

In Memory of Shanti Gupta

**GAMMA-DISTRIBUTION ORDER STATISTICS, MAXIMAL
MULTINOMIAL FREQUENCY and RANDOMIZATION
DESIGNS**

BY ANDREW. L. RUKHIN ¹

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

ABSTRACT

The paper discusses the relationship between the gamma-distribution order statistics, the maximal cell frequency in multinomial trials and the distribution arising in a commonly used balanced randomization scheme with several treatments. The latter is used in clinical trials, load balancing in computer files storage, parapsychological experiments, etc., where a randomization design starts with the uniform probability assignment of subjects to treatments. The limiting distributions of the waiting time until a treatment receives the given number of subjects are described. The relationship to classical occupancy problems, in particular, to the Banach match-box problem and to the birthday problem is discussed.

1 Introduction

To force balance in an assignment, among K treatments, sequentially arriving subjects, a common randomization design starts with the uniform probability of subjects to treatments until one of the treatments receives its quota of m subjects. It is of interest to evaluate the distribution of the waiting time.

Equal allocation to each of the treatments seems to be fairly standard practice in comparative studies, especially, in clinical trials. There are many sequential randomization schemes (Efron's biased coin design, Wei's adaptive urn design, etc) which are non-balanced, i.e. they do not guarantee the equal distribution of the subjects per treatment. An excellent survey of these

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, Clinical trials, Combinatorial extreme-value theory, Load balancing, Multinomial trials, Normal order statistics, Poisson approximation, Probability generating function, Renewal process

designs and their comparative merits in clinical trials is given in Rosenberger and Lachin (2002). The role of randomization in ESP experiments is discussed by Diaconis and Graham (1981). Several applications of the related maximal multinomial frequency are discussed by Levin (1983).

Another application of randomization designs is in storage of computer files (Avigdor, 1999). In a typical load balancing problem one has a file system, in which a computer disk is divided into K equal size zones, so that each zone can hold m fixed size data blocks. A file, which is divided into the same size blocks, must be stored in a disk which can hold mK blocks. When a file is written to the disk, each block is mapped into a disk zone via a hash function, so that after this mapping the blocks are randomly distributed into the disk zones. After a block has been mapped into a zone which already holds $m - 1$ other blocks, the disk is considered full. The utility of the disk for given m and K is the number of blocks which have been placed in all 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. Section 2 reviews the form of its limiting (non-normal) distribution under different assumptions. In Section 3 some of these results are extended to a more general allocation scheme where different treatment may have different quotas. Section 4 discusses the relationship to classical random variables involving the number of remaining matches in the Banach match-box problem and explores a continuous time setting. A summary of formulas for the average utility is given in Section 5. All proofs are collected in the Appendix.

2 Probability generating function for the waiting time τ

Let M_1, M_2, \dots be K -nomial random vectors 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 $\ell, \ell = 1, \dots, K$, and $M_i(\ell) = 0$, otherwise. For given K and m , denote by $\tau = \tau(K, m)$ the waiting time

$$\tau = \min \left\{ n : \max_{\ell} \sum_{i=1}^n M_i(\ell) \geq m \right\}. \quad (1)$$

The probability distribution of random variable τ has a rather compli-

cated form; the probability generating function of this random variable given in Theorem 1 is more tractable relating the behavior of τ to that of the minimum of a gamma-distributed random sample.

Let $G_{(1)} < G_{(2)} < \dots < G_{(K)}$ denote the order statistics of a random sample of size K from the gamma-distribution, $\Gamma(m, 1)$, with the density

$$f_m(u) = \frac{1}{\Gamma(m)} u^{m-1} e^{-u}, \quad u > 0.$$

The distribution of these order statistics has been studied in detail by Gupta (1960).

THEOREM 1. For any positive z

$$Ez^{-\tau} = Ee^{K(1-z)G_{(1)}}.$$

Theorem 1 is known even in more general settings like Proposition 1 below. See Anderson, Sobel, Uppuluri (1982) and Holst (1986). It can be used to derive the moments of τ as, for example, $E\tau = KEG_{(1)}$.

THEOREM 2. For $m \rightarrow \infty$, the distribution of $m^{-1/2}(Km - \tau)$ converges weakly to that of $K(\bar{Z}_K - Z_{(1)})$. 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 $\bar{Z}_K = (Z_1 + \dots + Z_K)/K$.

