Algorithm Foundations of Data Science and Engineering
Lecture 11: Sampling

MING GAO

DaSE @ ECNU
(for course related communications)
mgao@dase.ecnu.edu.cn
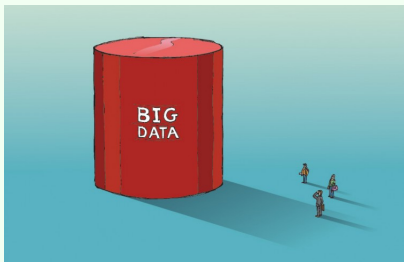
Jun. 10, 2019

Outline

Motivation of Sampling

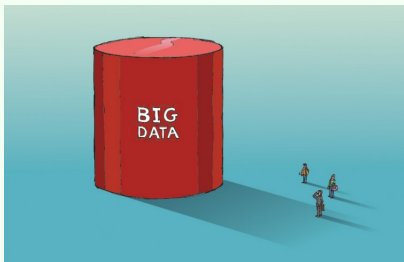Sampling
    Simple Random Sampling
    Systematic Sampling
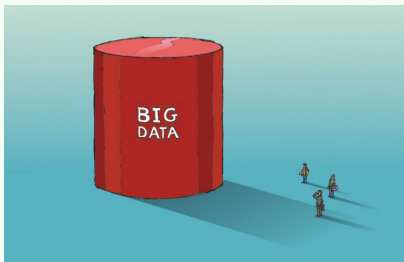   Stratified Sampling

Data volume in the world

- 1TB (Terabyte) $= 2^{10}$ GB $= 2^{40}$ B;
- 1PB (Petabyte) $= 2^{10}$ TB $= 2^{50}$ B;
- 1EB (Exabyte) $= 2^{10}$ PB $= 2^{60}$ B;
- 1ZB (Zettabyte) $= 2^{10}$ EB $= 2^{70}$ B;
- 1YB (YottaByte) $= 2^{10}$ ZB $= 2^{80}$ B;

Data volume in the world

- 1TB (Terabyte) $= 2^{10}$ GB $= 2^{40}$ B;
- 1PB (Petabyte) $= 2^{10}$ TB $= 2^{50}$ B;
- 1EB (Exabyte) $= 2^{10}$ PB $= 2^{60}$ B;
- 1ZB (Zettabyte) $= 2^{10}$ EB $= 2^{70}$ B;
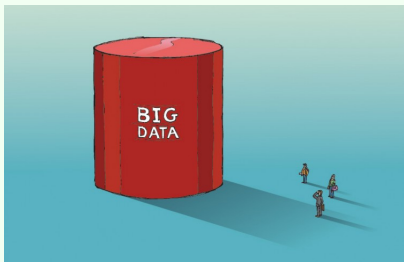- 1YB (YottaByte) $= 2^{10}$ ZB $= 2^{80}$ B;

Data volume in the world

World's data volume:

- 1TB (Terabyte) $= 2^{10}$ GB $= 2^{40}$ B;
- 1PB (Petabyte) $= 2^{10}$ TB $= 2^{50}$ B;
- 1EB (Exabyte) $= 2^{10}$ PB $= 2^{60}$ B;
- 1ZB (Zettabyte) $= 2^{10}$ EB $= 2^{70}$ B;
- 1YB (YottaByte) $= 2^{10}$ ZB $= 2^{80}$ B;
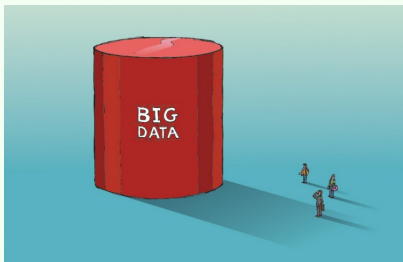
Data volume in the world

World's data volume:

- A full 90% of all the data in the world has been generated over the last two years;

- 1TB (Terabyte) $= 2^{10}$ GB $= 2^{40}$ B;
- 1PB (Petabyte) $= 2^{10}$ TB $= 2^{50}$ B;
- 1EB (Exabyte) $= 2^{10}$ PB $= 2^{60}$ B;
- 1ZB (Zettabyte) $= 2^{10}$ EB $= 2^{70}$ B;
- 1YB (YottaByte) $= 2^{10}$ ZB $= 2^{80}$ B;

