Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 5a: Simple and stratified random sampling

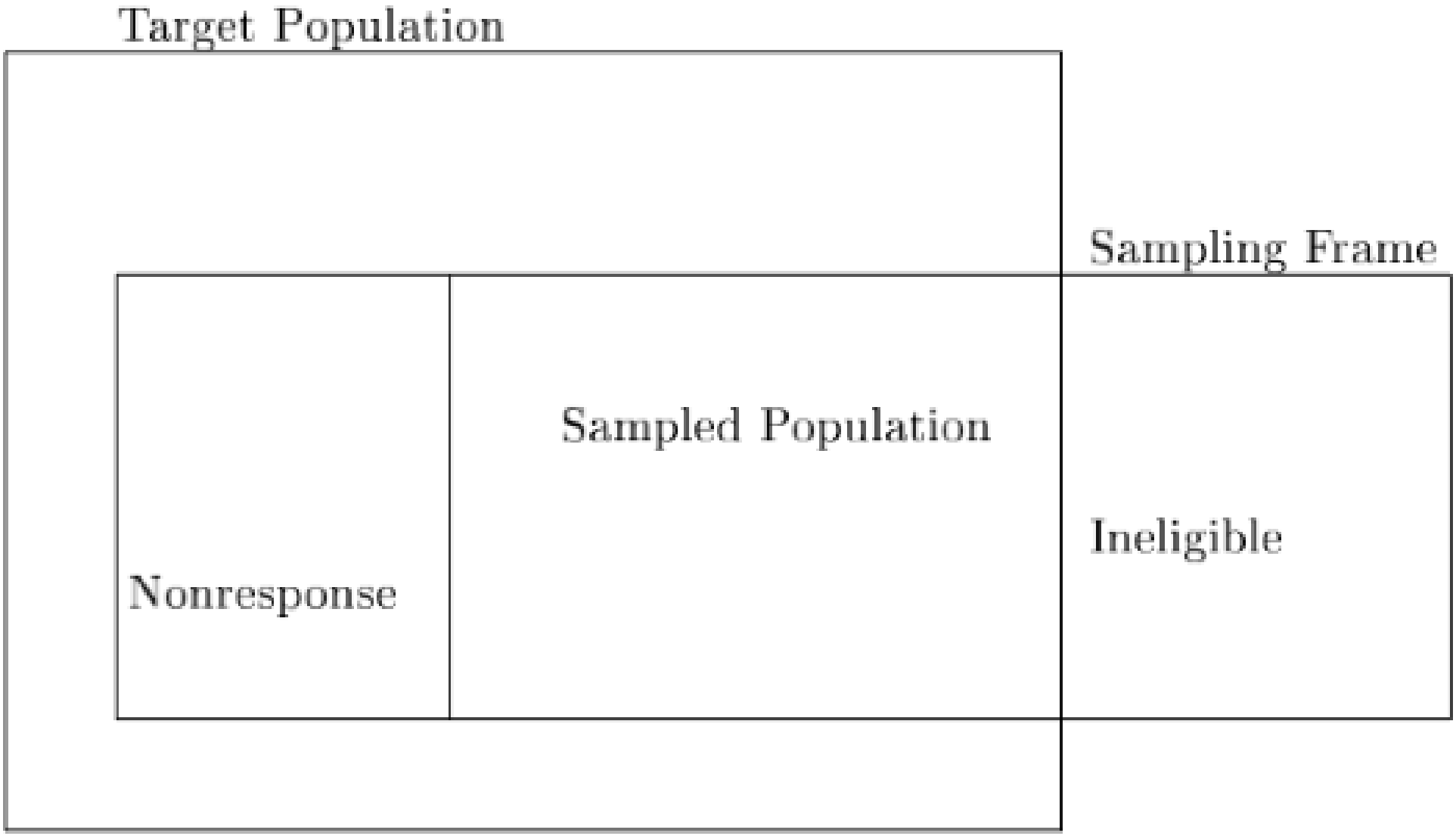
Simple random sampling

- Sampling from a list
- Missing and duplicate items

Simple random sampling in R

- Simulation
- Analysis using the "survey" package

The population and the sample



Sampling from the phone book

	Page	Column	Entry	Address #	Telephone #
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

