Balanced Randomization Designs and Classical Probability Distributions

Andrew Rukhin
rukhin@cam.nist.gov
Statistical Engineering Division
National Institute of Standards and Technology
Department of Mathematics and Statistics
University of Maryland, Baltimore County (UMBC)

May 22, 2003

In a clinical trial with n patients and several treatments (say, K), each patient must be assigned to one of the treatments at random. To force balance one of the two following randomization schemes is commonly used.

Let $n = mK$, so that each treatment must have exactly m patients.

The *truncated multinomial design*

Let $T_1, \ldots, T_n$ be K-nomial random vectors representing the allocation of the patients to K treatments, $T_i(\ell) = 1$ if at stage i the patient is assigned to treatment $\ell, \ell = 1, \ldots, K$, $T_i(\ell) = 0$ otherwise. The truncated multinomial design starts with the uniform probability assignment of patients to treatments until one of the treatments receives m patients. Denote by τ_1 the stopping rule of the form

$$\tau_1 = \min\Big\{ j : \max_{\ell} \sum_{i=1}^{j} T_i(\ell) \geq m \Big\}.$$

After the random moment τ_1 the probability of assignment to each of the remaining $K-1$ treatments changes to $1/(K-1)$, the allocation random vectors T_j get the uniform distribution over these treatments. The allocation process continues until the moment τ_2 when the second treatment receives exactly m patients. If $\sum_{i=1}^{\tau_1} T_i(k) = m$, then

$$\tau_2 = \min\left\{j : j > \tau_1, \max_{\ell:\ell\neq k} \sum_{i=1}^{j} T_i(\ell) \geq m\right\}.$$

After τ_2 the uniform allocation is performed among the remaining $K-2$ treatments with the uniform distribution among these treatments, and so on, until the moment τ_{K-1} after which only one uncompleted treatment is left. This treatment receives then all the remaining patients.

Equal allocation to each of the treatments seems to be fairly standard practice in comparative studies, especially, in clinical trials. Even if the trial is not designed for exactly n (known in advance) subjects, randomization is often applied in blocks or strata with known sizes.

Colditz, Miller and Mosteller (1989): a survey of real-life examples demonstrating the dangers of non-randomized assignments in medical studies. There are many sequential randomization schemes (Efron's biased coin design, Wei's adaptive urn design, etc) which are non-balanced.

Rosenberger and Lachin (2002): a survey of these designs and their comparative merits in clinical trials.

Comparative study of several randomized rules for the sequential allocation of treatments in a clinical trial Atkinson (2002).

Diaconis and Graham (1981): randomization and guessing strategies in ESP experiments.

Another application of randomization designs is in storage of computer files.

Typical load balancing problem: a file system in which a computer disk is divided into K equal size zones, 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 a total of $n = mK$ blocks.

When a file is written to the disk, each block is mapped into a disk zone via a hash function, after this mapping the blocks are randomly distributed into the disk zones. When a block is 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 the disk zones when a zone starts overflowing.

The joint probability distribution of random variables $\tau_1, \ldots, \tau_{K-1}$ has a rather complicated form, but the probability generating function of this vector is much more tractable.

Theorem 1. For $m \to \infty$, the distribution of

$$m^{-1/2}(\tau_1 - n, \tau_2 - \tau_1, \cdots, \tau_{K-1} - \tau_{K-2})$$

converges weakly to that of $(K(Z_{(1)} - \bar{Z}),$ $(K-1)(Z_{(2)} - Z_{(1)}), \cdots, 2(Z_{K-1)} - Z_{(K-2)}))$. $Z_{(1)} < Z_{(2)} < \cdots < Z_{(K)}$ are the order statistics of standard normal random sample of size K, $\bar{Z} = (Z_1 + \cdots + Z_K)/K$.

It follows that for $i = 1, \ldots, K-1$

$$\frac{n - \tau_i}{\sqrt{m}} \sim \sum_{i+1}^{K} Z_{(j)} - (K-r)Z_{(i)}$$

$$= \sum_{j=i}^{K} (Z_{(j)} - Z_{(i)}).$$

When $K = 2$,

$$\frac{n - \tau_1}{\sqrt{n}} \to \frac{Z_{(2)} - Z_{(1)}}{\sqrt{2}} = |Z|.$$

The truncated binomial design was introduced by Blackwell and Hodges (1957). They defined the selection bias to be proportional to the number of correct guesses of the experimenter whose goal is to assign patients with the largest expected response to a particular treatment (or by instructing the statistician in charge of allocation to administer treatments after selecting suitable subjects.)

