Estimating Population Coverage and Selection Probabilities

Application of a capture-recapture framework for the analysis of web panelist

Masahiko Aida, Civis Analytics

5/17 AAPOR Conference Toronto

CIVIS

Problems Public Opinion Research Faces

Problems of Research Options

- We all know telephone survey is ~~dead~~ on life support.
- In opt-in web survey, coverage bias is unobservable
- In opt-in web survey,response mechanism is unobservable

Mark and Recapture Framework I

In biology where population frame is not readily available, mark and capture method is used to estimate population size.

Lincoln–Petersen method

- n^t sample size for each time (t)
- $\hat{N}$ estimated population size
- r number of marked sample found in time 2 (recapture)

$$\frac{n^1}{N} = \frac{r}{n^2}$$

$$\hat{N} = n^1 \frac{n^2}{r}$$

Mark and Recapture Framework II

Key Considerations for good M-C model

- Animal density
- Are there territories?
- Trap cannot hurt the animal that impact survival
- Are they nocturnal or diurnal?

In Public Opinion Research context

- Geography
- Differential capture probability. Are there professional survey takers?
- Difference between panels.
- Some attitude and capture probability may be correlated

Example: Turtle in a pond

- Let us think of a scenario when we catch turtles from a pond. We assume there is no immigration, emigration, death nor birth.
- We caught 100 turtles. We mark the shell with durable paint then we release them.
- We sample again and capture another 75 turtles. We found $r = 15$ turtles had paint marks already! We can estimate a sampling fraction of π from this data.
- Now we know the value of π, the total number of turtle ($\hat{N}$) is 500.
- $\pi = \frac{100}{N}$
- $\pi = \frac{100}{N} = \frac{15}{75} = .2$
- $\hat{N} = 100\frac{75}{15} = 500$

Analysis of Capture Probability

Analysis to identify heterogeneous capture probability

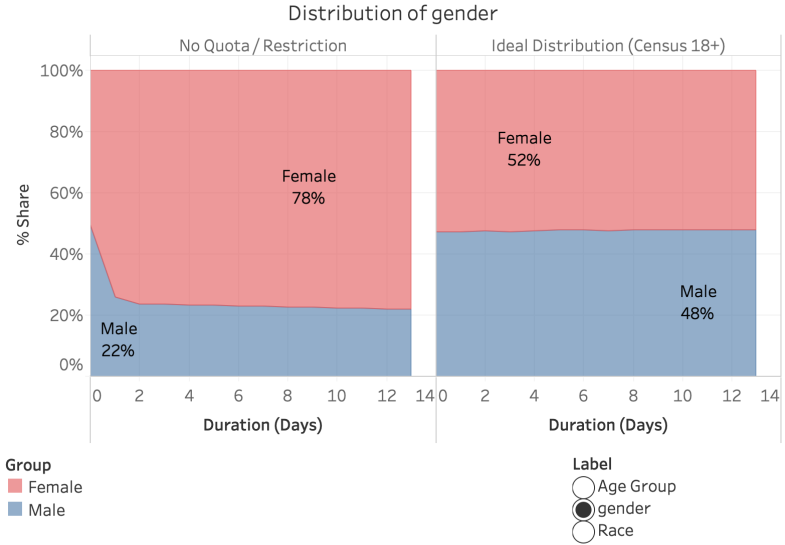
Objective

Identify strata that represent heterogeneous capture probability.

