

Sample design using the U.S. Census PDB

Stas Kolenikov

21 July 2021



BOLD
THINKERS
DRIVING
REAL-WORLD
IMPACT

Outline

1. The U.S. Census Planning Database, U.S. Census geography, race/ethnicity
2. Survey sampling design target
3. Stas' initial attempt
4. Actual workshop -- challenge to improve upon Stas' work!

Libraries

```
libs <- c('tidyverse', 'here', 'knitr', # 'vtable',  
          'xaringanthemr', 'kableExtra')  
for( l in libs) {  
  library(l, character.only = TRUE)  
}
```

The U.S. Census Bureau and its data

Planning Databases

<https://www.census.gov/topics/research/guidance/planning-databases.2021.html>

PDB data

```
if (file.exists(here('PDB','pdb2021trv3_ct.csv'))) {  
  PDB_CT <- read_csv(here('PDB','pdb2021trv3_ct.csv'))  
} else {  
  PDB_US <- read_csv(here('PDB','pdb2021trv3_us.csv'))  
}
```

US Census Tracts

- Tract \subset county \subset state
- Tract population: about 4000

<https://www2.census.gov/geo/pdfs/education/CensusTracts.pdf>

Example tracts

University of Michigan:

- ACS profile (MCDC)
- Tiger boundaries: <https://tigerweb.geo.census.gov/tigerweb/>, search for 500 S STATE ST, ANN ARBOR, MI, 48109

Stas' residence:

- ACS profile (MCDC)
- Tiger boundaries: <https://tigerweb.geo.census.gov/tigerweb/>, search for CT 11.08 COLUMBIA, MO, 65203

Sample design task

Sample design target

We need to create a sample of adults in the state of <https://en.wikipedia.org/wiki/Connecticut>, with the target of 2500, and oversample targets for racial/ethnic minorities:

- Black/African American: 500
- Hispanic: 500

Simplifications:

- disregard household size distributions between race/ethnicity groups
- disregard the age distributions between race/ethnicity groups

Connecticut

```
if (!exists("PDB_CT")) {  
  PDB_US %>% filter(State=='09', !is.na(Tot_Population_ACS_15_19),  
                    Tot_Population_ACS_15_19 > 0 ) -> PDB_CT  
  # export  
  write_csv(PDB_CT, here('PDB', 'pdb2021trv3_ct.csv'))  
}  
PDB_CT %>%  
  group_by(State, State_name) %>%  
  summarize(tracts=n(),  
            adult_pop = sum( Pop_18_24_ACS_15_19 + Pop_25_44_ACS_15_19 +  
                             Pop_45_64_ACS_15_19 + Pop_65plus_ACS_15_19) ) %>%  
  maybe_kable()
```

