In Defense of Randomization: a Subjectivist Bayesian Approach

Fernando V. Bonassi*, Raphael Nishimura† and Rafael B. Stern**

*Department of Statistical Science - Duke University - USA
fernando.bonassi@duke.edu
†Instituto de Matemática e Estatística - Universidade de São Paulo - Brasil
raphaeln@ime.usp.br
**Faculdade de Direito - Universidade de São Paulo - Brasil
stern@usp.br

Abstract. In research situations usually approached by Decision Theory, it is only considered one researcher who collects a sample and makes a decision based on it. It can be shown that randomization of the sample does not improve the utility of the obtained results. Nevertheless, we present situations in which this approach is not satisfactory. First, we present a case in which randomization can be an important tool in order to achieve agreement between people with different opinions. Next, we present another situation in which there are two agents: the researcher - a person who collects the sample; and the decision-maker - a person who makes decisions based on the sample collected. We show that problems emerge when the decision-maker allows the researcher to arbitrarily choose a sample. We also show that the decision-maker maximizes his expected utility requiring that the sample is collected randomly.

Keywords: Bayesian Statistics; Decision Theory; Game Theory; Intersubjectivism; Randomization
PACS: 02.50.Le

1. INTRODUCTION

It is common to study in Decision Theory a situation in which there is an unique agent. In this framework one wishes to choose the best among several possible decisions. It is well known that there is no random decision which provides the agent an expected utility better than that of the optimal deterministic decision [3]. Among the possible situations in which Decision Theory can be applied, a very important one is that of sample selection.

Based on the aforementioned conclusions, Lindley [5] argues against the use of randomization in sample selection. Many others also give strong arguments against its use on statistical inference [1], [6]. Nevertheless, it remains as a common sense that randomization plays an important role either to (I) accomodate intersubjective problems or (II) reduce the effects of unknown biases.

It is obvious that argument (I) is only reasonable when more than one decision-maker is involved. This kind of situation was first presented by Savage [9]: “The need for randomization presumably lies in the imperfection of actual people and, perhaps, in the fact that more than one person is ordinarily concerned with an investigation”. Kadane and Seidenfeld [6] reinforce the latter when stating that some experiments are made to prove hypotheses to others (and not only to learn).

Following argument (I), we present in Section 2 a model related to intersubjectivity that justifies the use of randomization. The individual rationality is given by utility functions, as in Decision Theory, while the intersubjective rationality is given by Cooperative Game Theory.

Most arguments in favor of randomization are due to (II). A possible justification was given in Berry and Kadane [2]. They present a probabilistic framework that supports this idea. However, their approach does not make explicit use of Decision Theory, since no utility function is clearly presented. Similar arguments are also proposed in [11].

Following [2], we present in Section 3 some models that justify argument (II), extending the idea by formally using an utility function and, consequently, a Decision Theoretic approach. We believe this is an important step to convince a subjectivist Bayesian that randomization is important. In this section we use ideas of Non-Cooperative Game Theory [4], such as dominated strategy.

2. COOPERATIVE GAME MODEL

In this section, we present situations in which different agents wish to reach an agreement on what to do. First, a simple model unrelated to Statistics is presented. Using Cooperative Game Theory we conclude that randomized decisions create new outcomes which are better for all agents. Next, the same methodology is used again to show that randomization can help agents with different prior distributions reach an agreement.

2.1. Illustrative Model

We consider a situation in which two persons, Alice (A) and Bob (B), decide what to do on the weekend. For simplicity, we suppose that there are only two places one can go to on the weekend: the cinema (C) or the theater (T). The possible outcomes are: Alice and Bob choose separately a place or they call it a date and choose a place for both. The following notation will be used: by $(\{X\},Y)$ we mean that X went alone to place Y and by $(\{A,B\},Y)$ we mean that A and B went on a date to place Y. Although both of them look forward to the date, they have opposite preferences on where to go to. The preferences are expressed by the following utility functions:

Outcome	Alice	Bob
$(\{A\},C),(\{B\},C)$	-1	1
$(\{A\},C),(\{B\},T)$	-1	-1
$(\{A\},T),(\{B\},C)$	1	1
$(\{A\},T),(\{B\},T)$	1	-1
$(\{A,B\},C)$	0	3
$(\{A,B\},T)$	3	0

The first column indicates the possible outcomes. The second and third columns indicate, respectively, Alice and Bob's satisfaction when such an outcome occurs.