(A bit of) the population

KASSOMBOLA—KATZ 509

KATOPIS Theodore 120 E 82 ..... 212 249-3047
KATOVITZ Michael 299 W 12 ..... 212 929-9511
KATOWSKY Marc 215 E 95 ..... 212 706-2855
KATRAGADDA Sireesha
31 E 31 ..... 212 532-6457
KATRANCI Elif 155 E 99 ..... 212 722-1951
KATRI Edmond 160 E 48 ..... 212 588-0118
KATRITSIS A 212 741-0174
KATROV Marat P 747 10 Av 212 757-4845
KATS Amir 531 W 48 ..... 212 333-5811
Ester 15 Willett 212 477-2490
Guyora 230 W 82 ..... 212 362-5351
I 212 588-1244
Inna 1277 3 Av 212 288-7739
Michael 345 E 93 ..... 212 987-2902
Victor 75 West St 212 385-1686
KATSAMAKIS Basil 315 E 69 ..... 212 628-9512
Basil 530 E 72 ..... 212 628-0312
KATSANOS Andrew 321 E 71 ..... 212 717-9393
Christina 417 W 47 ..... 212 459-2304
KATTULA Jennafer 409 E 69 ..... 212 327-2845
KATUN Mosammat 316 W 95 ..... 212 666-4817
KATUS B 210 W 89 ..... 212 362-9715
KATUSAK F J 176 E 77 ..... 212 737-8955
KATVAN Moshe 40 W 17 ..... 212 627-2169
Moshe 40 W 17 ..... 212 627-4362
Moshe 40 W 17 ..... 212 627-5035
Moshe & Rivka 117 W 17 ..... 212 627-5034
KATWAROO Dianna 434 W 163 ..... 212 568-0636
Errol 434 W 163 ..... 212 568-3629
KATYAL Monica 617 W 115 ..... 212 222-3669
KATYANG Keo 104 W 96 ..... 212 749-8386
KATZ A 212 721-3504
A 212 725-6758
A 268 E Bway 212 982-8619
A 737 Park Av 212 517-8897
A 2 5 Av 212 533-9692
A 148 10 Av 212 366-6487
A 315 E 86 ..... 212 831-7554
A D 433 W 21 ..... 212 255-1769

- How can you use random numbers to take a random sample of telephone households?
- 5 columns per page, 126 lines per column

A sample

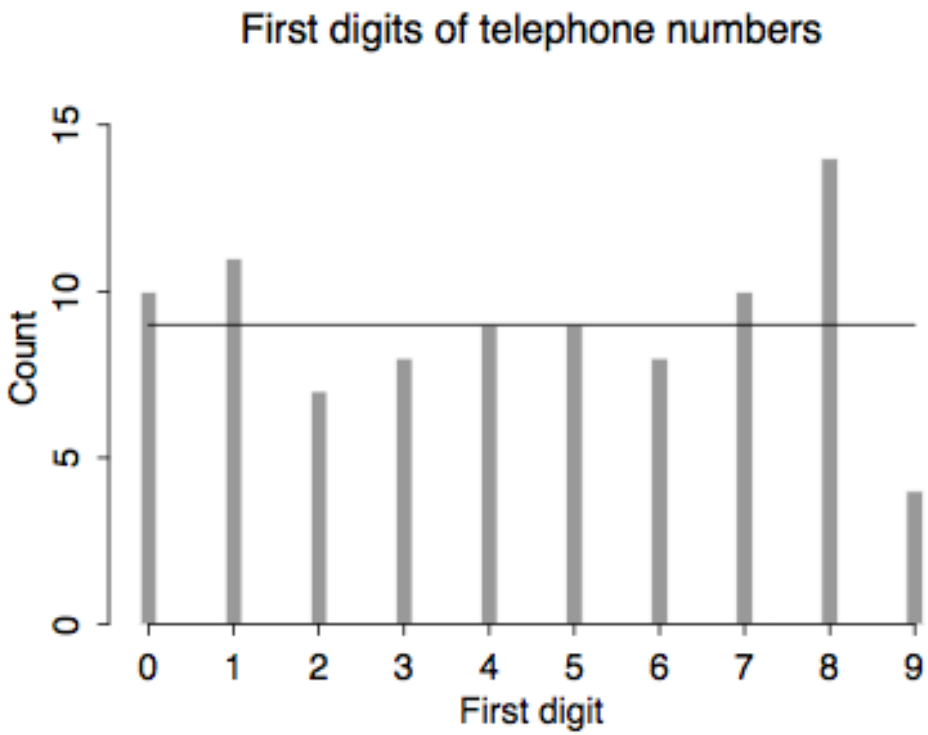
	Page	Column	Entry	Address #	Telephone #
1	520	5	100	15 W 53 St	586-7149
2	519	2	116	240 W 116 St	663-1076
3	519	4	087	710 West End Ave	749-2245
4	520	2	081	511 E 20 St	533-0614
5	519	4	115	2 Horatio St	206-7914
6	519	3	124	256 ...	304-2769
7	519	2	110	350 ...	308-4620
8	520	1	107	129 ...	xxx-2xxx
9	520	5	126	315 ...	xxx-2xxx
10	520	2	040	104 ...	xxx-1xxx