The distribution of τ is also related to that of the maximal coordinate $\eta_{max}(n)$ of a multinomial vector with K classes in n independent trials under equal probabilities $1/K$. Indeed,

$$P(\tau > n) = P(\eta_{max}(n) < m).$$

The generating function for probabilities $P(\eta_{max}(n) < m)$ is given in Riordan (1958), ex 6, Chapter 5. For a fixed m ,

$$\begin{aligned} \sum_{n=0}^{mK} \frac{(tK)^n}{n!} P(\eta_{max}(n) < m) &= \left(1 + t + \dots + \frac{t^{m-1}}{(m-1)!} \right)^K \\ &= \sum_{n=0}^{mK} \frac{(tK)^n}{n!} P(\tau > n). \end{aligned}$$

This function is also employed by David and Barton (1962) in Chapter 6 to derive the combinatorial extreme value distribution of $\eta_{max}(n)$ 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 n trials. Barbour, Holst and Janson (1992), Corollary 6.3.1, p 117, obtained an upper bound on the variation distance between the distribution of W and the approximating Poisson distribution. For this purpose W was represented as the sum of K dependent Bernoulli random variables with the parameter

$$\pi = \sum_{i=m}^n \binom{n}{i} \frac{(K-1)^{n-i}}{K^n},$$

so that $P(\eta_{max}(n) < n) = P(W = 0) \approx \exp\{-K\pi\}$.

Theorem 2 implies that for fixed K and x

$$\begin{aligned} \lim_{m \rightarrow \infty} P\left(\frac{mK - \tau}{K\sqrt{m}} < x\right) &= \lim_{n \rightarrow \infty} P\left(\frac{\eta_{max}(n) - n/K}{\sqrt{n/K}} < x\right) \\ &= P(Z_{(K)} - \bar{Z}_K < x). \end{aligned} \quad (2)$$

This approximation has been suggested by Johnson and Young (1960). The distribution of the extreme deviation $Z_{(K)} - \bar{Z}_K$ has been used for outliers detection (Gumbel, 1958, Section 4.2.4). This result allows, for example, to find for fixed K and α , $0 < \alpha < 1$, the critical value q_α , a percentile of the distribution of $Z_{(K)} - \bar{Z}_K$, such that for large m , $P(\tau \geq mK - q_\alpha K\sqrt{m}) \approx \alpha$. For a fixed K , with z_α denoting the critical point of standard normal distribution, $q_\alpha \approx z_{\alpha/K} \sqrt{(K-1)/K}$. Notice that according to the inequality in Mallows (1968),

$$P(\eta_{max}(n) < m) \leq \left[\sum_{i=0}^{m-1} \binom{n}{i} \frac{(K-1)^{n-i}}{K^n} \right]^K,$$

so that by (2) with $m = n/K + x\sqrt{n/K}$

$$P(Z_{(K)} - \bar{Z}_K < x) \leq \Phi^K\left(\frac{x}{\sqrt{1-1/K}}\right).$$

This gives an accurate approximation for the distribution function of the extreme deviation for large x , although for large K , (2) is not a good approximation for the distribution of τ .

In the computer storage applications K may be a large number, and the formula, $EZ_{(K)} = \sqrt{2 \log K} - \log \log K / (2\sqrt{2 \log K}) + O(1/\sqrt{\log K})$, can be used in the first approximation to the average relative utility,

$$\frac{E\tau}{Km} = 1 - \sqrt{\frac{2 \log K}{m}} + \frac{\log \log K}{2\sqrt{2m \log K}} + O\left(\frac{1}{\sqrt{m \log K}}\right).$$

It is worth noting that a fundamental study of the moments, $EZ_{(K)}$, has been performed by Bose and Gupta (1959). This study became a standard part of many textbooks on order statistics (see for example, Arnold, Balakrishnan and Nagaraja, 1993).

To investigate the situation when K tends to infinity, let us look first at the case of fixed m . The classical method of limit theorems for order statistics shows that as $K \rightarrow \infty$,

$$P(G_{(1)} > x(m!/K)^{1/m}) \rightarrow e^{-x^m},$$

i.e. $G_{(1)}$ asymptotically has a Weibull distribution. Lemma 1 in Appendix implies that the limiting distribution of τ/K is the same,

$$P(\tau > x(m!)^{1/m} K^{1-1/m}) \rightarrow e^{-x^m}. \quad (3)$$

The referee has noticed that this formula also can be obtained from Levin's (1981) representation of the multinomial cumulative distribution function which provides an efficient method to evaluate the probabilities $P(\eta_{max}(n) < m)$.

As was mentioned, the distribution of the number W of treatments which receive at least m subjects admits the Poisson approximation, and this number 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), where W is represented as the sum of $\binom{n}{m}$ Bernoulli random variables taking value 1 with probability $1/K^{m-1}$. This representation for $m > 2$ gives with $\lambda = \binom{n}{m} K^{1-m}$

$$|P(\tau > n) - e^{-\lambda}| = O\left(\frac{1}{n^{1/(m-1)}}\right).$$

When $m = 2$,

$$|P(\tau > n) - e^{-\lambda}| < \frac{16\lambda(1 - e^{-\lambda})}{n},$$

and this is an order sharp bound. Application of these inequalities to (3) shows that

$$|P(\tau > x(m!)^{1/m} K^{1-1/m}) - e^{-x^m}| = O\left(\frac{1}{K^{1/m}}\right).$$

Similar bounds in Section 6.2 in Barbour, Holst and Janson (1992) are sharp only when m is fixed.

