Design of Clinical Trials: A Mixed Intentional-Randomized Sampling Approach

Victor Fossaluza⋆, Marcelo S. Lauretto†, Carlos A. B. Pereira⋆ and Julio M. Stern⋆

⋆ IME–USP

† EACH–USP

1º Simpósio Regional de Matemática Industrial

São José do Rio Preto, 2015.

Table of Contents

Motivation

- 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).

Motivation

- *"The counterquestion 'How can you justify purposive sampling?' has a lot of force in it. The choice of a purposive plan will make a scientist vulnerable to all kinds of open and veiled criticisms. A way out of the dilemma is to make the plan very purposive, but to leave a tiny bit of randomization in the plan."*

  Basu (1987, ch.XIV, p.257) - Why to Randomize?

Motivation

- *"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."*

  Lindley (1982, p.438-439) - The Role of Randomization in Inference.

Motivation

- This work was motivated by a clinical trial with obsessive-compulsive disorder (OCD) patients at IPq–HCFMUSP:
    - Convenience sample: the patients come to the hospital looking for treatment;
    - Patients arriving at hospital need to be assigned quickly to one of the treatment groups;
    - It is not possible to determine the exact sample size (which could be smaller than expected).
- It is known that some variables may influence treatment response (such as gender or disease severity) and we would like that these variables are equally distributed among treatment groups.

Motivation

- This presentation focus on the sequential allocation approach described in Fossaluza et al. (2014), which is based on previous research in the field of intentional sampling developed in Fossaluza et al. (2009) and Lauretto et al. (2012).

Previous Research

- Datanexus (2002) – Survey institute for TV audience.
  - Our goal: to help in the selection of the *monitoring sample*
  - Budget constraint: $\beta = 250$ households
  - Preliminary survey provided detailed information of $m = 10,000$ candidate households.
    - Household monthly income
    - Household social-economic class
    - Individual sex
    - Individual age
    - Individual scholarity
    - Individual daily TV consumption

Previous Research

- Datanexus (2002) – Modelling:
  - x: binary decision vector: x_h indicates if household h belongs (or not) to the selected monitoring sample
  - b: vector representing the monitoring cost per household.
  - A: Sample information matrix. $A_{h,k}$ = number of individuals of class k living in household h
  - g: goal or target vector. g_k = expected number of individuals of class k in monitoring sample
  - r, s: non-negative surplus and slack variables.
  - Goal optimization:

$$\min ||r + s||_p$$

subject to

$$b'x \leq \beta \ , \quad A'x - r + s = g .$$

Previous Research

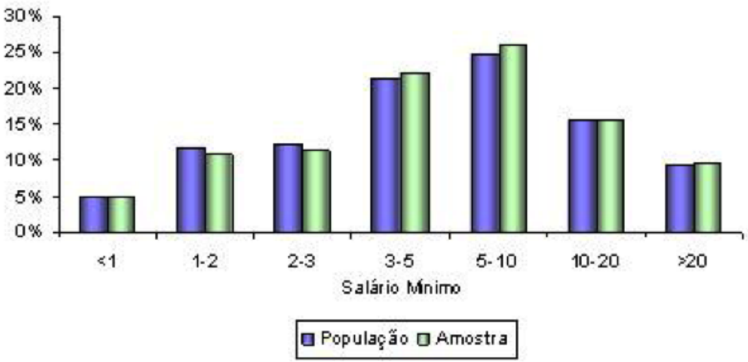

Figure 1. Household income: desired and actual sample frequencies

Previous Research

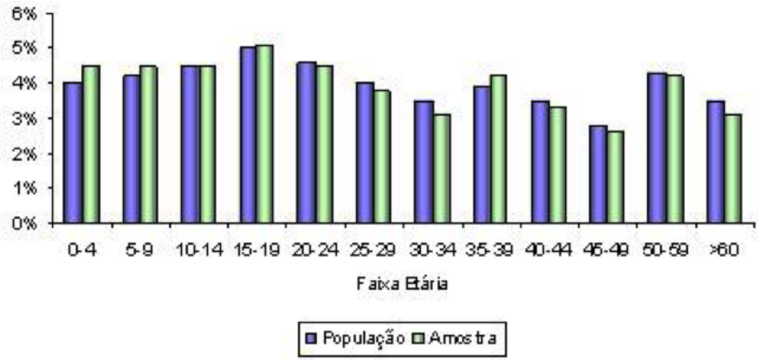

Figure 2. Age: desired and actual sample frequencies

Previous Research

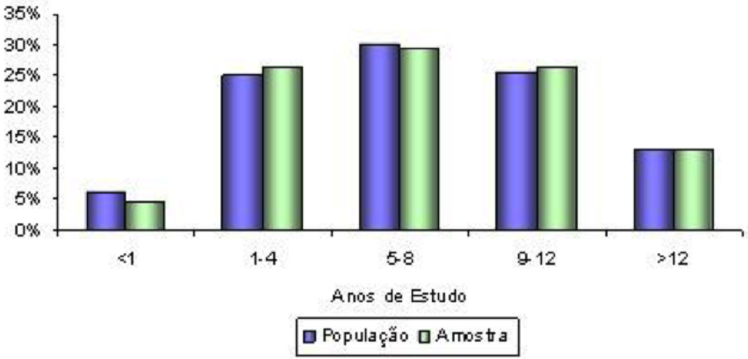

Figure 3. Scholarity: desired and actual sample frequencies

Previous Research

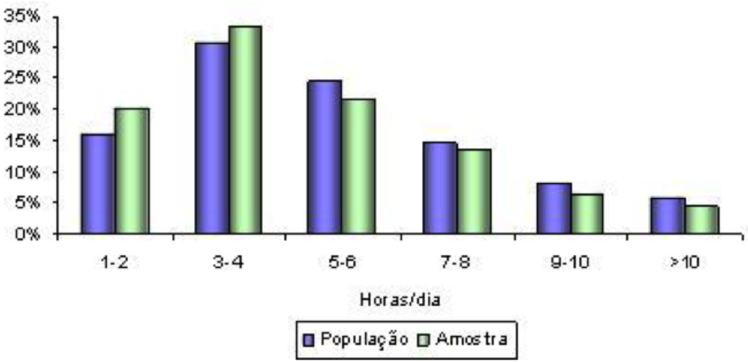

Figure 4. Daily TV consumption: desired and actual sample frequencies

Previous Research

- Lauretto et al. (2012): Goal programming problem extended with a randomization component:
  - Perturbations inspired by "negative results" concerning the instability of optimization problems.
  - β is replaced by $\tilde{\beta} = \beta + 1$
  - b is replaced by $\tilde{b} = b + z$ where $z = \epsilon(2/\beta)\text{rand}(m)$ and $\epsilon > 0$
- Fossaluza et al. (2009):
  - Deterministic sequential allocation
- Fossaluza et al. (2012):
  - Sequential allocation via mixed intentional/randomized sampling

Compositional Models and Simplex Geometry

- The open (m-1)-Simplex is the set

$$S^{m-1} = \left\{ \boldsymbol{x} \in R^m \mid \boldsymbol{x} > 0 \wedge \mathbf{1}'\boldsymbol{x} = 1 \right\},$$

  where $\mathbf{1}$ is the vector of ones of appropriate dimension.

- The *closure-to-unity* transformation, $clu : R^m_+ \rightarrow S^{m-1}$:

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

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

$$alr(\boldsymbol{x}) = \log\left((1/x_m)[x_1, \ldots, x_{m-1}]\right),$$

$$alr^{-1}(\boldsymbol{z}) = clu\left(\exp([z_1, \ldots, z_{m-1}, 0])\right).$$

We introduce the operators:

- *Power* (scalar multiplication):

$$\alpha * \boldsymbol{x} = clu([x_1^\alpha, \dots, x_m^\alpha])$$

  Interpreted as the α-times repeated effect of proportional decay rates.

- *Perturbation* (vector summation):

$$\boldsymbol{x} \oplus \boldsymbol{y} = clu([x_1 y_1, \dots, x_m y_m])$$

  Interpreted as the effect of proportional decay rates in $\boldsymbol{y}$ over the fractional composition in $\boldsymbol{x}$.

- *Difference*:

$$\boldsymbol{x} \ominus \boldsymbol{y} = clu([x_1/y_1, \dots, x_m/y_m])$$

Compositional Models and Simplex Geometry

We want a distance function on the Simplex, $D_S(\boldsymbol{x}, \boldsymbol{y})$, that exhibits the invariance properties that are most adequate for the purpose of compositional analysis, namely:

- *Perturbation invariance*: For any perturbation, $\boldsymbol{z}$,

$$D_S(\boldsymbol{x} \oplus \boldsymbol{z}, \boldsymbol{y} \oplus \boldsymbol{z}) = D_S(\boldsymbol{x}, \boldsymbol{y}).$$

- *Permutation invariance*: For any permutation matrix, $\boldsymbol{P}$,

$$D_S(\boldsymbol{P}\boldsymbol{x}, \boldsymbol{P}\boldsymbol{y}) = D_S(\boldsymbol{x}, \boldsymbol{y}).$$

