Continuous Survey Sample Optimization Using Ad Platform APIs

Nandan Rao [1] Dante Donati [2]

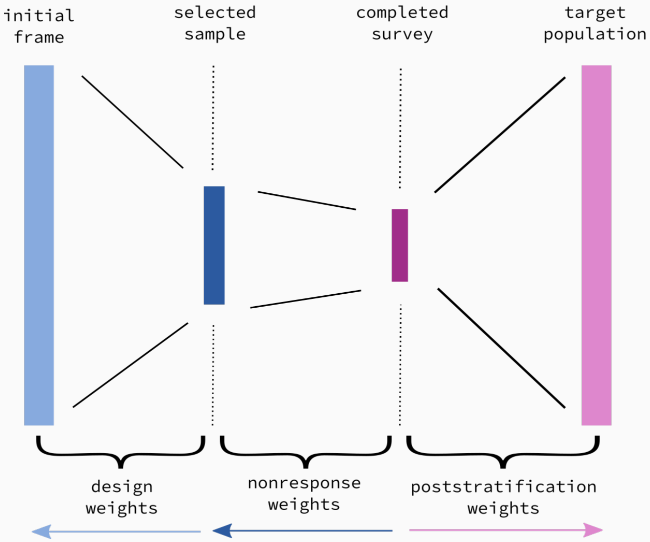
July 19, 2023

[1]Virtual Lab and UAB

[2]Columbia University

Introduction

Motivation - weighting



Setup

Poststratification weighting can be done in many ways, but we will consider the simplest case.

We want to estimate e a population parameter Y via sampling and measurement. We will assume that the researcher wishes to use the stratified mean for a set of strata $h \in H$ (mutually exclusive cells) with an assigned weight for each stratum W_h, which we will denote $\bar{y}$:

$$\hat{Y} := \sum_h W_h \bar{y}_h$$

Idea

1. Commit to poststratification weighting.
2. Measure response rate dynamically during the surveying process
3. Adjust the selected sample dynamically during the surveying process (dynamic over/under-sampling).
4. Make the adjustment to minimize variance subject to budget constraints.

Optimization

Setup

We assume that we are able to measure a set of additional survey responses which we will consider covariates and denote $x_i \in X$. We assume the existence of a mapping $X \to H$ such that the measured covariates are sufficient to assign each individual to one and only one stratum. In addition, we assume that x_i is measured during recruitment.

Setup

The variance of our sample estimate is thus given by:

$$\mathbb{V}[\hat{Y}] = \sum_h W_h^2 \frac{s_h^2}{n_h}$$

where s_h^2 denotes the variance of the population parameter of interest Y within stratum h. If the outcome was measured during recruitment, we can estimate this stratum-specific variance.

Setup

We simplify the problem by assuming that the variance of the outcome in each stratum is equal (i.e. $s_h^2 = s^2$). With that assumption, we have the following variance of our estimate:

$$\mathbb{V}[\hat{Y}] = s^2 \sum_h \frac{W_h^2}{n_h}$$

Note that, given a fixed n and the assumption of equal variance across strata, this quantity is minimized when $\frac{n_h}{n} = W_h$, known as the Neyman allocation.

Setup

But we don't have infinite moneys...

Setup

Denote the cost to recruit an individual from stratum h as P_h.

Denote total budget B, such that $B_h \leq P_h n_h$.

Denote desired maximum sample size N_d.

We can then frame the optimization problem of finding the best allocation of budget to minimize the variance of the final estimate as:

$$
\begin{aligned}
\underset{n_i,\ldots,n_h}{\text{argmin}} \sum_h \frac{W_h^2}{n_h} \\
s.t. \sum_h P_h n_h \leq B \\
\sum_h n_h \leq N_d
\end{aligned}
$$

Setup

But we don’t know the price per respondent...

How to measure cost?

We can model the inverse cost ($\frac{1}{P_h}$), the number of respondents recruited n_{ht} given budget spend B_h, as a Poisson random variable:

$$n_{ht}|B_{ht} \sim Poisson(\lambda)$$
$$\lambda \sim Gamma(\kappa, \beta)$$

We can use closed-form Bayesian updating to obtain a MAP estimator of λ and the implied mean of the predictive distribution $(1/\lambda)$.

Algorithm

How can we run this optimization problem?

We need an interface for recruitment that targets stratum h and allocates budget B_{ht} over a specific period of time t. We denote this interface Recruit(B_t) which accepts a budget allocation $B_t := \{B_{1t}, ..., B_{Ht}\}$.

Additionally, we require an interface GetResults(t) to collect information on the results of recruitment at time t given budget B_t. Results should be considered as the number of respondents recruited for each stratum h at time t and will be denoted n_{ht}.