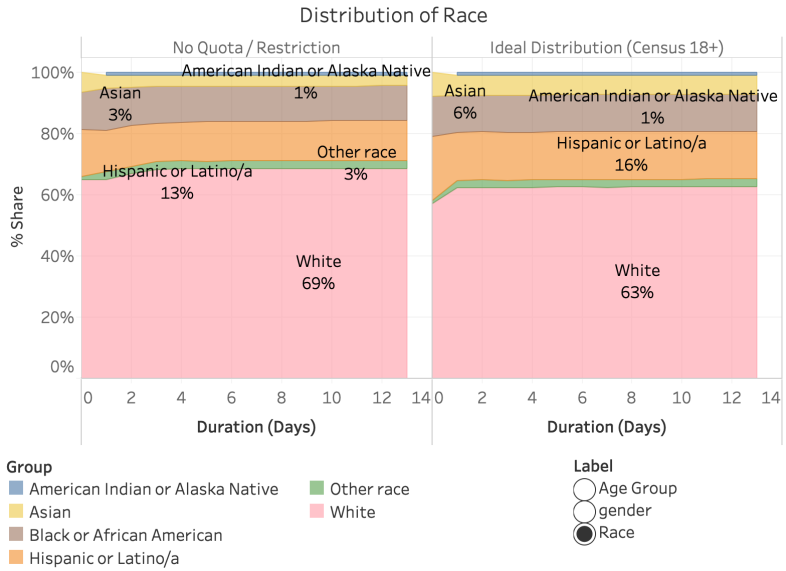
Data

- Data: Web survey conducted by Civis analytics
- n=63,535, collected across 13 days.
- Target audience was adults in the USA. There were no quota nor screening.
- Fit Poisson Regression model

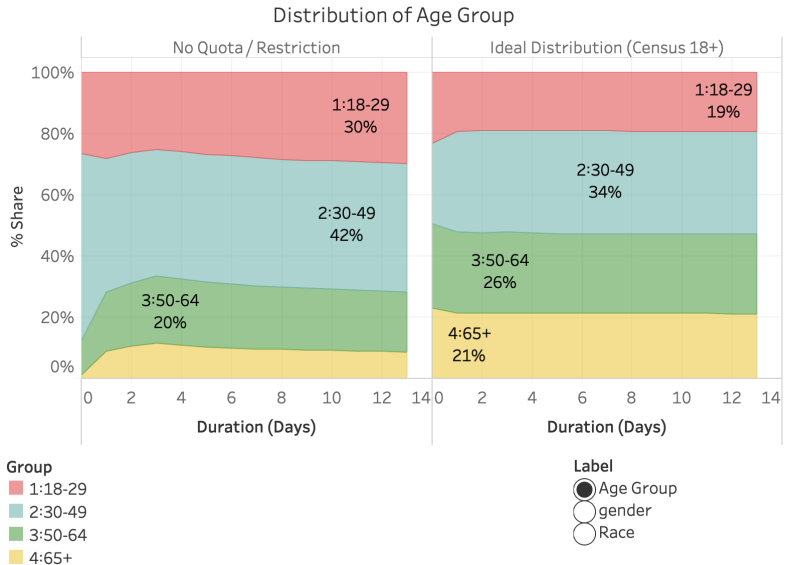
Gender


Distribution of gender
No Quota / Restriction
Ideal Distribution (Census 18+)
100%
80%
60%
40%
20%
0%
% Share
Female
78%
Male
22%
Female
52%
Male
48%
0
2
4
6
8
10
12
14
Duration (Days)
Duration (Days)
Group
Female
Male
Label
Age Group
gender
Race

