

LIMITING DISTRIBUTIONS in SEQUENTIAL OCCUPANCY PROBLEM

BY ANDREW L. RUKHIN ¹

UNIVERSITY OF MARYLAND, BALTIMORE COUNTY CAMPUS AND
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

ABSTRACT

In a general sequential allocation scheme the limiting distribution of the instant at which a treatment receives the given number of subjects is derived. Classical occupancy problems, in particular, the Banach match-box problem and the birthday problem are shown to be closely related and are discussed.

1 Introduction

In the classical occupancy problem there are K treatments (urns) and each of sequentially arriving subjects (balls) has to be assigned to one of these treatments. In a common randomization design the assignment probabilities are all equal to $1/K$, and the uniform probability assignment of subjects to treatments continues until one of the treatments receives its quota of subjects. In a more general allocation scheme different treatments may have different quotas and different probabilities to get a subject. This is one of the versions of sequential occupancy problem discussed by Johnson and Kotz (1977). Most relevant work is in Young (1961), Ivanov and Ivchenko (1978), and Anderson, Sobel and Uppuluri (1982). The later contributions (Boutsikas and Koutras, 2002) involve modifications and extensions of this scheme. Our goal in this paper is to obtain the approximate distribution of the waiting time, τ , until one of the quotas is fulfilled.

The main application of this problem is in sequential randomized allocation of patients to treatments in clinical trials. Rosenberger and Lachin (2002) survey these designs and their comparative merits. Siegmund (1985)

¹This research was supported by NSA grant #MDA904-00-1-0033.

MSC 2000 subject classifications. Primary 60E05, Secondary 60C05, 62E20, 62G30

Key words and phrases: Banach match-box problem, Birthday Problem, Combinatorial extreme-value theory, Load balancing, Multinomial trials, Poisson approximation, Probability generating function, Skew-normal distribution

in Chapter VI discusses balanced allocation rules for $K = 2$. Different randomization strategies in ESP experiments are studied by Diaconis and Graham (1981). Another application of these designs is in storage of computer files (see, for example, Azar et al. 1994, Diekman and Preis, 1999). In a typical load balancing problem a file system is formed by a computer disk divided into K , possibly unequal, size zones, so that zone i can hold the proportion p_i of fixed size data blocks. A file, which is divided into these blocks, must be stored in a disk which altogether can hold n blocks. After a block has been mapped into a zone i , which already holds $np_i - 1$ other blocks, the disk is considered full. If π_1, \dots, π_K are probabilities with which the zones are chosen, the utility of the disk for given p_1, \dots, p_K is the number of blocks which have been placed in all the disk zones when a zone starts overflowing.

Clearly this characteristic coincides with the instant τ , at which for the first time a treatment receives the prescribed number of subjects. In Section 2 we derive the form of its asymptotic distribution and give its heuristic motivation. Section 3 discusses the relationship to classical random variables involving the number of remaining matches in the Banach match-box problem. Relationship to the skew-normal distribution and some numerical results illustrating the accuracy of approximation in Theorem 1 are given in Section 4. All proofs are collected in Section 5.

2 Probability generating function for the stopping time τ and its asymptotic distribution

Let M_1, M_2, \dots be independent multinomial vectors with probabilities π_1, \dots, π_K representing the allocation of subjects to K treatments, so that $M_i(\ell) = 1$, if at stage $i, i = 1, 2, \dots$ the subject is assigned to treatment ℓ with probability $\pi_\ell, \ell = 1, \dots, K$, and $M_i(\ell) = 0$, otherwise. In a general allocation scheme k -th treatment has to obtain np_k subjects, (i.e. $n(p_1, \dots, p_K)$ is the composition vector in terminology of Diaconis and Graham, 1981, or the quota vector according to Anderson, Sobel and Uppuluri, 1982). The stopping time τ is the earliest instant at which one of the treatments fulfills its quota,

$$\tau = \min \left\{ n : \max_{\ell} \left[\sum_{i=1}^n M_i(\ell) - np_{\ell} \right] \geq 0 \right\}. \quad (1)$$

The probability distribution of τ has a rather complicated form; the prob-

ability generating function of this random variable given in Proposition 1 is more tractable relating its behavior to that of the minimum of a gamma-distributed random sample.

Let $G_{(1)}$ denote the minimum of K independent random variables G_1, \dots, G_K with gamma-distributions, $\Gamma(np_i, 1/\pi_i)$, i.e. G_i has the density

$$f_i(u) = \frac{\pi_i^{np_i}}{\Gamma(np_i)} u^{np_i-1} e^{-\pi_i u}, \quad u > 0.$$

The distribution function of $G_{(1)}$ has the form

$$F_{G_{(1)}}(u) = 1 - \prod_{j=1}^K \int_{\pi_j u}^{\infty} \frac{t^{np_j-1} e^{-t}}{\Gamma(np_j)} dt = 1 - \prod_{j=1}^K e^{-\pi_j u} \sum_{k=0}^{np_j-1} \frac{(\pi_j u)^k}{k!}.$$

If $\pi_k \equiv 1/K$, $G_{(1)}$ has the distribution of the minimum of K independent gamma-variables $\Gamma(np_i, 1)$ times K . Note that for any positive s ,

$$\int_0^{\infty} e^{(1-s)x} f_{G_1}(x) dx < \infty,$$

so that the moment generating function in Proposition 1 is well defined.

Proposition 1. If τ is defined by (1), then for any positive s ,

$$E s^{-\tau} = E e^{(1-s)G_{(1)}}.$$

Proposition 1 is known. See Anderson, Sobel and Uppuluri (1982), and Holst (1986). It can be used to derive the moments of τ as, for example, $E\tau = EG_{(1)}$.

