Introduction to Data Science

Live Session 10 - Unit 10

Samples, Populations, and Explorations

Future Plan

- Mar 28 – Unit 10
- Apr 04 – No class (watch unit 11)
- Apr 11 – Unit 11 & Unit 12 Videos (Python)
- April 18 – Python Presentations
- April 25 – Case Study II

Case study II

- Apr 18 (starting date)
- Apr 25 (submission date)
- Group assignment.
- Pick your partner! (Apr 11)

Last Live Session – (Apr 18)

- Find an "interesting" Python package. When you make your choice, post it to the wall. If someone posts the same library to the wall that you want to do – first come, first serve!
- Prepare no more than 10 PPT slides (6 minutes maximum) to describe your chosen Python package, much like we presented API in R.
- Slides should mention what the package does, how to install it, and give an example using the package.
- Post your slides to Live Session Unit 13 Assignment before class.

https://pypi.python.org/pypi

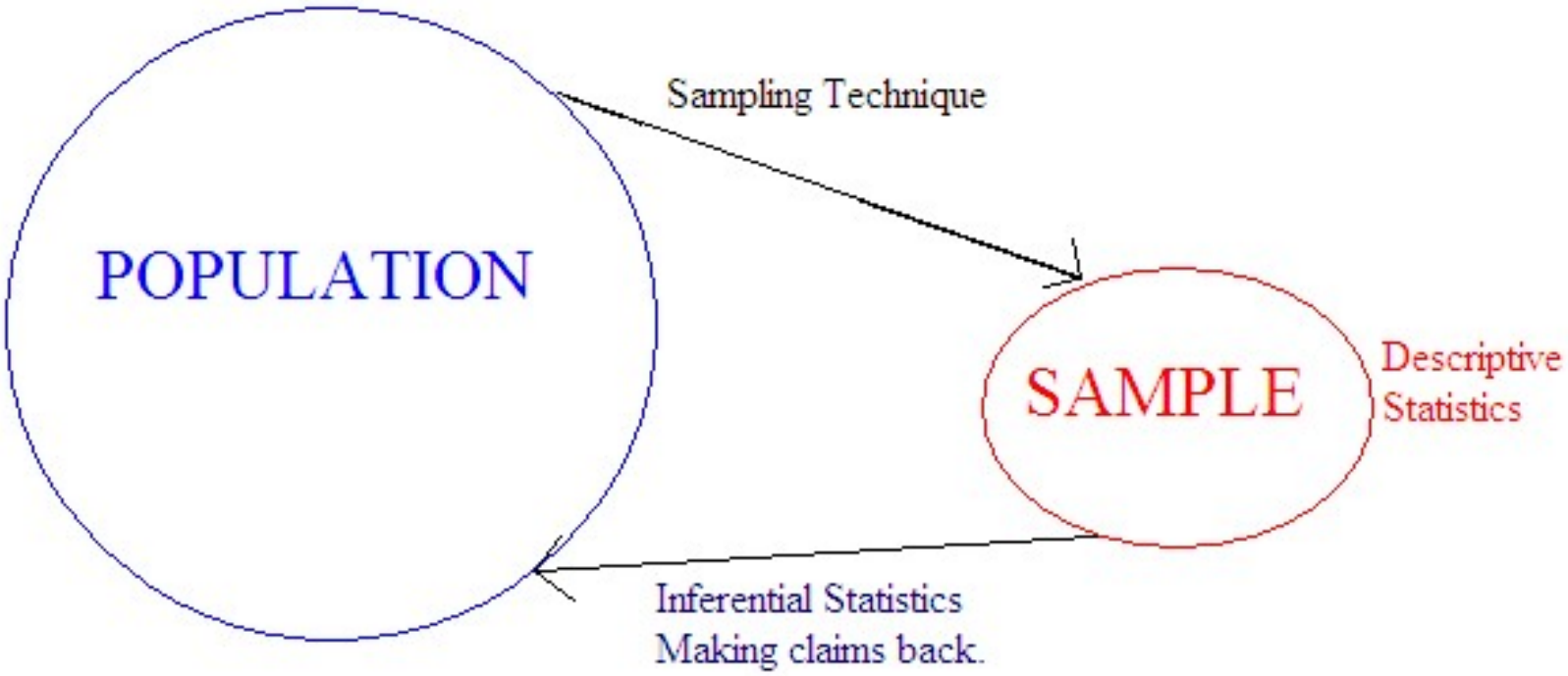
Office hours

- Raunak : Friday 6.30p.m.-7.30p.m. CT
- Chen Mo : Friday 8.30p.m. – 9.30p.m. CT

Sampling

- Sampling is an area of statistics that requires making a conceptual connection between
  - The research question
  - The data
  - The estimation methodology
- The conceptual connection is required for the analysis to be appropriate in that it achieves an answer to the research question

Parameters and Statistics



What are samples for?

- Definitions:
  - Population = all the units that you are interested in learning about
  - Sample = a subset of the population
  - Parameter = A numerical characteristic of a population (e.g., mean, total, proportion, difference in means between two parts)
  - Statistic = A numerical characteristic of a sample (e.g., sample mean, sample proportion, or other numerical summary)
- Goal of a sampling process is to estimate a parameter of a population from a sample statistic
- Thus the sample must be “representative” of the population

Health Insurance Rates

- The Health Insurance Marketplace Public Use Files contain data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace. Suppose we wanted to know how plan rates vary across states.
- What would be the population for this study? What would be the sample?