Data volume in the world

World's data volume:

- A full 90% of all the data in the world has been generated over the last two years;
- The world's data volume is expected to grow 40 per cent per year;

- 1TB (Terabyte) $= 2^{10}$ GB $= 2^{40}$ B;
- 1PB (Petabyte) $= 2^{10}$ TB $= 2^{50}$ B;
- 1EB (Exabyte) $= 2^{10}$ PB $= 2^{60}$ B;
- 1ZB (Zettabyte) $= 2^{10}$ EB $= 2^{70}$ B;
- 1YB (YottaByte) $= 2^{10}$ ZB $= 2^{80}$ B;
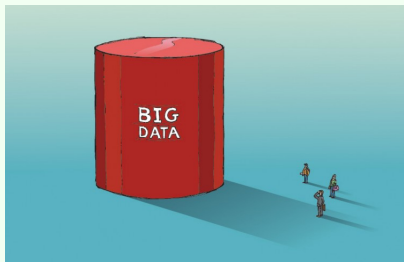
Data volume in the world

World's data volume:

- A full 90% of all the data in the world has been generated over the last two years;
- The world's data volume is expected to grow 40 per cent per year;
- Consumer IP internet traffic amounted to 100 EB per month in 2017;

- 1TB (Terabyte) $= 2^{10}$ GB $= 2^{40}$ B;
- 1PB (Petabyte) $= 2^{10}$ TB $= 2^{50}$ B;
- 1EB (Exabyte) $= 2^{10}$ PB $= 2^{60}$ B;
- 1ZB (Zettabyte) $= 2^{10}$ EB $= 2^{70}$ B;
- 1YB (YottaByte) $= 2^{10}$ ZB $= 2^{80}$ B;
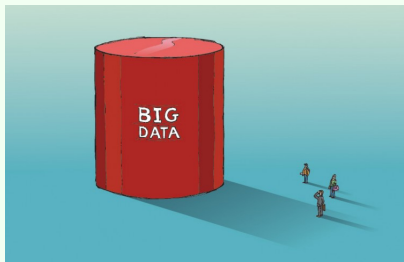
Data volume in the world

World's data volume:

- A full 90% of all the data in the world has been generated over the last two years;
- The world's data volume is expected to grow 40 per cent per year;
- Consumer IP internet traffic amounted to 100 EB per month in 2017;
- Data volume in the world will amount to 40 ZB by 2020.
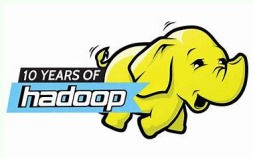
- 1TB (Terabyte) $= 2^{10}$ GB $= 2^{40}$ B;
- 1PB (Petabyte) $= 2^{10}$ TB $= 2^{50}$ B;
- 1EB (Exabyte) $= 2^{10}$ PB $= 2^{60}$ B;
- 1ZB (Zettabyte) $= 2^{10}$ EB $= 2^{70}$ B;
- 1YB (YottaByte) $= 2^{10}$ ZB $= 2^{80}$ B;

Hadoop: A brief history of big data processing

Hadoop was created by
Doug Cutting, the
creator of Apache
Lucene. Hadoop has its
origins in Apache Nutch,
which is a part of the
Lucene project.

Hadoop: A brief history of big data processing

Hadoop was created by Doug Cutting, the creator of Apache Lucene. Hadoop has its origins in Apache Nutch, which is a part of the Lucene project.

- Nutch was started in 2002. However, they realized that it wouldn't scale to the billions of pages on the Web;

Hadoop: A brief history of big data processing

Hadoop was created by Doug Cutting, the creator of Apache Lucene. Hadoop has its origins in Apache Nutch, which is a part of the Lucene project.
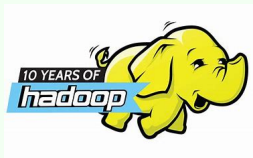
- Nutch was started in 2002. However, they realized that it wouldn't scale to the billions of pages on the Web;
- Google published papers in 2003 and 2004 that described GFS and MapReduce;

Hadoop: A brief history of big data processing

Hadoop was created by Doug Cutting, the creator of Apache Lucene. Hadoop has its origins in Apache Nutch, which is a part of the Lucene project.