To investigate further the situation when K is large, denote by $\alpha = \alpha_{Km}$ the solution of the equation

$$\frac{1}{\Gamma(m)} \int_0^\alpha u^{m-1} e^{-u} du = e^{-\alpha} \sum_{k=m}^{\infty} \frac{\alpha^k}{k!} = \frac{1}{K}. \quad (4)$$

In other terms α is the percentile of order K^{-1} of the gamma-distribution, $\Gamma(m, 1)$, or m is the percentile of the same order of the Poisson distribution with parameter α .

We also put

$$\beta_{Km} = K f_m(\alpha_{Km}) = \frac{K}{\Gamma(m)} \alpha_{Km}^{m-1} e^{-\alpha_{Km}}. \quad (5)$$

The asymptotic behavior of τ is determined by that of $m/\log K$. The following result, which is based on the asymptotic analysis of sequences α_{Km} and β_{Km} , deals with the situation when $m/\log K \rightarrow \infty$.

THEOREM 3. If for $K, m \rightarrow \infty$, $\frac{m}{\log K} \rightarrow \infty$, then for all real x

$$P\left(\frac{\tau}{K} > mw \left(\frac{2 \log K - \log \log K}{2m}\right) + \sqrt{\frac{m}{2 \log K}} \left(\frac{1}{2} \log 4\pi + x\right)\right) \rightarrow e^{-e^x}.$$

Here $w = w(u)$, $u > 0$, denotes the solution of the equation

$$w - \log w - 1 = u,$$

which is smaller than 1,

$$w(u) = 1 - \sqrt{2u} + O(u).$$

Under conditions of Theorem 4, $\alpha_{Km}/m \rightarrow 1$ and $\alpha_{Km}/\log K \rightarrow \infty$. If in addition to these conditions $m/(\log K)^3 \rightarrow \infty$, then

$$\lim_{K,m \rightarrow \infty} P\left(\frac{\tau}{K} > m - \sqrt{2m \log K} + \frac{\sqrt{m}(\log \log K + \log 4\pi + 2x)}{2\sqrt{2 \log K}}\right) = e^{-e^x},$$

as

$$w\left(\frac{1}{m}\left(\log K - \frac{\log \log K}{2}\right)\right) - 1 + \sqrt{\frac{2}{m}\left(\log K - \frac{\log \log K}{2}\right)} = o\left(\frac{1}{\sqrt{m \log K}}\right).$$

The remaining results of this Section pertain to the situation when $m/\log K$ is bounded.

THEOREM 4. Let $K, m \rightarrow \infty$, so that for $\gamma, 0 < \gamma < 1$, and a positive λ ,

$$K = \lambda(1 - \gamma)m!e^{\gamma m}(\gamma m)^{-m}[1 + o(1)].$$

Then

$$\lim_{K,m \rightarrow \infty} P(\tau > \gamma m K) = e^{-\lambda}.$$

Also for $k = 1, 2, \dots$

$$P(\tau(K, m + k) > \gamma m K) \rightarrow e^{-\lambda \gamma^k}.$$

Under conditions of Theorem 4,

$$m = \frac{\log K - 0.5 \log \log K + \log \lambda(1 - \gamma)\sqrt{2\pi}}{\gamma - \log \gamma - 1} + o(1),$$

so that $\frac{m}{\log K} \rightarrow \frac{1}{\gamma - \log \gamma - 1} > 0$ and

$$\frac{E\tau}{Km} = \gamma - \frac{\gamma \log m}{2m(\gamma - \log \gamma - 1)} + O\left(\frac{1}{m}\right).$$

The last result extends (3).

THEOREM 5. Assume that for $K, m \rightarrow \infty$, $\frac{m}{\log K} \rightarrow 0$. Then for a positive λ ,

$$P(\tau > K^{1-1/m}(\lambda m!)^{1/m}) \rightarrow e^{-\lambda}$$

and

$$P(\tau(K, m + 1) > K^{1-1/m}(\lambda m!)^{1/m}) \rightarrow 1.$$

Under conditions of Theorem 5, $E\tau = K^{1-1/m}(m!)^{1/m}[1 + O(K^{-1})]$.

