Intentional Sampling by Goal Optimization with Decoupling by Stochastic Perturbation

Why (not) to Randomize?

Marcelo de Souza Lauretto[+] **Fabio Nakano**[+]
Carlos Alberto de Bragança Pereira[*]
Julio Michael Stern[*,**]
[+]EACH-USP and [*]IME-USP
University of Sao Paulo
[**] jstern@ime.usp.br

EBEB 2012 - XI Brazilian Meeting on Bayesian Statistics
ICES 2013 - IV Simposium Edson Saad Institute - UFRJ
A.C. Camargo 2014 - Metodologia da Pesquisa Científica

1- The Datanexus Case (2002)

Monitoring sample: panel of $\beta = 250$ households for open TV watching habits in Metropolitan Region of São Paulo (MRSP).

Monitoring sample had to be chosen from a *Interview sample* of $m = 10,000$ households, where the head of each household answered a questionnaire about several features of interest*.

*Basic data for MRSP provided by IBGE, the *Brazilian Institute of Geography and Statistics* and *Brazilian Media Group*.

A "representative" monitoring sample should (approximately) reproduce the Interview sample frequencies for the following features:
- Household's income and socio-economical level;
- Individual's sex, age and scholarity;
- Daily hours of TV watching.

The project's tight budget ($\beta = 250$ households) precludes the use of traditional statistical randomized sampling techniques.

2a- Matrix Notation and Data Structure

Features of type $t \in \{1, 2, \ldots, u+v\}$.
$t \in \{1, 2, \ldots, u\}$, household's features;
$t \in \{u+1, u+2, \ldots, u+v\}$, individual's features.
Feature type t, entails a discrete, ordinal, $d(t)$-dimensional classification system, with classes $\{1, 2, \ldots, d(t)\}$.
The auxiliary vector $c(t)$ gives cumulative class dimensions, $c(0) = 0$ and $c(t) = d(t) + c(t-1)$.

Matrix A tabulates all the exploratory research.
$A(h, :)$, h-th row concerns household h and its individuals

For $1 \leq t \leq u$ and $c(t-1)+1 \leq k \leq c(t)$, $A(h,k) = 1$ if household h is of class k for feature type t (0 otherwise).

For $u+1 \leq t \leq u+v$ and $c(t-1)+1 \leq k \leq c(t)$, $A(h,k)$ counts individuals of class k for feature type t living in h.

2b- Matrix Notation and Data Structure

The following normalization conditions hold:

For the household's features,
$1 \leq t \leq u$, and $1 \leq k \leq c(u)$,

- $A(h, c(t-1)+1 : c(t))\mathbf{1} = 1$, h belongs to a single class.
- $\mathbf{1}'A(1 : m, c(t-1)+k)$ counts households of class k.

For the individual's feature,
$u+1 \leq t \leq u+v$ and $c(u)+1 \leq k \leq c(u+v)$,

- $A(h, c(t-1)+1 : c(t))\mathbf{1}$ counts individuals in house' h.
- $\mathbf{1}'A(1 : m, c(t-1)+k)$ counts individuals of class k.

Finally,

- $x'A$, same as $(A'x)'$, counts households or individuals of each class in the sample or "household selection" indicated by the Boolean vector x.

3a- Goal Optimization Sampling Problem

- $g(1 : c(u + v))$, *goal* or target vector for optimal panel representation;
- x, Boolean *decision variables*. x_h indicates if household h belongs (or not) to the selected monitoring sample;
- r, s, non-negative *surplus, r, and slack, s, variables*. In mathematical programming, these artificial variables measure departure from (idealized) constraints,

$$A'x - r + s = g \ ;$$

- b, the monitoring cost and β, the budget. Simplest case: Constant unitary monitoring cost, $b = \mathbf{1}$;
- w, positive *weights*. It may me convenient to write the weights as the ratio of importance and normalization vectors, $w = wm \oslash wn$, Romero (1991);

3b- Goal Optimization Sampling Problem

- *Knapsack constraint*;

$$b'x \leq \beta \ ,$$

- Goal (objective) function:

$$\min f(x) = \ \| \ w \odot (s + r) \ \|_p \ .$$

Milan Zeleny (1982, p.156) enunciates the following *"displaced ideal"* criterion for optimal choice:

- *Alternatives that are closer to the ideal are preferred to those that are farther. To be as close as possible to the perceived ideal is the rationale of human choice.*

For $p = 1$ and $p = \infty$, the absolute and minimax norms, or even a convex combination of the absolute and minimax norms, this GP Problem can be solved by the Simplex method (LP).