Population (left) & Sample (right)

- **All health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace.**

- **A subset of plans gathered from the marketplace**

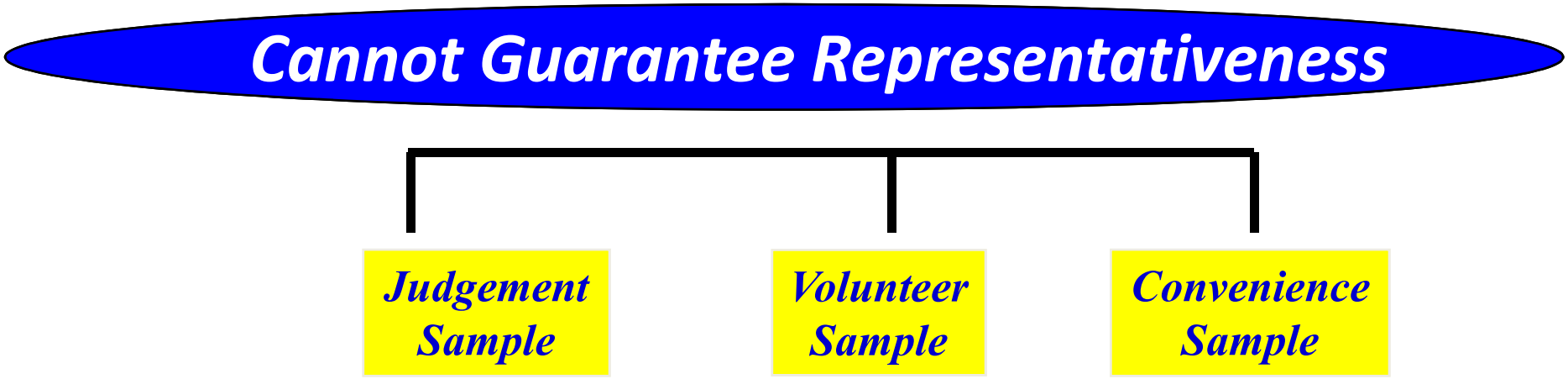
Sampling Error

- Sampling Error: The difference between the statistic and the parameter that is due to the fact that the estimate is made from only a subset of the population.

- The magnitude of the sampling error of an estimate can be assessed from the probability-based sample itself.

Non-Probability Samples

- Easy to Obtain, Usually Voluntary Responses
- Self-Selection is a Serious Problem
- Can Contain Useful Information

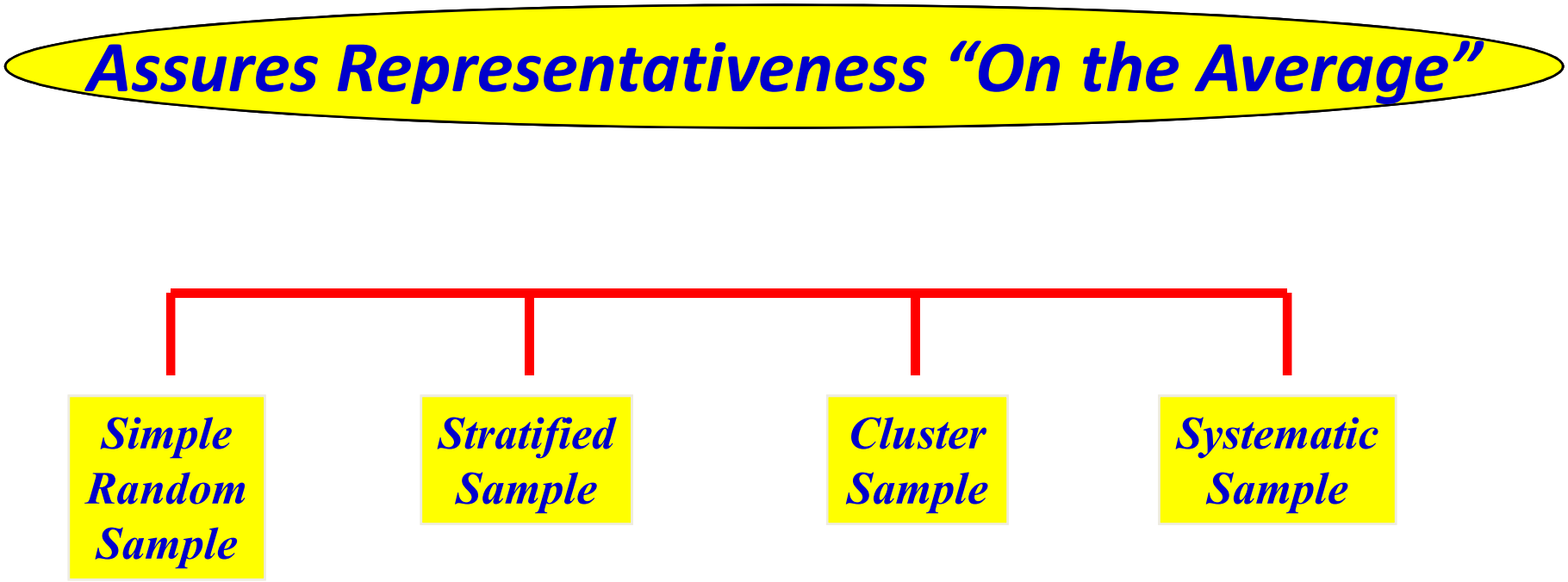


Probability Based Methods of Sampling

- A *probability sample* is one in which the probability of selection for every member of the sample is non-zero and known.
- Can Control for Known or Suspected Sources of Bias
  - Sampling Method
- **Randomization** Guards Against Unknown Sources of Bias
- Magnitudes of Possible Bias Can be Estimated, Final Results Adjusted
  - Census of Small Sub-populations, Historical Patterns
- Known Probabilities of Error Allow Uncertainty Estimates (Standard Errors)
  - Probability Distributions (e.g., Normal)