- Nutch was started in 2002. However, they realized that it wouldn't scale to the billions of pages on the Web;
- Google published papers in 2003 and 2004 that described GFS and MapReduce;
- In 2006, Doug Cutting joined Yahoo!, which provided a dedicated team and the resources to turn Hadoop into a system that ran at web scale;
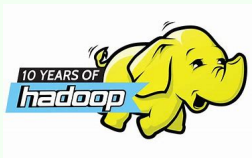
# Hadoop: A brief history of big data processing

Hadoop was created by Doug Cutting, the creator of Apache Lucene. Hadoop has its origins in Apache Nutch, which is a part of the Lucene project.

- Nutch was started in 2002. However, they realized that it wouldn't scale to the billions of pages on the Web;
- Google published papers in 2003 and 2004 that described GFS and MapReduce;
- In 2006, Doug Cutting joined Yahoo!, which provided a dedicated team and the resources to turn Hadoop into a system that ran at web scale;
- In February 2008, Yahoo! announced that its production search index was being generated by a 10,000-core Hadoop cluster.

Ecosystem for big data processing
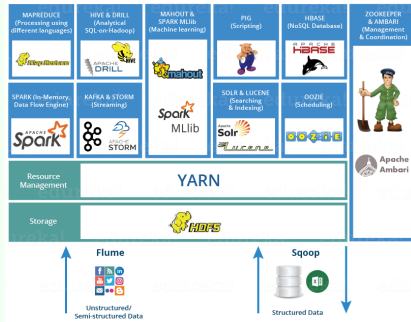
Ecosystem for big data processing
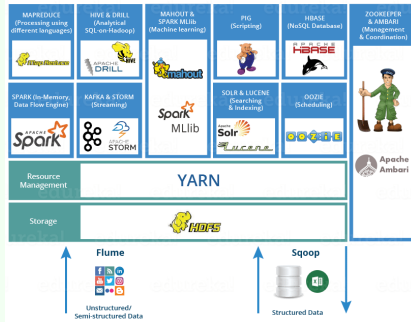
A number of large-scale data processing frameworks have thereby been developed, such as MapReduce, Spark, Storm, Flink, Dryad, Caffe, Tensorflow.

Ecosystem for big data processing

Infrastructural technologies are the core of the big data ecosystem.

A number of large-scale data processing frameworks have thereby been developed, such as MapReduce, Spark, Storm, Flink, Dryad, Caffe, Tensorflow.

Ecosystem for big data processing
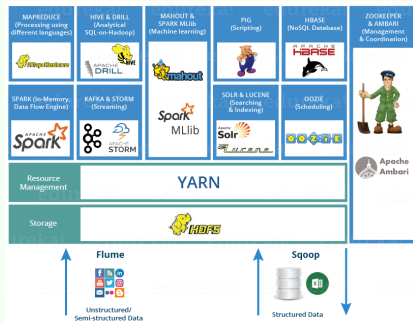

Infrastructural technologies are the core of the big data ecosystem.

- Hadoop;

A number of large-scale data processing frameworks have thereby been developed, such as MapReduce, Spark, Storm, Flink, Dryad, Caffe, Tensorflow.

Ecosystem for big data processing

Infrastructural technologies are the core of the big data ecosystem.

- Hadoop;
- NoSQL, like HBase;

A number of large-scale data processing frameworks have thereby been developed, such as MapReduce, Spark, Storm, Flink, Dryad, Caffe, Tensorflow.
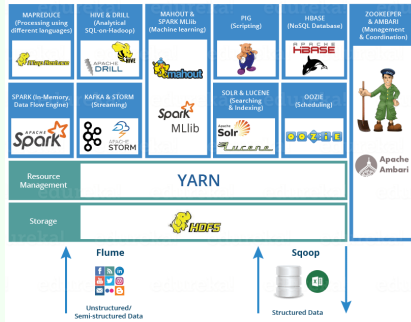
Ecosystem for big data processing

Infrastructural technologies are the core of the big data ecosystem.

