Statistical Methods and Data Analysis I
Lecture 11: Sampling Techniques

Oleg Goldshmidt
oleg.goldshmidt@post.idc.ac.il

Arison School of Business
Interdisciplinary Center (IDC)
Herzliya, Israel

April 28, 2019

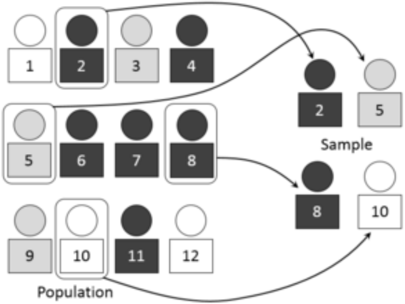
Simple Random Sampling

Algorithms:

Draw-by-draw: select element with equal probability, remove, add to sample. Requires random access!

Random sort: pass once, generate a random number in $(0, 1)$ as key, sort by the key, choose the first k items.

Reservoir sampling: keep the first k items. When item i ($i > k$) arrives choose one item from the sample with probability $1/k$, substitute the new one with probability k/i.
One can prove that out of $n \gg k$ items each will be included with probability k/n.



Exercise: Simple Random Sample

Scenario

You have to draw a random sample of 5 students out of 80.

Will a sequence of random digits 0516629305774827 *help?*

Exercise: Simple Random Sample

Scenario

You have to draw a random sample of 5 students out of 80.

Will a sequence of random digits 0516629305774827 *help?*

1. Number the students from 00 to 79.

Exercise: Simple Random Sample

Scenario

You have to draw a random sample of 5 students out of 80.
Will a sequence of random digits 0516629305774827 *help?*

1. Number the students from 00 to 79.
2. Take 2 digits at a time from the random sequence:

$$05 \quad 16 \quad 62 \quad 93 \quad 05 \quad 77 \quad 48 \quad 27$$

Exercise: Simple Random Sample

Scenario

You have to draw a random sample of 5 students out of 80.
Will a sequence of random digits 0516629305774827 *help?*

1. Number the students from 00 to 79.
2. Take 2 digits at a time from the random sequence:

   05 16 62 93 05 77 48 27

3. Ignore 93 as too large and 05 as a repetition:

   05 16 62 93 05 77 48 27

Exercise: Simple Random Sample

Scenario

You have to draw a random sample of 5 students out of 80.

Will a sequence of random digits 0516629305774827 *help?*

1. Number the students from 00 to 79.
2. Take 2 digits at a time from the random sequence:

   05 16 62 93 05 77 48 27

3. Ignore 93 as too large and 05 as a repetition:

   05 16 62 93 05 77 48 27

4. Choose students numbered as

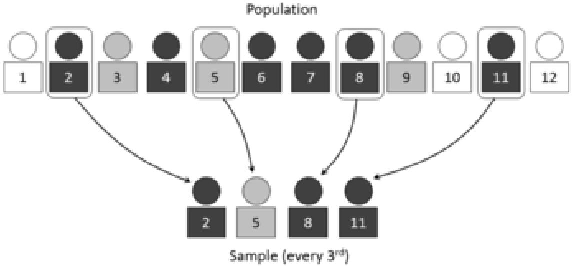
   05 16 62 77 48

   as your sample.

Systematic Random Sampling

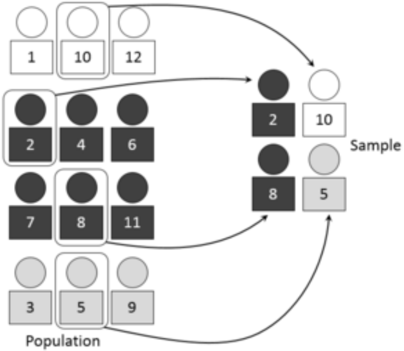
"Every 10th person from the telephone directory." "Every 10th customer." Start *at random*!
Sample a long street with low house numbers in a poor area and high numbers (up to 1000) in an affluent neighborhood.

- Random sample may be unrepresentative. Equally spaced sample may be better or worse.
- Always starting from either end of the street will be biased.
- Vulnerable to periodicity. What if odd numbers are in the North and even numbers in the South? If we know that we can choose an odd-sized step...



