Experimentally assessing the "missing middle" in studies of racial discrimination

Brian Libgober[*] Tirthankar Dasgupta[†]

September 17, 2018

[*] Department of Political Science, Yale University, 115 Prospect St, New Haven, CT 06511

[†] Department of Statistics and Biostatistics, Rutgers University, 110 Frelinghuysen Rd, Piscataway, NJ 08854

Abstract

Numerous studies have shown that applicants with black-sounding names receive fewer callbacks or e-mail replies than applicants with names that sound white. A troubling aspect of existing study designs is that the names used have a very strong racial valence. Highly identifiable names are both atypical and confounded with income and social status, which raises questions about experimental validity and substantive significance. We consider the problem of designing an email experiment to assess racially disparate outcomes for individuals with fairly or weakly racially identified names, which we call the "missing middle." The problem of discriminating between potential linear and non-linear effects of race is formulated as a statistical inference problem with the objective of inferring about a parameter that determines the shape of a specific function. Under resource constraints, the number of input levels and the total number of experimental units are determined on the basis of criteria associated with asymptotic properties of the maximum likelihood estimator.

1 Introduction

Correspondence studies have become a popular tool for studying racial disparities in outcomes. E-mail experiments have been used to investigate racial access barriers in many contexts, from employment (Bertrand & Mullainathan, 2004), to housing (Carpusor & Loges, 2006), to voting (White et al., 2017; Butler & Broockman, 2011), to getting advice on graduate school (Milkman et al., 2015), both in the American context and abroad (Carlsson & Rooth, 2007). An excellent, recent survey of this literature is (Bertrand & Duflo, 2017). In contrast with non-experimental studies of racial discrimination, or experimental studies that use actors posing as genuine applicants (e.g. Pager et al., 2006), an important benefit of e-mail experiments is that they raise much less concern about the effect of unobservable confounders. Though such studies are still not immune from criticism about whether they measure quantities that are economically meaningful (see, e.g., Heckman, 1998), many consider them as offering particularly clear evidence of racial discrimination (Quillian et al., 2017).

In such studies, the treatment instrument is an e-mail, job application, or some other written correspondence, which is sent to an appropriate set of recipients. The texts are left entirely identical, except for the sender's name which is assigned at random. Names are selected to signal race with near certainty, either based on data from birth-certificates, survey-evidence, or both. In Bertrand & Mullainathan (2004) for example, the names Lakisha Washington and Jamal Jones were used to signal that an applicant was black, while the names Emily Walsh or Greg Baker were used to signal that the applicant was white. The statistical question of interest is whether there are differential response rates by "race," with significance assessed via a simple t-test.

Correspondence studies on this template are straightforward to implement and analyze, yet their very simplicity masks significant interpretive challenges. In particular, difficulties arise because the assigned treatment is literally a name and not a race. On the one hand, there are issues of attribution. If the study finds differential response rates by name, are those due to the signal of racial-identity or to the other signals the name reveals? On the other hand, there are issues of substantive significance. How do the effects one finds about racially-disparate names map onto racial difference in the real world?

These two weaknesses are heightened by only using names with strong racial valence. Individuals with distinctively black names are more likely to come from communities that are racially segregated and poor (Fryer & Levitt, 2004). Hence, the difference in response rates to "Lakisha" and "Emily" may not be due to race *per se*, but rather the relative desirability of the individual's presumed socio-economic background. Moreover, most individuals have names that are only somewhat associated with a given race, so the problem of mapping between name and race is even worse than if one used names that were more typical.

Arguably the disparate treatment these studies observe provides an upper-bound on what the difference would be for the "typical African-American" and the "typical Caucasian." On the basis of existing scholarship, it is purely speculation whether the difference between representative members of each group would be closer to what was observed in these studies or to 0. Similarly, it is also hard to present results that are impactful for policymakers and advocates, who often prefer *lower bounds* on the extent of social problems. By assessing response rates in the "missing middle," or individuals whose names are not strongly identified with race, scholars can use correspondence studies to produce research that has less issues of internal validity, cleaner substantive interpretations, and greater policy-relevance.