4- Multiobjective Programming Sampling Problem

Vilfredo Pareto's (1896) criterion of *dominance*:

- In a Multiobjective Programming problem, a solution A dominates a solution B if and only if A is better than B with respect to at least one objective, and A is not worse than B with respect to the remaining objectives.

Zeleny (1982): GP may produce optimal solutions that are inefficient for an alternative, and better formulated, Multiobjective Programming problem, where only slack variables, s, not surplus, r, are explicitly penalized,

Multi-Objective function:

$$\min f(x) = \| w \odot s \|_p .$$

Notwithstanding apparent benefits of Multi-Objective Progr., Previously stipulated performance and evaluation metrics made Goal Optimization with $p = 1$ norm the formulation of choice.

5a- Debabrata Basu on Randomization

- The [sampling] plan S does not enter into the definition of [the posterior]. Thus, from the Bayesian (and likelihood principle) point of view, once the data x is before the statistician, he has nothing to do with the [sampling] plan S. He does not even need to know what the plan S was.
- Many eyebows were raised when I made the last remark in the opening section of Basu (1969.)... If, however, I know that the plan S is one of the set $\{S_1, S_2, \ldots S_k\}$*, every one of which I fully understand, then my Bayesian analysis of the data* $[x, S]$ *will not bepend on the exact nature of S. In this case I case reduce the data* $[x, S]$ *to the sample* x*.*
- The plan (S) *may be randomized or purposive, sequential or nonsequential. ...we should always be able to work out the corresponding likelihood function.*
Basu (1988, p.197,p.262,p.264)

5b- Debabrata Basu on Randomization

- The object of planning a survey [is a] "representative sampling". But no one has cared to give a precise definition of the term. It is taken for granted that the statistician with his biased mind is unable to select a representative sample.
So a simplistic solution is sought by turning to an unbiased die.
Thus, a deaf and dumb die is supposed to do the job of selecting a "representative sample" better than a trained statistician.
- (Why to randomize?) - The conterquestion '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;
for example, draw a systematic sample with a random start or a very extensive stratification and then draw samples of size 1...
Basu (1988, p.198,p.257) edited.

6a- Decoupling, Sparsity, Randomization, and *Objective* Bayesian Inference

The (false?) Bayesian - Subjective entaglement:
- A statistician who uses subjective probabilities is called a 'Bayesian'. Another name for a non-Bayesian is an objectivist.
I.G.Good (1983.p.87). *Objective* Bayesian?!

Cognitive Constructivism (Cog-Con) framework:

- Objects are tokens for eigen-solutions (behaviors). Eigen-values have been found ontologically to be (sharp) **discrete, stable, *separable* and composable,** *while ontogenetically to arise as equilibria that determine themselves through circular processes.* H.Foerster (2003,p.266).
- Objectivity means invariance with respect to the group of automorphisms. Hermann Weyl (1989, p.132).
- In the Cog-Con framework, model parameters converge to (invariant) eigen-solutions of the Bayesian learning process.
Stern (2011b,p.631).

6b- Decoupling, Sparsity, Randomization, and *Objective* Bayesian Inference

- **Decoupling** *is a general principle that allows us to* **separate** *simple components in a complex system.*
In statistics, decoupling is often expressed as zero covariance , no association, or independence relations.
These relations are sharp statistical hypotheses, that can be tested using the Full Bayesian Significance Test (FBST).
Decoupling relations can also be introduced by some techniques of Design of Statistical Experiments (DSEs), like randomization.
We discuss the concepts of decoupling, randomization and sparsely connected statistical models in the epistemological framework of Cognitive Constructivism (Cog-Con).
Stern (2005a, Abstract).

6c- Decoupling, Sparsity, Randomization, and *Objective* Bayesian Inference

- We change the natural behavior of the individual:
We assign treatment to some patients who, under normal circumstances, will not seek treatment, and give placebo to patients who otherwise would receive treatment.
We are severing one functional link & replacing it with another. Fisher's great insight was that connecting the new link to a random coin flip guarantees that the link we wish to break is actually broken. [Role of Charles Saunders Peirce?]
The reason is that a random coin is assumed to be unaffected by anything [at the] macroscopic level. Pearl (2000, p.348).