State	State_name	tracts	adult_pop
09	Connecticut	828	2 831 241

Stas' first steps

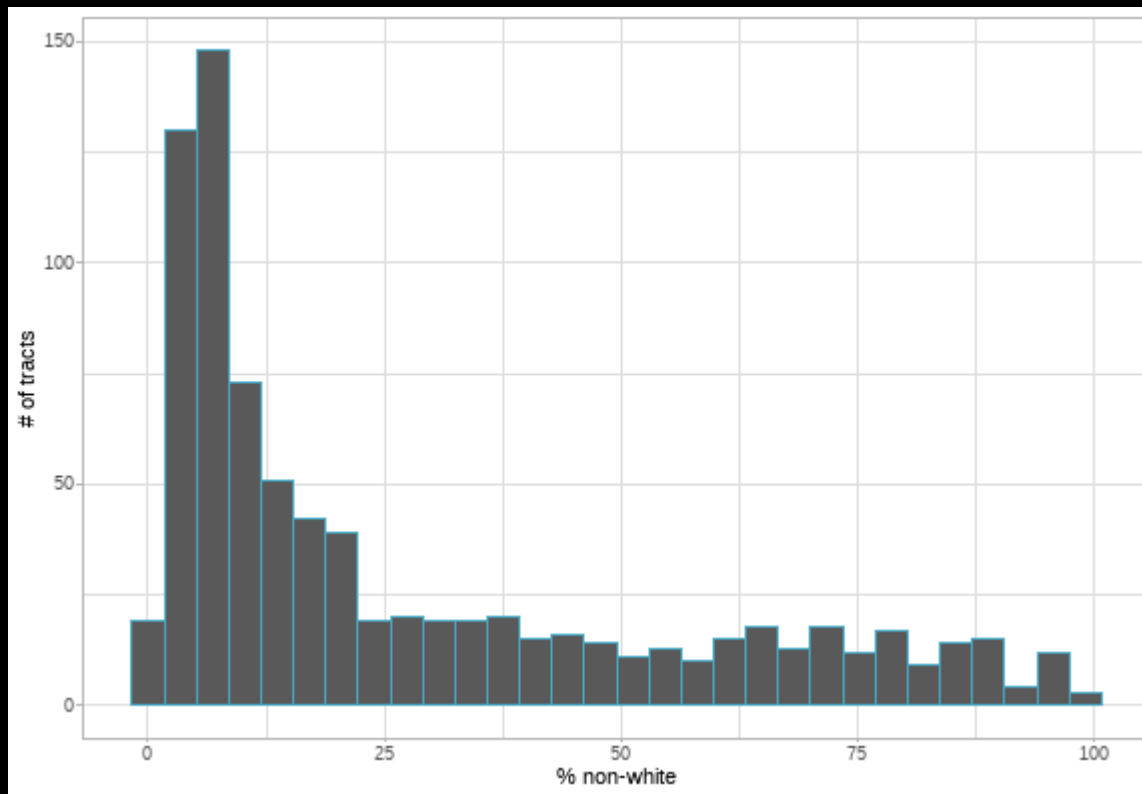
Solution: stratified design

Create several strata and vary sampling rates between them to achieve the target sample sizes.

```
PDB_CT %>% mutate(  
  pct_NH_black_alone = NH_Black_alone_ACS_15_19 / Tot_Population_ACS_15_19,  
  pct_hisp           = Hispanic_ACS_15_19 / Tot_Population_ACS_15_19,  
  pct_minority       = pct_NH_black_alone + pct_hisp  
) -> PDB_CT  
ggplot(data=PDB_CT) +  
  geom_histogram(aes(x=pct_minority), color='skyblue')
```

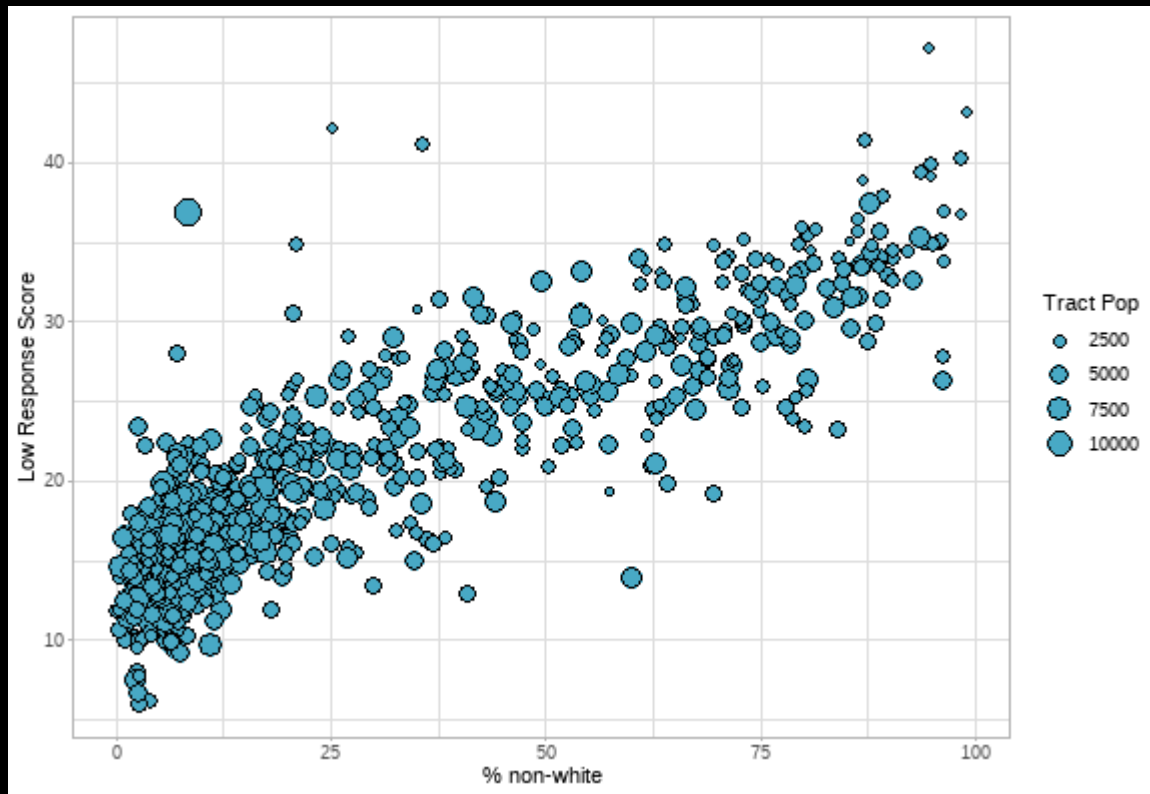
Solution: stratified design

Create several strata and vary sampling rates between them to achieve the target sample sizes.



Beware of nonresponse!

```
ggplot(data=(PDB_CT) +  
  geom_point(aes(x=pct_minority,y=Low_Response_Score,  
    size=Tot_Population_ACS_15_19) )
```



Overall anticipated response rate

```
PDB_CT %>% summarize(  
  LRS = weighted.mean(x=Low_Response_Score,  
                      w=Tot_Population_ACS_15_19,  
                      na.rm=TRUE) ) %>%  
  select(LRS) %>% unlist() -> LRS
```

The overall, population weighted low response score is 20.0746219.