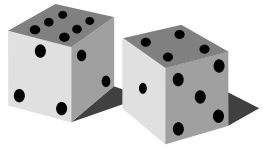
Probability Samples

Probability Samples



Simple Random Sample (SRS)

- An SRS of size n is taken when every possible subset of n units in the population has the same chance of being the sample.
- Selection may be with replacement or without replacement.
- One may use table of random numbers for obtaining samples.

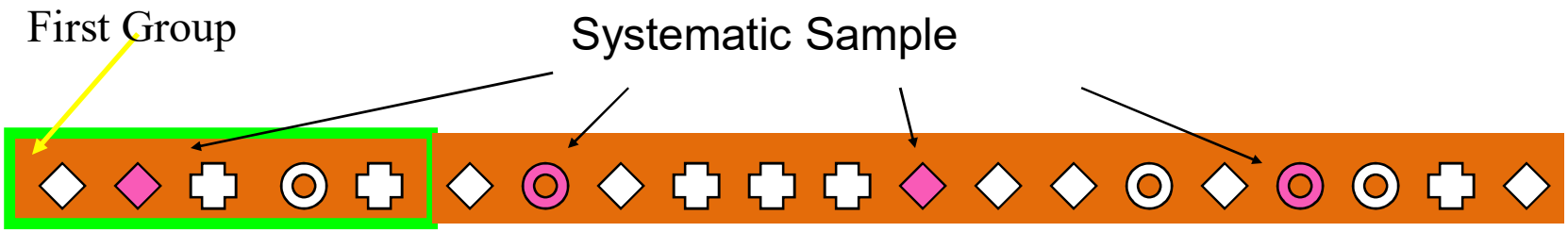


Can only be Guaranteed When a Sampling Frame is Available

Sampling Frame: List of Every Member or Item in a Population

Systematic Samples

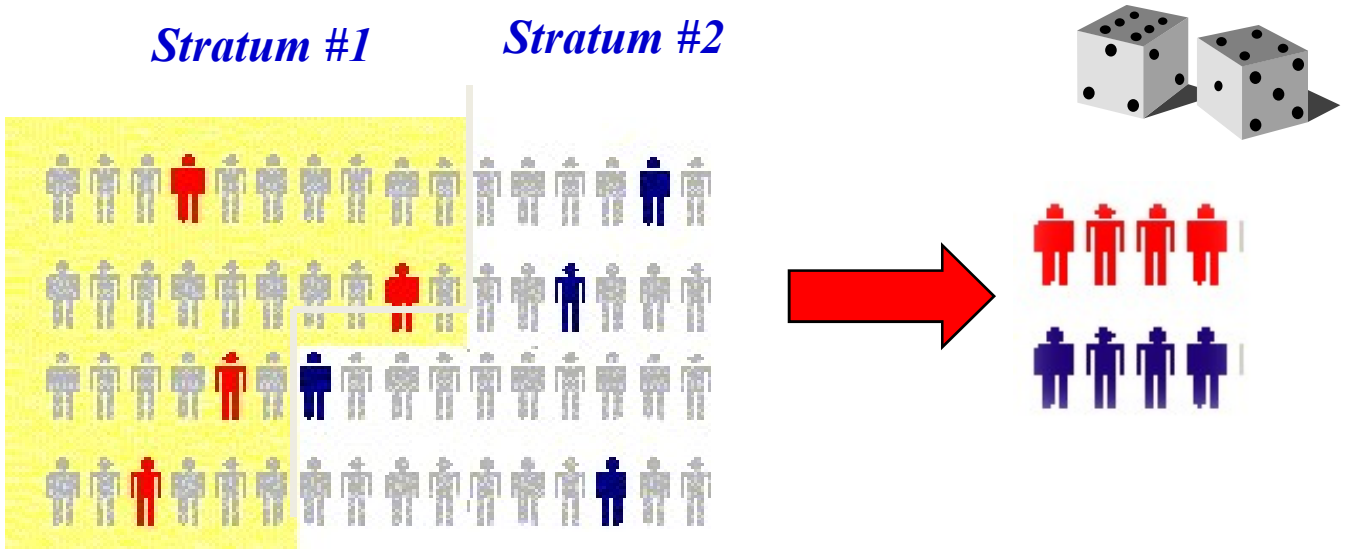
- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=N/n
- Randomly select one individual from the 1st group
- Select every k-th individual thereafter