Revising the correspondence study template to assess response rates in the "missing middle" creates applied statistical challenges. In particular, the empirical task changes from assessing the difference between two extreme, dichotomous treatments into estimating the response across multiple treatment levels. There are two important design questions that need to be answered: (i) how many levels of treatments do we need to consider? (ii) how many experimental units do we need to expose to each level of treatment? While it is implicitly understood that the more the number of levels of the treatment, the better will be the understanding about the effect of names less strongly associated with races, and higher the sample size, more precise will be the statistical inference, from a resource perspective, there are often resource constraints associated with such experiments. Identifying names with a specified level of association with a particular race is a challenging task and requires screening of large databases of names. Similarly, recipient databases must be assembled. Sending and receiving correspondence becomes more time consuming the more levels one uses, since each name requires its own inbox or mailing address. Thus from a resource perspective, the smaller the number of levels and sample size, the better. To answer questions (i) and (ii), we need a more precise statistical formulation of the problem. We do this in the following Section.

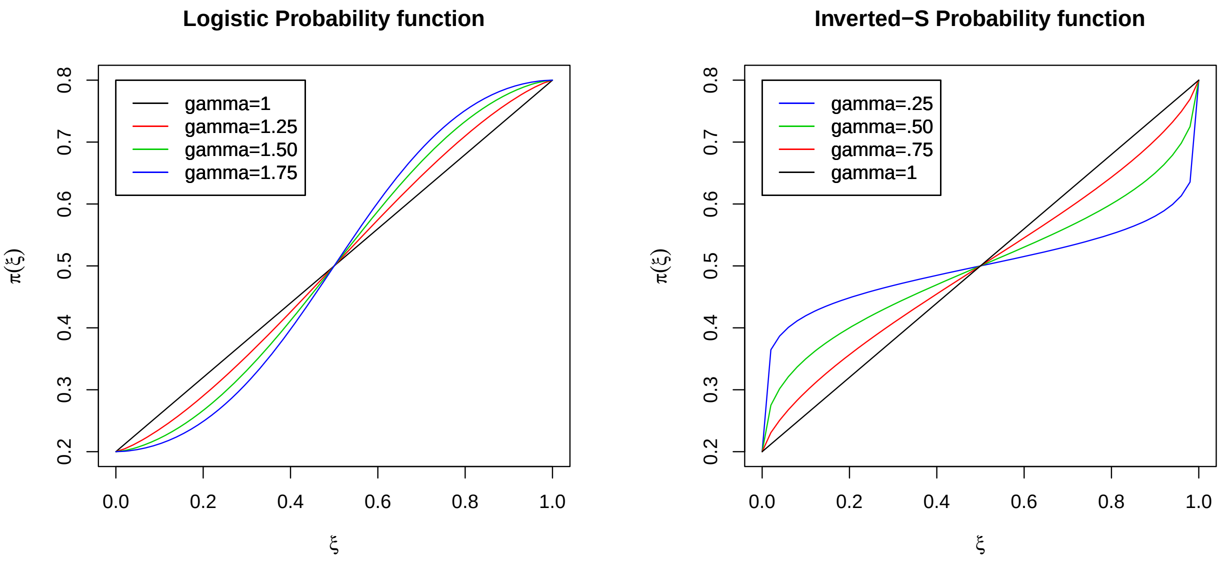
2 Statistical Formulation

Let $\xi \in [0, 1]$ be a continuous variable denoting the "degree of certainty" with which a name of a person can be associated with his/her race, with $\xi = 0$ representing names that can be identified as black with certainty, and $\xi = 1$ representing names that can be identified as white with certainty. We refer to ξ as the "race level", which is the intervention in our experiment. The motivation behind designing the experiment comes from the need for investigating how the probability $\pi(\xi)$ of a response to a sender behaves as a function of ξ. Is it linear, or non-linear? If it is non-linear, then the interest lies in discriminating primarily between the two types of functions shown in Figure 1. The left panel shows a set of logistic functions gradually converging to a linear function as the underlying parameter γ reduces to one. The right panel shows a set of inverted-S functions also approaching a linear function as γ increases to one. The functional form parametrized by γ will be introduced in equations (2)–(4).

Each response function $\pi(\xi)$ tells a different story about how racially disparate outcomes found in existing correspondence studies relate to response rates for those with less strongly identified names. In the inverted S model (right panel), individuals whose names are somewhat identified with a given race receive fairly similar outcomes to one another, regardless of whether their name is more likely white or more likely black. In the logistic model (left panel), it makes a bigger difference whether one is somewhat likely to be white or somewhat likely to be black. These differences between the two probability functions increase as γ moves further from one.

These differences also have real policy implications. If the logistic model is correct, the racial inequality identified by existing correspondence studies is larger and more pervasive than one might have assumed, for example via linearly interpolating between the two extremes. The difference in average or median black response rates is relatively closer to the upper bound estimates that come from existing event studies. If the inverted S model is correct, the inequalities that correspondence studies have found is smaller and more concentrated. The difference in response rates for the typical Caucasian and the typical African-American would be closer to zero.