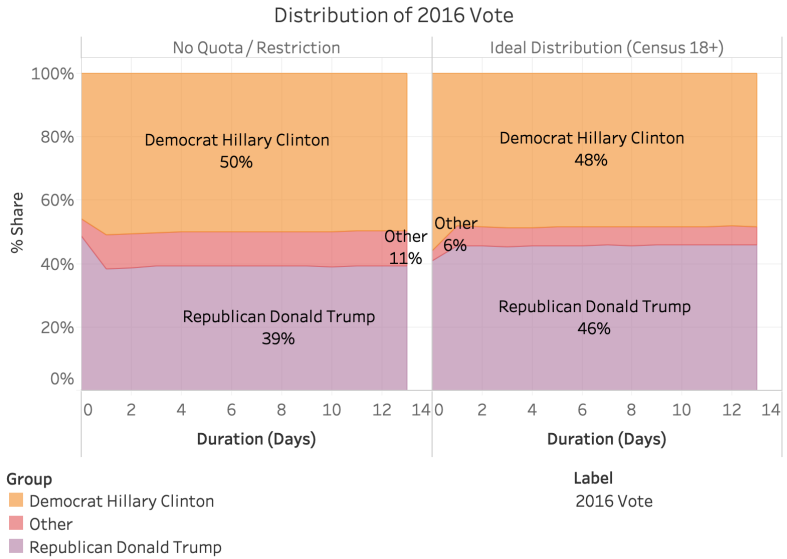
Race


Distribution of Race
No Quota / Restriction
Ideal Distribution (Census 18+)
% Share
100%
80%
60%
40%
20%
0%
American Indian or Alaska Native
1%
Asian
3%
Other race
3%
Hispanic or Latino/a
13%
White
69%
Asian
6%
American Indian or Alaska Native
1%
Hispanic or Latino/a
16%
White
63%
0
2
4
6
8
10
12
14
Duration (Days)
Duration (Days)
Group
American Indian or Alaska Native
Asian
Black or African American
Hispanic or Latino/a
Other race
White
Label
Age Group
gender
Race

Race


Distribution of Age Group
No Quota / Restriction
Ideal Distribution (Census 18+)
% Share
100%
80%
60%
40%
20%
0%
1:18-29
30%
2:30-49
42%
3:50-64
20%
4:65+
21%
1:18-29
19%
2:30-49
34%
3:50-64
26%
0
2
4
6
8
10
12
14
Duration (Days)
Group
1:18-29
2:30-49
3:50-64
4:65+
Label
Age Group
gender
Race

2016 Presidential Vote (voters only)



Summary

- Large capture probability difference by gender and age
- Important capture probability difference by 2016 vote, as this shows capture probability and partisanship is correlated.

Analysis of Panel Size

Estimating Panel Size

The key to obtaining good estimates of total seem:

- Use predictors of capture probability as strata
- Strata has homogeneous capture probability
- Monitor r_h so that it will not go below .1
- Use web survey data Civis collected from 2018 to 2019. n= 1,578,856 interviews.

Stratified Lincoln–Petersen Estimator

$$\hat{N} = \sum_{h=1}^{L} n_h^1 \frac{n_h^2}{r_h} \tag{1}$$

$$= n_1^1 \frac{n_1^2}{r_1} + n_2^1 \frac{n_2^2}{r_2} + \cdots + n_L^1 \frac{n_L^2}{r_L} \tag{2}$$

Result : LP Estimator

- Estimate population size using LP estimator with different stratification method identified in prior analysis.
- IP address is effective strata as it is tied to internet provider (territory).
- It appears our reachable panel size is between 3MM to 4MM.

Stratification	**Estimated Pop Size**
No strata	2,737,567
Age Only	2,748,248
State & Age	3,045,003
Age & Gender	2,762,103
IP address & 2016 vote	3,864,158
IP address & Age	3,975,413
IP address & Gender	4,152,487

Summary and Recommendation

- Our estimates of accessible panel size (4.1MM) for general population survey of 5 10 minutes.
- Large heterogeneity in capture probability by key characteristics exists, (some of them were highly correlated with vote choice).
- Using above characteristics as strata/quota should reduce bias in web respondents capture.
- Response model that encompass both demographic variable and static partisan indicator (ex. partisanship score or 2016 vote) should also help bias in web respondents capture.

Thank You

Correspondence: Masahiko Aida maida@civisanalytics.com

We are hiring.
https://www.civisanalytics.com/careers/

[1]

[1] illustration by https://www.irasutoya.com

Appendix

- Are there situations when people conduct scientific research, and sample/population frame is not available?
- Yes. Ecologist deal with this problem all the time.
- Hayashi (2004) summarises extensive effort (from the 1960s) in the sampling method of Hare.
- Footstep tracing method : create grids and counts grids with footsteps
- Count dropping : count hare poop in the area
- Full capture (census)
- Mark and recapture.

Impact of biased capture on Lincoln–Petersen Estimator

One of the critical assumptions of the mark-capture model is an equal probability of capture among units.

Let us say there are two classes of turtle, swimming turtle, and basking turtle.

- We captured 100 basking turtle and 100 swimming turtle in time 1.
- We then captured 100 swimming and 100 basking turtles in time 2. Recapture was 10 and 3 for the each.
- Estimate total population with and without consideration of class.

$$\hat{N}^{ignoreclass} = 200\frac{200}{13} = 3,076 \text{ False}$$

$$\hat{N}^{swimming} + \hat{N}^{basking} = 100\frac{100}{10} + 100\frac{100}{3} = 4,333 \text{ Truth}$$