Thus for the target sample size of 2500, one needs to field about 3128 cases.

Two-strata solution: high vs. low minority tracts

```
PDB_CT %>% mutate(strata2 = if_else(pct_minority > 0.5, 1, 2) ) %>%  
  group_by(strata2) %>%  
  summarize(  
    tract      = n(),  
    min_minority = min(pct_minority),  
    max_minority = max(pct_minority),  
    pop      = sum(Tot_Population_ACS_15_19),  
    black    = sum(NH_Black_alone_ACS_15_19),  
    hisp     = sum(Hispanic_ACS_15_19),  
    RR       = 1 - weighted.mean(x=Low_Response_Score,  
                                  w=Tot_Population_ACS_15_19,  
                                  na.rm=TRUE) / 100  
  ) -> CT_strata2
```

Two-strata solution: high vs. low minority tracts

```
CT_strata2 %>% kable()
```

strata2	tract	min_minority	max_minority	pop	black	hisp	RR
1	182	0.50134	0.9912892	704 096	213 395	294 119	0.7042099
2	646	0.00000	0.4997575	2 870 978	140 725	280 121	0.8225065

Trial-and-error allocation

Compute anticipated number of Black/AA interviews; number of Hispanic interviews; adjust inputs until the results are acceptable

```
CT_strata2 %>% mutate(  
  n_field = case_when(strata2 == 1 ~ 2100, strata2 == 2 ~ 1100),  
  n_total = floor(n_field * RR),  
  n_black = floor(n_field * RR * black / pop),  
  n_hisp = floor(n_field * RR * hisp / pop),  
  sampling_rate = n_field/pop*1e3  
) %>% select(strata2, sampling_rate, starts_with('n_')) -> CT_strata2_cc
```

strata2	sampling_rate	n_field	n_total	n_black	n_hisp
1	2.9825478	2 100	1 478	448	617
2	0.3831447	1 100	904	44	88

Overall sample size: 2382 vs. 2500, Black AA race and Hispanic ethnicity oversamples of 492 and 705 (vs. target 500 each).

Trial-and-error allocation

Compute anticipated number of Black/AA interviews; number of Hispanic interviews; adjust inputs until the results are acceptable (overall sample size 2500, Black AA race/Hispanic ethnicity oversamples of 500 each).

```
bind_rows( CT_strata2_completes %>% mutate(strata2=as.character(strata2)),  
           CT_strata2_completes %>% select(starts_with("n_")) %>%  
           summarize_all(sum) %>% mutate(strata2='Total') %>% kable()
```

strata2	sampling_rate	n_field	n_total	n_black	n_hisp
1	3.0038517	2 115	1 489	451	622
2	0.4284254	1 230	1 011	49	98
Total	NA	3 345	2 500	500	720

Simple weights

```
CT_strata2_completes %>%  
  full_join( CT_strata2 %>% select(strata2, pop), by='strata2') %>%  
  mutate(weight=pop/n_total) %>%  
  select(strata2, n_field, n_total, pop, weight) -> CT_strata2_weights  
CT_strata2_weights %>% maybe_kable()
```

strata2	n_field	n_total	pop	weight
1	2 115	1 489	704 096	472.865
2	1 230	1 011	2 870 978	2 839.741

Unequal weighting design effect

Unequal weighting design effect $1 + CV^2 \equiv 1 + L_{Kish}$ for this design is:

```
(CT_strata2_weights %>%  
  summarise( n_wgt = sum(n_total*weight),  
             n_wgt2 = sum(n_total*weight*weight),  
             n = sum(n_total) ) %>%  
  mutate(UWE_DEFF = n_wgt2 * n / (n_wgt*n_wgt) ) %>%  
  select(UWE_DEFF) %>% unlist() -> UWE_DEFF2)
```

```
## UWE_DEFF
```

```
## 1.659822
```

Can you do better??

Better solutions?

- Better choice of the threshold in a two-strata solution?
- Three strata?
 - two thresholds of minorities, combined?
 - separate thresholds for Black/African Americans vs. Hispanics?
- Four strata?
- Minimize design effect?
- Account for response rates at the tract level?

Your turn now!

- Breakout Zoom rooms, groups of ~4
- Create a *better* design:
 - the above one had too many Hispanics relative to the target (and relative to SRS; hence losses of efficiency)

Stas' best design has DEFF of about 1.23, but it involved heavy-handed numeric optimization with tons of fiddling with optimization parameters.

Further refinements

- Adult vs. total population
- Language barriers (speak English less than very well)
 - partially incorporated in the Low Response Score
 - limits the covered population
- Residential households (vs. group quarters)
- Vacant housing units
- Lower response rates to non-federal surveys

R Markdown

This is an R Markdown `library(xaringan)` presentation. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

R version: R version 4.0.2 (2020-06-22).

Package versions:

- `library(tidyverse)`: version 1.3.0
- `library(here)`: version 0.1
- `library(knitr)`: version 1.30
- `library(xaringanthemer)`: version 0.3.0
- `library(kableExtra)`: version 1.2.1