e.g., N=300, n=55, k = 300/55 ≈ 5; Random Starting Position (1 – 5) = 2

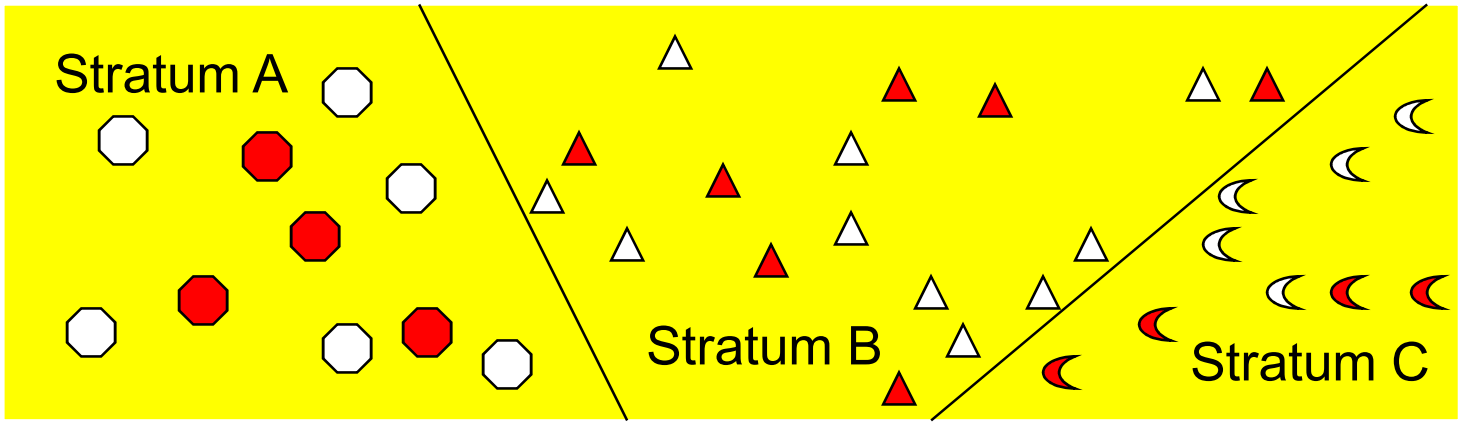
Stratified Samples

- Population divided into two or more groups according to some common characteristic
- Simple random sample selected from each group
- The two or more samples are combined into one



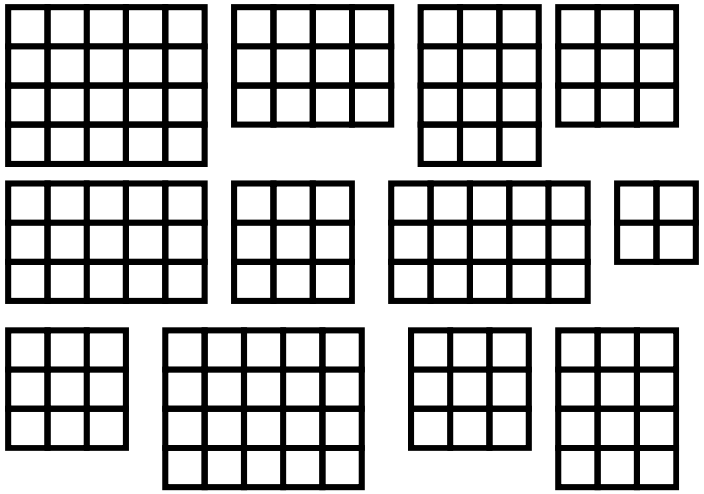
Stratified Sampling Details

- Units within stratum are similar
- Units in stratum A are different from units in stratum B and stratum C
- Use similarity within each stratum to obtain more precise information about population

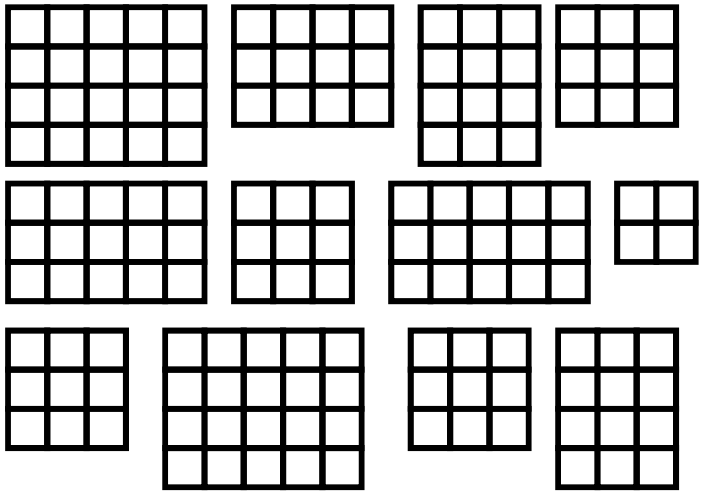


Note: Symbols Similar in Each Stratum
Different Colors Represent Different Responses