3 Different Treatment Quotas

Here M_1, M_2, \dots still are independent multinomial vectors with the uniform distribution, and p_1, \dots, p_K are positive probabilities. In a more 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). Let

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

be the analogue of the waiting time τ in (1). To elucidate the asymptotic behavior of $\hat{\tau}$, assume that for some $r, 1 \leq r \leq K$,

$$0 < p_1 = \dots = p_r < \min_{i:i>r} p_i. \quad (6)$$

THEOREM 6. Under condition (6) for $n \rightarrow \infty$,

$$P \left(\frac{Kn p_1 - \hat{\tau}}{K \sqrt{np_1}} \leq x \right) \rightarrow P \left(Z_{(r)} - \left(1 + \sqrt{1 - \frac{r}{K}} \right) \bar{Z}_r \leq x \right).$$

Now $Z_{(1)} < Z_{(2)} < \dots < Z_{(r)}$ are the order statistics of a standard normal random sample, Z_1, \dots, Z_r , of size r , and $\bar{Z}_r = (Z_1 + \dots + Z_r)/r$.

When $r = 1$, i.e. the minimal allocation proportion is unique,

$$P \left(\frac{Kn p_1 - \hat{\tau}}{\sqrt{nK(K-1)p_1}} \leq x \right) \rightarrow \Phi(x).$$

where Φ denotes the distribution function of standard normal distribution.

Thus, for unequal proportions of subjects per treatment, the limiting distribution of $\hat{\tau}$ can be different from that of τ in Theorem 2. The proof of Theorem 6 is based on the following extension of Theorem 1.

PROPOSITION 1. For any positive z

$$E z^{-\hat{\tau}} = E e^{K(1-z)G_{(1)}}.$$

Here $G_{(1)}$ denotes the minimum of K independent random variables G_1, \dots, G_K with gamma-distributions, $G_i \sim \Gamma(np_i, 1)$. The density of $G_{(1)}$ has the form

$$f_{G_{(1)}}(u) = \sum_{j=1}^K \frac{u^{np_j-1} e^{-u}}{\Gamma(np_j)} \prod_{\ell \neq j} \int_u^{\infty} \frac{t^{np_{\ell}-1} e^{-t}}{\Gamma(np_{\ell})} dt.$$

4 Classical Occupancy Problems and Continuous Time Setting

When $K = 2$, the quota fulfillment instant τ has the probability distribution

$$P(\tau = m + x) = \binom{m + x - 1}{x} \frac{1}{2^{m+x-1}}, \quad x = 0, 1, \dots, m - 1.$$

This formula has been derived first by Blackwell and Hodges (1957) who recognized it as the probability of $m - 1$ successes in $m + x - 1$ symmetric Bernoulli trials. 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 m . Indeed, $\tau = 2m - B$, if B denotes the random number of matches remaining in one match box at the instant when another box is emptied. As a matter of fact, the relationship of the Banach match-box problem and the inverse sampling procedure discussed here is well known (Cacoullos and Sobel, 1966). An extension of this distribution for unequal probabilities of choosing the match-boxes is given by Uppuluri and Blot (1970). The probability generating function in Theorem 1 can be interpreted as that in the version of the match-box problem with K match-boxes each containing m matches. Note that

$$E\tau = 2m - \frac{1}{2^{2m-2} B(m, m)}$$

with $B(m, n)$ denoting the beta-function.

The exact distribution of τ is also known when $m = 2$. Then the probability of the event, $\tau > n$, can be interpreted as the probability of no coincident birthdays in a year, which has K days, for a group of n people (Blom, Holst and Sandell, 1994, Sec 9.1). Thus,

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

and

$$E\tau = \sum_{n=0}^{K-1} \frac{K!}{(K-n)!K^n}.$$

Assume now that the subjects arrive according to a stationary Poisson process Π with intensity λK , and each of them is assigned at random to

one of K treatments. A similar model has been suggested by Holst (1989) who studied the match-box problem when the demand for the matches is described by a Poisson process. With $\Pi_\ell(t)$ denoting the number of subjects allocated to the treatment ℓ by time t ,

$$\Pi(t) = \Pi_1(t) + \cdots + \Pi_K(t)$$

with independent Poisson processes $\Pi_\ell(t), \ell = 1, \dots, K$. Let

$$\tau_c = \min \left\{ t : \max_{\ell} \Pi_\ell(t) \geq m \right\}$$

be the continuous time analogue of τ .

Then

$$\begin{aligned} P(\tau_c > t) &= P(\max_{\ell} \Pi_\ell(t) < m) = [P(\Pi_1(t) < m)]^K \\ &= \left[e^{-\lambda t} \sum_{k=0}^{m-1} \frac{(\lambda t)^k}{k!} \right]^K = \left[\frac{1}{\Gamma(m)} \int_{\lambda t}^{\infty} e^{-u} u^{m-1} du \right]^K. \end{aligned}$$

The Central Limit Theorem implies that

$$P \left(\frac{m - \lambda \tau_c}{\sqrt{m}} < x \right) \rightarrow [\Phi(x)]^K,$$

i.e. in this situation $(m - \lambda \tau_c)/\sqrt{m} \rightarrow Z_{(K)}$. A slightly more general result holds.