- The assertion that one does not need randomization in the context of the assumed model is an empty one because an intrinsic role of randomization is to insure against model inadequacies. Kempthorne (1977, p.16).

7- Pax: Goal Optimization w. Random Perturbations

Not - *Why (not) to randomize?* but - *How to randomize?*

New perturbed knapsack constraint:

$$\widetilde{b}'x \leq \widetilde{\beta}\ , \quad \text{where}$$

$\widetilde{\beta} = \beta + 1$, $\widetilde{b} = b + z$, and $z = \epsilon\,(2/\beta)\text{rand}(m)$.
Perturbation parameter ϵ not much greater than 1.

$E(z(i)) = \epsilon/\beta$. Hence, the expected increase in value for a boolean solution x with β non-zero elements is $E(z'x) = \epsilon$.

Inspirantions:

- Basu (1988). - *Make the plan very purposive, but to leave a tiny bit of randomization in it.*
- Blair (1998). Sensitivity Analysis for Knapsack Problems: A Negative Result. *Disc.Appl.Math.,* 81, 133-139.
  "Negative" $\Leftrightarrow$ Instability, chaos: It works (positively) for us!!

8a- Experimental results - Datanexus case

2,672 candidate households were eliminated from the interview sample due to missing values $\Rightarrow$ 7,328 candidate households.

Simulation experiment:

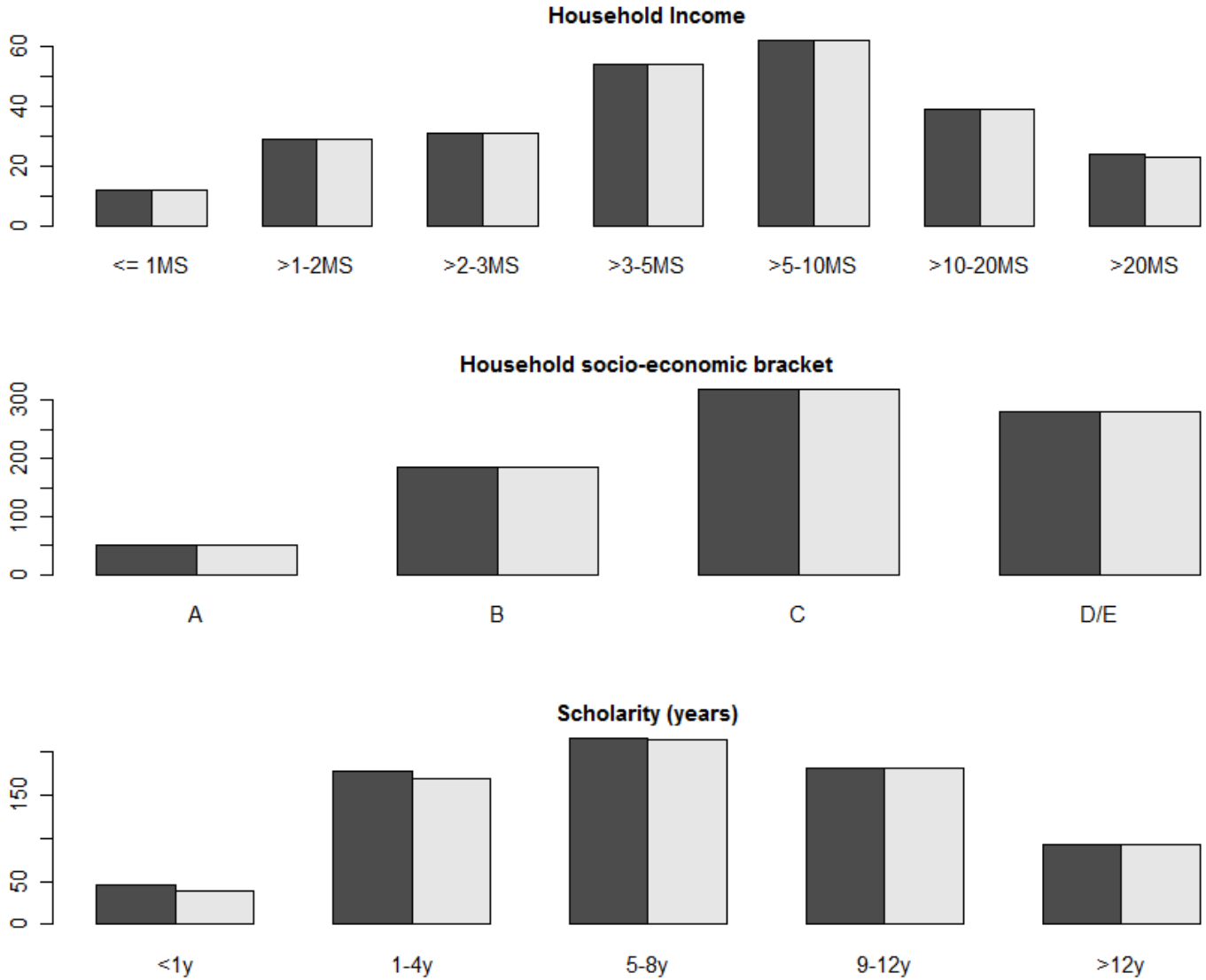
150 runs for each perturbation parameter $\epsilon \in \{1, 1.5, 2, 2.5, 3\}$

