a+\frac{6}{\sqrt{\pi}}a^{1/2}+O(a^{-1/2}). \quad (2\text{-}12)$$

Let L and $\mathrm{N_L}$ be the random variables defined in Introduction. The conditional mean and variance of $\mathrm{N_L}$ given $\mathrm{L}=l$ are, assuming $\mathrm{p}=\mathrm{q}=1/2$,

$$\mathrm{E}(\mathrm{N_L}|\mathrm{L}=l)=2l,\quad V_{ar}(\mathrm{N_L}|\mathrm{L}=l)=2l. \quad (2\text{-}13)$$

Since τ and L satisfy (1-1), it follows

$$\mathrm{E}(\tau)=2a+\mathrm{E}(\mathrm{L}), \quad (2\text{-}14)$$

$$V_{ar}(\tau)=2\,\mathrm{E}(\mathrm{L})+V_{ar}(\mathrm{L}). \quad (2\text{-}15)$$

Considering, (2-11), (2-12), (2-14) and (2-15) we have

$$\mathrm{E}(\mathrm{L})=\frac{2a}{2^{2a}}\binom{2a}{a}\sim\frac{2}{\sqrt{\pi}}a^{1/2}-\frac{1}{4\sqrt{\pi}}a^{-1/2}+O(a^{-3/2}), \quad (2\text{-}16)$$

$$V_{ar}(\mathrm{L})\sim(2-\frac{4}{\pi})a+\frac{4}{\sqrt{\pi}}a^{1/2}+O(a^{-1/2}), \quad (2\text{-}17)$$

which are slightly precise expressions found in Bradley[2].

We consider a simple asymptotic property of τ. Similar consideration done in

the proof of Lemma 2-2 tells us

$$\{\tau \leqq a+b+t\}=\{a \leqq Y_{1,a+b+t} \leqq a+t\}, \quad t=0,1,2,\ldots \tag{2-18}$$

where $Y_{1,a+b+t}$ has the binomial distribution $B(a+b+t,p)$. Thus we have

$$P(\tau \leqq a+b+t)=P\left[\frac{aq-(b+t)p}{\sqrt{(a+b+t)pq}} \leqq U_{a+b+t} \leqq \frac{(a+t)q-bp}{\sqrt{(a+b+t)pq}}\right] \tag{2-19}$$

where U_{a+b+t} is the normalized variable of $Y_{1,a+b+t}$. When $a=np$ and $b=nq$ namely, $a:b=p:q$, (2-19) becomes

$$P(\tau \leqq n+t)=P\left[-\sqrt{\frac{t^2}{n+t}\frac{p}{q}} \leqq U_{n+t} \leqq \sqrt{\frac{t^2}{n+t}\frac{q}{p}}\right] \tag{2-20}$$

Note that $a \to \infty \Leftrightarrow n \to \infty \Leftrightarrow b \to \infty$. Setting $t=x\sqrt{n}$, $x \geqq 0$, we have

Theorem 2-2.

$$\lim_{a\to\infty} P\left[\left(\tau-\frac{a}{p}\right)\Big/\sqrt{\frac{a}{p}} \leqq x\right]=\Phi\left(x\sqrt{\frac{q}{p}}\right)-\Phi\left(-x\sqrt{\frac{p}{q}}\right) \tag{2-21}$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. In case $p=q=1/2$ (necessarily $a=b$ since $a:b=p:q$),

$$\frac{\tau-2a}{\sqrt{2a}} \longrightarrow |N(0,1)|, \text{as } a \to \infty, \text{ in law} \tag{2-22}$$

where N(0,1) stands for the standard normal distribution.

References

[1] L.E. Baum and P. Billingsley, Asymptotic distributions for the coupon collector's problem. *Ann. Math. Statist.* 33(1965), 1835-1839.

[2] C. Bradley, The left-over matches. *Maths. Gaz.* 68(1984), 1-4.

[3] A.W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications.* Academic Press, 1979.

[4] H.B. Nath, On the collector's sequential sample size. *Trab. Estad.* 25(1974), 85-88.

[5] D.J. Newman and L. Shepp, The double dixie cup problem. *Amer. Math. Monthly.* 67(1960), 58-61.

[6] I. Olkin, Monotonicity properties of Dirichlet integrals with applications to the multinomial distribution and the analysis of variance. *Biometrika* 59(1972), 303-307.

[7] Y. Rinott, Multivariate majorization and rearrangement inequalities with some applications to probability and statistics. *Israel J. Math.* 15(1973), 60-77.

[8] E. Samuel-Cahn, Asymptotic distributions for occupancy and waiting time problems with probability of falling through the cells. *Ann. Probab.* 2(1974), 515-521.

(Received, September 30, 1985)