- Hadoop;
- NoSQL, like HBase;
- Spark;

A number of large-scale data processing frameworks have thereby been developed, such as MapReduce, Spark, Storm, Flink, Dryad, Caffe, Tensorflow.
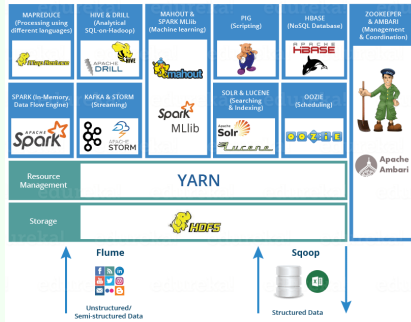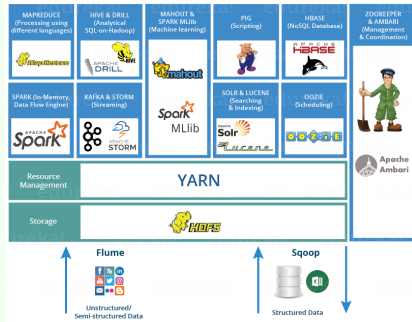
Ecosystem for big data processing

Infrastructural technologies are the core of the big data ecosystem.

- Hadoop;
- NoSQL, like HBase;
- Spark;
- Storm;
- $\cdots$;

A number of large-scale data processing frameworks have thereby been developed, such as MapReduce, Spark, Storm, Flink, Dryad, Caffe, Tensorflow.

Moore's Law is dead
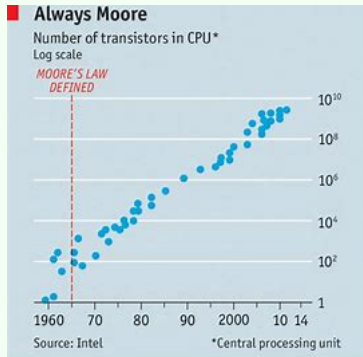
Always Moore
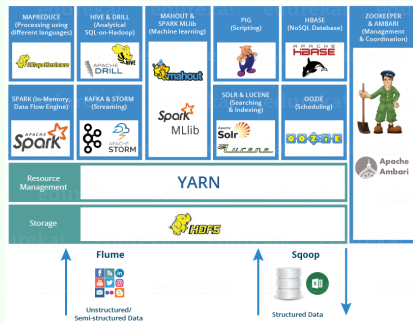Number of transistors in CPU*
Log scale

MOORE'S LAW
DEFINED

Source: Intel   *Central processing unit

Observing that the total number of
components in these circuits had
roughly doubled each year [Gordon E.
Moore 1965].

Moore's Law is dead

Always Moore
Number of transistors in CPU*
Log scale
MOORE'S LAW DEFINED

10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
1

1960 70 80 90 2000 10 14
Source: Intel   *Central processing unit
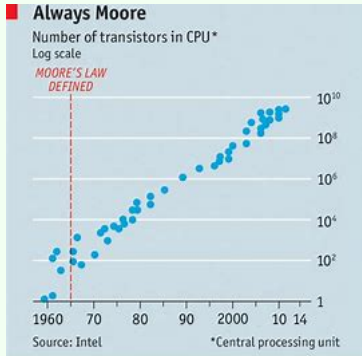
History proved Moore correct.

Observing that the total number of components in these circuits had roughly doubled each year [Gordon E. Moore 1965].

Moore's Law is dead

Always Moore
Number of transistors in CPU*
Log scale
MOORE'S LAW DEFINED

History proved Moore correct.

- Moore's lawthough its doubling period was lengthened to 18 months in the mid-1970s;

Observing that the total number of components in these circuits had roughly doubled each year [Gordon E. Moore 1965].

Moore's Law is dead

Always Moore
Number of transistors in CPU*
Log scale
MOORE'S LAW DEFINED
Source: Intel   *Central processing unit

Observing that the total number of components in these circuits had roughly doubled each year [Gordon E. Moore 1965].

