Sequential Intentional Sampling made Haphazard by Stochastic Perturbations

Victor Fossaluza*, Marcelo S. Lauretto+,
Carlos A. Bragança Pereira*, **Julio Michael Stern***

IME-USP* and EACH-USP+, University of São Paulo

jstern@ime.usp.br

33nd MaxEnt, 15-20 December 2013, Canberra, Australia
A.C. Camargo, 2 December 2014, São Paulo, Brazil

We describe a possible allocation that the experimenter judges to be free of covariate interference as haphazard. Randomization may be a convenient way of producing a haphazard design. We argue that it is the haphazard nature, and not the randomization, that is important. It seems therefore that a reasonable approximation to an optimal design would be to select a haphazard design. ...a detailed Bayesian consideration of possible covariates would almost certainly not be robust in that the analysis might be sensitive to small changes in judgments about covariates.
Lindley (1982, p.438-439) - The Role of Randomization in Inference.

Introduction

- Intentional sampling methods are non-randomized procedures that select or allocate groups of individuals with the purpose of meeting specific prescribed criteria.
- Such methods can overcome some of limitations of standard randomized designs for statistical experiments, when cost, ethical or inherent rarity constraints only admit the use of very small samples.
- However, intentional or purposive sampling methods pose several interesting questions concerning statistical inference, as extensively discussed in Basu and Ghosh (1988), see also Schreuder et al. (1993, Sec.6.2), Brewer and Särndal (1983) and following discussions in Madow et al. (1983).
- This paper focus on sequential allocation methods, and follows previous research in the field of intentional sampling presented in Fossaluza et al. (2009) and Lauretto et al. (2012).

Compositional Models and Simplex Geometry

- The open (m-1)-Simplex is the set $S^{m-1} = \left\{ x \in R^m \mid x > 0 \wedge \mathbf{1}'x = 1 \right\}$, where $\mathbf{1}$ is the vector of ones of appropriate dimension.
- The *closure-to-unity* transformation, $\mathrm{clu} : R^m_+ \to S^{m-1}$:

$$\mathrm{clu}(x) = (1/\mathbf{1}'x)x \ ,$$

- The *additive logratio transformation*, $\mathrm{alr} : S^{m-1} \to R^{m-1}$:

$$\mathrm{alr}(x) = \log\left((1/x_m)[x_1, \ldots x_{m-1}]\right), \ \mathrm{alr}^{-1}(z) = \mathrm{clu}\left(\exp([z_1, \ldots z_{m-1}, 0])\right).$$

- We introduce the operators:
  - *Power* (scalar multiplication): $\alpha \star x = \mathrm{clu}([x_1^\alpha, \ldots x_m^\alpha])$
    Interpreted as the α-times repeated effect of proportional decay rates.
  - *Perturbation* (vector summation): $x \oplus y = \mathrm{clu}([x_1 y_1, \ldots x_m y_m])$
    Interpreted as the effect of proportional decay rates in y over the fractional composition in x.
  - *Difference*: $x \ominus y = \mathrm{clu}([x_1/y_1, \ldots x_m/y_m])$

- We want a distance function on the Simplex, $D_S(x, y)$, that exhibits the invariance properties that are most adequate for the purpose of compositional analysis, namely:
  - Perturbation invariance: For any perturbation, z, $D_S(x \oplus z, y \oplus z) = D_S(x, y)$.
  - Permutation invariance: For any permutation matrix, P, $D_S(Px, Py) = D_S(x, y)$.
  - Power scaling: For any $\alpha > 0$, $(1/\alpha) D_S(\alpha \star x, \alpha \star y) = D_S(x, y)$.
- The following distance function exhibits all these desirable invariance properties, besides the standard properties for distance functions – positivity, symmetry and triangular inequality:

$$
\begin{aligned}
D_S^2(x, y) &= [\mathrm{alr}(x) - \mathrm{alr}(y)]' \, H^{-1} \, [\mathrm{alr}(x) - \mathrm{alr}(y)] \, , \\
H_{i,j} &= 2\delta_{i,j} + 1(1 - \delta_{i,j}) \, .
\end{aligned}
$$

Haphazard Intentional Allocation for Clinical Trials

- Case study: allocation of patients with Obsessive-compulsive disorder (OCD) between two treatment arms, see Fossaluza et al. (2009).

  Dataset: $T = 277$ patients

- Patients are enrolled sequentially, according to the order in which they start the treatment at the clinic or hospital.
- The allocation problem consists in assigning each new patient to one, and only one, of two alternative treatments (arms).
- Requisite: profiles in the alternative arms remain similar with respect to some relevant patients' factors:

  a) Current patient's *age* (a): under 30 years; between 30 and 45 years; over 45 years.
  b) Treatment *history* (h): $T0 =$ no previous appropriate treatment; $T1 =$ one previous appropriate treatment without response; $T2 =$ two or more appropriate treatments without response.
  c) OCD symptom *severity* (v): nine classes based on scores for each of the two symptom types (obsession and compulsion).
  d) *Gender* (g).