To elucidate the asymptotic behavior of τ , let for some $r, 1 \leq r \leq K$,

$$0 < \rho = \frac{p_1}{\pi_1} = \dots = \frac{p_r}{\pi_r} < \min_{i:i>r} \frac{p_i}{\pi_i}. \quad (2)$$

Theorem 1. Under condition (2) for $n \rightarrow \infty$,

$$P\left(\frac{n\rho - \tau}{\sqrt{n\rho}} \leq x\right) \rightarrow P\left(\left(1 + \sqrt{1 - \sum_1^r \pi_k}\right) \tilde{X} - X_{(1)} \leq x\right).$$

Here $X_{(1)}$ is the minimum of r independent normal random variables X_1, \dots, X_r with $X_k \sim N(0, 1/\pi_k)$, and $\tilde{X} = \sum_1^r \pi_k X_k / \sum_1^r \pi_k$.

When $r = 1$, with $\Phi(x)$ denoting the standard normal distribution function,

$$P \left(\frac{np_1 - \pi_1 \tau}{\sqrt{np_1(1 - \pi_1)}} \leq x \right) \rightarrow \Phi(x).$$

To motivate this result, notice that

$$\tau = \min [Y_{np_1}^{(1)}, \dots, Y_{np_K}^{(K)}].$$

Here $Y_j^{(k)}$ is the number of the multinomial trial resulting in the j -th occurrence of the outcome k , so that $Y_r^{(k)} = \sum_{j=0}^{r-1} U_j^{(k)}$, where for a fixed k , $U_j^{(k)}$ are independent random variables with the geometric, parameter π_k , distribution over positive integers, and $Y_j^{(k)}$ has negative binomial distribution with parameters j and π_k , $k = 1, \dots, K$. The meaning of $U_j^{(k)} = Y_{j+1}^{(k)} - Y_j^{(k)}$, $j = 1, 2, \dots$, is the duration time between two successive occurrences (j and $j+1$) of outcome k , while $U_0^{(k)} = Y_1^{(k)}$ is the number of the first trial resulting in the outcome k .

For $k = r+1, \dots, K$ with probability one, $n^{-1/2}(Y_{np_k}^{(k)} - n\rho) \rightarrow \infty$, so that

$$n^{-1/2}(\tau - n\rho) \sim n^{-1/2} \min [Y_{np_1}^{(1)} - np_1/\pi_1, \dots, Y_{np_r}^{(r)} - np_r/\pi_r].$$

It is proven in Rukhin (2003) that for $k \neq \ell$

$$\text{Cov}(U_i^{(k)}, U_j^{(\ell)}) = - \binom{i+j}{i} \frac{\pi_k^i \pi_\ell^j}{(\pi_k + \pi_\ell)^{i+j+1}},$$

which in terms of the incomplete beta-function $I_p(r, s)$ gives

$$\begin{aligned} \text{Cov}(Y_r^{(k)}, Y_s^{(\ell)}) &= - \sum_{0 \leq i < r, 0 \leq j < s} \binom{i+j}{i} \frac{\pi_k^i \pi_\ell^j}{(\pi_k + \pi_\ell)^{i+j+1}} \\ &= - \frac{r}{\pi_k} I_{\frac{\pi_k}{\pi_k + \pi_\ell}}(r+1, s) - \frac{s}{\pi_\ell} I_{\frac{\pi_\ell}{\pi_k + \pi_\ell}}(s+1, r). \end{aligned}$$

According to this formula the correlation coefficient between $Y_{np_k}^{(k)}$ and $Y_{np_\ell}^{(\ell)}$ has the form,

$$\mathbf{corr}(Y_{np_k}^{(k)}, Y_{np_\ell}^{(\ell)}) = - \frac{p_k \pi_\ell I_{\frac{\pi_k}{\pi_k + \pi_\ell}}(np_k + 1, np_\ell) + p_\ell \pi_k I_{\frac{\pi_\ell}{\pi_k + \pi_\ell}}(np_\ell + 1, np_k)}{[p_k p_\ell (1 - \pi_k)(1 - \pi_\ell)]^{1/2}}.$$

For any $z, 0 < z < 1$,

$$I_z(np_k + 1, np_\ell) = P(\text{Bin}(n(p_k + p_\ell), z) > np_k) \\ \sim 1 - \Phi\left(\frac{n(p_k(1 - z) - p_\ell z)}{\sqrt{n(p_k + p_\ell)z(1 - z)}}\right),$$

which tends to 1 if $p_k(1 - z) > p_\ell z$, tends to 0, if $p_k(1 - z) < p_\ell z$, and its limit is 1/2 if $p_k(1 - z) = p_\ell z$. This fact implies that for $1 \leq k, \ell \leq r$,

$$\text{corr}(Y_{np_k}^{(k)}, Y_{np_\ell}^{(\ell)}) \rightarrow -\frac{p_k \pi_\ell + p_\ell \pi_k}{[p_k p_\ell (1 - \pi_k)(1 - \pi_\ell)]^{1/2}} = -\left[\frac{\pi_k \pi_\ell}{(1 - \pi_k)(1 - \pi_\ell)}\right]^{1/2}.$$

In other terms, the limiting correlation matrix of $(Y_{np_1}^{(1)}, \dots, Y_{np_r}^{(r)})$ has the off-diagonal elements $-\left[\pi_k \pi_\ell / [(1 - \pi_k)(1 - \pi_\ell)]\right]^{1/2}$. This means that the limiting covariance matrix of this vector has the diagonal elements $(1 - \pi_k)/\pi_k$, and the off-diagonal elements all equal to -1 , which exactly is the covariance structure of the vector $[1 + (1 - \sum_1^r \pi_k)^{1/2}] \tilde{X} - X_k, k = 1, \dots, r$.

3 Classical Occupancy Problems

When $K = 2$, the random variable τ appears in the classical Banach match-box problem (Feller, 1968, p 166) in which the original number of matches in each box is $n/2$. Indeed, $\tau = n - B$, if B denotes the random number of matches remaining in one match box at the instant when another box is emptied. Our problem can be interpreted as the version of the match-box problem with K match-boxes when the i th of them contains np_i matches and is used with probability π_i .

If $\pi_1 + \pi_2 = 1$, the quota fulfillment instant τ has the probability distribution

$$P(\tau = t) = \binom{t-1}{np_1-1} \pi_1^{np_1} \pi_2^{t-np_1} + \binom{t-1}{np_2-1} \pi_1^{t-np_2} \pi_2^{np_2}, \quad (3)$$

$t = n \min(p_1, p_2), \dots, n - 1$. For $p_1 = p_2 = \pi_1 = \pi_2 = 1/2$, (3) has been derived by Blackwell and Hodges (1957).