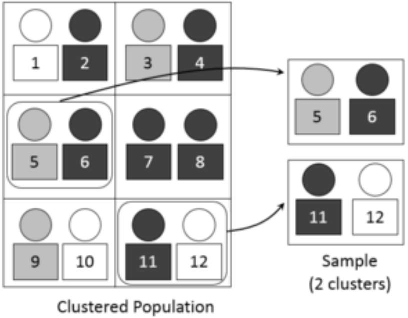
Stratified Sampling

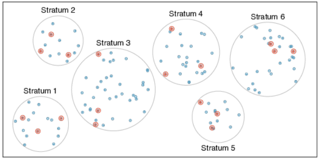
- Divide population into distinct categories (“strata”).
- Sample each stratum as an independent population.
- Identifying strata may be difficult and may require larger frame overall.
- Efficacy criteria:
  - low variability *within* the strata
  - high variability *between* the strata
  - stratification variables are strongly correlated with the response variable



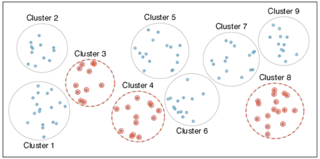
Cluster Sampling

- Divide the population into groups ("clusters"). Typical case: by geography. Pick n clusters *at random*. Sample the chosen clusters.
- Often reduces sampling costs. E.g., start with a block level map of the city. Choose n blocks at random. Sample the selected blocks.
- Multistage sampling.

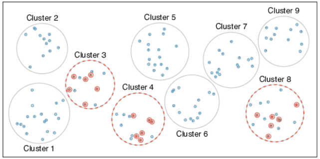




Stratified. *Strata* are made up of similar observations. We take a *simple random sample* from *each* stratum.



Cluster. *Clusters* are usually not made up of homogeneous observations. We take a *simple random sample* of clusters, and then sample *all* observations in that cluster. Usually preferred for economical reasons.



Multistage. *Clusters* are usually not made up of homogeneous observations. We take a *simple random sample* of clusters, and then take a *simple random sample* of observations from the sampled clusters.

Example: Tasting Boston Cream Pie

Does the following picture correspond to *cluster* or *stratified* sampling?

What would the *other* sampling technique correspond to?

Pop Quiz

Scenario

We need to draw a random sample of the 300 passengers on a flight from Tel Aviv to Hong Kong.
What sampling methods do the following procedures correspond to?

1. Pick every 10th passenger as people board the plane
2. From the passenger manifest, randomly choose 5 people flying business class and 25 from economy
3. Randomly generate 30 seat numbers and survey the passengers who sit there
4. Randomly select a seat position (right window, right center, right isle, etc. — there are 10 seats abreast in a wide-body aircraft such as Boeing 747 or Airbus A380) and survey those passengers

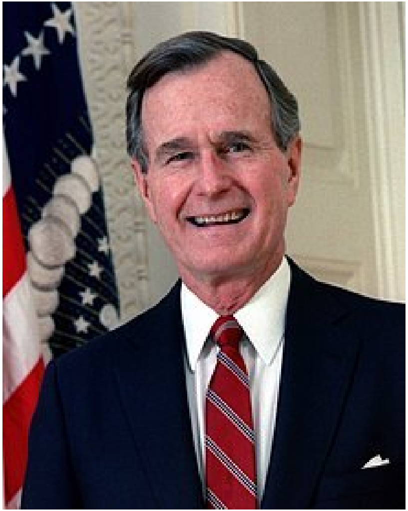
Pop Quiz

Scenario

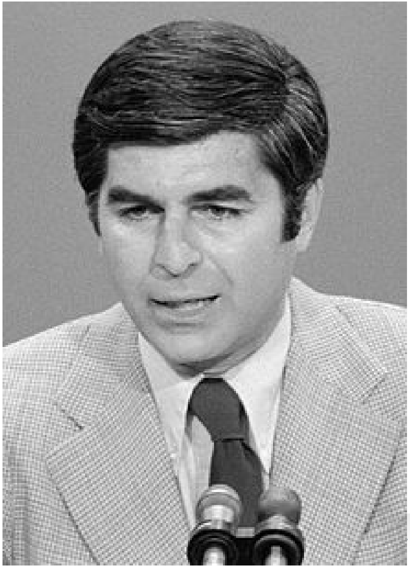
We need to draw a random sample of the 300 passengers on a flight from Tel Aviv to Hong Kong. *What sampling methods do the following procedures correspond to?*

1. Pick every 10th passenger as people board the plane *(systematic)*
2. From the passenger manifest, randomly choose 5 people flying business class and 25 from economy *(stratified)*
3. Randomly generate 30 seat numbers and survey the passengers who sit there *(simple)*
4. Randomly select a seat position (right window, right center, right isle, etc. — there are 10 seats abreast in a wide-body aircraft such as Boeing 747 or Airbus A380) and survey those passengers *(cluster)*

Election prediction in the past: 1988



In 1988, the presidential candidates were Michael Stanley Dukakis (D) and George Herbert Walker Bush (R).