Statistical Decision Problem

Statistician vs Experimenter
Decisions: Randomization Devices
States of Nature: Guessing Strategies

The minimaxity of the truncated binomial design proven by Blackwell and Hodges and the loss function determined as the total number of the correct guesses G_{TM}, holds for any K.

Independently of the guessing strategy

$$EG_{TM} = \frac{1}{K}E\tau_1 + \frac{1}{K-1}E(\tau_2 - \tau_1) + \cdots$$

$$+\frac{1}{2}E(\tau_{K-1} - \tau_{K-2}) + E(n - \tau_{K-1})$$

$$\sim m + \sqrt{m}EZ_{(K)}.$$

The *random allocation rule*

Randomly select one of the

$$\binom{Km}{m \quad \cdots \quad m}$$

assignments of exactly m patients per treatment.

It is described by the allocation vectors $S_i, i = 1, \ldots, n$, so that

$$P(S_{j+1} = e_\ell | \sum_{i=1}^{j} S_i = s) = \frac{m - s_\ell}{n - j}.$$

The convergence strategy under this design is to guess the treatment which has been previously less used, i.e. the experimenter predicts that the patient $j+1$ will be assigned to any treatment ℓ such that

$$\sum_{i=1}^{j} S_i(\ell) = \min_k \sum_{i=1}^{j} S_i(k).$$

Theorem 2. The expected number of the correct guesses EG_{RA} has the form

$$EG_{RA} \sim m + \frac{\pi\sqrt{m}}{2} EZ_{(K)}.$$

The truncated multinomial design has smaller selection bias than the random allocation rule, for any K

$$\frac{EG_{RA} - m}{EG_{TM} - m} \sim \frac{\pi}{2}.$$

The distribution of $\tau = \tau_1$ in the definition of the the truncated multinomial design is related to that of the maximum of coordinates of a multinomial vector $\eta_{max}(N)$

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

The generating function for these probabilities Riordan (1958). David and Barton (1962) derive the combinatorial extreme value distribution of the largest multinomial frequency when $K \to \infty$. A method to evaluate these probabilities by Levin (1981), Kolchin, Sevast'yanov and Chistyakov (1978).

Theorem 1 implies that for a fixed K and $N \to \infty$,

$$P\left(\frac{\eta_{max}(N) - N/K}{\sqrt{N/K}} < x\right) \to P\left(Z_{(K)} - \bar{Z} < x\right).$$

The distribution of the extreme deviation $Z_{(K)} - \bar{Z}_K$ has been used for outliers detection

Fixed K,$\alpha, 0 < \alpha < 1$, can find the critical value q_α, such that for large m,

$$P(\tau \geq mK - q_\alpha K\sqrt{m}) \approx \alpha.$$

$$q_\alpha \approx z_{\alpha/K}\sqrt{(K-1)/K}$$

z_α the critical point of standard normal distribution.

$K \to \infty$, m fixed, τ asymptotically has a Weibull distribution,

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

The distribution of the number W of treatments which receive at least m subjects admits the Poisson approximation. Barbour, Holst and Janson (1992) represented W as the sum of K dependent Bernoulli random variables with the parameter

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

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

W 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. Arratia, Goldstein and Gordon (1989)

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

a sharp bound.

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

Similar bounds in Barbour, Holst and Janson (1992).

Let $\alpha = \alpha_{Km}$ be the solution of the equation

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

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.

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

The asymptotic behavior of $\tau = \tau(K,m)$ is determined by that of $\frac{m}{\log K}$.

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

$$\lim_{K,m\to\infty} P\Bigg(\frac{\tau}{K} > m - \sqrt{2m\log K}$$

$$+\frac{\sqrt{m}(\log\log K + \log 4\pi + 2x)}{2\sqrt{2\log K}}$$

$$+\frac{2\log K - \log\log K}{3}\Bigg) = e^{-e^{x}}.$$

Theorem 4. Let $K, m \to \infty$, 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

$$\lim_{K,m\to\infty} P\left(\tau(K,m) > \gamma m K\right) = \exp\left\{-\lambda\right\}.$$

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

$$P\left(\tau(K, m+k) > \gamma m K\right) \to \exp\left\{-\lambda\gamma^{k}\right\}.$$

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

$$P\left(\tau(K,m) > K^{1-1/m}(\lambda m!)^{1/m}\right) \to e^{-\lambda}$$

and

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

When $K = 2$, the stopping rule τ has the probability distribution

$$P\left(\tau = m + x\right)$$

$$= \binom{m+x-1}{x} \frac{1}{2^{m+x-1}}, \quad x = 0, 1, \ldots, m-1.$$

i.e. the probability of $m-1$ successes in $m+x-1$ Bernoulli trials.