In this situation the probability generating function of τ can be expressed in terms of the incomplete beta-function,

$$Ez^\tau = \left(\frac{\pi_1 z}{1 - \pi_2 z}\right)^{np_1} I_{1-\pi_2 z}(np_1, np_2) + \left(\frac{\pi_2 z}{1 - \pi_1 z}\right)^{np_2} I_{1-\pi_1 z}(np_2, np_1),$$

which extends the formula for unequal probabilities of choosing the match-boxes given by Uppuluri and Blot (1970) (who assumed that $p_1 = p_2 = 1/2$). It follows that

$$\begin{aligned} E\tau &= \frac{np_1}{\pi_1} I_{\pi_1}(np_1, np_2) + \frac{np_2}{\pi_2} I_{\pi_2}(np_2, np_1) - \frac{\pi_1^{np_1-1} \pi_2^{np_2-1}}{B(np_1, np_2)} \\ &= \frac{np_1}{\pi_1} I_{\pi_1}(np_1 + 1, np_2) + \frac{np_2}{\pi_2} I_{\pi_2}(np_2 + 1, np_1). \end{aligned}$$

The exact distribution of τ is also known when $np_j \equiv 2$. Then the probability of the event, $\tau > t$, can be interpreted as the probability of no coincident birthdays in a K -day year, with the k th day having probability π_k , for a group of t people. Then for $t \leq K$,

$$P(\tau > t) = t! \sum_{i_1 < i_2 < \dots < i_t} \pi_{i_1} \cdots \pi_{i_t},$$

which reduces to

$$P(\tau > t) = \frac{K(K-1) \cdots (K-t+1)}{K^t},$$

when $\pi_k \equiv 1/K$.

If $p_k \equiv 1/K$, the distribution of the stopping rule τ is related to that of the maximal coordinate $\nu_{max}(t)$ of a multinomial vector with K classes in t independent trials. Indeed, with $m = n/K$,

$$P(\tau > t) = P(\nu_{max}(t) < m).$$

The generating function for the distribution of $\nu_{max}(t)$ is given in Johnson and Kotz (1977), Section 3.1.2. This function is also employed by David and Barton (1962) in Chapter 6 to derive the combinatorial extreme value distribution of $\nu_{max}(t)$ when $K \rightarrow \infty$. These authors used the Poisson approximation to the distribution of the number of treatments W which receive at least m subjects in t trials. Barbour, Holst and Janson (1992), Corollary 6.3.1, derived an upper bound on the variation distance between the distribution of W and the approximating Poisson distribution. The distribution of W also appears in the version of the birthday problem where K denotes the number of days in a year and m -way coincidences are of interest. See Example 2 in Arratia, Goldstein and Gordon (1989),

The use of $\nu_{max}(t)$ in testing uniformity of a multinomial distribution was investigated by Yusas (1972). A method to evaluate these probabilities was given by Freeman (1979), and Levin (1981) has found an expression for the cumulative multinomial distribution function through that of a sum of truncated Poisson random variables. See also Lemma 1, Chapter 2 of Kolchin, Sevast'yanov and Chistyakov (1978), who showed that the asymptotic distribution of $\nu_{max}(t)$ for $K \rightarrow \infty$ is related to that of the maximum of a Poisson random sample. Further results in the situation when $K \rightarrow \infty$ and $n \rightarrow \infty$ are given by Rukhin (2004).

Theorem 1 implies that for fixed K and x ,

$$\begin{aligned} \lim_{n \rightarrow \infty} P\left(\frac{n - \tau}{\sqrt{nK}} < x\right) &= \lim_{n \rightarrow \infty} P\left(\frac{\nu_{max}(n) - n/K}{\sqrt{n/K}} < x\right) \\ &= P(\bar{Z}_K - Z_{(1)} < x) = P(Z_{(K)} - \bar{Z}_K < x). \end{aligned}$$

Here $Z_{(1)} < Z_{(2)} < \dots < Z_{(K)}$ denote the order statistics of a standard normal random sample, Z_1, \dots, Z_K , of size K , and $\bar{Z}_K = (Z_1 + \dots + Z_K)/K$. This approximation has been suggested by Johnson and Young (1960). The distribution of the extreme deviation $\bar{Z}_K - Z_{(1)}$ can be used for outliers detection (Gumbel, 1958, Section 4.2.4). According to this result, if q_α , $0 < \alpha < 1$, is the $(1-\alpha)$ -percentile of the distribution of $\bar{Z}_K - Z_{(1)}$, or of $Z_{(K)} - \bar{Z}_K$, then for fixed K and large n , $P(\tau \geq n - q_\alpha K \sqrt{Kn}) \approx \alpha$. An approximate formula for the critical value q_α , $\alpha \rightarrow 0$, is $q_\alpha \approx z_{\alpha/K} \sqrt{(K-1)/K}$, with z_α denoting the critical point of the standard normal distribution.

For large n , the average utility can be found from Theorem 1 as

$$E\tau \sim n\rho + \sqrt{n\rho}EX_{(1)} = n\rho - \sqrt{n\rho}EX_{(r)}.$$

In computer storage applications K and r can be large numbers, in which case the classical method of limit theorems for order statistics (e. g. Leadbetter, Lindgren and Rootzen, 1983) shows that as $r \rightarrow \infty$,

$$P(X_{(r)} \leq \ell_r) \rightarrow e^{-\lambda},$$

provided that

$$\sum_{j=1}^r [1 - \Phi(\sqrt{\pi_j} \ell_r)] \rightarrow \lambda > 0. \quad (4)$$