Each run consists in generating a random vector z and solving the perturbed goal optimization problem.

Indicators of interest:

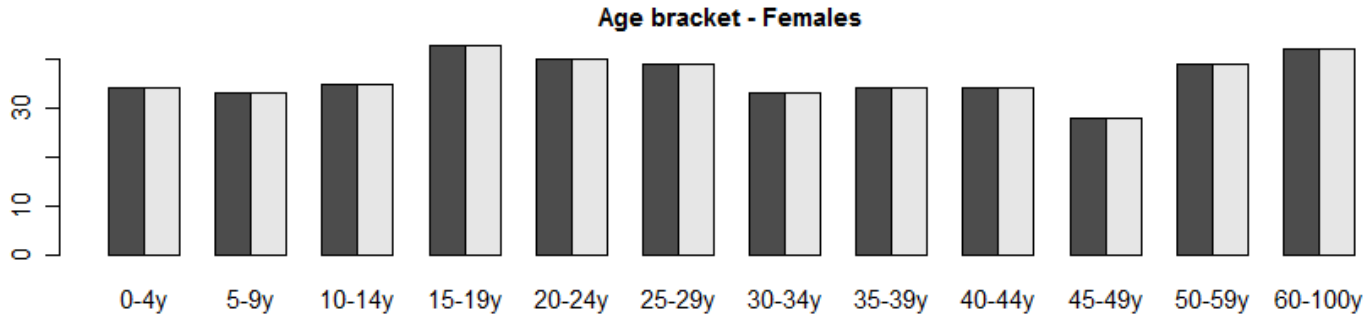
- Optimum objective function value for each run, w.r.t. the optimum solution of non-perturbed goal progr. problem.
- Unbiasedness in sample frequencies of interest.
- Decoupling in household choices among different simulations: # times each household (or pairs of) was (were) chosen in the 150 simulations, for each value of ϵ.

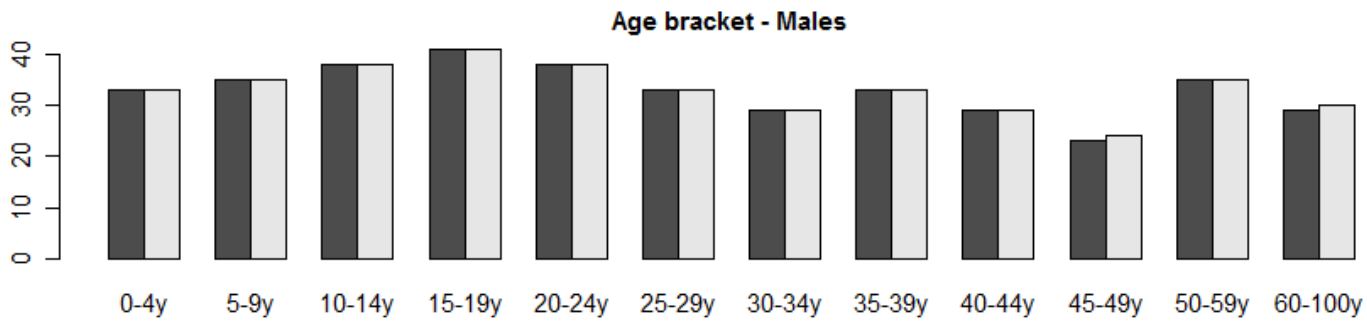
8b- Experimental results - Datanexus case