History proved Moore correct.
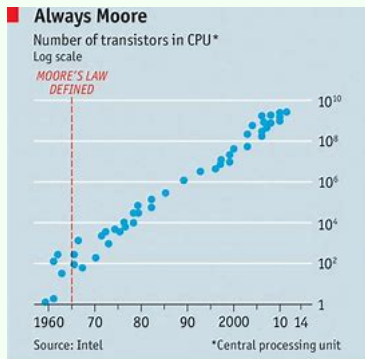
- Moore's lawthough its doubling period was lengthened to 18 months in the mid-1970s;

- It continued into the second decade of the 21st century with tens of nanometres in size;

Moore's Law is dead

Always Moore
Number of transistors in CPU*
Log scale
MOORE'S LAW DEFINED

Source: Intel   *Central processing unit

Observing that the total number of components in these circuits had roughly doubled each year [Gordon E. Moore 1965].

History proved Moore correct.
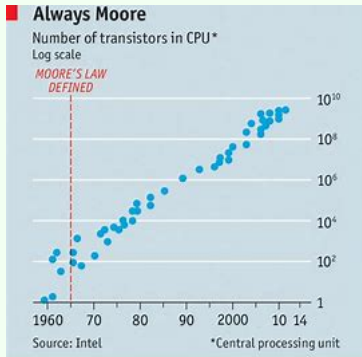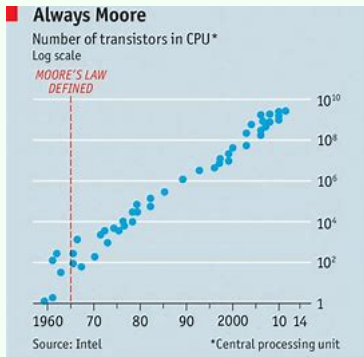
- Moore's lawthough its doubling period was lengthened to 18 months in the mid-1970s;

- It continued into the second decade of the 21st century with tens of nanometres in size;

- The processing power of computers increases exponentially every couple of years has hit its limit, according to Jensen Huang, CEO of Nvidia.

Choices of big data processing

Choices of big data processing

Given the characteristics of the unprecedented amount of data, the speed of data production, and the multiple of the structure of data, large-scale data processing is essential to analyzing and mining such big data timely. However, it is not the only choice.

Choices of big data processing

Given the characteristics of the unprecedented amount of data, the speed of data production, and the multiple of the structure of data, large-scale data processing is essential to analyzing and mining such big data timely. However, it is not the only choice.

We don't need to poll every single American of voting age whenever these events occur. We may need only polling 1,100 people randomly selected.

Choices of big data processing

Given the characteristics of the unprecedented amount of data, the speed of data production, and the multiple of the structure of data, large-scale data processing is essential to analyzing and mining such big data timely. However, it is not the only choice.

We don't need to poll every single American of voting age whenever these events occur. We may need only polling 1,100 people randomly selected. With that (relatively) small set of data, they are extrapolating how the rest of the public feels and is likely to vote.

Sampling

> **Sample** is a subset of the population on which observations are taken for obtaining information about the population.

Sampling

Sample is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;

Sampling

Sample is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;
- The most common sampling techniques used for official surveys are:

Sampling

> **Sample** is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;
- The most common sampling techniques used for official surveys are:
  - Simple random sampling;

Sampling

Sample is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;
- The most common sampling techniques used for official surveys are:
  - Simple random sampling;
  - Systematic sampling;

Sampling

Sample is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;
- The most common sampling techniques used for official surveys are:
  - Simple random sampling;
  - Systematic sampling;
  - Stratified sampling;

Sampling

Sample is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;
- The most common sampling techniques used for official surveys are:
  - □ Simple random sampling;
  - □ Systematic sampling;
  - □ Stratified sampling;
  - □ Cluster sampling;

Sampling

Sample is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;
- The most common sampling techniques used for official surveys are:
  - □ Simple random sampling;
  - □ Systematic sampling;
  - □ Stratified sampling;
  - □ Cluster sampling;
  - □ Multi-stage sampling;

Sampling

Sample is a subset of the population on which observations are taken for obtaining information about the population.

- Since studying a sample we wish to draw valid conclusions about the population, sample should desirably be "representative" of the target population;
- The most common sampling techniques used for official surveys are:
  - □ Simple random sampling;
  - □ Systematic sampling;
  - □ Stratified sampling;
  - □ Cluster sampling;
  - □ Multi-stage sampling;
- All are examples of probability sampling.