Let $q = q(r)$ denote the number of smallest π 's, $\pi_1 = \dots = \pi_q < \min_{i:q \leq i \leq r} \pi_i$, so that $\pi_1 = \min \pi_i$. It is easy to see that (4) holds if q is large, $q \rightarrow \infty$. Then one can put $\lambda = e^{-x}$, and

$$\ell_r = \frac{1}{\sqrt{\pi_1}} \left[\sqrt{2 \log q} - \frac{\log \log q + \log 4\pi}{2\sqrt{2 \log q}} + \frac{x}{\sqrt{2 \log q}} \right].$$

The resulting formula,

$$\sqrt{\pi_1} E X_{(r)} = \sqrt{2 \log q} - \frac{\log \log q}{2\sqrt{2 \log q}} + O\left(\frac{1}{\sqrt{\log q}}\right),$$

can be used to approximate the average utility,

$$E\tau = n\rho - \sqrt{\frac{2n\rho \log q}{\pi_1}} + \frac{\sqrt{n\rho} \log \log q}{2\sqrt{2\pi_1 \log q}} + O\left(\frac{\sqrt{n}}{\sqrt{\log q}}\right).$$

4 Skew-Normal Distribution and Numerical Examples

As was discussed in Section 2, the distribution of the vector $((1 + \sqrt{1 - \sum_1^r \pi_k})\tilde{X} - X_1, \dots, (1 + \sqrt{1 - \sum_1^r \pi_k})\tilde{X} - X_r)$ is normal with the mean vector 0, and the covariance matrix Σ of this distribution has the diagonal elements $(1 - \pi_j)/\pi_j$ and the off-diagonal elements all equal to -1 . Therefore, these random variables are negatively associated (cf. Joag-Dev and Proshan, 1983), so that the distribution function,

$$G_r(x) = G_r^{\pi_1, \dots, \pi_r}(x) = P\left(\left[1 + \sqrt{1 - \sum_1^r \pi_k}\right] \tilde{X} - X_{(1)} \leq x\right),$$

satisfies the following inequality,

$$\begin{aligned} G_r(x) &= P\left(\left[1 + \sqrt{1 - \sum_1^r \pi_k}\right] \tilde{X} - X_j \leq x, \quad j = 1, \dots, r\right) \\ &\leq \prod_{j=1}^r \Phi\left(\frac{x\sqrt{\pi_j}}{\sqrt{1 - \pi_j}}\right) \leq \Phi^q\left(\frac{x\sqrt{\pi_1}}{\sqrt{1 - \pi_1}}\right). \end{aligned}$$

As we will see from (5), in the notation of Section 2 for $x \rightarrow \infty$,

$$1 - G_r(x) \sim q \left[1 - \Phi \left(\frac{x}{\sqrt{1 - \pi_1}} \right) \right],$$

so that the approximate formula for the critical value g_α^r of G_r is $g_\alpha^r \approx z_{\alpha/q} \sqrt{1 - \pi_1}$. If $\sum_1^r \pi_i = 1$, then G_r is supported by $(0, \infty)$. In this case for $r \geq 2$ and small $x > 0$,

$$G_r(x) \sim \frac{x^{r-1}}{(2\pi)^{(r-1)/2} (r-1)! \sqrt{\pi_1 \cdots \pi_r}}.$$

When $\sum_1^r \pi_i < 1$, then for $x \rightarrow -\infty$,

$$G_r(x) \sim \frac{C e^{-\frac{\sum \pi_i x^2}{2(1-\sum \pi_i)}}}{|x|^r}$$

with $C = C(\pi_1, \dots, \pi_r)$, which can be calculated recursively. These formulas, which follow from (5), give approximate values of the lower quantiles of G_r .

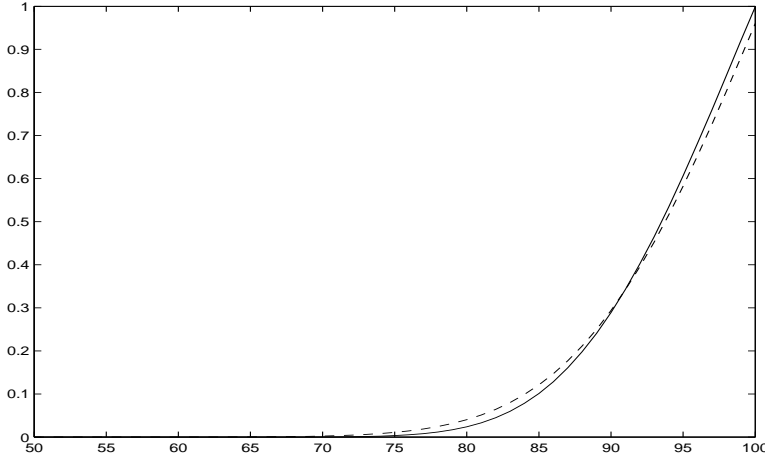


Figure 1: The plot of the cumulative distribution function of τ (solid line) and its approximation G_2 (dashed line) when $r = 2$, $\rho = 1$ and $\pi_1 = 1/2$.

The range of the ratio $(n\rho - \tau)\sqrt{n\rho}$ is from $(K - n(1 - \rho))/\sqrt{n\rho}$ to $\sqrt{n}(1 - \min p_i)/\sqrt{\rho}$. This implies that when $\rho < 1$, Theorem 1 can provide an adequate approximation only if $n > K/(1 - \rho)$, and, when $\rho = 1$, if $n > K^2$.

If one wants to match the critical α -points of two distributions, then, as $\rho\pi_1 \leq \min p_i$, $n \geq z_{\alpha/q}^2 / [\rho(1 - \pi_1)]$. Keeping in mind these restrictions on n , let us look at small values of r .