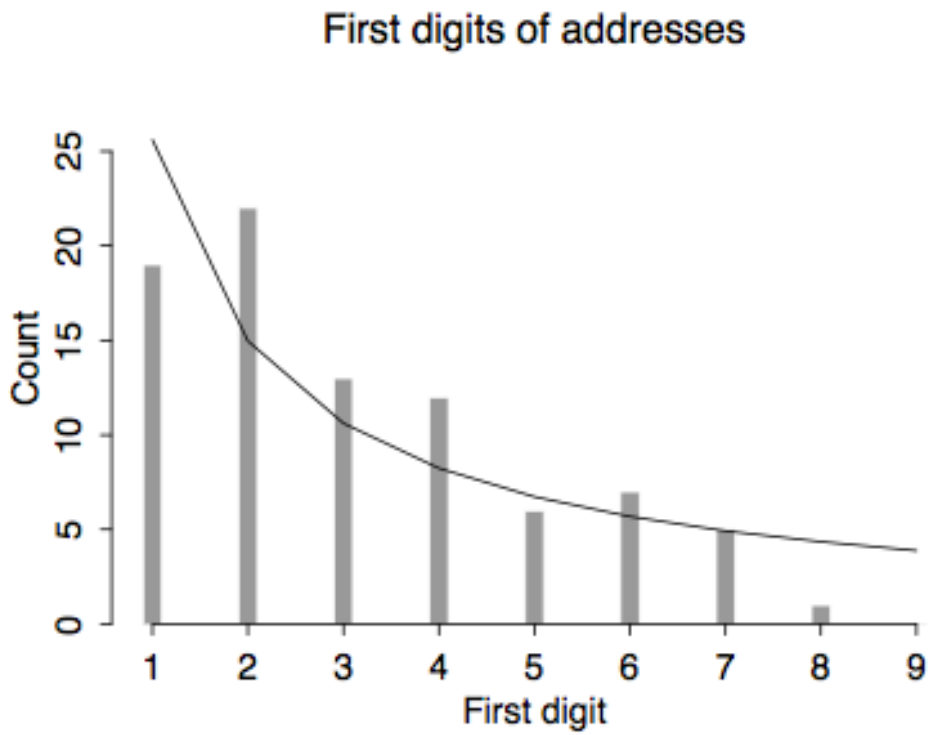
- This is not an equal-probability sample
- What did they do wrong?

First digits of addresses and phone number suffixes

First digits of phone number suffixes	First digits of addresses
0	
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9