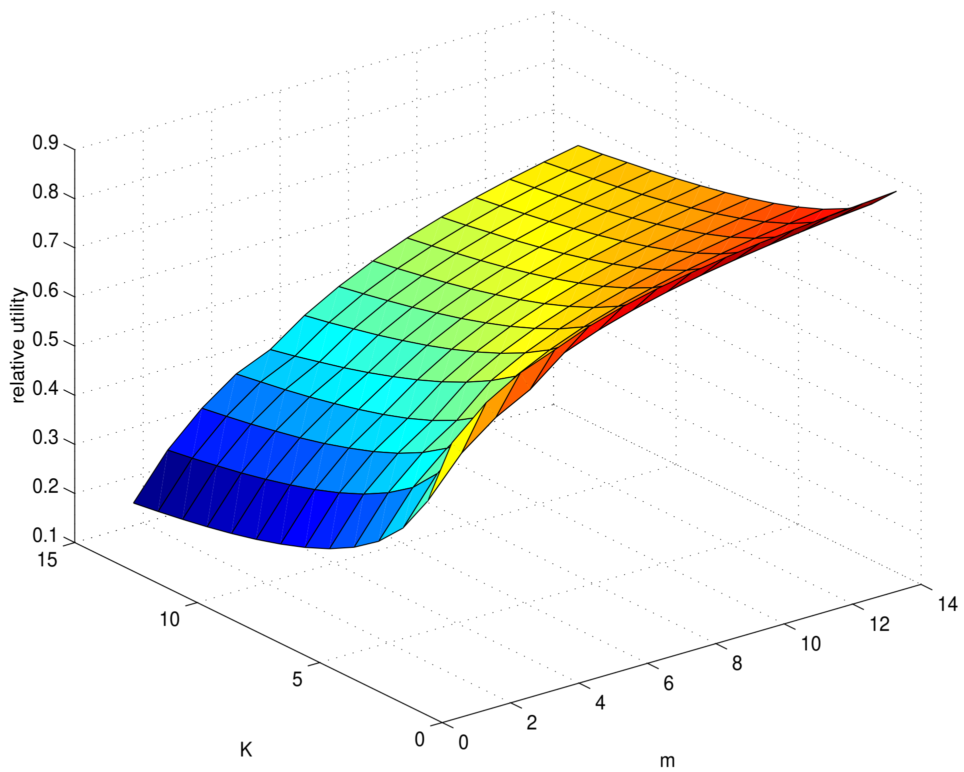
τ appears in the classical Banach match-box problem. If the original number of matches in each box is m, B - the random variable equal to the number of matches remaining in one match box when another box is emptied, $\tau = 2m - B$.

When $m = 2$, the event, $\tau > n$, means no coincident birthdays in a year which has K days for a group of n people.

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

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

Table 1: Approximate formulas for the average relative utility



To design an efficient randomization scheme for load balancing, it is not enough merely to increase the number K of zones (treatments). It is much more important to increase the capacity of each zone m, to attain larger values of the average relative utility.

Independent multinomial trials with possible outcomes $k = 1, \ldots, K$, whose probabilities are $\pi_1, \ldots, \pi_K$; $p_1, \ldots, p_K$ positive probabilities. In a more general allocation scheme, k-th treatment has to obtain np_k subjects, and is chosen with probability π_k. 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 stopping time τ_1.

$$0 < \rho = \frac{p_1}{\pi_1} = \cdots = \frac{p_r}{\pi_r} < \min_{i: i \geq r} \frac{p_i}{\pi_i}.$$

Theorem 6. For $n \to \infty$,

$$P\left(\frac{n\rho - \widehat{\tau}}{\sqrt{n\rho}} \leq x\right)$$

$$\to P\left(\left(1 + \sqrt{1 - \sum_1^r \pi_k}\right)\tilde{Y} - Y_{(1)} \leq x\right).$$

$Y_{(1)}$ is the minimum of r independent normal random variables $Y_1, \ldots, Y_r$ with $Y_k \sim N(0, 1/\pi_k)$, $\tilde{Y} = \sum_1^r \pi_k Y_k / \sum_1^r \pi_k$.