When $r = 2$, the joint distribution of $((1 + \sqrt{1 - \pi_1 - \pi_2})\tilde{X} - X_1, (1 + \sqrt{1 - \pi_1 - \pi_2})\tilde{X} - X_2)$ is singular if $\pi_1 + \pi_2 = 1$. Then for $x > 0$,

$$G_2(x) = P(\tilde{X} - \min(X_1, X_2) \leq x) = \Phi\left(\sqrt{\frac{\pi_1}{\pi_2}}z\right) - \Phi\left(-\sqrt{\frac{\pi_2}{\pi_1}}z\right)$$

and

$$E \max(\tilde{X} - X_1, \tilde{X} - X_2) = \frac{1}{\sqrt{2\pi\pi_1\pi_2}}.$$

For large n (≥ 50) the approximation in Theorem 1 is fairly good if $n\pi_1 > 5$, and it benefits from the continuity correction,

$$P(\tau > t) \approx \Phi\left(\frac{\sqrt{\frac{\pi_1}{\pi_2}}(n\rho - t - 0.5)}{\sqrt{n\rho}}\right) - \Phi\left(-\frac{\sqrt{\frac{\pi_2}{\pi_1}}(n\rho - t - 0.5)}{\sqrt{n\rho}}\right).$$

Figure 1 gives for illustration the exact distribution function of τ derived from (3) and its approximation G_2 as given above when $n = 100$, $r = K = 2$, $\rho = 1$ and $\pi_1 = 1/2$. The maximal deviation of these cumulative distribution functions is about 0.04, and this deviation decreases as n increases. When $n = 100$, $\pi_1 = 0.1$, the maximal deviation is 0.065, and it increases as π_1 decreases.

If $\pi_1 + \pi_2 < 1$, denote by α and β the diagonal elements of the square root of the covariance matrix Σ and by $-\gamma$ its off-diagonal element. Thus, $\alpha^2 + \gamma^2 = (1 - \pi_1)/\pi_1$, $\beta^2 + \gamma^2 = (1 - \pi_2)/\pi_2$ and $\gamma = 1/(\alpha + \beta)$. Then a straightforward calculation shows that with independent standard normal variables Z_1 and Z_2 ,

$$\begin{aligned} G_2(x) &= P\left(\sqrt{\frac{\alpha + \gamma}{\beta + \gamma}}Z_2 \leq Z_1 \leq \frac{1}{\beta}(x + \gamma Z_2)\right) \\ &+ P\left(\sqrt{\frac{\beta + \gamma}{\alpha + \gamma}}Z_2 \leq Z_1 \leq \frac{1}{\alpha}(x + \gamma Z_2)\right). \end{aligned}$$

It follows that the density of this distribution is

$$g_2(x) = \eta_1\varphi(\eta_1x)\Phi(\eta_1\lambda_1x) + \eta_2\varphi(\eta_2x)\Phi(\eta_2\lambda_2x).$$

Here $\varphi(x)$ is the density of the standard normal distribution, $\eta_1 = \sqrt{\frac{\pi_1}{1-\pi_1}}$, $\lambda_1 = \sqrt{\frac{\pi_1}{\pi_2(1-\pi_1-\pi_2)}}$ and $\eta_2 = \sqrt{\frac{\pi_2}{1-\pi_2}}$, $\lambda_2 = \sqrt{\frac{\pi_2}{\pi_1(1-\pi_1-\pi_2)}}$.
The distribution with the density of the form,

$$2\eta\varphi(\eta x)\Phi(\eta\lambda x), \quad \eta > 0$$

is known as the *skew-normal* distribution with the scale parameter $1/\eta$ and the skew parameter λ (Azzalini, 1985). Thus, the distribution in Theorem 1 when $r = 2$ and $\pi_1 + \pi_2 < 1$, is the mixture with equal weights of two skew-normal distributions. The skew parameters λ_1 and λ_2 are always positive. In particular, when $\pi_1 = \pi_2$, G_2 is the skew-normal distribution with the scale parameter $\sqrt{\frac{1-\pi_1}{\pi_1}}$ and the skew parameter $\lambda = \frac{1}{\sqrt{1-2\pi_1}}$. If $\pi_1 + \pi_2 \rightarrow 1$, then $\lambda_1 \rightarrow \infty$, $\eta_1 \rightarrow \sqrt{\frac{\pi_1}{\pi_2}}$, and $\lambda_2 \rightarrow \infty$, $\eta_2 \rightarrow \sqrt{\frac{\pi_2}{\pi_1}}$. Thus, $g_2(x)$ converges to $\eta_1\varphi(\eta_1 x) + \eta_2\varphi(\eta_2 x)$, $x > 0$, which is the density of $G_2(x)$ when $\pi_1 + \pi_2 = 1$.

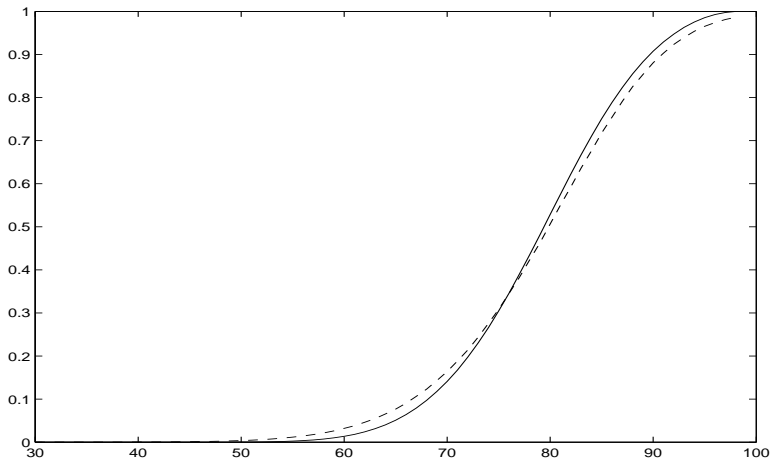


Figure 2: The plot of the cumulative distribution function of τ (solid line) and its approximation G_2 (dashed line) when $n = 100$, $K = 3$, $r = 2$, $p_1 = p_2 = 0.3$, $\rho = 8/9$ and $\pi_1 = \pi_2 = 0.3375$.

The formula for the expected value,

$$E \max((1 + \sqrt{1 - \pi_1 + \pi_2})\tilde{X} - X_1, (1 + \sqrt{1 - \pi_1 + \pi_2})\tilde{X} - X_2) = \frac{\sqrt{\pi_1 + \pi_2}}{\sqrt{2\pi\pi_1\pi_2}},$$

follows from the known formula for the skew-normal distribution. For large n , according to Theorem 1,

$$E\tau \approx n\rho - \frac{\sqrt{n\rho(\pi_1 + \pi_2)}}{\sqrt{2\pi\pi_1\pi_2}}.$$

Figures 2 and 3 show the distribution function of τ (derived as the sum of multinomial probabilities for $K = 3$ as in (6)) and the distribution functions G_2 or G_3 (the latter obtained via numerical integration from (5)) for $n = 100$, $\rho = 8/9$ or $\rho = 1$ and $p_1 = p_2 = 0.3$.