Figure 1: Plausible non-linear behaviors of $\pi(\xi)$



To further formulate this inference objective from this experiment conducted with N recipients (units), with the ith receiving treatment level ξ_i (that is, a letter from an individual whose name has a race-level ξ), we assume that the ith unit generates a binary response Y_i, which is one if he/she responds to the email and zero if not. Further, we assume that $Y_1, \ldots Y_N$ are independent Bernoulli random variables with $P[Y_i = 1] = \pi(\xi_i)$ for $i = 1, \ldots N$, where the function $\pi(\cdot)$ satisfies

$$0 \leq \pi(0) = \alpha \leq \pi(\xi) \leq \beta = \pi(1) \leq 1, \text{ for all } 0 \leq \xi \leq 1 \tag{1}$$

To discriminate between the logistic and inverted-S functional forms of $\pi(\xi)$, we assume the following functional form of $\pi(\xi)$:

$$\pi(\xi) = \frac{a + g(\gamma, \xi)}{a + b}, \tag{2}$$

where

$$g(\gamma, \xi) = \xi^\gamma / \{\xi^\gamma + (1 - \xi)^\gamma\}, \ \gamma > 0, \text{ and} \tag{3}$$

$$a = \alpha/(\beta - \alpha), \ b = (1 - \alpha)/(\beta - \alpha). \tag{4}$$

Note that (4) ensures that $\pi(\cdot)$ in (2) satisfies (1).

The function $g(\gamma, \xi)$, its variants and more generalized forms have found applications in psychology literature for weighting probability functions (see, for example Goldstein & Einhorn , 1987; Gonzalez & Wu , 1997). The function is inverted-S for $\gamma < 1$, linear for $\gamma = 1$ and logistic for $\gamma > 1$. The shapes in Figure 1 were generated from (2)–(4) by setting $\alpha = 0.2$, $\beta = 0.8$ and by substituting $\gamma = 1, 1.25, 1.50, 1.75$ (left panel) and $\gamma = 0.25, .50, .75, 1$ (right panel).

The problem of discriminating between the linear, logistic and inverted-S forms of $\pi(\cdot)$ can be considered equivalent to the problem of inferring the parameter γ. The design question therefore boils down to determining (a) the number k of levels of ξ, (b)the total sample size N, and (c) the distribution $r_1, \ldots, r_k$ of the total sample size N to the k groups with the objective of estimating γ efficiently.

A common way to address such problems is through an optimal design formulation (Atkinson et al., 2007). For a single-parameter inference problem as in this case, one can consider an information-based design criterion, such as the asymptotic variance of the maximum likelihood estimator (MLE) of γ, and determine an "approximate" optimal design, which means a probability distribution defined on the domain of ξ that optimizes the criterion. However, we choose not to adopt such a "hard" optimization route for three reasons.

First, the model under consideration is non-linear, making the design criterion dependent on the parameter. Therefore, the optimal design solution is dependent on the true value of the parameter, typically known as the local optimal design (Chernoff, 1953). In such cases to find the optimal design, one has to substitute a "guess" of the true parameter value, which defeats the basic purpose of our study, which is to discriminate between models. A popular alternative to this approach is to use a sequential strategy, which may be either frequentist (Chaudhuri & Mykland, 1993) or Bayesian (see, for example Chaloner & Verdinelli, 1995; Zhu et al., 2014). However, sequential experiments are harder to implement and need to be conducted with a reasonably large number of recipients. *Second*, although we have set our goal as discriminating between competing models, in reality the data may reveal a functional form of $\pi(\xi)$ that is completely different from the one described by (2)–(4). For example, there is no guarantee that the actual function form will not be quadratic or cubic, in which case the optimal design may actually turn out perform poorly in estimating the parameters of such a model. *Third*, the optimal design strategy does not answer the question about what N will be adequate for our purpose - rather it addresses the question of distributing N over different levels of ξ.

Due to these reasons, we decide to adopt a "softer" space-filling design strategy while retaining spirit of optimal designs described above, but is more flexible in view of the uncertainty associated with the true model. We decide to conduct the experiment with k equispaced levels of ξ, each of which will be assigned to equal number $r = N/k$ of experimental units. The problem is to determine N and k such that the asymptotic variance of the MLE of γ is within acceptable limits. We describe this approach in the following two Sections.