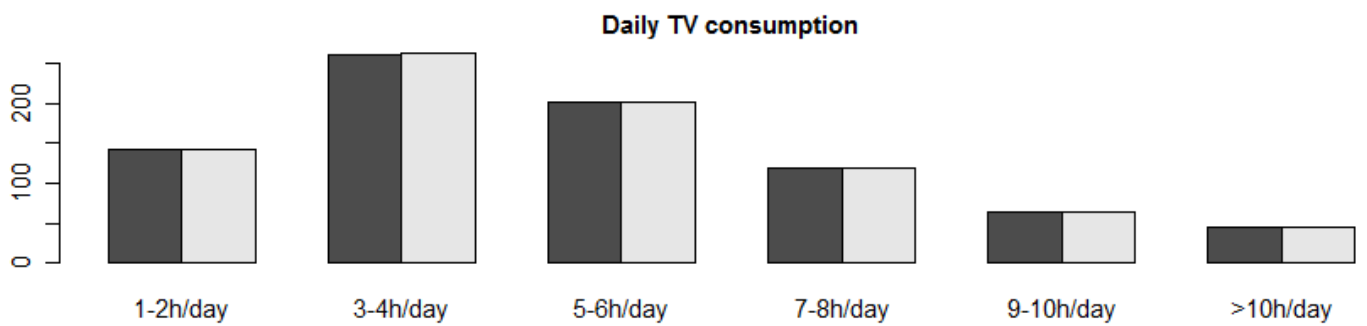


Expected (dark) and actual (light) sample frequencies, $\epsilon = 3$.