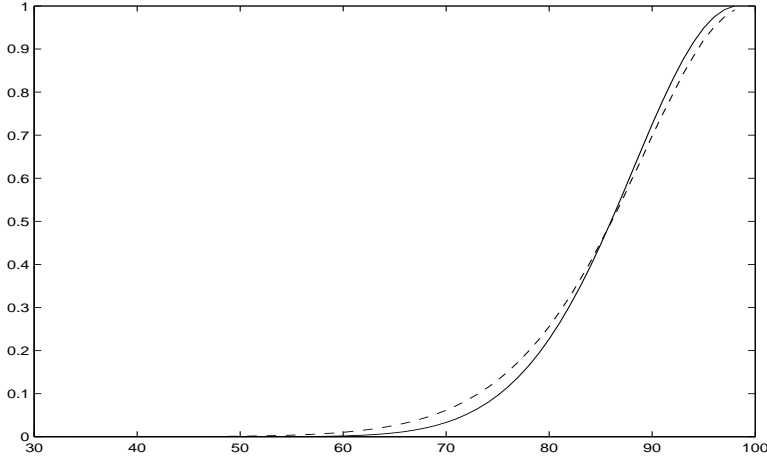


Figure 3: The plot of the cumulative distribution function of τ (solid line) and its approximation G_3 (dashed line) when $n = 100$, $r = K = 3$, $\rho = 1$ and $\pi_1 = \pi_2 = 0.3$, $\pi_3 = 0.4$.

For $r \geq 2$, one can use the recurrent formula relating the distribution function $G_r(x) = G_r^{\pi_1, \dots, \pi_r}(x)$ to $G_{r-1}(x) = G_r^{\pi_1/(1-\pi_r), \dots, \pi_{r-1}/(1-\pi_r)}(x)$,

$$\begin{aligned} G_r(x) &= \sqrt{\frac{\pi_r}{1-\pi_r}} \int_{-\infty}^x G_{r-1} \left(x\sqrt{1-\pi_r} + \frac{\pi_r y}{\sqrt{1-\pi_r}} \right) \varphi \left(y\sqrt{\frac{\pi_r}{1-\pi_r}} \right) dy \\ &= \int_{-\infty}^x \sqrt{\frac{\pi_r}{1-\pi_r}} G_{r-1}(\sqrt{1-\pi_r}x + \sqrt{\pi_r}y) \varphi(y) dy. \end{aligned} \quad (5)$$

This formula is proved by conditioning on $(1 + \sqrt{1 - \sum_1^r \pi_k})\tilde{X} - X_r$. It shows

that for $x \rightarrow \infty$,

$$1 - G_r(x) \sim \sum_{j=1}^K \left[1 - \Phi \left(\frac{x}{\sqrt{1 - \pi_j}} \right) \right].$$

Clearly,

$$G_1^\pi(x) = \Phi \left(\sqrt{\frac{\pi}{1 - \pi}} x \right),$$

and G_1^1 is the distribution function of a point mass at 0. Thus, the approximate formulas for quantiles given in the beginning of this Section are also true when $r = 1$.

If $\nu = (\nu_1, \dots, \nu_K)$, $\sum \nu_i = t$, is a multinomial vector, with probabilities (π_1, \dots, π_K) , $n > t$, then the identity

$$P(\tau > t) = P(\nu_1 < np_1, \dots, \nu_K < np_K)$$

gives an approximate formula for the cumulative distribution function of multinomial distribution via Theorem 1. This approximation can be good only if $K - r$ is fairly small. In an example, let $K = 3$, $t = 70$, $\pi_1 = \pi_2 = 0.3$, $\pi_3 = 0.4$ and assume that the value of the probability $P(\nu_1 < 30, \nu_2 < 30, \nu_3 < 40)$ is desired. Writing this probability as $P(\tau > 70)$ for $n = 100$, $p_1 = p_2 = 0.375$, $p_3 = 0.25$, so that $r = 3$, $\rho = 1$ one obtains from the approximation of Theorem 1 the value 0.9557 while the exact value is 0.9668. However, if for the same t and the same π_1, π_2, π_3 , one needs the probability $P(\nu_1 < 30, \nu_2 < 50, \nu_3 < 20)$, this approximation with $n = 100$, $p_1 = 0.3, p_2 = 0.5, p_3 = 0.2$, $r = 1, \rho = 0.75$ gives less accurate answer 0.6543 instead of 0.5950.

The approximation in Theorem 1 is too crude for large K and small values of p_i . It gives a poor answer in the case when $t = K$ and $p_i = \pi_i \equiv 1/K$, considered by Mallows (1968) and by Levin (1981). However, it leads to a good approximation when $t \approx n\rho$ is large, in which situation the Bonferroni-Mallows bounds can be wide and the calculations needed in the first-order normal approximation formula (and especially the four term Edgeworth expansion) in Levin's (1981) formula may be difficult. Notice that the exact multinomial probabilities can be obtained from the tables in Sobel, Uppuluri and Frankowski (1985). These probabilities can be also derived from the recursive formulas for the multinomial distribution, and from the formula for the exponential probability generating function, $R(z)$, given in the next Section.