THEOREM 7. Assume that the arrival process can be represented as

$$R(t) = R_1(t) + \cdots + R_K(t) \tag{7}$$

with independent renewal processes R_k , $R_k(t) = \sup\{n : V_1^{(k)} + \cdots + V_n^{(k)} \leq t\}$, $k = 1, \dots, K$, determined by a sequence of *iid* random variables $V_1^{(k)}, V_2^{(k)}, \dots$ with the mean μ and variance σ^2 . Then for

$$\tau_c = \min \left\{ t : \max_{\ell} R_\ell(t) \geq m \right\},$$

one has

$$\frac{m\mu - \tau_c}{\sqrt{m}\sigma} \rightarrow Z_{(K)}.$$

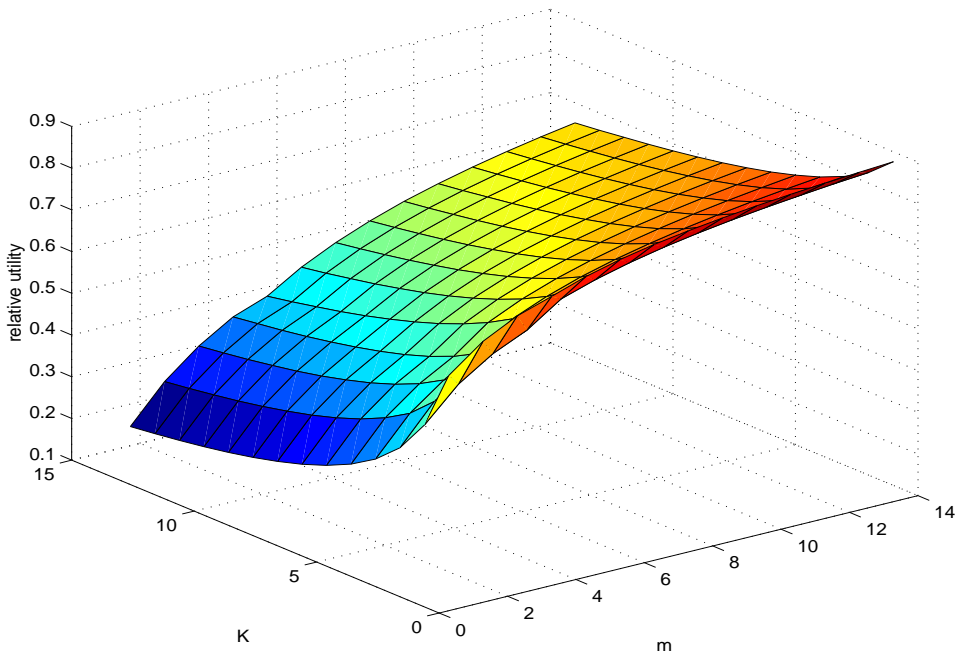


Figure 1: Plot of the average relative utility for $K = 2 : 14, m = 2 : 14$.

5 Conclusions

Table 1 summarizes the approximate formulas for the average relative utility $E\tau/(Km)$, and Figure 1 displays the plot of this characteristic. It is evident that to design an efficient randomization scheme for load balancing, it is not enough merely to increase the number K of zones (treatments). To attain larger values of the average relative utility, it is much more important to increase the capacity of each zone m .

Notice that exact formulas for $E\tau$ exist when $K = 2$ or when $m = 2$ (see Section 4).

Table 1: Approximate formulas for the average relative utility

	m fixed	$m \rightarrow \infty$
K fixed	$\frac{EG_{(1)}}{m}$	$1 - \frac{EZ_{(K)}}{\sqrt{m}}$
		$1 - \sqrt{\frac{2 \log K}{m}}$ if $\frac{\log K}{m} \rightarrow 0$
$K \rightarrow \infty$	$\Gamma\left(1 + \frac{1}{m}\right) \frac{m^{1/m}}{mK^{1/m}}$	$\gamma - \frac{\gamma \log m}{2ma}$ if $\frac{\log K}{m} \rightarrow a$
		$\left(\frac{\log K}{K}\right)^{1/m}$ if $\frac{\log K}{m} \rightarrow \infty$

Here $a = \gamma - 1 - \log \gamma > 0$.

6 Appendix

6.1 Theorems 1-2

The statement of Theorem 1 follows from that of Proposition 1, and Theorem 2 is a corollary of Theorem 6.