Outline

Simple random sampling

Simple random sampling

Let N be the volume of population, and X_i be a r.v.

$$X_i = \begin{cases} 1, & \text{the i-th unit is selected;} \\ 0, & \text{otherwise.} \end{cases}$$

Simple random sampling

Let N be the volume of population, and X_i be a r.v.

$$X_i = \begin{cases} 1, & \text{the i-th unit is selected;} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$Pr[X_i = 1] = \frac{1}{N}.$$

Simple random sampling

The procedure of sampling in which the units are selected with probability proportional to a given measure of size.

Let N be the volume of population, and X_i be a r.v.

$$X_i = \begin{cases} 1, & \text{the i-th unit is selected;} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$Pr[X_i = 1] = \frac{1}{N}.$$

Simple random sampling

Let N be the volume of population, and X_i be a r.v.

$$X_i = \begin{cases} 1, & \text{the i-th unit is selected;} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$Pr[X_i = 1] = \frac{1}{N}.$$

The procedure of sampling in which the units are selected with probability proportional to a given measure of size.

- Simple random sampling is simplest method of probability sampling;

Simple random sampling

Let N be the volume of population, and X_i be a r.v.

$X_i = \begin{cases} 1, & \text{the i-th unit is selected;} \\ 0, & \text{otherwise.} \end{cases}$

Then

$$Pr[X_i = 1] = \frac{1}{N}.$$

The procedure of sampling in which the units are selected with probability proportional to a given measure of size.

- Simple random sampling is simplest method of probability sampling;
- Simple random sampling is special type of equal probability selection method;

Simple random sampling

Let N be the volume of population, and X_i be a r.v.

$$X_i = \begin{cases} 1, & \text{the i-th unit is selected;} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$Pr[X_i = 1] = \frac{1}{N}.$$

The procedure of sampling in which the units are selected with probability proportional to a given measure of size.

- Simple random sampling is simplest method of probability sampling;
- Simple random sampling is special type of equal probability selection method;

  □ With replacement;

Simple random sampling

Let N be the volume of population, and X_i be a r.v.

$$X_i = \begin{cases} 1, & \text{the i-th unit is selected;} \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$Pr[X_i = 1] = \frac{1}{N}.$$

The procedure of sampling in which the units are selected with probability proportional to a given measure of size.

- Simple random sampling is simplest method of probability sampling;
- Simple random sampling is special type of equal probability selection method;
  - □ With replacement;
  - □ Without replacement.

Outline

Systematic sampling

Systematic sampling

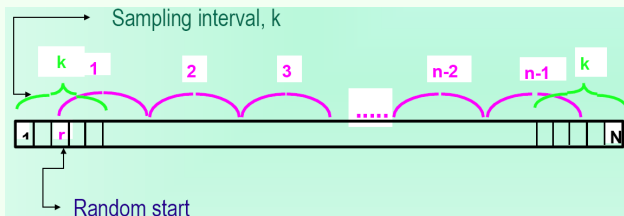- Systematic sampling involves selecting *n* sample units from a population of *N* units;

Systematic sampling

Sampling interval, k

k 1 2 3 n-2 n-1 k

r

N

Random start

- Systematic sampling involves selecting *n* sample units from a population of *N* units;

- Instead of randomly choosing the *n* units in the sample, a skip pattern is run through a list (frame) of the *N* units to select the sample;

Systematic sampling

Sampling interval, k

k   **1**   **2**   **3**   .....   **n-2**   **n-1**   k
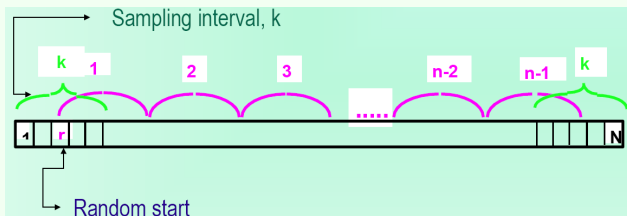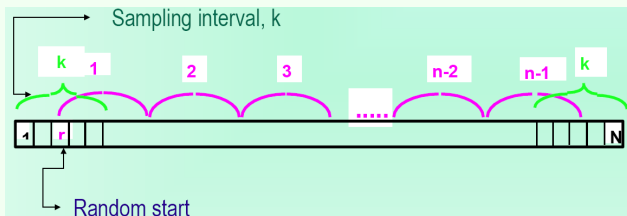
r