5 Proofs of Proposition 1 and Theorem 1

Although the statement of Proposition 1 is known, for completeness sake we give here a short proof. Notice that

$$P(\tau > t) = \sum_{\substack{\ell_1 + \dots + \ell_K = t \\ \ell_1 < np_1, \dots, \ell_K < np_K}} \binom{t}{\ell_1 \dots \ell_K} \pi_1^{\ell_1} \dots \pi_K^{\ell_K}, \quad (6)$$

so that

$$\begin{aligned} R(z) &= \sum_t P(\tau > t) \frac{z^t}{t!} = \sum_{\ell_1 < np_1, \dots, \ell_K < np_K} \frac{z^{\ell_1 + \dots + \ell_K} \pi_1^{\ell_1} \dots \pi_K^{\ell_K}}{\ell_1! \dots \ell_K!} \\ &= \prod_1^K \left(\sum_{\ell_j < np_j} \frac{(\pi_j z)^{\ell_j}}{\ell_j!} \right). \end{aligned}$$

Therefore,

$$\sum_t P(\tau > t) \frac{z^{t-1}}{(t-1)!} = \sum_t P(\tau > t+1) \frac{z^t}{t!} = R'(z),$$

and

$$\sum_t P(\tau = t) \frac{z^{t-1}}{(t-1)!} = R(z) - R'(z).$$

Multiplying this identity by e^{-sz} , $s > 0$, and integrating over positive z , one obtains

$$\begin{aligned} E s^{-\tau} &= \sum_t P(\tau = t) s^{-t} = \sum_t P(\tau = t) \int_0^\infty e^{-sz} z^{t-1} dz / (t-1)! \\ &= \int_0^\infty e^{-sz} [R(z) - R'(z)] dz = 1 + (1-s) \int_0^\infty e^{-sz} R(z) dz. \end{aligned}$$

The last formula here follows from integration by parts as $R(z)$ is a polynomial of degree less than n , so that $R(z)e^{-sz} \rightarrow 0$ as $z \rightarrow \infty$, and $R(0) = 1$. Since

$$R(z) = \prod_{j=1}^K e^{z\pi_j} \int_{z\pi_j}^\infty \frac{u^{np_j-1} e^{-u}}{\Gamma(np_j)} du = e^z \prod_{j=1}^K \int_{z\pi_j}^\infty \frac{u^{np_j-1} e^{-u}}{\Gamma(np_j)} du = e^z P(G_{(1)} > z),$$

another integration by parts shows that

$$\begin{aligned} E s^{-\tau} &= 1 + (1-s) \int_0^\infty e^{(1-s)z} P(G_{(1)} > z) dz \\ &= \int_0^\infty e^{(1-s)z} f_{G_1}(z) dz = E e^{(1-s)G_{(1)}}. \end{aligned} \quad (7)$$

PROOF OF THEOREM 1. Rewrite (7) in the form

$$\begin{aligned} & E s^{(n\rho-\tau)/\sqrt{n\rho}} \\ &= s^{\sqrt{n\rho}} \sum_{j=1}^K \frac{\pi_j^{np_j}}{\Gamma(np_j)} \int_0^\infty e^{z(1-s^{1/\sqrt{n\rho}})} e^{-\pi_j z} z^{np_j-1} \prod_{\ell \neq j} \left[\int_{z\pi_\ell}^\infty \frac{v^{np_\ell-1} e^{-v\pi_\ell}}{\Gamma(np_\ell)} dv \right] dz \\ &= s^{\sqrt{n\rho}} \sum_{j=1}^K \frac{1}{\Gamma(np_j)} \int_0^\infty e^{u(1-s^{1/\sqrt{n\rho}})/\pi_j} e^{-u} u^{np_j-1} \prod_{\ell \neq j} \left[\int_{u\pi_\ell/\pi_j}^\infty \frac{v^{np_\ell-1} e^{-v\pi_\ell}}{\Gamma(np_\ell)} dv \right] du, \end{aligned} \quad (8)$$

and make a transformation of variables in the last integral, $u = np_j + \sqrt{np_j}y$. Then

$$\begin{aligned} \frac{u(1-s^{1/(\sqrt{n\rho})})}{\pi_j} &= -\frac{p_j \sqrt{n} \log s}{\pi_j \sqrt{\rho}} - \frac{y \sqrt{p_j} \log s}{\pi_j \sqrt{\rho}} - \frac{p_j}{2\pi_j \rho} (\log s)^2 \\ &\quad + (y^2 + 1)(\log s)^2 O\left(\frac{1}{n^{1/2}}\right). \end{aligned}$$

According to the local limit theorem for exponential random variables

$$\begin{aligned} & \exp\{-(np_j + \sqrt{np_j}y)\} \frac{\sqrt{np_j}}{\Gamma(np_j)} (np_j + \sqrt{np_j}y)^{np_j-1} \\ &= \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \left[1 + (y^3 - 3y) O\left(\frac{1}{n^{1/2}}\right) \right]. \end{aligned}$$

The Central Limit Theorem for these random variables shows that

$$\int_{np_j + \sqrt{np_j}y}^\infty \frac{v^{np_\ell-1} e^{-v}}{\Gamma(np_\ell)} dv \sim P\left(Z > \frac{n(p_j - p_\ell) + \sqrt{np_j}y}{\sqrt{np_\ell}}\right),$$

which tends to zero if $p_j\pi_\ell > p_\ell\pi_j$, tends to 1 if $p_j\pi_\ell < p_\ell\pi_j$, and its limit is $1 - \Phi(y)$, if $p_j\pi_\ell = p_\ell\pi_j$. Thus, all terms in (8) for $j > r$ tend to zero. By combining the formulas above, one gets

$$\begin{aligned} E_S^{(n\rho-\tau)/\sqrt{n\rho}} &\rightarrow e^{-(\log s)^2/2K} \sum_1^r \frac{\sqrt{\pi_j}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} s^{-y} e^{-\pi_j y^2/2} \prod_{\ell \neq j} [1 - \Phi(\sqrt{\pi_\ell} y)] dy \\ &= e^{-(\log s)^2/2} E_S^{-X_{(1)}}. \end{aligned}$$