- *Power scaling*: For any $\alpha > 0$,

$$(1/\alpha) D_S(\alpha * \boldsymbol{x}, \alpha * \boldsymbol{y}) = D_S(\boldsymbol{x}, \boldsymbol{y}).$$

Aitchison's distance

- The **Aitchison's distance** exhibits all these desirable invariance properties, besides the standard properties for distance functions – *positivity*, *symmetry* and *triangular inequality*:

$$D_S^2(\boldsymbol{x}, \boldsymbol{y}) = [alr(\boldsymbol{x}) - alr(\boldsymbol{y})]' \boldsymbol{H}^{-1} [alr(\boldsymbol{x}) - alr(\boldsymbol{y})],$$

where $H_{i,j} = 2\delta_{i,j} \ + \ 1(1 - \delta_{i,j})$.

- Alternatively, we can write

$$D_S(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} \left[log\left(\frac{x_i}{y_i}\right) - \bar{L}\right]^2},$$

where $\bar{L} = \frac{\sum log(x_i/y_i)}{m}$.

Haphazard Intentional Allocation

- 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:

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

Haphazard Intentional Allocation

- We denote by $\boldsymbol{n}_i^a$, $\boldsymbol{n}_i^h$, $\boldsymbol{n}_i^v$ and $\boldsymbol{n}_i^g$ the quantities of patients already allocated to arm i belonging to each category of factors *age*, *history*, *severity* and *gender*.
  - For example, $\boldsymbol{n}_1^a = [n_{11}^a \ , n_{12}^a \ , n_{13}^a]$ 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* ($\boldsymbol{z}$) in each arm.
  - Purpose: to yield allocations with approximately the same number of patients in each arm.
  - We denote by q_i the total number of patients allocated to arm i, and by $\boldsymbol{n}_i^z = [q_i \ , \ (q_1 + q_2 - q_i)]$ 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 $\boldsymbol{n}_i = [\boldsymbol{n}_i^a \ , \boldsymbol{n}_i^h \ , \boldsymbol{n}_i^v \ , \boldsymbol{n}_i^g \ , \boldsymbol{n}_i^z]$.

Haphazard Intentional Allocation

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

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

- From vectors $\boldsymbol{n}$ and $\boldsymbol{w}$ we obtain the *regularized proportions* vector:

$$\boldsymbol{p}_i = [\boldsymbol{p}_i^a, \boldsymbol{p}_i^h, \boldsymbol{p}_i^v, \boldsymbol{p}_i^g, \boldsymbol{p}_i^z] \ ,$$

  where $\boldsymbol{p}_i^\xi = clu(\boldsymbol{n}_i^\xi + \boldsymbol{w}_i^\xi)$, $\xi \in \{a, h, v, g, z\}$.

Haphazard Intentional Allocation

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

$$\Delta(\boldsymbol{p}_1, \boldsymbol{p}_2) = \frac{1}{5}\Big\{ D_s(\boldsymbol{p}_1^a, \boldsymbol{p}_2^a) + D_s(\boldsymbol{p}_1^h, \boldsymbol{p}_2^h) + D_s(\boldsymbol{p}_1^v, \boldsymbol{p}_2^v) + D_s(\boldsymbol{p}_1^g, \boldsymbol{p}_2^g) + D_s(\boldsymbol{p}_1^z, \boldsymbol{p}_2^z) \Big\}$$

- Consider a new patient that enrolls the study and must be allocated to one of arms 1 or 2. We denote by $\boldsymbol{x} = [\boldsymbol{x}^a, \boldsymbol{x}^h, \boldsymbol{x}^v, \boldsymbol{x}^g, \boldsymbol{x}^z]$ the binary vector indicating to which categories the new patient belongs in each factor.

Allocation Algorithm

- For $j = 1, 2$, consider the allocation of the new patient, $\boldsymbol{x}$, in arm j, that is, for $i = 1, 2$, make $\boldsymbol{m}_i = \boldsymbol{n}_i + \delta(i, j)\boldsymbol{x}$ and perform the following steps:
  1. For $i = 1, 2$ and $\xi \in \{a, h, v, g, z\}$, compute the regularized proportions
  $$\boldsymbol{p}_i^\xi = clu(\boldsymbol{m}_i^\xi + \boldsymbol{w}_i^\xi) \ ;$$
  2. For $i = 1, 2$, set $\boldsymbol{p}_i = [\boldsymbol{p}_i^a, \boldsymbol{p}_i^h, \boldsymbol{p}_i^v, \boldsymbol{p}_i^g, \boldsymbol{p}_i^z] \ ;$
  3. For $i = 1, 2$, set $\boldsymbol{b}_i = [u_i, 1 - u_i]$, where u_i are independently generated from $Uniform(0, 1)$ distribution;
  4. For $\varepsilon \in [0, 1]$, compute the ε-perturbed distance
  $$d_\varepsilon(j) = (1 - \varepsilon)\Delta(\boldsymbol{p}_1, \boldsymbol{p}_2) + \varepsilon D_s(\boldsymbol{b}_1, \boldsymbol{b}_2).$$
- Choose the allocation j that minimizes $d_\varepsilon(j)$, assign the new patient to the corresponding arm, and update vector $\boldsymbol{n}$ accordingly.