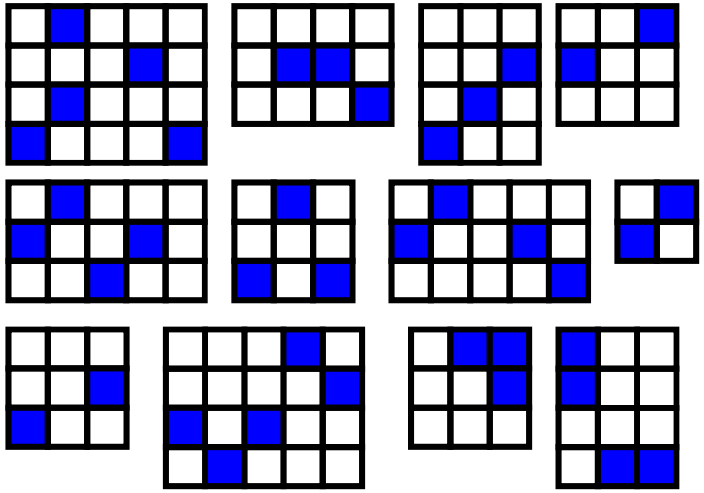
Cluster and Stratified Sampling



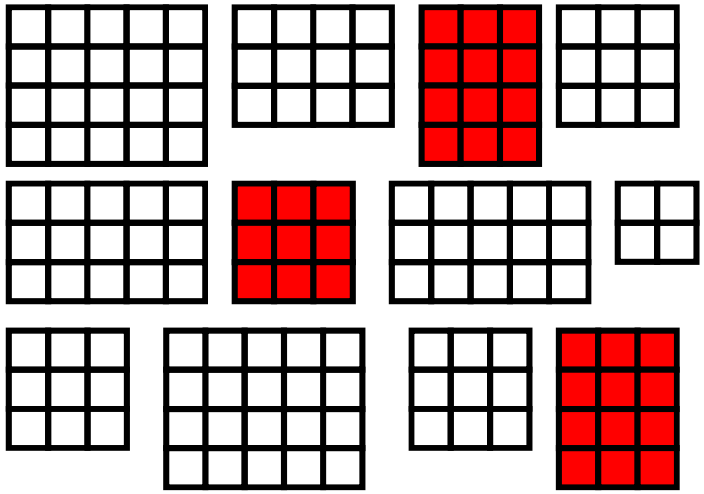
Population of H strata, stratum h contains n_h units