To recognize this integral as the probability generating function of $\left(1 + \sqrt{1 - \sum \pi_k}\right) \tilde{X} - X_{(1)}$ with $X_j \sim N(0, 1/\pi_j)$, we show first that $\tilde{X} - X_{(1)}$ and \tilde{X} are independent. Indeed, introduce a new model with observations $Y_j = \theta + X_j$, so that θ is the location parameter. Then \tilde{Y} is a complete sufficient statistic for this parameter. In this model, $\tilde{Y} - Y_{(1)} = \tilde{X} - X_{(1)}$ is a similar (ancillary) statistic, so that the Basu Lemma (Casella and Berger, 2002) implies independence of \tilde{Y} (or \tilde{X}) and $\tilde{X} - X_{(1)}$. Since \tilde{X} has a normal distribution with variance $1/\sum_k \pi_k$,

$$E_S^{-X_{(1)}} = E_S^{(\tilde{X} - X_{(1)}) - \tilde{X}} = E_S^{\tilde{X} - X_{(1)}} E_S^{-\tilde{X}} = E_S^{\tilde{X} - X_{(1)}} e^{(\log s)^2/(2\sum \pi_k)}.$$

Therefore,

$$\begin{aligned} e^{-(\log s)^2/2} E_S^{-X_{(1)}} &= \exp \left\{ [1 - (\sum_k \pi_k)^{-1}] (\log s)^2/2 \right\} E_S^{\tilde{X} - X_{(1)}} \\ &= E_S^{\sqrt{1 - \sum_k \pi_k} \tilde{X}} E_S^{\tilde{X} - X_{(1)}} = E_S^{(1 + \sqrt{1 - \sum_k \pi_k}) \tilde{X} - X_{(1)}}, \end{aligned}$$

which proves Theorem 1.

REFERENCES

- ANDERSON, K., SOBEL, M, AND UPPULURI, R. R., 1982. Quota fulfillment times, *Can. J. Statist.* 10, 73–88.
- ARRATIA, R., GOLDSTEIN, L., AND GORDON, L., 1990. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.* 17, 9–25.
- AZAR, Y., BRODER, A. Z., KARLIN, A. R., AND UPFAL, E. 1994. Balanced allocations. in *Proceedings of the 26th ACM Symposium on Theory of Computing (STOC)* Montreal, Canada.

- AZZALINI, A. 1985. A class of distributions including the normal ones, *Scand. J. Statist.* 12, 171–178.
- BARBOUR A. D., HOLST, L. AND JANSON, S., 1992. *Poisson Approximation*, Clarendon Press, Oxford.
- BOUTSIKAS M. V., AND KOUTRAS, M. V. 2002. On the number of overflowed urns and excess balls in an allocation model with limited urn capacity, *J. Statist. Plann. Inference* 104, 259–286.
- CASELLA G., AND BERGER, R., 2002. *Statistical Inference*. 2nd edition, Duxbury.
- DAVID, F. N., AND BARTON D. E., 1962. *Combinatorial Chance*. New York, Hafner Publishing Co.
- DIACONIS, P., AND GRAHAM, R.L., 1981. The analysis of sequential experiments with feedback to subjects, *Ann. Statist.* 9, 3-23.
- DIEKMAN, R., AND PREIS, R., 1999. Load balancing strategies for distributed memory machines, Chapter 7 In TOPPING, B. H. V., ED. *Parallel and Distributed Processing for Computational Mechanics: Systems and Tools*. Saxe-Coburg Publications, Edinburgh, UK.
- FELLER, W., 1968. *An Introduction to Probability Theory and Its Applications, Vol I*. Wiley, New York.
- FREEMAN, P. R., 1960. Exact distribution of the largest multinomial frequency, *Appl. Statist.* 28, 333-336.
- GUMBEL, E. J., 1958. *Statistics of Extremes*. Columbia University Press, New York.
- HOLST, L., 1986. On birthday, collectors', occupancy and other classical urn problems. *Internat. Statist. Rev.* 54, 15-27.
- IVANOV A. V., AND IVCHENKO, G. I., 1978. On some boundary functionals for Markov walks, *Math. Notes* 23, 315–326.
- JOAG-DEV, K., AND PROSHAN. F., 1983. Negative association of random variables, with applications. *Ann. Statist.* 11, 286-295.

- JOHNSON, N. L., AND KOTZ, S., 1977. *Urn Models and Their Applications. An Approach to Modern Discrete Probability Theory*. Wiley, New York.
- JOHNSON, N. L., AND YOUNG, D. H., 1960. Some applications of two approximations to the multinomial distribution, *Biometrika* 47, 463-469.
- KOLCHIN, V. F., SEVAST'YANOV, B. A., AND CHISTYAKOV, V. P., 1978. *Random Allocations*. Whinston Sons, Washington, DC.
- LEADBETTER, M. R., LINDGREN, G. AND ROOTZEN, H., 1983. *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.
- LEVIN, B., 1981. A representation for multinomial distribution function. *Ann. Statist.* 9, 1123-1126.
- ROSENBERGER, W. F. AND LACHIN, J. M., 2002. *Randomization in Clinical Trials: Theory and Practice*. Wiley, New York.
- RUKHIN, A. L., 2003. Covariance identity for multinomial trials, *Statist&Probab. Letters* **63**, 107-112.
- RUKHIN, A. L., 2004. Gamma-distribution order statistics, maximal multinomial frequency and randomization designs, *J. Statist. Plann. Inference* to appear.
- SIEGMUND, D. 1985. *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York.
- SOBEL, M, UPPULURI, R. R. AND FRANKOWSKI, K. 1985. *Tables in Mathematical Statistics*. Vol 9, American Mathematical Society, Providence RI.
- UPPULURI, R. R. AND BLOT, J., 1970. A probability distribution arising in a riff-shuffle. In PATIL, G. P., ED. *Random Counts in Scientific Work*. Pennsylvania State University Press, University Park.
- YOUNG, D. H., 1961. Quota fulfillment using unrestricted random sampling, *Biometrika* 48, 333-342.

YUSAS, I. S., 1972. On the distribution of the maximum frequency of a multinomial distribution. *Theor. Probab. Appl.* 17, 712-717.

University of Maryland, Baltimore County Campus
1000 Hilltop Circle, Baltimore, Maryland 21250
E-mail: rukhin@math.umbc.edu and
Statistical Engineering Division
National Institute of Standards and Technology
Bldg 820, Gaithersburg, MD 20899
E-mail: rukhin@cam.nist.gov