- We denote by n_i^a, n_i^h, n_i^v and n_i^g the quantities of patients already allocated to arm i belonging to each category of factors *age*, *history*, *severity* and *gender*.
  - For example, $n_1^a = [n_{1,1}^a, n_{1,2}^a, n_{1,3}^a]$ denotes the quantity vector of patients in arm 1 belonging to the three age classes.
- Besides the previous factors, we also consider the sample *size* (z) in each arm.

  Purpose: to yield allocations with approximately the same number of patients in each arm.

  We denote by q_i as the total number of patients allocated to arm i, and by $n_i^z = [q_i, (q_1 + q_2 - q_i)]$ the vector of total allocation to arm i and its complement.
- The complete profile of arm i, $i = 1, 2$ is stored in the concatenated vector $n_i = [n_i^a, n_i^h, n_i^v, n_i^g, n_i^z]$.

- In order to avoid empty categories in the allocation process, we may add to vector n a *ground-state* or *weak-prior*, see Pereira and Stern (2008), in the form of vector $w = [w^a, w^h, w^v, w^g, w^z]$.

  For any character ξ in the set $\{a, h, v, g, z\}$, where factor w^ξ has $\kappa(\xi)$ categories, we take $w^\xi = [1/\kappa(\xi), \ldots 1/\kappa(\xi)]$.

- From vectors n and w we obtain the *regularized proportions* vector: $p_i = [p_i^a, p_i^h, p_i^v, p_i^g, p_i^z]$,
  where $p_i^\xi = clu(n_i^\xi + w_i^\xi)$, $\xi \in \{a, h, v, g, z\}$.

- We define the heterogeneity measure between arms 1 and 2 by the function:

$$\begin{aligned} \Delta(p_1, p_2) \quad = \quad & [\, D_s(p_1^a, p_2^a) + D_s(p_1^h, p_2^h) + D_s(p_1^v, p_2^v) + \\ & D_s(p_1^g, p_2^g) + D_s(p_1^z, p_2^z)\,]/5. \end{aligned}$$

- Let us consider a new patient that enrolls the study and must be allocated to one of arms 1 or 2.

- We denote by $x = [x^a, x^h, x^v, x^g, x^z]$ the binary vector indicating to which categories the new patient belongs in each factor.

Allocation Algorithm

1. For each factor $\xi \in \{a, h, v, g, z\}$ and arm $i = 1, 2$, generate a random vector r_i^{ξ}, with uniform distribution in the $(\kappa(\xi)$-1)-simplex.
2. For $j = 1, 2$ consider the allocation of the new patient x in arm j, that is,

   – For $i = 1, 2$, make $m_i = n_i + \delta(i, j)x$ and perform the following steps:

   a) For $i = 1, 2$ and $\xi \in \{a, h, v, g, z\}$, compute
      - The regularized proportions: $p_i^{\xi} = clu(m_i^{\xi} + w_i^{\xi})$ and
      - The ε-perturbed proportions: $b_i^{\xi} = clu(p_i^{\xi} + \varepsilon r_i^{\xi})$.

   b) For $i = 1, 2$, set $b_i = [b_i^a, b_i^h, b_i^v, b_i^g, b_i^z]$.

   c) Compute the distance $d(j) = \Delta(b_1, b_2)$.
3. Choose the allocation j that minimizes $d(j)$, assign the new patient to the corresponding arm, and update vector n accordingly.

- Perturbation parameter ε: introduces a random component in the allocation method.

  For $\varepsilon = 0$: deterministic intentional allocation scheme.

Numerical Experiments

- We analyse the performance of our haphazard intentional allocation procedure, for $\varepsilon \in \{0, 0.1, 0.5, 1.0, 2.0\}$.
- We generated $P = 200$ random permutations of the original data – each one representing a possible sequence of patients arriving to the hospital or clinic.

  For each permutation, we ran the pure random method and the haphazard intentional allocation method $H = 200$ times.
- Performance criteria:
  - *Optimality*: based on the distance Δ; concerns the difference among the relative frequencies of patients in the several categories for both arms;
    Benchmark: deterministic intentional allocation scheme ($\varepsilon = 0$).
  - *Decoupling*: based on the Yule's Q coefficient of association (Yule, 1912); concerns the absence of a tendency to allocate each pair patients to the same arm.
    Benchmark: pure random allocation method.