When $r = 1$,

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

$\tilde{Y} - Y_{(1)}$ and $\tilde{Y}$ independent. $\tilde{Y}$ is a complete sufficient statistic for the location parameter in the model when Y's have normal distribution with the mean being this location parameter. $\tilde{Y} - Y_{(1)}$ is a similar (ancillary) statistic, Joint distribution of $(1 + \sqrt{1 - \sum_1^r \pi_k})\tilde{Y} - Y_k$, $k = 1, \ldots, r$ multivariate normal, mean 0, diagonal elements $(1 - \pi_k)/\pi_k$, the off-diagonal elements all equal to -1.

$X_j^{(k)}$ is the number of j-th trial resulting in the outcome k,

$$\hat{\tau} = \min\left[X_{np_1}^{(1)}, \ldots, X_{np_K}^{(K)}\right].$$

$X_j^{(k)}$ has negative binomial distribution with parameters j, π_k,

$$X_i^{(k)} = \sum_{j=0}^{i-1} U_j^{(k)}, \;\; i = 1, 2, \ldots$$

$U_j^{(k)}$ are independent random variables with the geometric, parameter π_k, distribution, $U_0^{(k)} = X_1^{(k)}$ is the number of the first trial resulting in the outcome k; $U_i^{(k)} = X_{i+1}^{(k)} - X_i^{(k)}, i = 1, 2, \ldots$.

A remarkably beautiful formula (unknown or well forgotten?)

$$\mathrm{Cov}(U_i^{(k)}, U_j^{(\ell)})$$

$$= -\binom{i+j}{i} \frac{\pi_k^i \pi_\ell^j}{(\pi_k + \pi_\ell)^{i+j+1}}$$

$U_i^{(k)}$ is the duration time between two successive occurrences (i and $i+1$) of outcome k.

When $K = 2$, $\pi_k + \pi_\ell = 1$,

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

For $k \neq \ell$

$$\text{Cov}\left(X_i^{(k)}, X_j^{(\ell)}\right)$$

$$= -\frac{1}{\pi_k \pi_\ell}\left[i\pi_\ell I_{\frac{\pi_k}{\pi_k+\pi_\ell}}(i,j) + j\pi_k I_{\frac{\pi_\ell}{\pi_k+\pi_\ell}}(j,i)\right]$$

$$+\frac{\pi_k^{i-1}\pi_\ell^{j-1}}{(\pi_k+\pi_\ell)^{i+j}B(i,j)}.$$

$I_p(i,j)$ the incomplete beta-function.

The correlation coefficient for $1 \leq k, \ell \leq r$,

$$\mathbf{corr}\left(X_{np_k}^{(k)}, X_{np_\ell}^{(\ell)}\right) \to -\sqrt{\frac{\pi_k \pi_\ell}{(1-\pi_k)(1-\pi_\ell)}}.$$

The limiting covariance matrix of

$$\frac{\max\left[X_{np_1}^{(1)} - np_1/\pi_1, \cdots, X_{np_r}^{(r)} - np_r/\pi_r\right]}{\sqrt{n}}.$$

is that of $(1+\sqrt{1-\sum_1^r \pi_k})\tilde{Y} - Y_k,\ k = 1, \ldots, r$

For the random allocation rule the stopping time κ

$$\kappa = \min\left\{ j : \max_{\ell} \sum_{i=1}^{j} S_i(\ell) \geq m \right\}$$

with S_i denoting the allocation vectors after this rule.

The limiting behavior of κ?

Let $\left\{ \begin{array}{c} r \\ K \end{array} \right\}$ denote Stirling's number of the second kind, i.e. the number of partitions of an r element set in K nonempty subsets.
(Knuth 1992, r "subset" K).