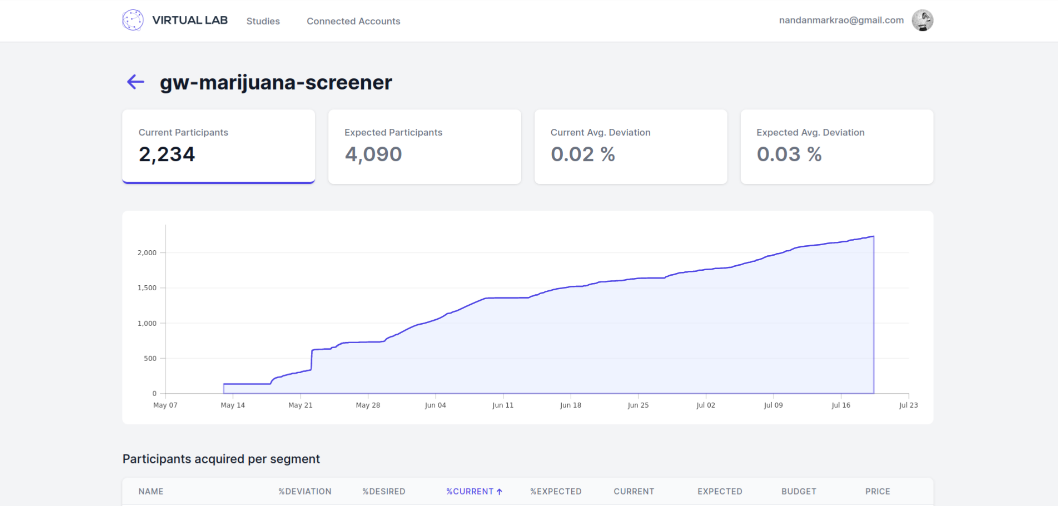
Algorithm

Algorithm 1 Optimizing Stratified Recruitment with Unknown Costs

```
procedure AdOptimization(W, B, N_d, κ, θ)
    B_0 := [1, ..., 1]                                  ▷ Budget indexed by H strata
    n := [0, ..., 0]                                    ▷ Results indexed by H strata
    for t ∈ T do
        p_t := []                                       ▷ Price estimates indexed by H strata
        for h ∈ H do
            n_ht := GetResults(h, t)
            n_h := n_h + n_ht
            p_ht := EstimatePrice(κ, θ, B_ht, n_ht)
        end for
        B_{t+1} := Optimize(W, B, N_d, n, p_t)
        Recruit(B_{t+1})
    end for
end procedure
```

Software

Dashboard


VIRTUAL LAB
Studies
Connected Accounts
nandanmarkrao@gmail.com
gw-marijuana-screener
Current Participants
2,234
Expected Participants
4,090
Current Avg. Deviation
0.02 %
Expected Avg. Deviation
0.03 %
2,000
1,500
1,000
500
0
May 07
May 14
May 21
May 28
Jun 04
Jun 11
Jun 18
Jun 25
Jul 02
Jul 09
Jul 16
Jul 23
Participants acquired per segment
NAME
%DEVIATION
%DESIRED
%CURRENT
%EXPECTED
CURRENT
EXPECTED
BUDGET
PRICE

Dashboard

VIRTUAL LAB   Studies   Connected Accounts   nandanmarkrao@gmail.com

Participants acquired per segment

NAME	%DEVIATION	%DESIRED	%CURRENT ↑	%EXPECTED	CURRENT	EXPECTED	BUDGET	PRICE
location-other_rec-gender-1	0.02	0.07	0.05	0.04	109	158	497.937	10
location-midwest-gender-1	0.03	0.1	0.07	0.06	161	250	641.329	7.2
location-ne_rec-gender-1	0.03	0.1	0.07	0.07	165	296	674.218	5.14
location-focal_rec-gender-1	0.02	0.1	0.08	0.09	181	362	623.442	3.43
location-other_rec-gender-2	0.01	0.07	0.08	0.11	187	454	329.176	1.23
location-focal_rec-gender-2	0	0.1	0.1	0.07	227	294	350.771	5.2
location-midwest-gender-2	0.01	0.1	0.11	0.12	254	491	444.551	1.87
location-ne_rec-gender-2	0.01	0.1	0.11	0.11	246	455	456.279	2.18

Software

Software is fully open source :)

But requires a server cluster to run :(

But is easily installable on kubernetes with helm :)

github.com/vlab-research

Governance Structure

How to make a SaaS sustainable for research purposes?

Results

Results

	Country	Strata	Max CTR	Reach	Respondents	Average Cost
0	India	24	5.09%	873653	9130	$0.24
1	Libya	176	27.1%	1000294	8338	$0.3
2	Labanon	48	15.03%	1370234	17399	$0.57
3	Jordan	192	10.05%	2223793	17223	$0.6
4	Iraq	144	8.58%	4280255	16015	$0.61
5	Serbia	48	13.58%	985109	13995	$0.67
6	Nigeria	23	6.42%	145437	2393	$0.75
7	US	10	13.61%	16849	1118	$0.85
8	US	4	6.9%	10616	316	$0.87
9	Haiti	16	10.0%	1218842	10491	$0.94
10	Honduras	144	7.4%	909597	4922	$0.95

Results

	Country	Strata	Max CTR	Reach	Respondents	Average Cost
11	Lebanon	48	13.82%	1650052	14354	$1.03
12	Papa NG	64	8.71%	118980	1825	$1.46
13	Iraq	80	8.27%	5616663	6520	$1.55
14	US	14	6.14%	120647	2553	$1.66
15	Ukraine	64	5.17%	648512	2394	$1.69
16	Kyrgyzstan	24	9.63%	489054	3004	$1.76
17	Djibouti	16	10.84%	313493	2252	$2.19
18	Kosovo	56	19.93%	663059	6084	$2.46
19	Chad	32	7.59%	327048	2305	$2.64
20	Jamaica	16	15.03%	482097	4105	$2.72
21	Belize	24	9.1%	43684	264	$2.89

Results

	Country	Strata	Max CTR	Reach	Respondents	Average Cost
22	Serbia	1	6.75%	342403	1737	$2.99
23	Macedonia	32	24.58%	384956	3156	$3.19
24	Romania	200	9.31%	785811	1863	$3.23
25	Macedonia	32	25.17%	538295	4565	$3.26
26	Jordan	96	7.14%	1304979	2146	$3.72
27	US	16	3.93%	241134	1429	$4.55
28	Nigeria	120	7.75%	563939	177	$5.55
29	Bulgaria	8	3.91%	170205	170	$6.16
30	Cameroon	80	11.85%	1427793	1712	$6.72
31	India	160	5.77%	3526970	1639	$7.85
32	Gambia	2	5.86%	279008	697	$11.07

Thank you!

https://vlab.digital