A natural question to ask is: What will Alice and Bob do on the weekend? Cooperative Game Theory gives an answer using the concept of core of a game [10], [8]. An outcome is in the core of this game whenever three conditions are satisfied:

1. Alice can not get on her own a better outcome.
2. Bob can not get on his own a better outcome.
3. There is no date which is a better outcome for both Alice and Bob.

Outcomes which are not on the core of the game are unstable because there is some person or group that can get, on its own, a better outcome for all members of the group. For example, $(\{A,B\},C)$ is not in the core of the game since in $(\{A\},T),(\{B\},C)$ Alice can get on her own a better outcome. Therefore, it seems reasonable to expect that, if Alice and Bob are rational, then only outcomes on the core of the game will happen.

It is easy to see that the only outcome on the core of this game is $(\{A\},T),(\{B\},C)$. This raises another question: Should Alice and Bob give up on the date because of their divergence on where to go to?

We consider a broader space of outcomes in which the place is randomly chosen. There are two possible outcomes. First, by $(\{A\},p_1*C+(1-p_1)*T),(\{B\},p_2*C+(1-p_2)*T)$ we mean that Alice and Bob do not go on a date, Alice goes with probability p_1 to the cinema and with $(1-p_1)$ to the theater and Bob goes with probability p_2 to the cinema and with $(1-p_2)$ to the theater. Second, by $(\{A,B\},p*C+(1-p)*T)$ we mean that Alice and Bob go on a date and both go to the cinema with probability p and to the theater with probability $1-p$.

In this space it is natural to consider that the utility of an outcome R for some person is that person's expected utility for R. According to Decision Theory this is actually a necessary condition for Alice and Bob to be rational. For example, Alice's utility for $(\{A\},p_1*C+(1-p_1)*T),(\{B\},p_2*C+(1-p_2)*T)$ is $1-2p_1$ and Bob's is $2p_2-1$.

Now one can ask once again: Which outcome will happen? Which outcomes are in the core of the game? Firstly we observe that if $p_1 \neq 0$ or $p_2 \neq 1$, then $(\{A\},p_1*C+(1-p_1)*T),(\{B\},p_2*C+(1-p_2)*T)$ can be improved for Alice on her own or for Bob on his own. Therefore, the only outcome which is not a date which can possibly be on the core of the game is $(\{A\},T),(\{B\},C)$. Nevertheless, for instance, in the outcome $(\{A,B\},0.5*C+0.5*T)$ both Alice and Bob get an expected utility equal to 1.5, greater than in $(\{A\},T),(\{B\},C)$. Therefore, all outcomes in the core of this game are dates. It can be shown the core of this game is $\{\{A,B\},p*C+(1-p)*T : p \in [1/3,2/3]\}$ [1].

2.2. Funding Model

In the illustrative model, randomization provides a way for all players to get a better utility. Nevertheless, this still has little to do with Statistics [2]. Next, we will consider a

---

[1] Even though the symmetry of the utilities might suggest that $p = 1/2$ is the "fairest", Game Theory states that all outcomes in the core might happen. The chosen outcome is a consequence of Alice and Bob's bargaining skills. Although this process is important, it won't be approached in this work.

[2] In section 2.1, there was no uncertainty about the state of nature, a very important feature in Statistics.

sampling model which is permeated by the same ideas presented in 2.1. By doing so, we show that randomization can be an effective way of dealing with inter-subjectivity.