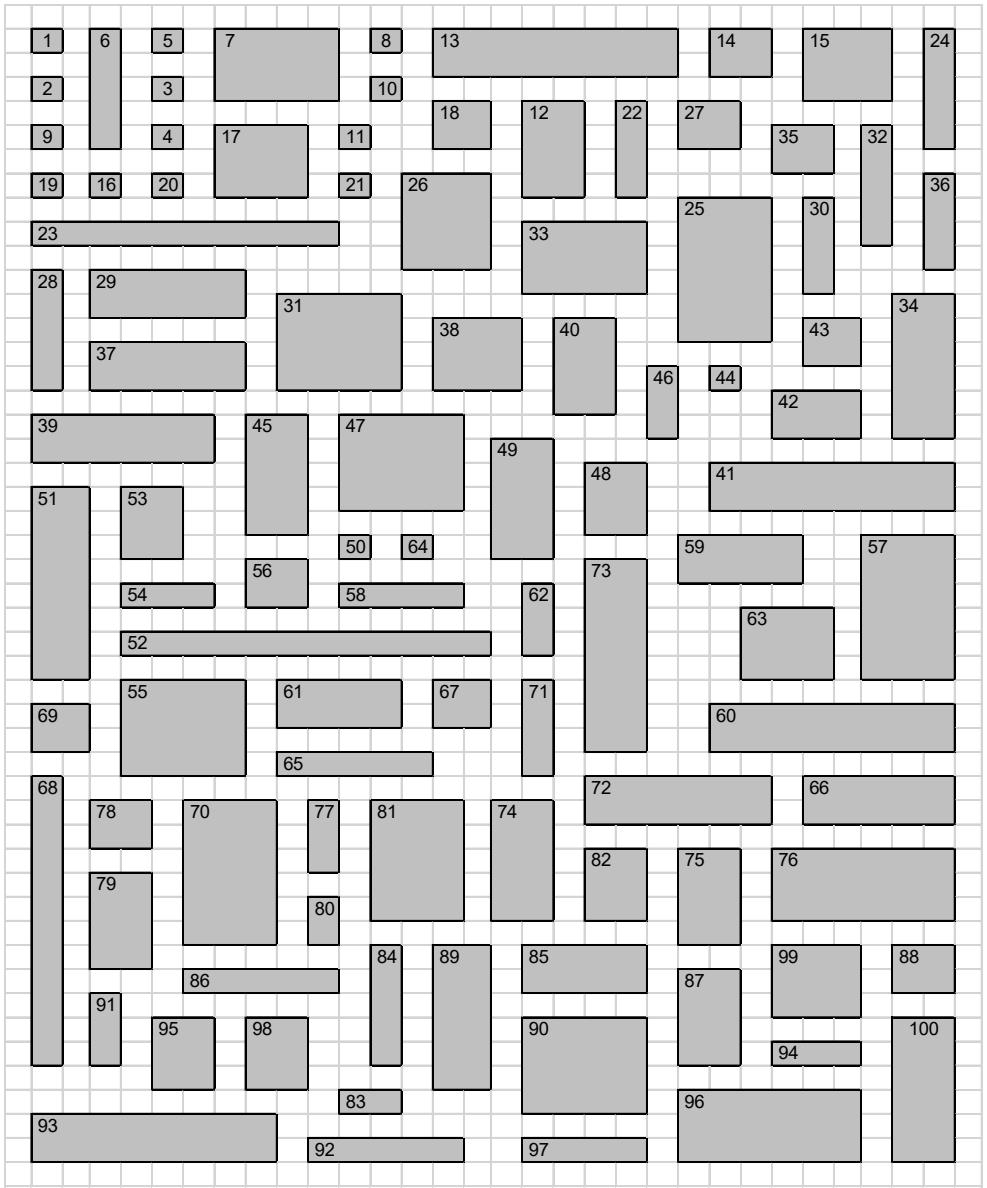
Population of M clusters

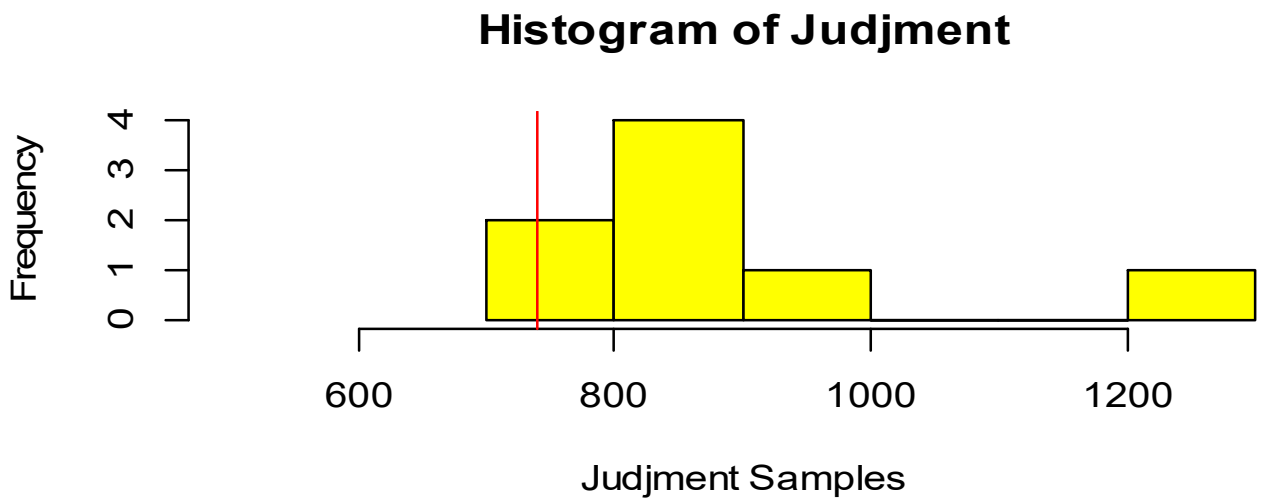


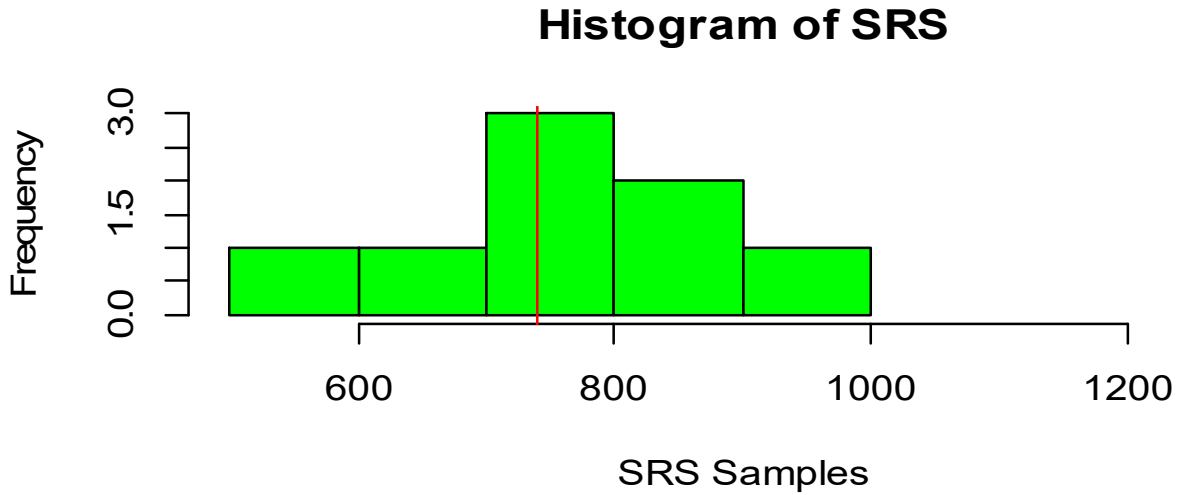
Take simple random sample in *every* stratum



Take srs of m clusters, sample every unit in chosen clusters


1
6
5
7
8
13
14
15
24
2
3
10
18
12
22
27
9
4
17
11
35
32
19
16
20
21
26
36
25
30
23
33
28
29
31
38
40
34
43
37
46
44
42
39
45
47
49
48
41
51
53
50
64
59
57
56
73
54
58
62
63
52
55
61
67
71
69
60
65
68
78
70
77
81
74
72
66
82
75
76
79
80
84
89
85
99
88
86
87
91
95
98
90
100
94
83
96
93
92
97