N

Random start

- Systematic sampling involves selecting n sample units from a population of N units;

- Instead of randomly choosing the n units in the sample, a skip pattern is run through a list (frame) of the N units to select the sample;

  □ Decide on a sample size n and compute the skip, $k = \frac{N}{n}$;

Systematic sampling

- Systematic sampling involves selecting *n* sample units from a population of *N* units;

- Instead of randomly choosing the *n* units in the sample, a skip pattern is run through a list (frame) of the *N* units to select the sample;

  □ Decide on a sample size *n* and compute the skip, $k = \frac{N}{n}$;
  □ Choose a random start *r* between 1 and *k* (inclusive);

Systematic sampling

Sampling interval, k

Random start

- Systematic sampling involves selecting *n* sample units from a population of *N* units;

- Instead of randomly choosing the *n* units in the sample, a skip pattern is run through a list (frame) of the *N* units to select the sample;

  □ Decide on a sample size *n* and compute the skip, $k = \frac{N}{n}$;
  □ Choose a random start *r* between 1 and *k* (inclusive);
  □ Add "k" to selected random number to select the next unit repeatedly.

Problem of linear systematic sampling

- If N is a multiple of n, then the number of units in each of the k possible systematic samples is n;

Problem of linear systematic sampling

- If N is a multiple of n, then the number of units in each of the k possible systematic samples is n;
  - In this case systematic sampling amounts to grouping the N units into k samples of exactly n units each in a systematic manner and selecting one of them with probability $\frac{1}{k}$

Problem of linear systematic sampling

- If N is a multiple of n, then the number of units in each of the k possible systematic samples is n;
  - In this case systematic sampling amounts to grouping the N units into k samples of exactly n units each in a systematic manner and selecting one of them with probability $\frac{1}{k}$
 - In this case, the sampling scheme is equal probability selection method (epsem);

Problem of linear systematic sampling

- If N is a multiple of n, then the number of units in each of the k possible systematic samples is n;
  - In this case systematic sampling amounts to grouping the N units into k samples of exactly n units each in a systematic manner and selecting one of them with probability $\frac{1}{k}$
  - In this case, the sampling scheme is equal probability selection method (epsem);
- But, if $\frac{N}{n}$ is not an integer, then the number of units selected systematically with the sampling interval

$$k \approx \text{ nearest integer to } \frac{N}{n}$$

Problem of linear systematic sampling

- If N is a multiple of n, then the number of units in each of the k possible systematic samples is n;
  - In this case systematic sampling amounts to grouping the N units into k samples of exactly n units each in a systematic manner and selecting one of them with probability $\frac{1}{k}$
  - In this case, the sampling scheme is equal probability selection method (epsem);
- But, if $\frac{N}{n}$ is not an integer, then the number of units selected systematically with the sampling interval
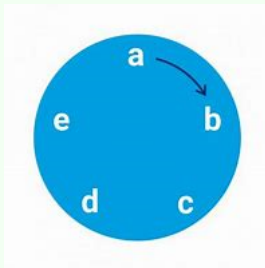
$$k \approx \text{ nearest integer to } \frac{N}{n}$$

- This problem may be overcome by adopting a device, known as circular systematic sampling;
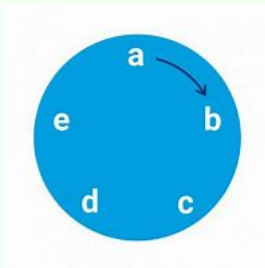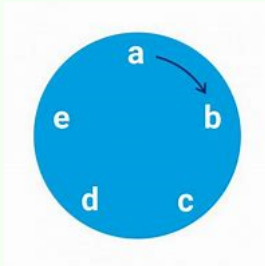
Circular systematic sampling

Circular systematic sampling
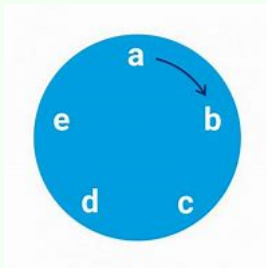

Useful when $\frac{N}{n}$ is not integer.