3 Maximum Likelihood Estimator and its Asymptotic Variance

The results derived in this Section are based on the asymptotic theory of maximum likelihood estimation, that can be found in most textbooks on statistical inference (see, for example Boos & Stefanski , 2013, Chapter 6). Let ξ_i denote the level of the treatment assigned to unit $i \in \{1, \ldots, N\}$, and let y_i denote the binary response obtained from unit i. Recall from Section 2, that the y_i's are assumed to be realizations of independent Bernoulli random variables Y_i's with $P[Y_i = 1] = \pi(\xi_i)$ where $\pi(\xi_i)$ is given by (2)–(4). The likelihood function of the parameter γ given the data $(\boldsymbol{\xi}, \boldsymbol{y})$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ is

$$L(\gamma|\boldsymbol{\xi}, \boldsymbol{y}) \propto \prod_{i=1}^{N} \{\pi(\xi_i)\}_i^y \{1 - \pi(\xi_i)\}^{1-y_i} .$$

Substituting the expression for $\pi(\xi_i)$ from (2), taking logarithm, and dropping terms not involving γ, the log-likelihood function is obtained as

$$\ell(\gamma|\boldsymbol{\xi}, \boldsymbol{y}) = \sum_{i=1}^{N} \left[y_i \log\left(a + g(\gamma, \xi_i)\right) + (1 - y_i) \log\left(b - g(\gamma, \xi_i)\right)\right], \qquad (5)$$

where $g(\gamma, \xi)$ is given by (3), and a, b are given by (4). The MLE $\hat{\gamma}_N$ of the parameter γ based on a sample size of N can be obtained by maximizing (5). The following proposition gives the asymptotic distribution of $\hat{\gamma}_N$. Note that in the proposition, we use the notation $\dot{\sim}$ to denote "approximately distributed as".

Proposition 1. For large N,

$$\hat{\gamma}_N \dot{\sim} \mathcal{N}\left(0, \{I_N(\gamma)\}^{-1}\right), \tag{6}$$

where

$$I_N(\gamma) = \sum_{i=1}^{N} \frac{\xi_i^{2\gamma}(1-\xi_i)^{2\gamma}}{\{\xi_i^{\gamma} + (1-\xi_i)^{\gamma}\}^4} \left[\log\left(\frac{\xi_i}{1-\xi_i}\right)\right]^2 \{a + g(\gamma, \xi_i)\}^{-1} \{b - g(\gamma, \xi_i)\}^{-1} \tag{7}$$

is the expected Fisher information of γ obtained from N data points $(\boldsymbol{\xi}, \boldsymbol{y})$.

The proof of Proposition 1 is in the Appendix. Note that Proposition 1 cannot be directly used to construct asymptotic confidence intervals or conduct tests of hypothesis for γ as the variance term depends on γ. However, we can use a plug-in estimator of the approximate variance by substituting the MLE $\hat{\gamma}_N$ in place of γ in the expected Fisher information $I_N(\gamma)$. Consequently, an approximate $100(1-\alpha)\%$ confidence interval for γ can be obtained as

$$\hat{\gamma}_N \pm z_{\alpha_2}\sqrt{I_N^{-1}(\hat{\gamma}_N)}, \tag{8}$$

where z_{α_2} denotes the $(1-\alpha/2)$th quantile of the standard normal distribution.

To determine appropriate choices for N and k, we consider the following three asymptotic criteria:

1. Coverage of the confidence interval (8), that is, the probability that this interval contains the true value of γ.

2. Width of the confidence interval (8), which is proportional to the asymptotic standard deviation $\sqrt{I_N^{-1}(\hat{\gamma}_N)}$ of the MLE.

3. The probability that the confidence interval (8) contains the value 1. If the interval contains the value 1, it means the null hypothesis $H_0 : \gamma = 1$ cannot be rejected against

the two-sided alternative $H_0 : \gamma \neq 1$ using the Wald test (Wald , 1943). In other words, the data cannot distinguish a linear effect of the treatment from a non-linear effect.

Among these three measures, the width of the confidence interval should be as small as possible. The probability that the confidence interval contains one should be as small when γ is distant from one and increase as γ gets closer to one. The coverage should be as close to 95% as possible. The simulation studies conducted to determine N and k on the basis of the above four criteria are reported in the following section.