Histogram of Judjment
Frequency
0 1 2 3 4
600
800
1000
1200
Judjment Samples


Histogram of SRS
Frequency
0.0 1.5 3.0
600
800
1000
1200
SRS Samples

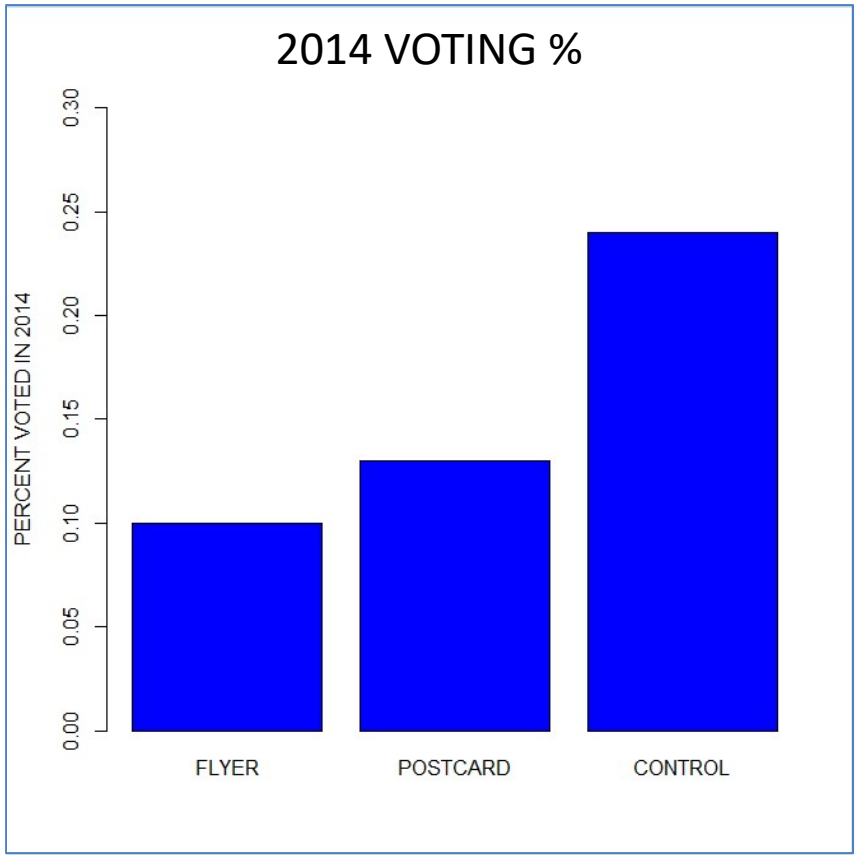
How do we assure representativeness?

- In the rectangle exercise, you tried to assure representativeness using human judgement.
  - Quota sampling
- People are not very good at this
- Random selection is a dependable method of assuring representativeness. Its advantage is that…
  - It has a high chance of getting a sample that is close to representative
  - We can compute the probability that it is will NOT be representative.
  - More specifically, we can use the mathematics of probability to compute the probability that the estimate is within a certain distance of the parameter.
- This is NOT true of a nonprobability sample

HW7:League of Women Voters Data

- Sample consists of 24K randomly selected individuals from a population of 531,735.
- The population was "low propensity voters"
- 8000 each assigned to receive
  - Postcard reminder to vote
  - Flyer with reminder and voting instructions
  - Nothing
- After the 2014 election, information regarding voter participation was collected for all voters.

Results from Study



- What happened?
- And yes, something really did go wrong.
- It has to do with sampling.
- Use plots and descriptive statistics to figure out the problem.

Percent of Each treatment group actually voting in 2014

Sampling Method/Group Assignment

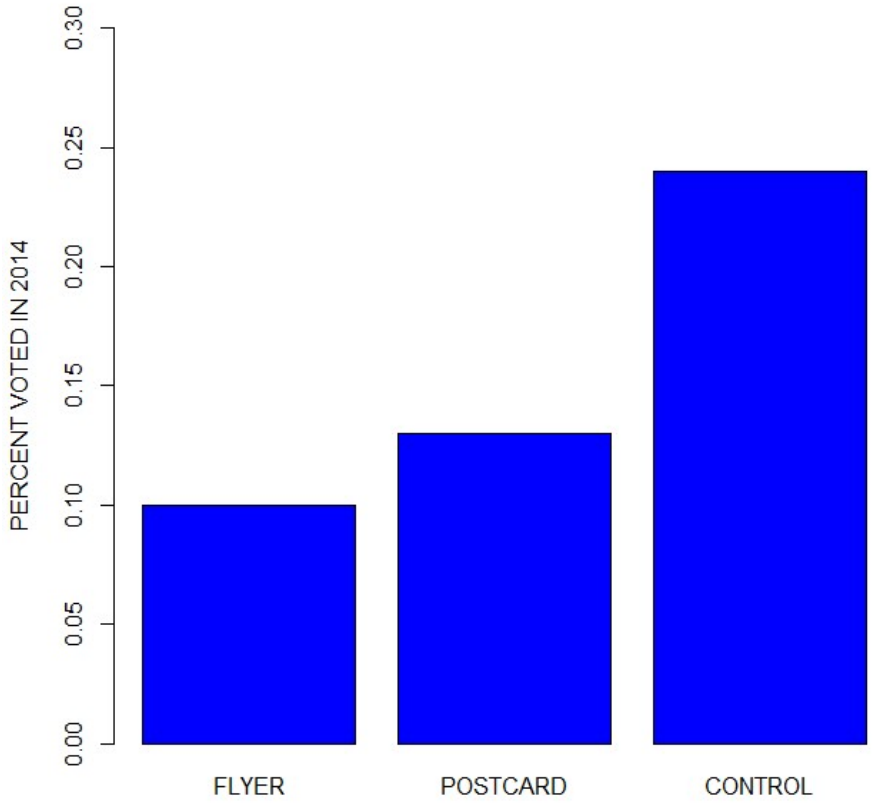
- Randomly selected 24,000 individuals from a population of 531,735 low-propensity voters ahead of the 2014 election.
  - "LOW PROPENSITY" = ONLY VOTED IN 0 OR 1 of THE LAST 3 ELECTIONS
  - 24,000 were partitioned into three treatment groups
    - 8,000 received a postcard.
    - 8,000 received a flyer.
    - 8,000 did not receive any LWV mailing (i.e., control).

