# Samples, Populations, and Explorations

Unit 8 Slides

# 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

# 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**

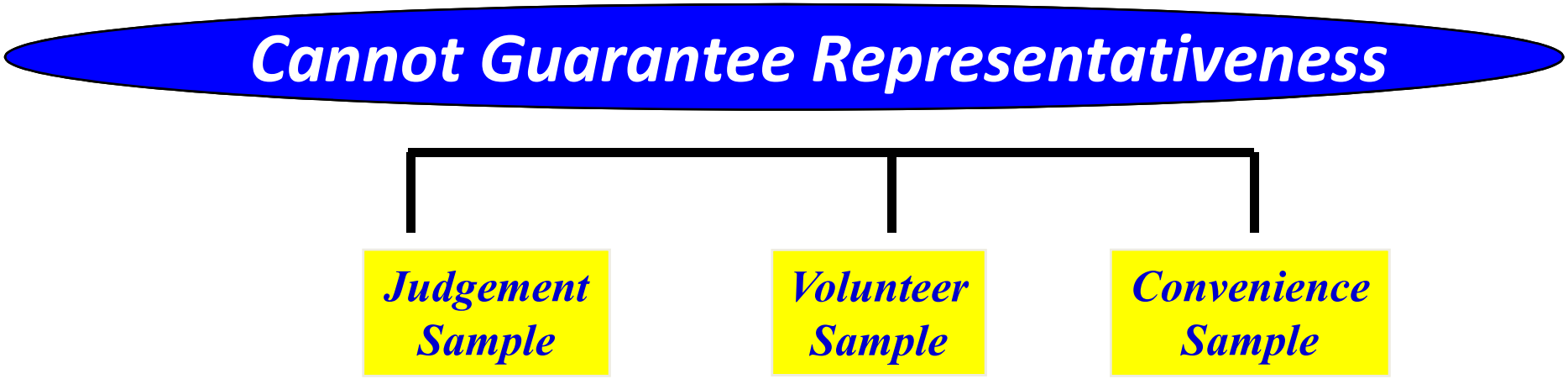
# Sampling Error

- 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

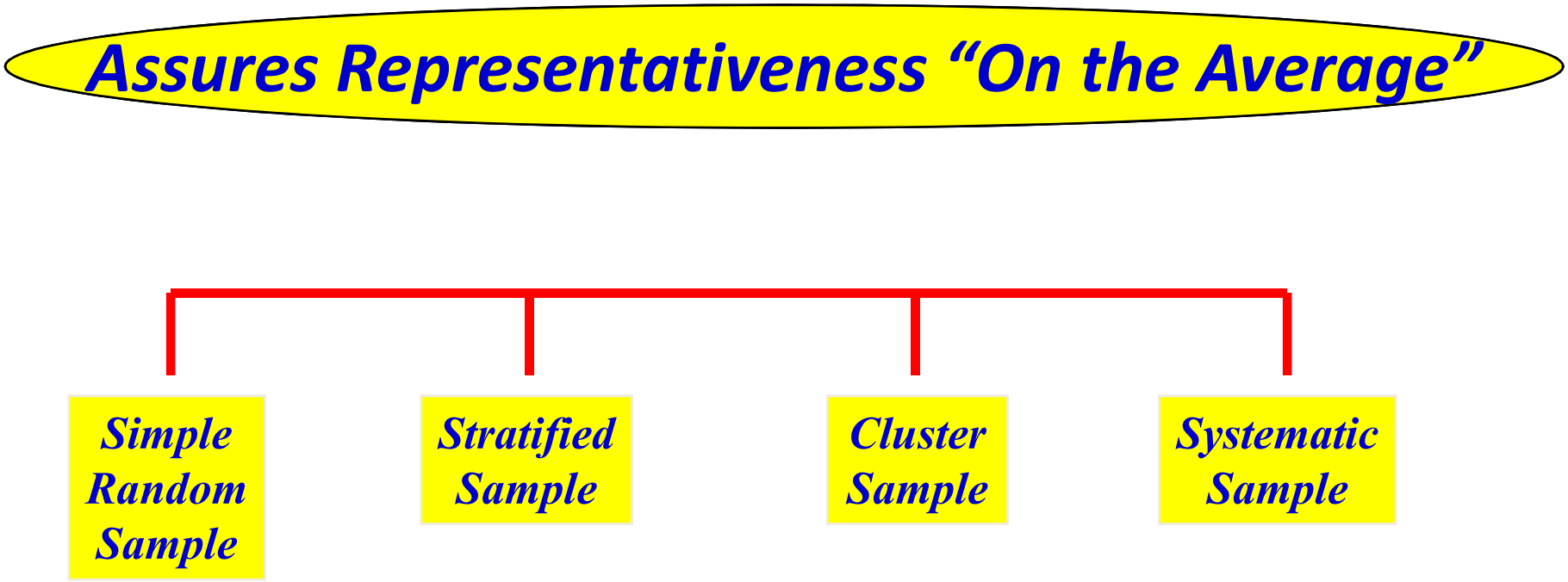


# Probability Based Methods of Sampling

- 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)

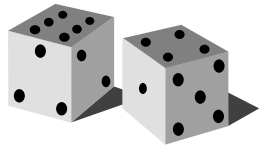
# Probability Samples

Probability Samples



# Simple Random Samples (SRS)

- Every individual or item from the sampling frame has an equal chance of being selected.
- Selection may be with replacement or without replacement.
- One may use table of random numbers for obtaining samples.