4 Simulation studies to determine the design parameters

Armed with the results in Section 3 we now conduct extensive simulation studies to identify our design parameters N and k. We set parameters α and β at 0.2 and 0.8 respectively and consider true values of γ ranging from 0.25 to 1.75 in increments of 0.25. For each value of γ, we consider sample size N from 600 to 2000 in increments of 200. Again, for each N, we consider eight possible choices of k (number of treatment levels): $3, 4, \ldots, 10$. The number of units r assigned to each level is taken to be the largest integer contained in N/k. For each fixed value of γ, N, k and r, the following steps are executed:

Step 1. The treatment levels are obtained as $\xi^j = (j-1)/(k-1)$ for $j = 1, \ldots, k$. The boundary levels ξ^1 and ξ^k are slightly adjusted to 0.001 and 0.999 to avoid computational errors.

Step 2. The probability of obtaining a response of 1 at each treatment level ξ^j for $j = 1, \ldots, k$ is computed by substituting $\xi = \xi^j$ in (2)–(4).

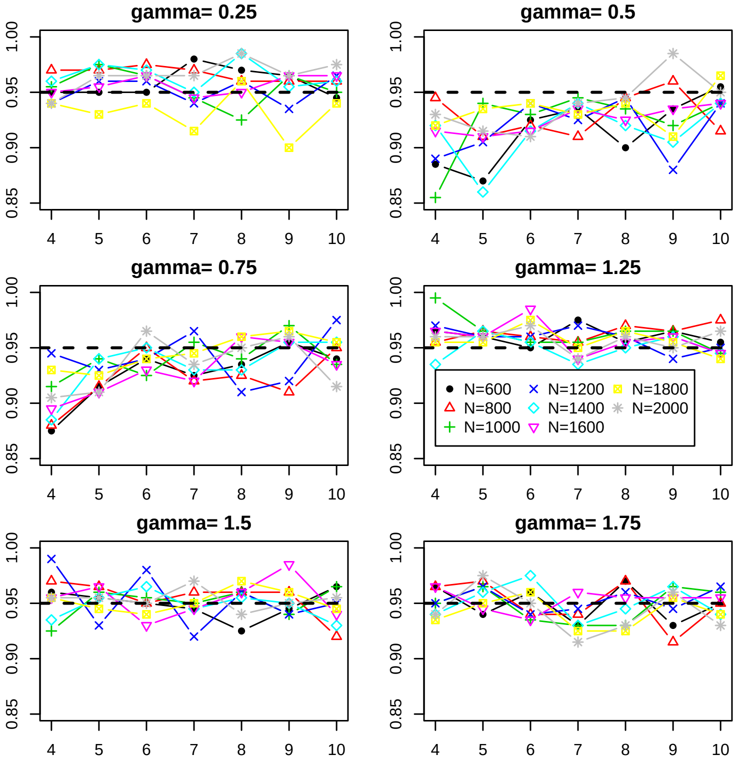
Step 3. The responses for r units receiving treatment level ξ^j are simulated as independent Bernoulli($\pi(\xi^j)$) variables. Generating responses for $j = 1, \ldots, k$ treatment levels yields the complete data $(\boldsymbol{\xi}, \boldsymbol{y})$.

Step 4. The MLE $\hat{\gamma}_N$ is computed by maximizing (5) using the "optim" function in R and its asymptotic variance $I_N^{-1}(\hat{\gamma}_N)$ is computed by substituting $\hat{\gamma}_N$ in (7). A 95% asymptotic confidence interval for γ is computed using (8). Three measures associated with the interval – its width, whether it contains the true value of γ, and whether it contains 1 are recorded.

Step 5. Steps 3–4 are repeated 200 times to estimate (a) the median width of the confidence interval, (b) coverage of the confidence interval and (c) proportion of times the interval includes 1 from 200 datasets.

