Maximizing experimental power through dynamic treatment assignment

Stephen Morgan and Bachir Kassas
Food and Resource Economics Department
University of Florida
snmorgan@ufl.edu

DRAFT
Please do not circulate or cite without author permission
Selected paper prepared for presentation at the 2019 Agricultural and Applied Economics Association Annual Meeting, Atlanta, Georgia, July 21-23, 2019

Copyright 2019 by Stephen Morgan and Bachir Kassas. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.
1. INTRODUCTION

Widespread use of laboratory and lab-in-the-field experiments to inform economic policymaking has intensified the focus on replicability as a key metric for successful experimental research (Camerer et al. 2016; Maniadis, Tufano and List 2017). Camerer et al. (2016) find a strong correlation between statistical power and multiple indicators of replicability, suggesting that both sample size and the design of adequately powered experiments should be a priority for experimental economics research. Although the literature carries an extensive account on power analysis for the design of randomly controlled trials in development economics (e.g. Barrett and Carter 2010; Duflo, Glennerster and Kremer 2007; McKenzie 2012), tools for applied economic researchers conducting laboratory experiments are fragmented and relatively ad hoc. As a result, many experimental researchers rely on common rules of thumb (e.g.“allocate 50 subjects to each treatment arm”) that overestimate statistical power to detect a significant average treatment effect (ATE) when designing laboratory experiments or applying for grant funding. The goal of this investigation is to develop a flexible method for empirical researchers seeking to maximize experimental power given the constraints of project funding and subject availability. Focusing on updating experimenter beliefs during the data collection process, this research seeks to answer the question: how can adaptive treatment assignment during data collection increase the power of experimental studies?

In dealing with optimal treatment assignment, we propose leveraging the rollout of an experimental design to update experimenter beliefs about the variance of outcomes under different treatments and dynamically assign remaining subjects to treatment status. This is especially useful for laboratory experiments conducted across multiple sessions. Suppose an experimenter wants to conduct a study with 200 subjects and test for a significant difference in consumer WTP for a food
product between two groups: control (C) and treatment (T). Consider the following iterative procedure using an updating method. First, the researcher can run an experimental session with 20 participants, naïvely allocating an equal number of subjects to the C and T conditions (10 subjects). Once the session is complete, the researcher can estimate the variance of WTP for each treatment group and update their belief about the ratio of the variances between the two arms of the experiment. In the next experimental session with 20 subjects, the researcher can now allocate subjects to T and C using the updated variance data, which will be pooled with the observations from the previous session to re-estimate the ratio of variances between treatment arms. The updating process would continue until the total sample size of 200 participants is reached.

We argue that this procedure has two main advantages over other design methods. First, it economizes on the resources available to run the experiment by preserving what would normally be discarded as pilot data. One of the main uses of pilot experiments (in addition to pretesting the experimental design) is to generate estimates of the treatment outcomes and variances to conduct the power analysis and calculate optimal sample size and treatment assignment. This can be a costly process in terms of participant incentives and committing a fraction of the subject pool to the pilot study. Additionally, pilot studies are often run with convenience samples (e.g. undergraduate students) that may or may not represent the target population of interest (e.g. specific consumers or resource users). In this case, relying on such samples to determine optimal treatment assignment might be misleading as the magnitude and variance of the outcome variable might differ significantly between the pilot sample and target population. Our adaptive process addresses this issue by providing a way to iteratively optimize power without the need for a separate pilot study, thus economizing on experimental resources. Second, this method is accessible for researchers who tend to follow the convention of allocating an equal number of
subjects to each treatment arm with little or no consideration for differences in outcome variances. With equal allocation as a starting point (in the first session), our method allows researchers to update their beliefs about the underlying outcome variances across treatments and adjust subject assignment to maximize power during the data collection process.

This paper proceeds as follows. Section 2 discusses previous literature on experimental power and method for calculating the optimal sample size. Section 3 describes the simulation process and section 4 presents preliminary results. Finally, we conclude and discuss potential implications of this work in section 5.

2. CONTRIBUTION TO EXISTING LITERATURE

This paper builds on the framework of List, Sadoff and Wagner (2011) to analyze the relationship between variance of the outcome variable, treatment assignment, and experimental power. List, Sadoff and Wagner (2011) show that the standard practice of assigning the same number of subjects to each treatment arm maximizes power only when the variance of outcomes under different treatments is equal. Instead, optimal power is achieved when subjects are allocated directly proportional to the standard deviation of the outcomes. For researchers, this means that forming an accurate belief about the variance of outcome variables is critical for maximizing the power of a given study. There are two primary methods of forming this belief: previous literature and pilot studies. While closely related studies can provide good estimates of outcome variance under different treatments, these estimates are necessarily biased due to different contexts of the experimental environment (e.g. monetary incentives, experimenters, products, etc.). On the other hand, pilot studies can provide more direct estimates of outcome variance in a similar context but
are costly to perform and may suffer from having too small of a sample size to get a reliable estimate (Albers and Lakens 2018).