Proposition For $n \to \infty$

$$n - \kappa \to U,$$

where $U = U_K$ is a discrete random variable with the distribution over $K-1, K, \ldots$ with probabilities

$$P(U = r) = \frac{K!}{K^{r+1}} \left\{ \begin{matrix} r \\ K-1 \end{matrix} \right\};$$

$$\sum_r t^r P(U = r)$$

$$= \frac{(K-1)!\; t^{K-1}}{(K-t)(K-2t)\cdots(K-(K-1)t)},$$

$$EU = \sum_{j=1}^{K-1} \frac{K}{K-j},$$

$$\mathrm{Var}(U) = \sum_{j=1}^{K-1} \frac{j^2+j}{(K-j)^2}.$$

For $K = 2$,

$$P(U = k) = \frac{1}{2^k},\ k = 1, 2, \ldots,$$

i.e. U_2 is a geometric random variable with parameter 1/2.

For any K and $r = K, K+1, \ldots$

$$P(U < r) = \frac{K!}{K^r}\left\{ \begin{array}{c} r \\ K \end{array} \right\}$$

$$= \sum_{j=0}^{K} (-1)^j \binom{K}{j} \left(1 - \frac{j}{K}\right)^r$$

is the probability that in r independent trials with K equally likely outcomes all K outcomes occur, or, in the problem of distribution of r balls in K cells, the probability that all cells are occupied. Feller (1968), Sec IV.2; Blom, Holst and Sandell (1994), Sec 5.6, discuss these probabilities in the context of the classical occupancy problem.

$\mathbf{e} = \mathbf{e}_K$ generic vectors with unit coordinates.

Proposition For the truncated multinomial design, $i = 1, \ldots, n$

$$ET_i = \frac{1}{K}\mathbf{e},$$

$$\mathrm{Var}(T_i) = \frac{1}{K}\left[\mathbf{I} - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right] = \Sigma_0,$$

and for $i < j$

$$\mathrm{Cov}(T_i, T_j) = \sum_{\ell=1}^{K-1} \frac{\ell P(\tau_\ell < i \wedge j \le \tau_{\ell+1})}{(K-1)(K-\ell)}\Sigma_0.$$

For the random allocation scheme vectors S_i

$$ES_i = \frac{1}{K}\mathbf{e},$$

$$\mathrm{Var}(S_i) = \frac{1}{K}\left[I - \frac{1}{K}\mathbf{e}\mathbf{e}^T\right] = \Sigma_0,$$

and for $i \ne j$

$$\mathrm{Cov}(S_i, S_j) = -\frac{1}{K(n-1)}\Sigma_0.$$

Efron (1971):the concept of *accidental bias* which is a measure of the expected bias of the treatment effect from a linear regression model if important covariates are ignored. The degree of susceptibility to accidental bias is given by $E(z^T T)^2 = z^T \Sigma_T z$,

$$z^T \Sigma_T z \leq \lambda_{\max}(\Sigma_T),$$

$z^T \mathbf{e} = 0$, $z^T z = 1$.

The accidental bias of an allocation rule with the covariance matrix Σ as

$$\mathbf{AB} = \max_{z : z^T \mathbf{e}=0} \frac{z^T \Sigma z}{z^T z}.$$

The largest eigenvalue $\lambda_{\max}(\mathbf{C})$ has order $m^{1/2}$.

$$w_r = \frac{K}{(K-1)(K-r)(K-r+1)},$$

$r = 1, \ldots, K-1$, weights which add up to one.

$\phi_{rK}(t)$ is the density of $\sum_{j=r}^{K}(Z_{(j)} - Z_{(r)})$,

$$\phi_K(t) = \sum_{r=1}^{K-1} w_r \phi_{rK}(t).$$

Proposition For sufficiently large m

$$\frac{(K-1)\int_0^\infty \int_0^\infty \min[x,y]\phi_K(x)\phi_K(y)\,dx\,dy}{KEZ_{(K)}}$$

$$\leq \frac{\lambda_{\max}(\mathbf{C})}{\sqrt{m}} \leq \frac{KEZ_{(K)}}{K-1}.$$

The largest eigenvalue $\lambda_{\max}(\mathbf{B})$ of $\mathbf{B}$ is

$$\lambda_{\max}(\mathbf{B}) = \frac{n}{n-1}.$$

Conclusions

- The truncated multinomial design has smaller selection bias than the random allocation rule.
- The truncated multinomial design, unlike the random allocation rule, can be susceptible to a high degree of accidental bias.
- The truncated multinomial design behaves really different from the random allocation rule.