8c- Experimental results - Datanexus case

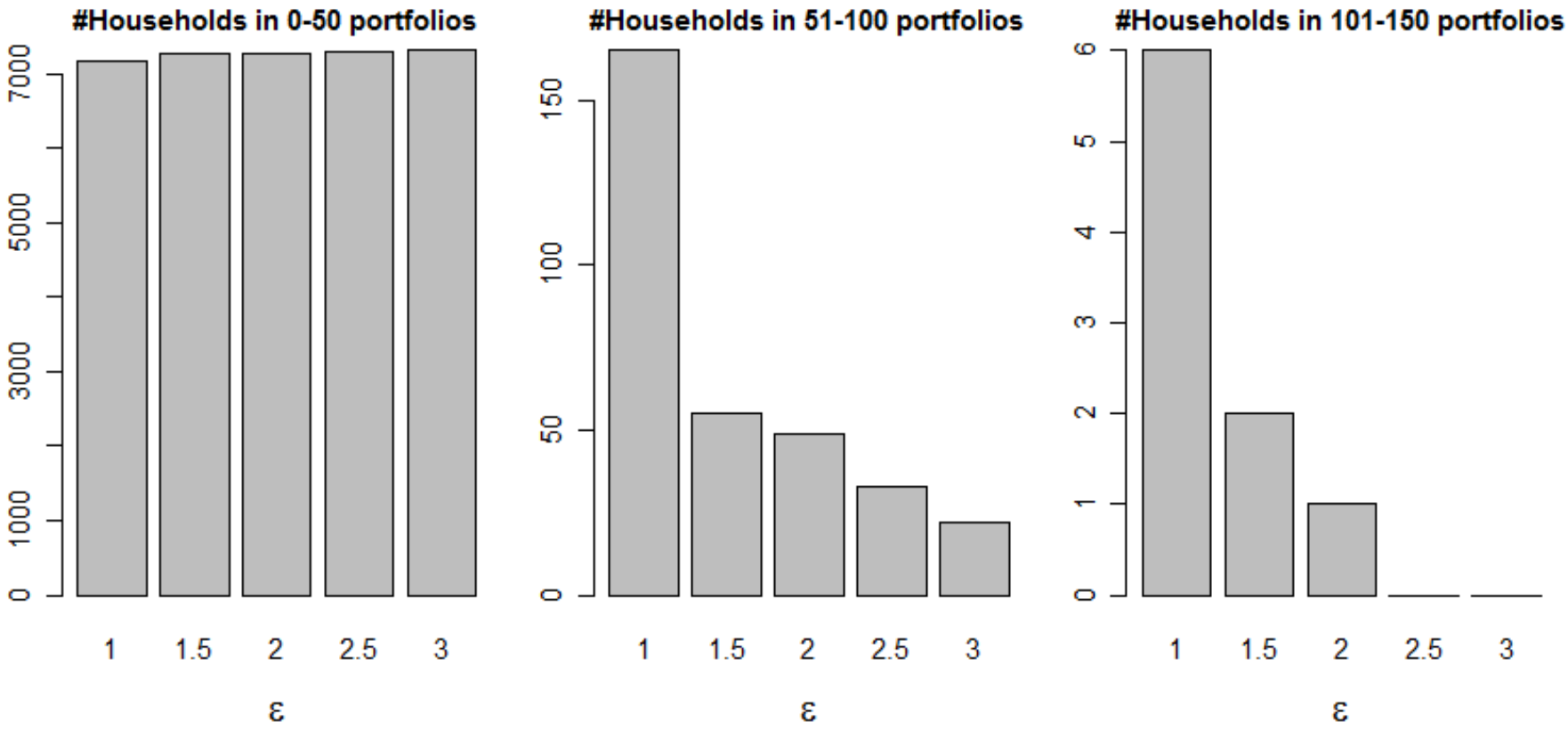






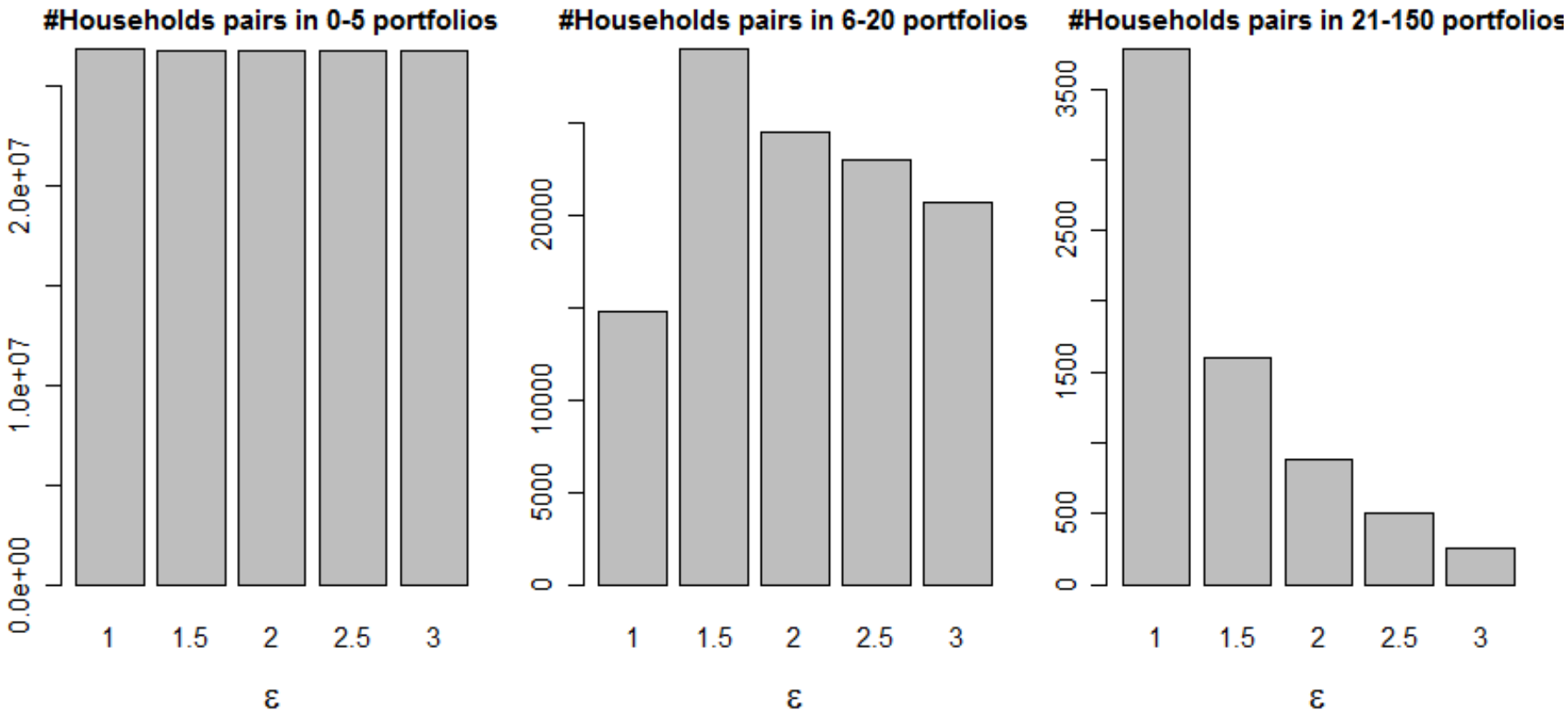
Expected (dark) and actual (light) sample frequencies, $\epsilon = 3$.