Hahn, Hirano and Karlan (2011) use a similar sequential updating method for treatment assignment focusing on assigning subjects to treatment using propensity scores. Using data from a first experimental session, the authors propose to estimate the conditional variance of outcomes. Propensity scores are chosen to minimize the asymptotic variance associated with estimating the ATE and applied to treatment assignment in the second stage. Applying their method to data on charitable giving (Karlan and List 2007) with an unconstrained overall treatment probability, the authors find efficiency gains of 9.1% when the propensity score is calculated using one covariate and 9.3% when using two covariates (Hahn et al. 2011). However, because the authors rely on asymptotic approximations with large N in both stages of the experimental design, this adaptive process has not been widely adopted for use in laboratory experiments. We contribute to this work by focusing on gains in statistical power under prevalent conditions in laboratory experiments, namely, relatively small sample sizes and data collection stages without treatment assignment based on covariates.

This investigation also relates to a large literature on clinical trials and the assignment of individuals to different treatment regimens over time to address issues related to statistical power, optimal treatment dosage, and ethical concerns about the provision of care (Coffey and Muller 1999; Shen and Fisher 1999; Gould 2001; Berry 2004; Friede and Kieser 2006; Brown et al. 2009; Zhang et al. 2018). For example, Coffey and Muller (1999) show that collection of a specified proportion of a study sample for use as an internal pilot can allow researchers to generate a better estimate of the variance and increase power at the expense of an inflated test size.
3. SIMULATION DESIGN, PARAMETERIZATION, AND IMPLEMENTATION

To test the expected gains in statistical power from using dynamic treatment assignment, we simulate experimental data collection under different updating rules. Similar to the example described above, we construct an environment where we measure the difference in outcome variables between two groups: C and T. Outcomes (Y) are drawn from normal distributions with a known mean and variance. However, we structure the data collection process such that the variance in outcomes is not equal between C and T, which implies that equal subject allocation results in suboptimal statistical power.

Assuming a fixed total sample size, we simulate experimental results and estimate power using Monte-Carlo simulation under three different scenarios. In the first scenario, we follow the naïve convention of assigning an equal number of participants to C and T. In the second scenario, we apply a slow updating process, where the initial beliefs about the true variance ratio between treatments are updated via a weighting process. Suppose we allow $\tilde{\sigma}_i$ to represent the estimated ratio of the variances between the treatment and control group using all available data in period $i$. Equation (1) then illustrates how the slow updating process involves assigning equal weights to the variance ratio calculated at the end of the previous session.

$$\tilde{\sigma}_{i+1} = 0.5 \times \tilde{\sigma}_{i-1} + 0.5 \times \tilde{\sigma}_i$$  \hspace{1cm} (1)

After the ratio is estimated, the next session allocates subjects to C and T according to $\tilde{\sigma}_{i+1}$. Given the fact that laboratory sessions may involve a small sample size (e.g. $n < 30$), the slow updating process should be fairly robust to problems where a single session might result in imprecise estimates of the ratio of the variances. As more sessions are run, the estimated ratio of the treatment variances will converge from the naïve prior towards the sample variance. We should note that in the first session of any experiment, we begin with the assumption of equal variances, or $\tilde{\sigma}_{i-1} = 1$. 
The third scenario involves a faster updating process, where the calculated variance ratio in the current session is used to inform subject assignment in the subsequent session.

\[ \hat{\sigma}_{i+1} = \hat{\sigma}_i \]  

(2)

Equation (2) reflects this updating process whereby new data is collected and zero weight is placed on the previous variance estimates. We will ultimately compare the experimental power achieved under each scenario illustrating the possible gains from dynamic treatment assignment as opposed to the conventional naïve approach.

Our baseline model is parameterized as follows. We set an overall fixed sample size of 200 subjects and collect data in sessions of 20 subjects each. Outcome variables are drawn from distributions with a known mean and variance where \( Y_C \sim N(60, 30) \) and \( Y_T \sim N(70, 60) \). This means that the true ATE is 10 and the ratio of standard deviations \( \frac{\sigma_C}{\sigma_T} = 0.5 \), suggesting that at optimal power there should be two subjects assigned to T for each subject assigned to C. Each data collection process is conducted 1000 times, and we estimate the ATE using OLS. Statistical significance is calculated at the 5% level.

In addition to comparing these three dynamic treatment assignment procedures under the baseline parameter setting, we also vary the parameters across two dimensions. First, we investigate how the fixed sample size affects the performance of the dynamic treatment assignment process under sample sizes ranging from 100 – 300. Second, we investigate the performance of the metric as the true ratio of the outcome variance changes. We do this by changing the ratio of the standard deviations to 0.75 and 0.25 such that the standard deviation in the treatment group is represented by \( Y_T \sim N(70, 90) \) and \( Y_T \sim N(70, 45) \) respectively (standard deviation in the control group remains fixed).
4. PRELIMINARY RESULTS

We present the preliminary results from our simulation exercise in Table 1. We first compare the results of the naïve and dynamic treatment assignment methods under the baseline scenario with a sample size of 200 and $\frac{\sigma_C}{\sigma_T} = 0.5$. Given these parameters, we find that the naïve treatment results in 32.8% power compared to 34.3% power under both the slow and the fast updating methods. Overall this is a relatively small gain suggesting that the assignment of equal samples to C and T performs on par with the proposed updating procedures.