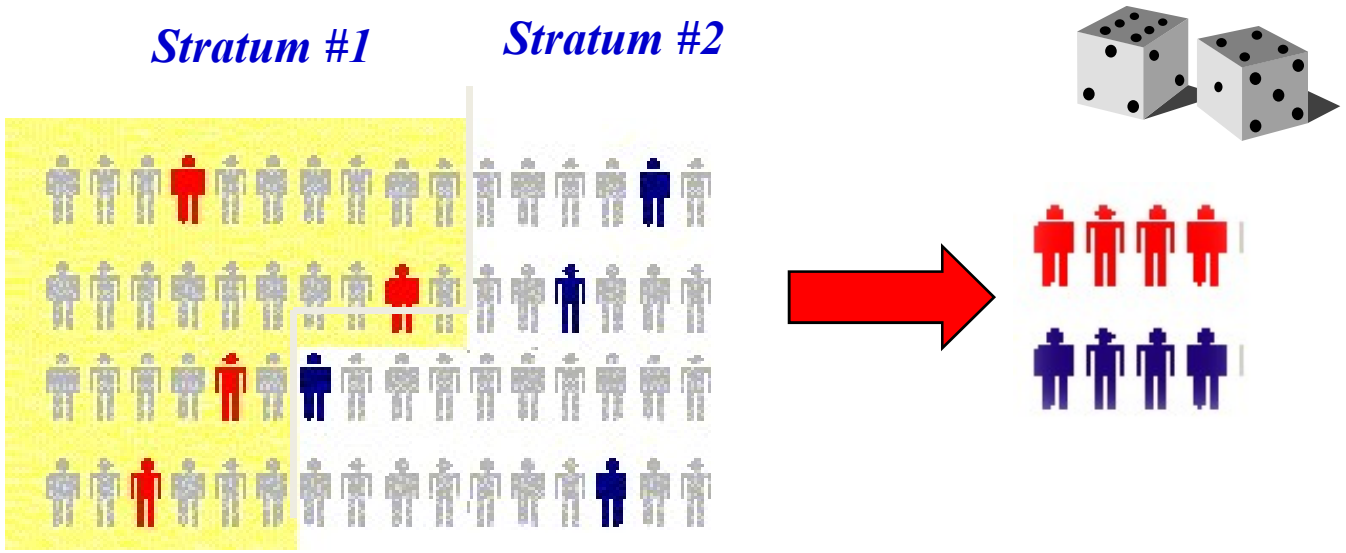


***Can only be Guaranteed When a Sampling Frame is Available***

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

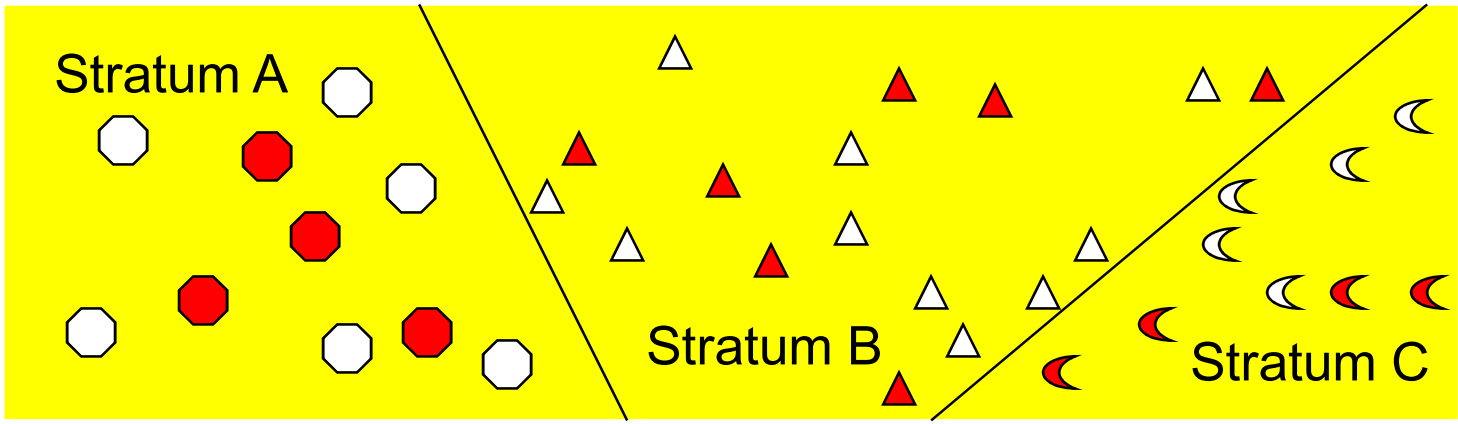
# Stratified Samples

- 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



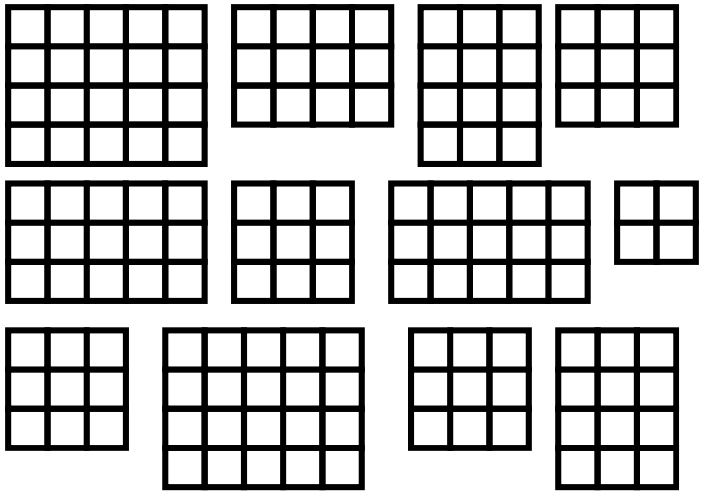