We consider a situation in which there are two research funding agencies, *FAPESP* and *CNPq*. Both of them wish to conduct an experiment in order to learn about a parameter $\theta \in \{0,1\}$. This experiment consists of observing one of two unknown quantites, X_1 and X_2, which also assume values in $\{0,1\}$. This observation is so costly that it is only possible if both agencies cooperate.

Nevertheless, *FAPESP* and *CNPq* disagree on the distribution of these unknown quantities. By P_F we mean *FAPESP*'s probability distribution and by P_C, *CNPq*'s probability distribution. They are given by the following table:

P	P_F	P_C
$P(X_1 = 1 \mid \theta = 0)$	0.1	0.5
$P(X_1 = 1 \mid \theta = 1)$	0.9	0.5
$P(X_2 = 1 \mid \theta = 0)$	0.5	0.1
$P(X_2 = 1 \mid \theta = 1)$	0.5	0.9
$P(\theta = 1)$	0.5	0.5

Therefore, it is possible to say that *FAPESP* believes X_1 is informative for θ and X_2 is not. On the other hand, *CNPq* believes that X_2 is informative for θ and X_2 is not. Although both agencies have the same prior distributions for θ, they have different likelihood functions.

There are three possible deterministic allocations in this situation: by $(\{F\},\{C\})$ we mean that *FAPESP* and *CNPq* do not cooperate and, therefore, no experiment is realized; by $(\{F,C\},X_1)$ we mean that *FAPESP* and *CNPq* cooperate and X_1 is observed and by $(\{F,C\},X_2)$ we mean that *FAPESP* and *CNPq* cooperate and X_2 is observed. We consider that in $(\{F\},\{C\})$ the utility for both agencies is 0. On the other hand, when an experiment is realized the utility for an agency is the expected Kullback-Leibler divergence between posterior and prior distributions for θ, how much is learned, minus 0.1, the utility of the amount invested.

We procede calculating *FAPESP*'s expected utility in each outcome. First, in $(\{F\},\{C\})$, *FAPESP*'s utility is always 0 and, therefore, that is the expected utility. Next, in $(\{F,C\},X_2)$ *FAPESP*'s posterior distribution for θ is equal to its prior distribution with probability 1. Therefore, *FAPESP*'s expected utility in $(\{F,C\},X_2)$ is 0 minus 0.1. Last, in $(\{F,C\},X_1)$ the Kullback-Leibler divergence between *FAPESP*'s posterior distribution for θ and its prior is 0.36 with probability 1. We conclude that *FAPESP*'s expected utility in $(\{F,C\},X_1)$ is 0.26.

Both agency's expected utilities are given by the following table:

Outcome	*FAPESP*	*CNPq*
$(\{F\},\{C\})$	0	0
$(\{F,C\},X_1)$	0.26	-0.1
$(\{F,C\},X_2)$	-0.1	0.26

Next, we can once again make use of Cooperative Game Theory in order to analyze this situation. When only those three outcomes are considered the core of the game is $(\{F\},\{C\})$, since one agency always gets less than 0 in the other ones. Nevertheless,

we can consider outcomes of the form $(\{F,C\}, p*X_1 + (1-p)*X_2)$, that is, in which *FAPESP* and *CNPq* cooperate to observe X_1 with probability p and X_2 with probability $(1-p)$ - a random selection. It can be shown that, when these outcomes are considered, the core of the game is $\{(\{F,C\}, p*X_1 + (1-p)*X_2) : p \in]0.27, 0.73[\}$. Therefore, collecting the sample randomly is best in this situation.

3. NON-COOPERATIVE GAME MODEL

In the last section we studied problems in which people were trying to reach an agreement. Cooperative Game Theory was used to understand the group rationality in these situations. In this section we present situations in which there is a stronger opposition of interests. The agents which are presented do not talk to each other or try to reach an agreement. They try to maximize their expected utility on their own. To analyze these situations we will make use of Non-cooperative Game Theory.