Turning towards our sensitivity analysis, we find that both the fixed sample size and the ratio of the standard deviations affect the performance of the dynamic updating procedure. First, we find that dynamic allocation of subjects to treatment offers the largest performance gains when there is a larger difference in the standard deviations of outcomes under the C and T. Specifically, when $\frac{\sigma_C}{\sigma_T} = 0.25$, the dynamic treatment assignment algorithm results in uniformly higher power when compared to a naïve allocation process. This ratio suggests that 20% of the sample should be allocated to C and 80% to T to maximize power, and the updating process drives the realized assignment closer to this split. When $\frac{\sigma_C}{\sigma_T} = 0.75$, the power maximizing allocation is much closer to equal assignment and the naïve allocation procedure outperforms the dynamic process. Second, we find that dynamic updating offers the largest gains in power when the total sample size is small. As shown in Table 1, we find that with a fixed sample size of 100, the slow updating procedure has 211% ($\frac{\sigma_C}{\sigma_T} = 0.25$) and 61% ($\frac{\sigma_C}{\sigma_T} = 0.75$) higher power compared to the naïve process. At larger sample sizes the benefits of dynamic updating are reduced, except where the ratio of standard deviations is small. This is mainly due to the fact that power is significantly enhanced, and treatment assignment is less sensitive, when sample size is large. It is important to note that this
rarely applies to laboratory experiments, which are usually conducted using a relatively small sample size. Furthermore, this demonstrates the value of our dynamic treatment assignment procedure in maximizing power while economizing on limited resources related to financial budget and subject availability.

Interestingly, there is no difference in power based on the two updating procedures we test in this simulation. This suggests that the estimates of the standard deviations are quickly converging towards the true value after the first session of data is collected. One implication of this finding is that it may be possible to update the allocation of the sessions to treatment arm at a single point rather than doing so after every session. Alternatively, following this result, the researcher can conduct the experiment over a smaller number of sessions. This would be less computationally expensive and may allow researchers to achieve the same gain in power.

5. DISCUSSION

Dynamic assignment of subjects to treatment and control groups can improve the power of existing experimental designs, especially when little is known about the variance of different treatment arms. Using simulation methods, we investigate how the use of already collected data can inform treatment assignment in future experimental sessions to increase experimental power. We find some evidence to suggest that dynamic treatment assignment can be beneficial, especially when the total sample size is low and the difference between outcome variances is large.

These preliminary results face several important limitations that deserve highlighting and further investigation. First, more simulations are required to better understand the interaction of different nuisance parameters and their effect on experimental power. For example, additional simulations should include a broader range of treatment variance ratios, a range of ATEs,
additional mean/variance combinations, and additional distributions to describe the data generating process (e.g. binomial, Poisson, truncated normal, etc.). Second, additional updating methods should be investigated to ensure that if updating methods are used, power is at least as high as the naïve treatment assignment. Third, updating introduces dependence into the treatment assignment process that likely inflates the probability of a Type I error occurring in the final estimation. More work is needed to integrate these considerations into the simulation process and adjust the power calculation accordingly.

Shedding new light on the relationship between treatment assignment and statistical power is important for several reasons. First, given the increased focus on pre-analysis plans, which often impose commitments to a fixed sample size (Olken 2015), our method can be used to help experimental researchers optimize the statistical power of their analysis by iteratively updating subject allocation to the different treatment arms. Second, this study calls into question the argument that increasing power is synonymous with an increase in sample size. Working in the context of limited research budgets, increased attention to other parameters affecting power can save resources and increase the replicability and credibility of laboratory findings.

REFERENCES:


<table>
<thead>
<tr>
<th>Total Sample Size</th>
<th>Naïve</th>
<th>Slow</th>
<th>Fast</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>12.3</td>
<td>38.3</td>
<td>38.3</td>
</tr>
<tr>
<td>120</td>
<td>12.7</td>
<td>39.2</td>
<td>39.2</td>
</tr>
<tr>
<td>140</td>
<td>15.1</td>
<td>42.6</td>
<td>42.6</td>
</tr>
<tr>
<td>160</td>
<td>15.1</td>
<td>46.3</td>
<td>46.3</td>
</tr>
<tr>
<td>180</td>
<td>17.8</td>
<td>47.3</td>
<td>47.3</td>
</tr>
<tr>
<td>200</td>
<td>18.3</td>
<td>46.6</td>
<td>46.6</td>
</tr>
<tr>
<td>220</td>
<td>20.1</td>
<td>46.5</td>
<td>46.5</td>
</tr>
<tr>
<td>240</td>
<td>22.5</td>
<td>45.6</td>
<td>45.6</td>
</tr>
<tr>
<td>260</td>
<td>24.1</td>
<td>47.2</td>
<td>47.2</td>
</tr>
<tr>
<td>280</td>
<td>25.0</td>
<td>45.0</td>
<td>45.0</td>
</tr>
<tr>
<td>300</td>
<td>24.1</td>
<td>51.0</td>
<td>51.0</td>
</tr>
</tbody>
</table>