# Stratified Sampling Details

- 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

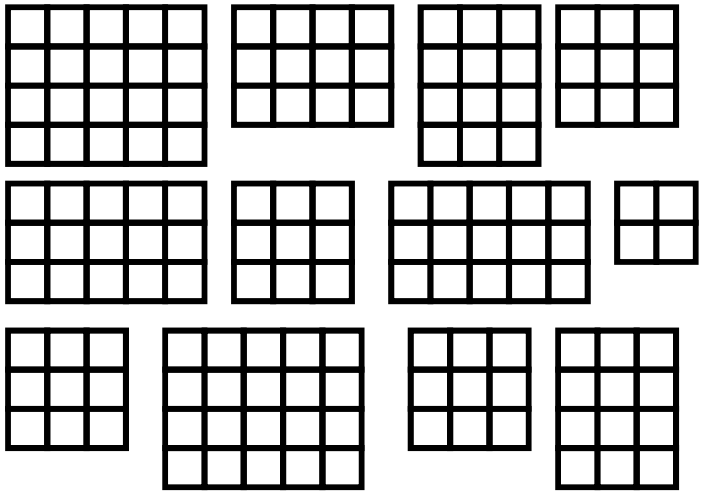



***Note: Symbols Similar in Each Stratum***
***Different Colors Represent Different Responses***

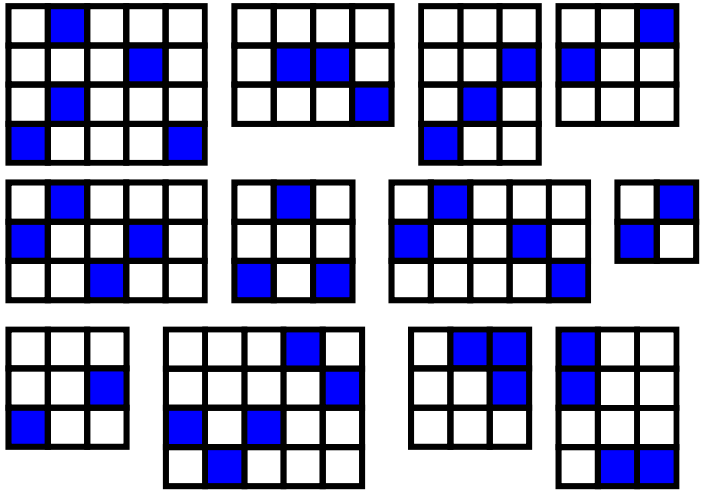
# Cluster and Stratified Sampling



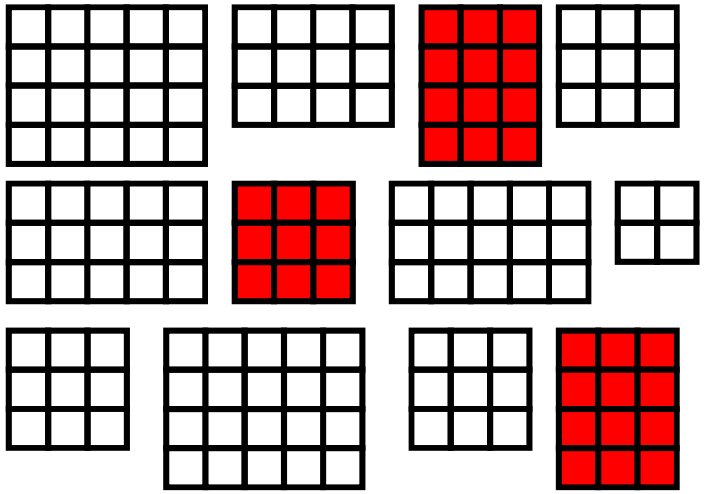
Population of $H$ strata, stratum $h$ contains $n_h$ units