Numerical Experiments

- We analyse the performance of our haphazard intentional allocation procedure, for $\varepsilon \in \{0, 0.005, 0.01, 0.05, 0.25, 1\}$.
- We generated $P = 300$ 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 $H = 300$ 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.
    - *Decoupling*: based on the Yule's Q coefficient of association (Yule, 1912); concerns the absence of a tendency to allocate each pair of patients to the same arm.
      Benchmark: pure random allocation method.

Heterogeneity Measure Δ

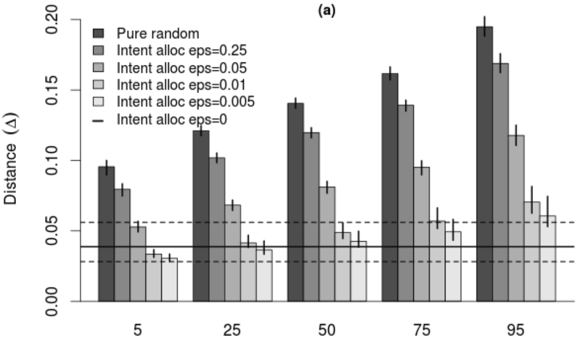

Figure 2. $5\%, 25\%, 50\%, 75\%, 95\%$ empirical percentiles of Δ computed from the H haphazard allocations.

- 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.

Yule's Q Coefficient

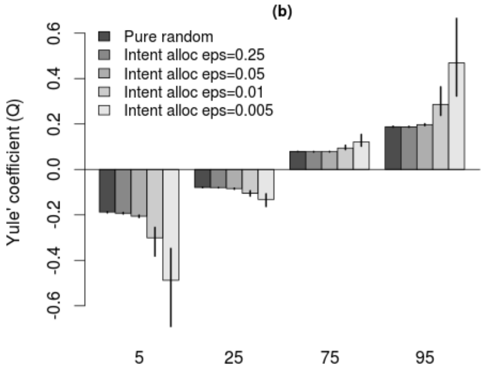

Figure 3. $5\%, 25\%, 75\%, 95\%$ empirical percentiles of Yule's Q coefficient.

- 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

- 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 with the performance on decoupling achieved by the pure random allocation method.
    - For $\epsilon \leq 0.01$, the optimality achieved by the haphazard intentional method comes close to the optimality achieved by the deterministic method; and
    - For $\epsilon \geq 0.05$, the decoupling achieved by the haphazard intentional method is very close to the pure random method – which provides our benchmark for decoupling performance.

References

V. Fossaluza, J. B. Diniz, B. B. Pereira, E. C. Miguel, C. A. B. Pereira
Sequential Allocation to Balance Prognostic Factors in a Psychiatric Clinical Trial. *Clinics, 64, 511-518*, 2009.

V. Fossaluza, M. S. Lauretto, C. A. B. Pereira, J. M. Stern
Combining Optimization and Randomizatin Approaches for the Design of Clinical Trials. In: A.Polpo *et al.* (eds.), *Interdisciplinary Bayesian Statistics*,, Springer Proceedings in Mathematics & Statistics118, 173–184, 2014.

M. S. Lauretto, F. Nakano, C. A. B. Pereira, J. M. Stern
Intentional Sampling by Goal Optimization with Decoupling by Stochastic Perturbation. *AIP Conference Proceedings, 1490, 189-201*, 2012.

D. Lindley
The Role of Randomization in Inference. *Proceedings of the Biennial Meeting of the Philosophy of Science Association, V.2, 431-446*, 1982.

K. R. W. Brewer, C. E. Särndal
Six approaches to enumerative survey sampling. *Incomplete Data in Sample Surveys, 3, 363-368*, 1983.

Basu, D., Ghosh, J. K. (eds.)
Statistical Information and Likelihood, A Collection of Essays by Dr. Debabrata Basu Lecture Notes in Statistics, 45, Springer, 1988.

H. T. Schreuder, T. G. Gregoire, G. B. Wood
Sampling Methods for Multiresource Forest Inventory Wiley, New York, 1993.

J. Aitchison
The statistical analysis of compositional data, Blackburn Press, Caldwell, NJ, USA, 2003.