DATA COLLECTION AND CLEANING

OBS	VOTER ID	DOB	GENDER	HISPANIC	VOTED	……
1	2013748	3/8/1975	M	0	1	……
2	2375934	9/2/1956	F	0	0	……
3	2386483	1/3/1993	F	1	0	……
4	2475233	4/9/1987	F	1	0	……
	…	…	…	…	…	…
23,998	6121293	12/3/1938	M	0	1	……
23,999	6385385	5/3/1991	F	1	0	……
24,000	6836571	2/9/1965	M	1	0	……

Information regarding voter participation was collected for all voters in the 2014 election.

ANALYZE THE DATA!!!



Ahhh!!!

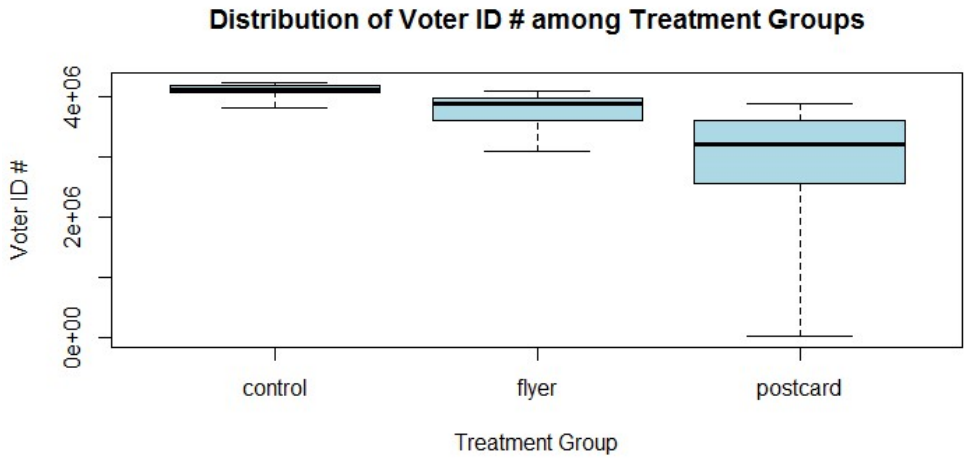
	OBS	VOTER ID	DOB	GENDER	HISPANIC	VOTED	……
POSTCARD	1	2013748	3/8/1975	M	0	1	……
	2	2375934	9/2/1956	F	0	0	……
	3	2386483	1/3/1993	F	1	0	……
	4	2475233	4/9/1987	F	1	0	……
FLYER		…	…	…	…	…	…
CONTROL	23,998	6121293	12/3/1938	M	0	1	……
	23,999	6385385	5/3/1991	F	1	0	……
	24,000	6836571	2/9/1965	M	1	0	……

Information regarding voter participation was collected for all voters in the 2014 election.

Concerns with Group Assignment

- Sample of 24,000 individuals were not randomly assigned to postcard, flyer, control groups with respect to voter ID #.

- Individuals with large voter ID #s registered to vote more recently.
- Three treatment groups are very different in their distributions of voter ID numbers.
- May influence the propensity to vote in 2014.



Discussion

- Why is exploratory data analysis (EDA) necessary?
- Give an example in a real setting where EDA was important in uncovering issues with the data.

EDA Exercise (10.8)

- There are 31 data sets named: nyt1.csv, nyt2.csv, …, nyt31.csv.
- Each set represents one simulated day's worth of ads shown and clicks recorded on *New York Times* home page in May 2012.
- Each row represents a user.
- Five columns represent: age, gender (0 = female and 1 = male), number impressions, number of clicks, and whether user logged in.

Live Session Unit 10 Assignment

Module 10.8

- (http://stat.columbia.edu/~rachel/datasets/nyt1.csv)
- Create a new variable *ageGroup* that categorizes age into following groups:
  < 18, 18–24, 25–34, 35–44, 45–54, 55–64 and 65+.
- Use sub set of data called “ImpSub” where Impressions > 0) in your data set.
- Create a new variable called click-through-rate (CTR = click/impression).
- Use this ImpSub data set to do further analysis.

Analysis of Click Stream Data

- For a single day:

  (Use sub set of data “ImpSub” where Impressions > 0)

  - Plot distributions of number impressions and click-through-rate (CTR = click/impression) for the age groups.
  - Define a new variable to segment users based on click -through-rate (CTR) behavior.

    CTR< 0.2, 0.2<=CTR <0.4, 0.4<= CTR<0.6, 0.6<=CTR<0.8, CTR>0.8
  - Get the total number of Male, Impressions, Clicks and Signed_In

    (0=Female, 1=Male)
  - Get the mean of Age, Impressions, Clicks, CTR and percentage of males and signed_In
  - Get the means of Impressions, Clicks, CTR and percentage of males and signed_In by AgeGroup.

Analysis of Click Stream Data

- For a single day:

  (Use sub set of data “ImpSub” where Impressions > 0)

  - Create a table of CTRGroup vs AgeGroup counts.
  - Plot distributions of number impressions and click-through-rate (CTR = click/impression) for the age groups
  - One more plot you think which is important to look at.
  - Submit your file in to Live session Unit 10 Assignment