1988 Gallup Poll

- Sample size: 1500 *(multistage sampling)*
- Sample proportions: 840 for Bush, 660 for Dukakis
- Calculate the *95% confidence interval* of the election result.

©2019 Oleg Goldshmidt

1988 Gallup Poll

- Sample size: 1500 *(multistage sampling)*
- Sample proportions: 840 for Bush, 660 for Dukakis
- Calculate the *95% confidence interval* of the election result.

95% confidence interval for simple random sampling:

$$P_p = P_s \pm \text{"sampling allowance"} = ?$$

1988 Gallup Poll

- Sample size: 1500 *(multistage sampling)*
- Sample proportions: 840 for Bush, 660 for Dukakis
- Calculate the *95% confidence interval* of the election result.

95% confidence interval for simple random sampling:

$$P_p = P_s \pm \text{"sampling allowance"} = P_s \pm z_{95} \times SEM = ?$$

1988 Gallup Poll

- Sample size: 1500 *(multistage sampling)*
- Sample proportions: 840 for Bush, 660 for Dukakis
- Calculate the *95% confidence interval* of the election result.

95% confidence interval for simple random sampling:

$$P_p = P_s \pm \text{"sampling allowance"} = P_s \pm z_{95} \times SEM = P_s \pm 1.96\sqrt{\frac{P_s(1-P_s)}{n}}$$

where P_p and P_s are the population and sample proportions, n is the sample size, $z_{95} = 1.96$ is the 95% z-score (NB: normal — see CLT!), and we substituted the SEM of a binomial distribution.

1988 Gallup Poll

- Sample size: 1500 *(multistage sampling)*
- Sample proportions: 840 for Bush, 660 for Dukakis
- Calculate the *95% confidence interval* of the election result.

95% confidence interval for simple random sampling:

$$P_p = P_s \pm \text{"sampling allowance"} = P_s \pm z_{95} \times SEM = P_s \pm 1.96\sqrt{\frac{P_s(1-P_s)}{n}}$$

where P_p and P_s are the population and sample proportions, n is the sample size, $z_{95} = 1.96$ is the 95% z-score (NB: normal — see CLT!), and we substituted the SEM of a binomial distribution.

$$Sample: \quad P_s = \frac{840}{1500} = 0.56$$

1988 Gallup Poll

- Sample size: 1500 *(multistage sampling)*
- Sample proportions: 840 for Bush, 660 for Dukakis
- Calculate the *95% confidence interval* of the election result.

95% confidence interval for simple random sampling:

$$P_p = P_s \pm \text{"sampling allowance"} = P_s \pm z_{95} \times SEM = P_s \pm 1.96\sqrt{\frac{P_s(1-P_s)}{n}}$$

where P_p and P_s are the population and sample proportions, n is the sample size, $z_{95} = 1.96$ is the 95% z-score (NB: normal — see CLT!), and we substituted the SEM of a binomial distribution.

$$\begin{aligned} Sample: \quad P_s &= \tfrac{840}{1500} = 0.56 \\ Population: \quad P_p &= 0.56 \pm 1.96\sqrt{\tfrac{0.56\times(1-0.56)}{1500}} = 0.56 \pm 0.03 = (53\%, 59\%) \end{aligned}$$

1988 Gallup Poll

- Sample size: 1500 *(multistage sampling)*
- Sample proportions: 840 for Bush, 660 for Dukakis
- Calculate the *95% confidence interval* of the election result.

95% confidence interval for simple random sampling:

$$P_p = P_s \pm \text{"sampling allowance"} = P_s \pm z_{95} \times SEM = P_s \pm 1.96\sqrt{\frac{P_s(1-P_s)}{n}}$$

where P_p and P_s are the population and sample proportions, n is the sample size, $z_{95} = 1.96$ is the 95% z-score (NB: normal — see CLT!), and we substituted the SEM of a binomial distribution.

$$\begin{aligned} Sample: \quad & P_s &=& \frac{840}{1500} = 0.56 \\ Population: \quad & P_p &=& 0.56 \pm 1.96\sqrt{\frac{0.56\times(1-0.56)}{1500}} = 0.56 \pm 0.03 = (53\%, 59\%) \end{aligned}$$

- Actual election result: 53.4% for Bush, 45.6% for Dukakis.