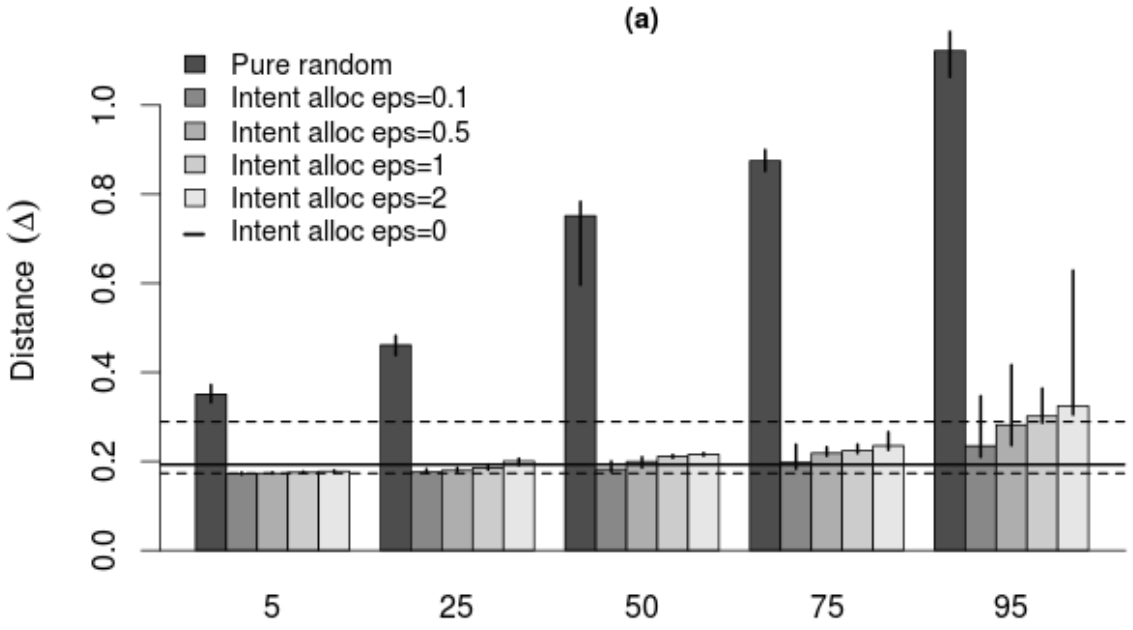

Figure 1. 5%, 25%, 50%, 75%, 95% empirical percentiles of Δ computed from the H haphazard allocations.

- Bar height: median over the P random permutations;
- Vertical line in each bar: corresponding (5%, 95%) percentiles.
- Continuous and dashed horizontal lines represent, respectively, the median of distance Δ for the deterministic intentional allocation method, $\varepsilon = 0$, and the (5%, 95%) percentiles over P random permutations.

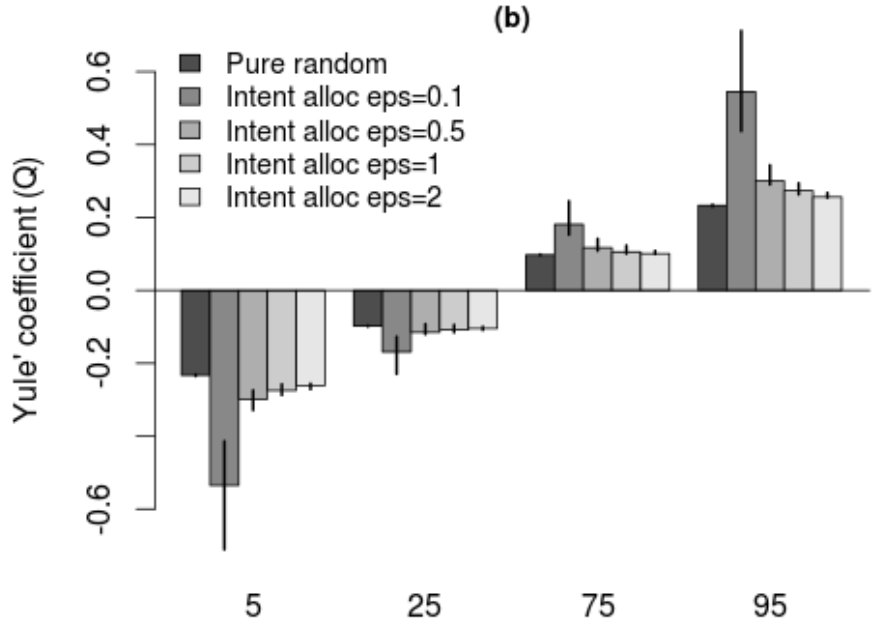

Figure 2. 5%, 25%, 75%, 95% empirical percentiles of Yule's Q coefficient.

- Quantiles for Q span the $T(T-1)/2$ pairs of patients, where the Q for each pair is computed over the H haphazard allocations.
- Bar height: median over the P random permutations;
- Vertical line in each bar: corresponding (5%, 95%) percentiles.

Final Remark and References

- Under an appropriate calibration of the perturbation parameter ε, the haphazard intentional allocation method proposed in this work has the remarkable property of being able to conciliate:
  - the performance on optimality achieved by the deterministic intentional allocation; and
  - the performance on decoupling achieved by the pure random allocation method.
- References:

  - V.Fossaluza, J.B.Diniz, B.B.Pereira, E.C.Miguel, C.A.B.Pereira (2009). Sequential Allocation to Balance Prognostic Factors in a Psychiatric Clinical Trial. *Clinics,* 64, 511-518.
  - M.S.Lauretto, F.Nakano, C.A.B.Pereira, J.M.Stern (2012). Intentional Sampling by Goal Optimization with Decoupling by Stochastic Perturbation. *AIP Conf.Proc.*, 1490, 189-201.
  - D.Lindley (1982), The Role of Randomization in Inference. in: P.Asquith, T.Nickles eds. *Proceedings of the Biennial Meeting of the Philosophy of Science Association,* V.2, 431-446. Univ.of Chicago Press.