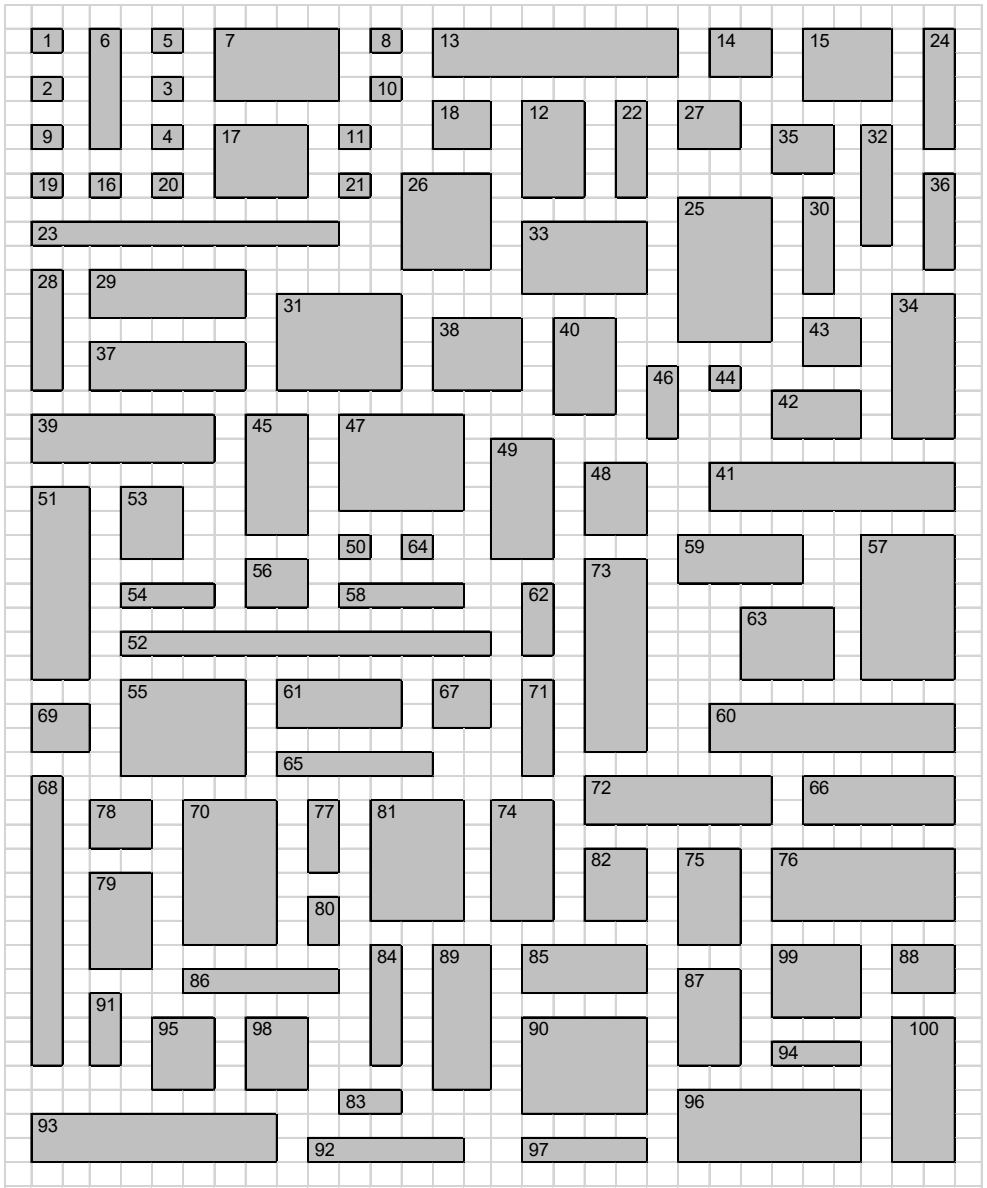
Population of $M$ clusters

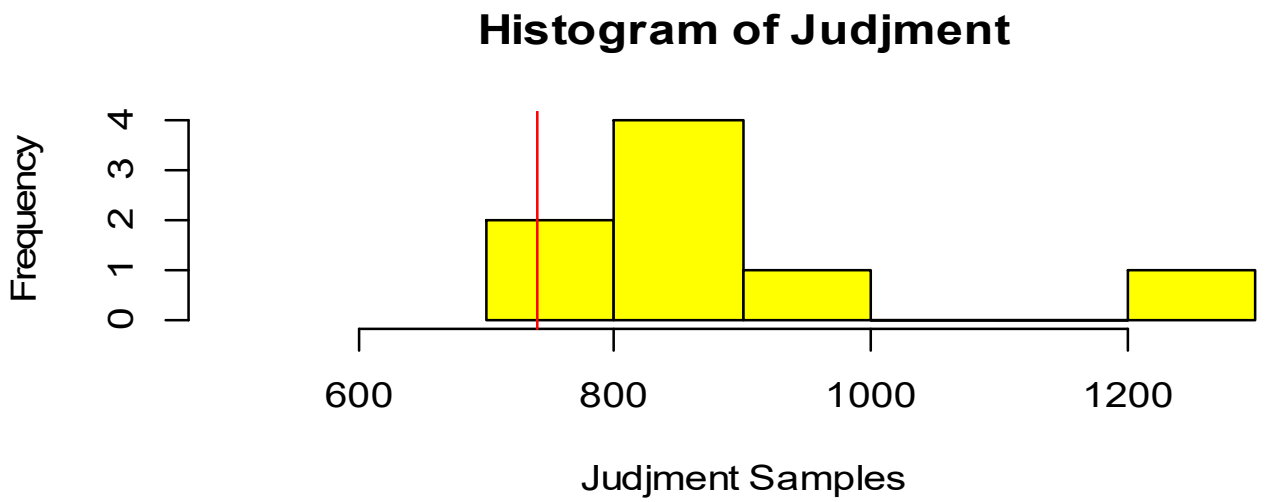


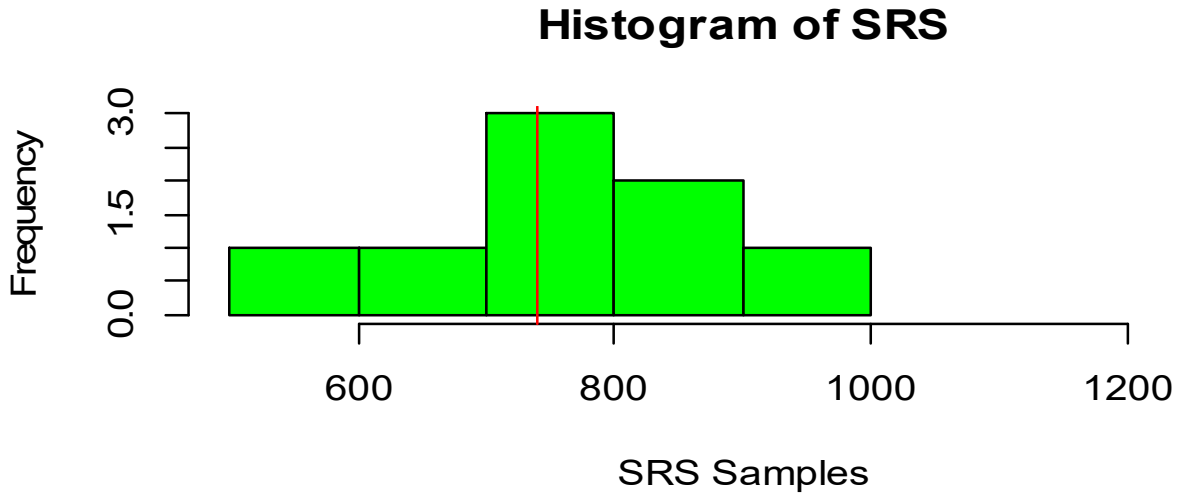
Take simple random sample in *every* stratum



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

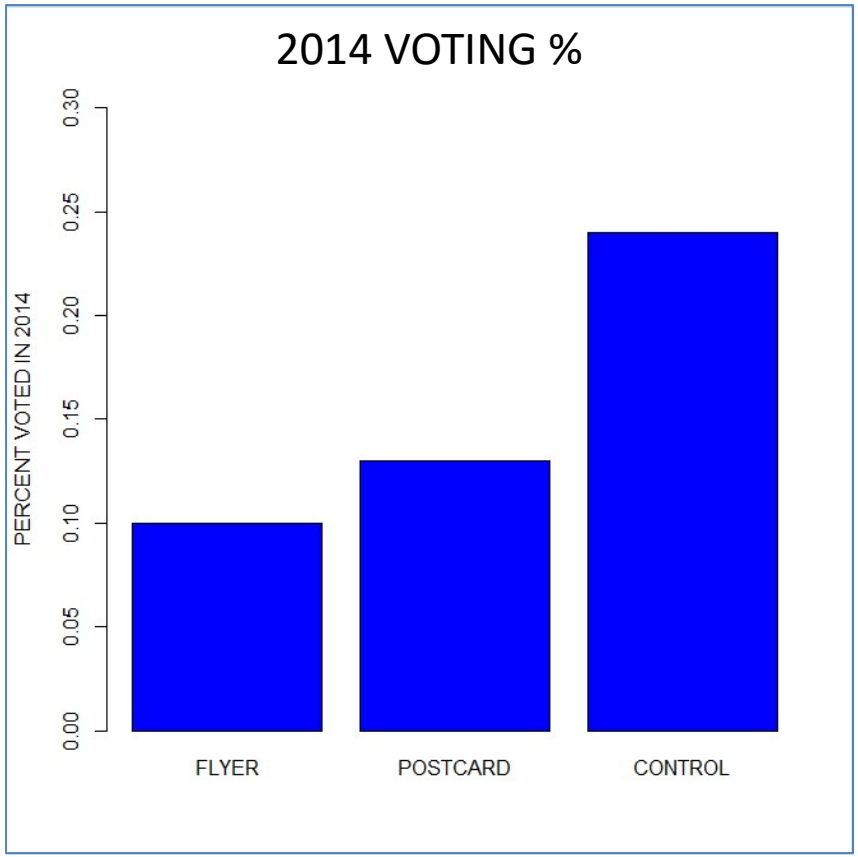
# How do we assure representativeness?

- 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

# HW8: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

# Unit 9 Live Session – July 13

- No live session on July 6

- Homework due before live session
  - Homework originally assigned for Unit 8 on League of Women Voters data
  - You may also work in groups, as long as there are no more than three people in a group.
  - You have to submit your own answer!

- live session 9
  - Watch week 9 videos!

# Upcoming Classes

- July 6 (No live session - Live session Week 8 assignment)
- July 13 (HW8)
- July 20 (Unit 10 – have case study rough draft ready)
- Due date for Unit 10 Case Study is July 27

# 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.
- Why is it important to understand that data values are realizations of random variables?

# Data Prep for Live Session 8

Module 8.9

- (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 8 Assignment