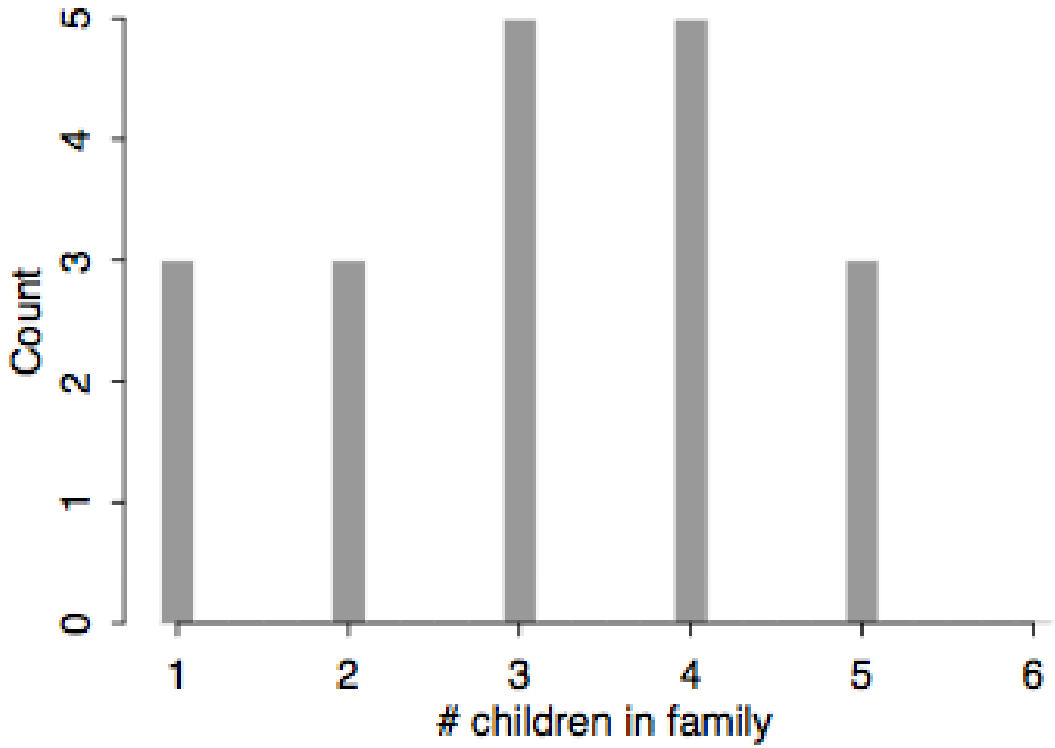
First digits of addresses and phone number suffixes





How many siblings are in your family?

Family size (# of siblings, including self)	Count
1	3
2	3
3	5
4	5
5	3
6 or more	0



Simple random sampling

- Define SRS
- Sampling with and without replacement
- Give examples of equal-probability sampling that are not SRS
- The sampling frame

Problems with the sampling frame

- Items not on the list
- Blanks in the list
- Duplicates
- Clusters
- Potential solutions
  - Ignore the problems
  - Redefine the population to fit the frame
  - Separate stratum for elements not in the frame
  - Treat clusters as individual units
  - Rejecting blanks that are sampled

Systematic sampling

- Sample every 10th unit in the list
- Actually a form of cluster sampling
- Will discuss analysis later

Stratified sampling in R

- Simulation
- Analysis using the "survey" package

Stratified sampling: population

- Strata $h = 1, \ldots, H$, with $N_1, \ldots, N_H$ units in each stratum
- Population size $N = \sum_{h=1}^{H} N_h$
- Within each stratum h: population stratum mean $\overline{Y}_h$ and population stratum variance S_h^2
- Parameter (population quantity) of interest: $\overline{Y}_W = \sum_{h=1}^{H} W_h \overline{Y}_h$
- Sum of weights $\sum_{i=1}^{H} W_h = 1$
- Usually, $W_h = N_h/N$. If so, then $\overline{Y}_W = \overline{Y}$, the population mean

Stratified sampling: data

- Independent sampling from each stratum
- Sample sizes $n_1, \ldots, n_H$, total sample size $n = \sum_{h=1}^{H} n_h$
- Within each stratum h: sample stratum mean $\bar{y}_h$ and sample stratum variance s_h^2
- Within each stratum, any kind of sampling might be done; for now, assume simple random sampling

Stratified sampling: concepts

- Interpretation as a regression estimate
- Connection to survey weights
- Why stratify?
  - Practicality/cost
  - Bias reduction
  - Variance reduction

Stratified sampling: weights

- Relation between W_h and N_h
- Usually $W_h = N_h/N$.
- When does $W_h \neq N_h/N$?
  - You are interested in a larger superpopulation (for example, age adjustment of death rates)
  - N_h's are unknown. Then you must create weights based on estimates

Stratified sampling: design and analysis

- Standard error and design effect
- Tasks of stratified sampling
  - Setting up the strata
  - Allocate the sample size across strata
  - Perform the sampling
  - Do the analysis
- Poststratification