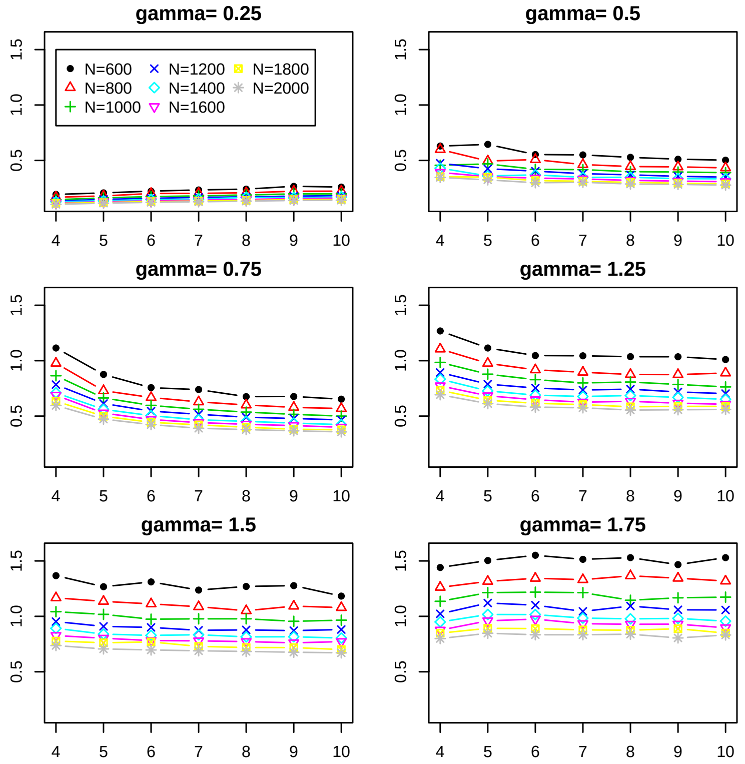
Steps 1-5 are repeated for the seven set values of γ, each for 64 combinations of N and k.

Figure 2: Estimated coverage of asymptotic confidence intervals



Three large-sample measures, estimated coverage and median width of the confidence interval, and the estimated proportion of interval that include one, are plotted against k for different values of N in Figures 2, 3 and 4 respectively. Each sub-plot within each plot represents the results for a specific true value of γ.

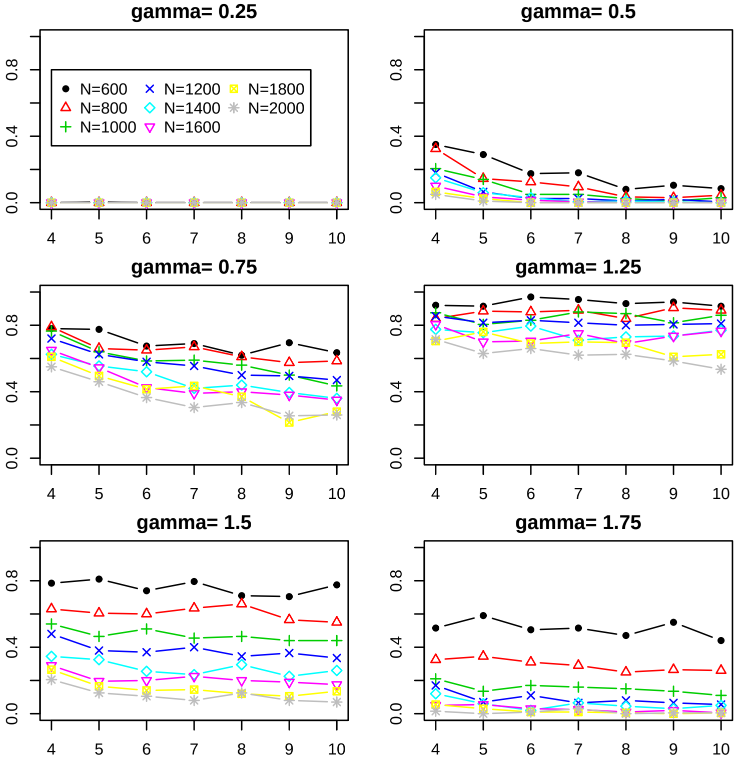
Figure 3: Median width of asymptotic confidence intervals



A close examination of Figure 2 does not reveal any distinct pattern, although the confidence intervals appear to slightly undercover the true parameter values for $\gamma = 0.5$ and $\gamma = 0.75$. Overall, the coverage properties of the intervals appear to be fairly satisfactory for all choices of N and k within the region of exploration, and does not favor any specific choice of these design parameters.

As expected, Figure 3 is informative about the choices of k and N. In each sub-figure, the plots of width versus k are almost parallel for different choices of N, indicating (a) a negative effect of N on width (as expected), and (b) the absence of any interaction between k and N with respect to their effects on width. However, it is interesting to note that while the relationship between k and width is consistent across values of N within each sub-figures, they appear to vary across sub-figures. Specifically, for $\gamma = 0.25$, increasing the number of levels appears to increase the width making $k = 3$ the most desirable choice, although the magnitude of the effect is very small. However, For $\gamma = 0.50$, 0.75 and 1.25, k appears to have a negative effect on width making it advantageous to choose a higher number of levels. The gain, however, does not appear to be substantial for more than six levels. For $\gamma = 1.50$ and 1.75 changing the number of levels do not appear to change the width. From these observations, we tentatively select $k = 6$ as the number of levels. Also, the gains achieved by increasing the sample size appear to be marginal after $N = 1200$, suggesting that this can be a reasonable choice for the sample size.