In all the models presented in this section there will always be two agents: the researcher and the decision-maker. The researcher is a person who collects a sample based on his utility function. The decision-maker is a person who makes a decision based on the sample collected. In all the models, the decision-maker is able to decide if he will let the researcher arbitrarily choose a sample or if only a sample collected randomly is acceptable.

3.1. The Convenience Sampling Model

We consider a model in which there is a population with only two units, $\{t_1, t_2\}$. The decision-maker is interested in some populational parameter, here denoted by θ, in $\Theta = \{\theta_0, \theta_1\}$. In order to learn about this parameter, the decision-maker performs an experiment which consists of observing a characteristic of one populational unit, $Y_i \in \{0,1\}$, $i \in \{t_1, t_2\}$. In addition, θ is related to a random feature of each populational unit, $X_i \in \{0,1\}$, $i \in \{t_1, t_2\}$. However, the decision-maker cannot observe this feature. On the other hand the researcher does observe it before collecting the sample. It is more convenient for the researcher to collect a unit with $X_i = 1$.

We illustrate this model with a situation which might occur in a clinic. The researcher might have two options: collect the information from a patient who has just arrived in his clinic or from another patient, who is very ill and unable to leave his house. In the latter, the researcher would have to go to the patient's house, which might be considered incovenient. If the populational parameter of interest is related to the illness, then it would be certainly correlated with the convenience feature of this example.

The decision-maker's prior distribution for θ is $P(\theta = \theta_0) = P(\theta = \theta_1)$. He also believes $P(X_i = 0) = P(X_i = 1)$, $\forall i \in \{t_1, t_2\}$ and X_{t_1} is independent of X_{t_2}. At last, he believes the association between θ, X_i and Y_i is given by $P(Y_i = 1 | \theta = \theta_0, X_i = 1) = 0.3$, $P(Y_i = 1 | \theta = \theta_0, X_i = 0) = 0.2$, $P(Y_i = 1 | \theta = \theta_1, X_i = 1) = 0.7$ and $P(Y_i = 1 | \theta = \theta_1, X_i = 0) = 0.8$, $\forall i \in \{t_1, t_2\}$.

The researcher must choose a unit from $\{t_1, t_2\}$ to collect. This unit will be named d. The researcher's utility function is $U(t_i) = X_{t_i}$. Therefore, it is reasonable that the

decision-maker assumes the researcher will collect an unit with $X_i = 1$, whenever it is available in the population. As a matter of fact, this is a dominant strategy for the researcher in this model.

The decision-maker must choose between d_0, deciding that $\theta = \theta_0$, and d_1, deciding that $\theta = \theta_1$, based on the observed variable Y_d. His utility function is given by the following table:

Utility	d_0	d_1
$\theta = \theta_0$	1	-2
$\theta = \theta_1$	-2	1

It is easy to see that the best decision when observing $Y_d = 0$ is d_0 and when observing $Y_d = 1$ is d_1. It is now possible to calculate the expected utility of this decision rule.

Using the dominant strategy for the researcher, we have:

$$P(Y_d = 1|\theta = \theta_0) = P(Y_d = 1|\theta = \theta_0, X_d = 1)P(X_1 = 1 \cup X_2 = 1) +$$
$$P(Y_d = 1|\theta = \theta_0, X_d = 0)P(X_1 = 0 \cap X_2 = 0) = 0.3 * 0.75 + 0.2 * 0.25 = 11/40$$

$$P(Y_d = 1|\theta = \theta_1) = P(Y_d = 1|\theta = \theta_1, X_d = 1)P(X_1 = 1 \cup X_2 = 1) +$$
$$P(Y_d = 1|\theta = \theta_1, X_d = 0)P(X_1 = 0 \cap X_2 = 0) = 0.7 * 0.75 + 0.8 * 0.25 = 29/40$$

Thus, the expected utility of this decision rule is:

$E(U(Y_d, \theta)) = P(\theta = \theta_0)P(Y_d = 0|\theta = \theta_0) - 2P(\theta = \theta_1)P(Y_d = 0|\theta = \theta_1) -$
$2P(\theta = \theta_0)P(Y_d = 1|\theta = \theta_0) + P(\theta = \theta_0)P(Y_d = 1|\theta = \theta_1) = 0.5 * (29/40 - 22/40 - 22/40 + 29/40) =$
0.175

We can also calculate the expected utility of the decision-maker when he collects a sample randomly instead of giving the researcher freedom to decide. We consider π a uniform random variable in $\{t_1, t_2\}$ and independent of all others. In this situation we have:

$$P(Y_\pi = 1|\theta = \theta_0) = P(Y_\pi = 1|\theta = \theta_0, X_d = 1)P(X_\pi = 1) +$$
$$P(Y_\pi = 1|\theta = \theta_0, X_\pi = 0)P(X_\pi = 0) = 0.3 * 0.5 + 0.2 * 0.5 = 10/40$$

$$P(Y_\pi = 1|\theta = \theta_1) = P(Y_\pi = 1|\theta = \theta_1, X_d = 1)P(X_\pi = 1) +$$
$$P(Y_\pi = 1|\theta = \theta_1, X_\pi = 0)P(X_\pi = 0) = 0.7 * 0.5 + 0.8 * 0.5 = 30/40$$

And the expected utility for the decision-maker is:

$E(U(Y_d, \theta)) = P(\theta = \theta_0)P(Y_d = 0|\theta = \theta_0) - 2P(\theta = \theta_1)P(Y_d = 0|\theta = \theta_1) -$
$2P(\theta = \theta_0)P(Y_d = 1|\theta = \theta_0) + P(\theta = \theta_0)P(Y_d = 1|\theta = \theta_1) = 0.5 * (30/40 - 10/40 - 10/40 + 30/40) =$
0.25

Therefore, it is possible to conclude that, in this case, it is better for the decision-maker to require that the sample is collected randomly than to let the researcher choose one on his own. The problem with this conclusion is that this procedure yields the same expected utility as any sample chosen by the decision-maker. This problem can be approached in two ways.

First, one can consider that the decision-maker and the researcher are the same person. If a person believes all samples yield the same expected utility, can he be certain that other aspects, such as the convenience in collecting the sample, will not intuitively be taken into account? And can this person be sure that this intuitive factor does not depend on the parameter in some manner? If so, the intuitive factor should be taken into account in the model. Nevertheless, as presented by [7], it is difficult to model such a factor.

Next, it is also possible to analyze cases in which the decision-maker must have a known algorithm, which will be used to decide the sample for all researchers eventually working with him. Such is the case of governmental agencies which cannot present *ad hoc* criteria because of the equality principle. Two such algorithms are: requiring that the sample is selected randomly or letting the researcher arbitrarily choose the sample. It is difficult to imagine another general algorithm which cannot be exploited by any researcher.

We conclude that randomization can be an important tool for the decision-maker to prevent the disturbance generated by convenience sampling.

3.2. The Ethical Sampling Model

We can also consider another model in which the variables are the same as those presented in the Convenience Sampling and both the decision-maker and the researcher believe in the same probabilities as those described. They believe that $P(Y_i = 1|\theta = \theta_0, X_i = 1) = 0.3$, $P(Y_i = 1|\theta = \theta_0, X_i = 0) = 0.2$, $P(Y_i = 1|\theta = \theta_1, X_i = 1) = 0.7$ and $P(Y_i = 1|\theta = \theta_1, X_i = 0) = 0.8$, $\forall i \in \{t_1, t_2\}$, among others. Nevertheless, in contrast to the model of Section 3.1 we now consider the researcher's utility function as $U(d) = P(Y_d = 1)$.