6.2 Theorem 3

LEMMA 1. Assume that $K \rightarrow \infty$, so that $\beta_{Km}^2/K \rightarrow 0$. Then $\beta_{Km}[\tau/K - \alpha_{Km}]$ converges in distribution if and only if $\beta_{Km}[G_{(1)} - \alpha_{Km}]$ converges to the same distribution.

Indeed, by Theorem 1

$$\begin{aligned}
 E z^{\beta_{Km}[\tau/K - \alpha_{Km}]} &= z^{-\beta_{Km}\alpha_{Km}} E \exp\{K(1 - z^{-\beta_{Km}/K})G_{(1)}\} \\
 &= z^{-\beta_{Km}\alpha_{Km}} E \exp\{\beta_{Km} \log z G_{(1)}\} \left[1 + O\left(\frac{\beta_{Km}^2}{K}\right)\right] \\
 &= z^{-\beta_{Km}\alpha_{Km}} E z^{\beta_{Km}G_{(1)}} \left[1 + O\left(\frac{\beta_{Km}^2}{K}\right)\right]. \blacksquare
 \end{aligned}$$

Lemma 1 with $\beta_{Km} = (m!)^{-1/m} K^{1/m}$ implies (3).

PROOF OF THEOREM 3. With α_{Km} and β_{Km} defined by (4) and (5) when $m, K \rightarrow \infty$, so that $m/\log K \rightarrow \infty$,

$$\lim_{K, m \rightarrow \infty} P(\beta_{Km} [G_{(1)} - \alpha_{mK}] > x) = e^{-e^x}.$$

The proof of this convergence is based on the classical methods of limit theorems for order statistics. Indeed,

$$P(G_{(1)} > \alpha_{Km}) = \left(1 - \frac{1}{K}\right)^K \rightarrow e^{-1},$$

and for real x

$$P(G_{(1)} > \alpha_{Km} + x/\beta_{Km}) = \left[1 - \frac{P(G_1 \leq \alpha_{Km} + x/\beta_{Km})}{KP(G_1 \leq \alpha_{Km})}\right]^K.$$

Thus,

$$\begin{aligned} \log P(G_{(1)} > \alpha_{Km} + x/\beta_{Km}) &= -\frac{P(G_1 \leq \alpha_{Km} + x/\beta_{Km})}{P(G_1 \leq \alpha_{Km})} + O\left(\frac{1}{K}\right) \\ &= -KP(G_1 \leq \alpha_{Km} + x/\beta_{Km}) + O\left(\frac{1}{K}\right) \end{aligned}$$

provided that the ratio of the probabilities in the right-hand side is bounded. It follows that

$$\begin{aligned} \lim_{K, m \rightarrow \infty} \log(-\log P(G_{(1)} > \alpha_{Km} + x/\beta_{Km})) \\ = \lim_{K, m \rightarrow \infty} \log(KP(G_1 \leq \alpha_{Km} + x/\beta_{Km})). \end{aligned}$$

By the definition of β_{Km} , with $\bar{\alpha}_{Km}$, $|\bar{\alpha}_{Km} - \alpha_{Km}| \leq |x|/\beta_{Km}$,

$$\begin{aligned} \log\left(KP\left(G_1 \leq \alpha_{Km} + \frac{x}{\beta_{Km}}\right)\right) &= x + \frac{Kx^2}{2\beta_{Km}} [f'_m(\bar{\alpha}_{Km}) - Kf_m^2(\bar{\alpha}_{Km})] \\ &= x + \frac{x^2}{2} \left[\frac{m-1}{\bar{\alpha}_{Km}} - 1 - \beta_{Km}\right] + O\left(\frac{\log K}{m}\right). \end{aligned}$$

The following analysis shows that the second term has the order $\sqrt{\log K/m} = o(1)$, so that $P(G_{(1)} \leq \alpha_{Km} + x/\beta_{Km}) \rightarrow e^{-e^x}$, i.e. the asymptotic distribution for $G_{(1)}$ is $-\Lambda_3$. This result along with the asymptotic distribution for the maximum $G_{(K)}$ was obtained by Ivchenko (1973).