Figure 4: Estimated proportions of intervals containaing one



We now use Figure 4 to check whether the choices of $N = 1200$ and $k = 6$ can distinguish a non-linear treatment effect from a linear one with a reasonable power. For $\gamma = 0.25$, the estimated probability of the confidence interval including one is negligibly small for all choices of k and N. For $\gamma = 0.50$, 0.75, 1, 1.25, 1.50 and 1.75, this estimated proportion is 0.025, 0.58, 0.96, 0.83, 0.37 and 0.11 respectively. This means, for the above values of γ, the Wald test will reject the null hypothesis of a linear effect with probabilities 0.975, 0.42, 0.04, 0.17, 0.63 and 0.89 respectively. The probability for $\gamma = 1$ is the type-I error, which is seen to be well-controlled by the design. Although the power of the test appears to be very small (0.17) for $\gamma = 1.25$, Figure 1 shows that for $\gamma = 1.25$, the logistic function is barely distinguishable from the linear function. The situation is slightly better when $\gamma = .75$, in which case the nul hypothesis is rejected 42% of the times.

5 Discussion

Although email experiments have been gaining in popularity in the social sciences, careful planning of such experiments from a statistical perspective is rarely used. This paper gives an example of how principles of statistical inference, sampling and design of experiments can be collectively used to plan such experiments, so that there is a trade-off between resource constraints and statistical operating characteristics of the study. The approach adopted is different from a traditional experimental design approach in the sense that it revolves around a tentative model (such as the $\pi(\cdot)$ function in (2)), but is flexible enough to fit other plausible models (such as cubic or higher degree polynomials). Finally, it should be mentioned that the problem addressed in this paper is just one of the complications associated with the type of experiments described here. We have assumed that the input level ξ is a constant and can be chosen without noise. However, in reality the "level" applied will only be an approximate estimate of an underlying "true level". Further, there are several confounding factors associated with the experimental recipients that need to be taken into consideration while designing the experiment. More research is necessary to address these problems.

Appendix: Proofs

Proof of Proposition 1

Differentiating both sides of (5) yields

$$\frac{\partial \ell(\gamma|\boldsymbol{\xi},\boldsymbol{y})}{\partial\gamma} = \sum_{i=1}^{N} \frac{\partial g(\gamma,\xi_i)}{\partial\gamma}\left\{\frac{y_i}{a+g(\gamma,\xi_i)} - \frac{1-y_i}{b-g(\gamma,\xi_i)}\right\}, \tag{9}$$

where

$$\frac{\partial g(\gamma,\xi_i)}{\partial\gamma} = \frac{\partial}{\partial\gamma}\left\{\frac{\xi_i^\gamma}{\xi_i^\gamma+(1-\xi_i)^\gamma}\right\} = \frac{\{\xi_i(1-\xi_i)\}^\gamma}{\{\xi_i^\gamma+(1-\xi_i)^\gamma\}^2}\log\left(\frac{\xi_i}{1-\xi_i}\right). \tag{10}$$

Again, differentiating both sides of (9), after a little algebra, we get

$$\begin{aligned}&\frac{\partial^2}{\partial\gamma^2}\ell(\gamma|\boldsymbol{\xi},\boldsymbol{y})\\ =\;& \sum_{i=1}^{N}\left[\frac{\partial^2 g(\gamma,\xi_i)}{\partial\gamma^2}\left\{\frac{y_i}{a+g(\gamma,\xi_i)} - \frac{1-y_i}{b-g(\gamma,\xi_i)}\right\} - \left\{\frac{\partial g(\gamma,\xi_i)}{\partial\gamma}\right\}^2\left\{\frac{y_i}{(a+g(\gamma,\xi_i))^2} + \frac{1-y_i}{(b-g(\gamma,\xi_i))^2}\right\}\right]\end{aligned} \tag{11}$$

Taking expectation of both sides of (11) and substituting $\mathbb{E}[Y_i] = (a+b)^{-1}\{a+g(\gamma,\xi_i)\}$, $\mathbb{E}[1-Y_i] = (a+b)^{-1}\{b-g(\gamma,\xi_i)\}$, the first term within square brackets in the RHS vanishes, and we have that

$$\begin{aligned}\mathbb{E}\left\{\frac{\partial^2}{\partial\gamma^2}\ell(\gamma|\boldsymbol{\xi},\boldsymbol{y})\right\} &= -\sum_{i=1}^{N}\left\{\frac{\partial g(\gamma,\xi_i)}{\partial\gamma}\right\}^2\left[\frac{1}{(a+g(\gamma,\xi_i))(a+b)} + \frac{1}{(b-g(\gamma,\xi_i))(a+b)}\right]\\ &= -\sum_{i=1}^{N}\left\{\frac{\partial g(\gamma,\xi_i)}{\partial\gamma}\right\}^2(a+g(\gamma,\xi))^{-1}(b-g(\gamma,\xi))^{-1}.\end{aligned}$$