8d- Experimental results - Datanexus case



Household repetitions in (sub)optimal panels w.random perturbations for $\epsilon \in \{1, 1.5, 2, 2.5, 3\}$.

8e- Experimental results - Datanexus case



Household repetitions of **pairs** in (sub)optimal panels w.random perturbations for $\epsilon \in \{1, 1.5, 2, 2.5, 3\}$.

9a- Conclusions and Final Remarks

Once the single perturbation parameter, ϵ, is calibrated, the method worked admirably well in the following senses:

- The (objective function) optimal value of the perturbed problem usually decreases only slightly.
- Distinct perturbation vectors, $z = \epsilon(2/\beta)\text{rand}(m)$, produce very different optimal solutions, that is, distinctively diverse (decoupled?) samples;

Box et al. (1978, pp. 102-103) stated:
Control what you can, and randomize what you can not.

- Control by Goal Optimization or Multiobjective Programming;
- Decouple (randomize) by stochastic perturbations on the unstable (chaotic) knapsack constraint.

9b- Conclusions and Final Remarks

Future Research Concerning Inference:

- We hope to extend this paper's approach of integrating intentional sampling with decoupling by stochastic perturbation to sequential designs, see for example Fossaluza et al. (2009).
- We hope to develop more empirical tests for effective decoupling (and some theory?)

Future Research Concerning Optimization:

- We hope to use Blair's results to obtain useful analytical bounds for calibrating the perturbation parameter ϵ.
- We hope to extend our approach using formulations based on non-linear integer programming problems, see for example Skorin-Kapov and Granot (1987).

Bibliography

- D.Basu, J.K.Ghosh (ed.) (1988), Statistical Information and Likelihood. *Lecture Notes in Statistics,* 45, Springer.
- C.Blair (1997). Integer and Mixed-Integer Programming. p.9.1-9.25 in T.Gal, H.J.Greenberg (1997).
- C.E.Blair (1998). Sensitivity analysis for knapsack problems: A negative result. *Discrete Applied Mathematics,* 81, 133-139.
- G.E.P.Box, W.G.Hunter, J.S.Hunter (1978). *Statistics for Experimenters: An introduction to design, data analysis and model building.* NY: Wiley.
- K.R.W.Brewer (2002). *Combined Survey Sampling Inference: Weighing of Basu's Elephants.* Hodder Education Publishers.
- I.J.Good (1983). *Good Thinking: The Foundations of Probability and its Applications.* Minneapolis: Univ.of Minnesota Press.
- 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.
- H.von Foerster (2003). *Understanding Understanding: Essays on Cybernetics and Cognition.* NY: Springer.
- T.Gal, H.J.Greenberg (1997). *Advances in Sensitivity Analysis and parametric programming.* Dordrecht: Kluwer.

- I.Hacking (1988). Telepathy: Origins of Randomization in Experimental Design. *Isis,* 79, 3, 427-451.
- O.Kempthorne (1977). Why Randomize? *J. of Statistical Planning and Inference,* 1, 1-25
- J.Pearl (2000). *Causality: Models, reasoning, and inference.* Cambridge University Press.
- C.S.Peirce, J.Jastrow (1884). On small differences of sensation. *Memoirs of the National Academy of Sciences,* 3, 75-83.
- C.Romero (1991). *Handbook of Critical Issues in Goal Programming.* Oxford: Pergamon Press.
- J.Skorin-Kapov, F.Granot (1987). Non-linear Integer Programming: Sensitivity Analysis for Branch and Bound. *Operations Research Letters,* 6, 269-274.
- J.M.Stern (2005). Decoupling, Sparsity, Randomization, and Objective Bayesian Inference. *Cybernetics and Human Knowing*, 15, 49-68.
- J.M.Stern (2011a). Spencer-Brown vs. Probability and Statistics: Entropy's Testimony on Subjective and Objective Randomness. *Information,* 2, 2, 277-301.
- J.M.Stern (2011b). Symmetry, Invariance and Ontology in Physics and Statistics. *Symmetry*, 3, 3, 611-635.
- M.Zeleny (1982). *Multiple Criteria Decision Making.* McGrawHill