To obtain the needed asymptotic expansions of α_{Km} and β_{Km} in (4) and (5), notice that the first expansion means inversion of the incomplete gamma-function, an extensively studied problem. See for example Temme (1992). Introduce an intermediate variable η_0 as the solution of the equation

$$\Phi(\eta_0\sqrt{m}) = 1 - \frac{1}{K},$$

so that

$$\eta_0 = \frac{1}{\sqrt{m}} \left[-\sqrt{2\log K} + \frac{\log \log K + \log 4\pi}{2\sqrt{2\log K}} + O\left(\frac{\log \log K}{\log K^{3/2}}\right) \right].$$

Then the percentile α_{Km} can be found by solving the equation

$$\frac{\alpha_{Km}}{m} - 1 + \log\left(\frac{\alpha_{Km}}{m}\right) = \frac{1}{2} \left(\eta_0 + \sum_k \frac{\epsilon_k}{m^k} \right)^2$$

with coefficients ϵ_k explicitly given on pp 758-759 of Temme (1992). Thus,

$$\begin{aligned} & \frac{\alpha_{Km}}{m} \\ &= w \left(\frac{\log K - 0.5 \log \log K}{m} \right) + \frac{\log 4\pi}{2\sqrt{2m \log K}} + o\left(\frac{1}{\sqrt{m \log K}}\right). \end{aligned}$$

These results imply that $\alpha_{Km}/m \rightarrow 1$, and

$$\beta_{Km} = \sqrt{\frac{2 \log K}{m}} [1 + o(1)].$$

The proof of Theorem 3 is concluded now by application of Lemma 1 according to which

$$\lim_{K,m \rightarrow \infty} P(\beta_{Km} [\tau/K - \alpha_{Km}] > x) = e^{-e^x}.$$

6.3 Theorems 4 and 5

These theorems can be derived from Theorems 1 and 2 in Section 6, Chapter 2 of Kolchin, Sevast'yanov and Chistyakov (1978).

When $\frac{\log K}{m} \rightarrow \gamma - 1 - \log \gamma > 0$, all conditions of Theorem 2 Section 6, Chapter 2 of Kolchin, Sevast'yanov and Chistyakov (1978) are met with $\alpha = \gamma m$, $n = \alpha K = \gamma K m$. As $\alpha / \log K \rightarrow \gamma / (\gamma - \log \gamma - 1)$, this theorem implies that

$$P(\eta_{max}(n) < m + k) \rightarrow \exp\{-\lambda \gamma^k\}.$$

Similarly, the conditions of Theorem 1 are satisfied when $m / \log K \rightarrow \infty$, in which case $\alpha = K^{1-1/m} (\lambda m!)^{1/m}$.

6.4 Theorem 6

According to Proposition 1

$$\begin{aligned} & E_S^{(nKp_1 - \hat{\tau}) / (K\sqrt{np_1})} \\ &= s^{\sqrt{np_1}} \sum_{j=1}^K \frac{1}{\Gamma(np_j)} \int_0^\infty e^{-u} u^{np_j-1} e^{Ku(1-s^{1/(K\sqrt{np_1})})} \prod_{\ell \neq j} \left[\int_u^\infty \frac{v^{np_\ell-1} e^{-v}}{\Gamma(np_\ell)} dv \right] du. \end{aligned} \quad (8)$$

Make a transformation of variables in the integral in the right-hand side, $u = np_j + \sqrt{np_j}y$. Then

$$\begin{aligned} Ku(1 - s^{1/(K\sqrt{np_1})}) &= -p_j \sqrt{np_1} \log s - y \sqrt{p_j/p_1} \log s \\ &\quad - \frac{p_j}{2Kp_1} (\log s)^2 + (y^2 + 1)(\log s)^2 O\left(\frac{1}{n^{1/2}}\right). \end{aligned}$$

The Central Limit Theorem for exponentially distributed 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 $p_j > p_\ell$, tends to 1 if $p_j < p_\ell$, and its limit is $1 - \Phi(y)$ if $p_j = p_\ell$. According to the local limit theorem for exponential random variables

$$\exp\{-(np_j + \sqrt{np_j}y)\} \frac{\sqrt{np_j}}{\Gamma(np_j)} (np_j + \sqrt{np_j}y)^{np_j-1}$$

$$= \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \left[1 + (y^3 - 3y) O\left(\frac{1}{n^{1/2}}\right) \right].$$

Thus, all terms in the sum in (8) for $j > r$ tend to zero, and combination of the formulas above gives

$$\begin{aligned} E_{S^{(nKp_1 - \hat{\tau})/(K\sqrt{np_1})}} &\rightarrow \frac{r e^{-(\log s)^2/(2K)}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} s^{-y} e^{-y^2/2} [1 - \Phi(y)]^{r-1} dy \\ &= e^{-(\log s)^2/(2K)} E_{S^{-Z_{(1)}}} = e^{-(\log s)^2/(2K)} E_{S^{Z_{(r)}}}. \end{aligned}$$

To recognize this integral as the probability generating function of $Z_{(r)} - \left(1 + \sqrt{1 - r/K}\right) \bar{Z}_r$, notice that by independence of \bar{Z}_r and $Z_{(r)} - \bar{Z}_r$,

$$E_{S^{Z_{(r)}}} = E_{S^{(Z_{(r)} - \bar{Z}_r) + \bar{Z}_r}} = E_{S^{(Z_{(r)} - \bar{Z}_r)}} E_{S^{\bar{Z}_r}} = E_{S^{(Z_{(r)} - \bar{Z}_r)}} e^{(\log s)^2/(2r)}.$$