Thus, the expected Fisher information is

$$-\mathbb{E}\left\{\frac{\partial^2}{\partial\gamma^2}\ell(\gamma|\boldsymbol{\xi},\boldsymbol{y})\right\} = \sum_{i=1}^{N}\left\{\frac{\partial g(\gamma,\xi_i)}{\partial\gamma}\right\}^2(a+g(\gamma,\xi_i))^{-1}(b-g(\gamma,\xi_i))^{-1},$$

where $\partial g(\gamma, \xi_i)/\partial\gamma$ is given by (10). The result follows by the asymptotics of the MLE.

References

Atkinson, A, Donev, A & Tobias, R (2007), *Optimum Experimental Designs, With SAS*, Oxford University Press, Oxford.

Bertrand, M & Mullainathan, S (2004), 'Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination', *The American Economic Review*, **94**, 991–1013.

Bertrand, M & Duflo, E (2017), 'Field Experiments on Discrimination', *Handbook of Field Experiments*, North Holland, 309–383.

Boos, D & Stefanski, L (2013), *Essential Statistical Inference*, Springer, NY.

Butler, DM. & Broockman, DE (2011), 'Do politicians racially discriminate against constituents? A field experiment on state legislators', *American Journal of Political Science*, **55**, 463–477.

Carlsson, M & Rooth, DO (2007), 'Evidence of ethnic discrimination in the Swedish labor market using experimental data', *Labour Economics*, **14**, 716–729.

Carpusor, AG & Loges, WE (2006), 'Rental discrimination and ethnicity in names', *Journal of Applied Social Psychology*, **36**, 934–952.

Chaloner, K & Verdinelli, I (1995), 'Bayesian Experimental Design: A Review', *Statistical Science*, **10**, 273–304.

Chaudhuri, P & Mykland, PA (1993), 'Nonlinear Experiments: Optimal Design and Inference Based on Likelihood,' *Journal of the American Statistical Association*, **88**, 538–546.

Chernoff, H (1953), 'Locally Optimal Designs for Estimating Parameters', *Annals of Mathematical Statistics*, **30**, 586–602.

Fryer, R & Levitt, S (2004), 'The causes and consequences of distinctively black names', *Quarterly Journal of Economics*, **119**, 767–805.

Goldstein, WM & Einhorn, HJ (1987), 'Expression theory and the preference reversal phenomena', *Psychological Review*, **94**, 236–254.

Gonzalez, R & Wu, G (1999), 'On the Shape of the Probability Weighting Function', *Cognitive Psychology*, **38**, 129–166.

Heckman, James J (1998), 'Detecting Discrimination', *Journal of Economic Perspectives*, **12**, 101–116.

Milkman, KL, Akinola, M & Chugh, D (2015), 'What Happens Before? A Field Experiment Exploring How Pay and Representation Differentially Shape Bias on the Pathway into Organizations', *Journal of Applied Psychology*, **100**, 1678–1712.

Pager, D, Western, B & Bonokowski, B (2006), 'Targeting, Universalism, and Single-Mother Poverty: A Multilevel Analysis Across 18 Affluent Democracies', *Demography*, **49**, 719–746.

Quillian, L, Pager, D, Hexel, O & Midtbøen, AH (2017), 'Meta-analysis of field experiments shows no change in racial discrimination in hiring over time', *Proceedings of the National Academy of Sciences*, 201706255.

Wald, A (1943), ' Tests of statistical hypotheses concerning several parameters when the number of observations is large', *Transactions of the American Mathematical Society*, **54**, 426–482.

White, AR, Nathan, NL & Faller, JK (2015), 'What Do I Need to Vote? Bureaucratic

Discretion and Discrimination by Local Election Officials', *American Political Science Review*, **109**, 129–142.

Zhu, L, Dasgupta, T & Huang, Q (2014), 'A D-Optimal Design for Estimation of Parameters of an Exponential-Linear Growth Curve of Nanostructures', *Technometrics*, **56**, 432–442.