We illustrate this model with a situation which might occur in a clinical trial. The variable Y_i corresponds to the survival of the patient submitted to the clinical trial. The variable θ corresponds to the efficiency of the drug being tested in the population. Finally X_i corresponds to a factor which, when present, increases the chance of survival of a person. It is reasonable to assume that a physician, based on ethical reasons, wishes to maximize the probability of survival of the patient.

It is easy to show that the dominant strategy for the researcher is the same as that presented in the Convenience Sampling model. Therefore, the same analysis applies and collecting the sample randomly is still better for the decision-maker than the one collected by the researcher.

The interpretation in which we consider the decision-maker as being the same person as the physician is specially appealing in this case. It is reasonable to assume that the physician cannot put into the model all of his intuitive knowledge about medicine. On the other hand, that knowledge might be used when choosing the sample. As shown, this can be prejudicial for the statistical experiment.

On the other hand, this approach also brings new insight into the benefits of agencies, such as the *FDA*, in requiring the samples to be selected randomly.

4. CONCLUSIONS

In a subjectivist bayesian perspective, we present models to support two of the most common reasons on why to randomize: accommodate intersubjective problems and reduce the effects of unknown biases.

The intersubjective problems are presented in a context of cooperative games. In the model analyzed, it is impossible for the agents to reach an agreement without the use of randomization. On the other hand, when random decisions are possible the agents are able to reach an agreement which is in the best interest of all of them.

The unknown biases are presented in a context of non-cooperative games. Agents play different roles in the process of collecting and analyzing data. Each one of them has a different utility function. It is shown that the agent collecting the sample can introduce prejudicial biases. Therefore, it is best for the analyst to have the sample collected randomly. Opposed to previous presentations, we make explicit use of decision theory through utility functions.

ACKNOWLEDGMENTS

We are grateful to Professor Sergio Wechsler for the opportunity of exposing and discussing this work in his group of Bayesian studies. We also thank André Katsurada, Carlos Pereira, Fabio Niski, Fernando Montera, Fernando Rozenblit, Julio Stern, Luis Gustavo Esteves, Marcelo Arruda, Paulo Marques and Rafael Izbicki for their ideas and suggestions regarding the text. We also thank the comments of both anonymous referees. The authors of this paper have benefited from the support of CNPq and FAPESP.

REFERENCES

1. BASU, D. (1978). Randomization in Statistical Experiments. *FSU Statistics Report M466*, pp 2-8.
2. BERRY, S.M. AND KADANE, J.B (1997). Optimal Bayesian Randomization. *Journal of the Royal Statistical Society. Series B*, **59(4)**, pp 813-819.
3. DEGROOT, M. (2004). *Optimal Statistical Decisions*. Wiley-Interscience, pp 86-106.
4. FUDENBERG, D. AND TIROLE, J.. (1991). *Game Theory*. MIT Press, Cambridge, pp. 1-28.
5. LINDLEY, D.V. (1982). The Role of Randomization in Inference. *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, **2**, pp 431-446.
6. KADANE, J.B AND SEIDENFELD T. (1990). Randomization in a bayesian perspective. *Journal of statistical planning and inference*, **25(3)**, pp 329-345.
7. KAHNEMANN, D., SLOVIC P. AND TVERSKY, A. (1982). *Judgement under uncertainty: Heuristics and biases*. Cambridge University Press, Cambridge, pp 3-22.
8. PELEG, B. (1985) An axiomatization of the core of cooperative games without side payments. *Journal of Mathematical Economics*, **14**, 203-214.
9. SAVAGE, L.J. (1962). Subjective probability and statistical practice. *The Foundations of Statistical Inference*. Methuen, London, pp 33-34.
10. SCARF, H. (1967). The core of an N person game. *Econometrica*, **38**, pp. 50-69.
11. STERN, J.M. (2008). Decoupling, Sparsity, Randomization and Objective Bayesian Inference. *Cybernetics and Human Knowing*, **15(2)**, pp. 49-68.