Therefore,

$$\begin{aligned} e^{-(\log s)^2/(2K)} E_{S^{Z_{(r)}}} &= \exp\left\{ \frac{(\log s)^2}{2} \left[\frac{1}{r} - \frac{1}{K} \right] \right\} E_{S^{(Z_{(r)} - \bar{Z}_r)}} \\ &= E_{S^{\sqrt{(K-r)/K} \bar{Z}_r}} E_{S^{(Z_{(r)} - \bar{Z}_r)}} = E_{S^{Z_{(r)} - (1 + \sqrt{(K-r)/K}) \bar{Z}_r}}. \end{aligned}$$

6.5 Theorem 7

As in the Poisson process case, with $t = m\mu - v\sqrt{m}\sigma$,

$$\begin{aligned} P(\tau_c > t) &= P(\max_{\ell} R_{\ell}(t) < m) = [P(\Pi_1(t) < m)]^K \\ &= \left[P(V_1^{(k)} + V_2^{(k)} \dots + V_m^{(k)} > t) \right]^K \\ &= \left[P\left(\frac{V_1^{(k)} + V_2^{(k)} \dots + V_m^{(k)} - m\mu}{\sqrt{m}\sigma} > -v \right) \right]^K. \end{aligned}$$

The Central Limit Theorem shows that

$$P\left(\frac{m\mu - \tau_c}{\sqrt{m}\sigma} < v \right) \rightarrow [\Phi(v)]^K.$$

REFERENCES

- ANDERSON, K., SOBEL, M, AND UPPULURI, R. R., 1982. Quota fulfillment times. *Can. J. Statist.* 10, 73–88.
- ARNOLD, B. C., BALAKRISHNAN, N., AND NAGARAJA, N.H., 1993. *A First Course in Order Statistics*, Wiley, New York.
- ARRATIA, R., GOLDSTEIN, L., AND GORDON, L., 1990. Poisson approximation and the Chen-Stein Method. *Statist. Sci.* 5, 403-434.
- AVIGDOR, N. A., 2002. Building a scaleable and reliable parallel file system using commodity computers, Ph. D. thesis, Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County Campus.
- BARBOUR A. D., HOLST, L. AND AND JANSON, S., 1992. *Poisson Approximation*. Clarendon Press, Oxford.
- BLACKWELL, D. AND HODGES, J. L., JR., 1957. Design for the control of selection bias. *Ann. Math. Statist.* 28, 449-460.
- BLOM, G., HOLST, L., AND SANDELL, D., 1994. *Problems and Snapshots from the World of Probability*. Springer, New York.
- BOSE, R.C., AND GUPTA, S. S., 1959. Moments of order statistics from a normal population. *Biometrika* 46, 433-440.
- CACOULOS, T., AND SOBEL, M., 1966. An inverse sampling procedure for selecting the most probable event in a multinomial distribution. in *Proceedings International Symposium on Multivariate Analysis*. Ed. P. R. Krishnaiah, 423-444, Academic Press, NY.
- 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.
- FELLER, W., 1968. *An Introduction to Probability Theory and Its Applications, Vol I*. Wiley, New York.

- GUMBEL, E. J., 1958. *Statistics of Extremes*. Columbia University Press, New York.
- GUPTA, S. S., 1960. Order statistics from the Gamma distribution. *Technometrics* 2, 243-262.
- HOLST, L., 1986. On birthday, collectors', occupancy and other classical urn problems. *Int. Statist. Rev.* 54, 15-27.
- HOLST, L., 1989. A note on Banach match box problem. *Statist&Probab. Letters* 8, 441-443.
- IVCHENKO, G. I., 1973. On extremal values of samples, *Trudy Moscow Instituta Elektron. Machinostr.* 32, 12-31.
- 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.
- LEVIN, B., 1981. A representation for multinomial distribution function. *Ann. Statist.* 9, 1123-1126.
- LEVIN, B., 1983. On calculations involving the maximum cell frequency. *Commun. Statist. Theory Methods* 12, 1299-1327.
- MALLOWS, C. L., 1968. An inequality involving multinomial probabilities. *Biometrika* 55, 422-424.
- RIORDAN, J., 1958. *An Introduction to Combinatorial Analysis*. Wiley, New York.
- ROSENBERGER, W. F. AND LACHIN, J. M., 2002. *Randomization in Clinical Trials: Theory and Practice*. Wiley, New York.
- TEMME, N. M., 1992. Asymptotic inversion of incomplete gamma function. *Math. Comp.* 68, 755-764.
- 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.

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