Circular systematic sampling

- Determine the interval k C rounding down to the integer nearest to $\frac{N}{n}$, e.g., If $N = 15$ and $n = 4$, then k is taken as 3 and not 4;

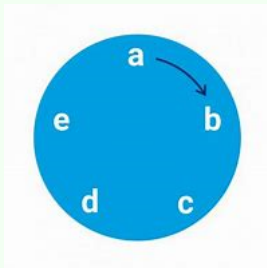Useful when $\frac{N}{n}$ is not integer.

Circular systematic sampling

Useful when $\frac{N}{n}$ is not integer.

- Determine the interval k C rounding down to the integer nearest to $\frac{N}{n}$, e.g., If $N = 15$ and $n = 4$, then k is taken as 3 and not 4;

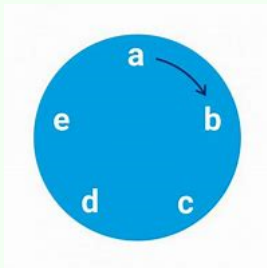- Take a random start between 1 and N;

Circular systematic sampling

Useful when $\frac{N}{n}$ is not integer.

- Determine the interval k C rounding down to the integer nearest to $\frac{N}{n}$, e.g., If $N = 15$ and $n = 4$, then k is taken as 3 and not 4;

- Take a random start between 1 and N;

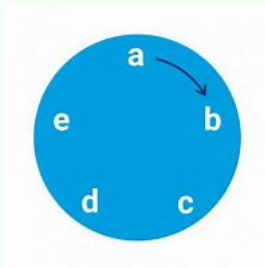- Skip through the circle by k units each time to select the next unit until n units are selected;

Circular systematic sampling

Useful when $\frac{N}{n}$ is not integer.

- Determine the interval k C rounding down to the integer nearest to $\frac{N}{n}$, e.g., If $N = 15$ and $n = 4$, then k is taken as 3 and not 4;

- Take a random start between 1 and N;

- Skip through the circle by k units each time to select the next unit until n units are selected;

- Thus there could be N possible distinct samples instead of k;

Circular systematic sampling

Useful when $\frac{N}{n}$ is not integer.

- Determine the interval k C rounding down to the integer nearest to $\frac{N}{n}$, e.g., If $N = 15$ and $n = 4$, then k is taken as 3 and not 4;

- Take a random start between 1 and N;

- Skip through the circle by k units each time to select the next unit until n units are selected;

- Thus there could be N possible distinct samples instead of k;

Ensures each unit equal chance of being selected into sample.

Pros and cons of circular systematic sampling

- Pros

Pros and cons of circular systematic sampling

- Pros
  - Operationally convenient - easier to draw a sample;

Pros and cons of circular systematic sampling

- Pros
  - □ Operationally convenient - easier to draw a sample;
  - □ It distributes the sample more evenly over the population.

Pros and cons of circular systematic sampling

- Pros
  - □ Operationally convenient - easier to draw a sample;
  - □ It distributes the sample more evenly over the population. Thus it is likely to be more efficient than SRSWOR, particularly when the ordering of the units in the list is related to characteristics of the variable of interest;

Pros and cons of circular systematic sampling

- Pros
  - Operationally convenient - easier to draw a sample;
  - It distributes the sample more evenly over the population. Thus it is likely to be more efficient than SRSWOR, particularly when the ordering of the units in the list is related to characteristics of the variable of interest;
- Cons

Pros and cons of circular systematic sampling

- Pros
  - □ Operationally convenient - easier to draw a sample;
  - □ It distributes the sample more evenly over the population. Thus it is likely to be more efficient than SRSWOR, particularly when the ordering of the units in the list is related to characteristics of the variable of interest;
- Cons
  - □ Requires complete list of the population;

Pros and cons of circular systematic sampling

- Pros
  - □ Operationally convenient - easier to draw a sample;
  - □ It distributes the sample more evenly over the population. Thus it is likely to be more efficient than SRSWOR, particularly when the ordering of the units in the list is related to characteristics of the variable of interest;
- Cons
  - □ Requires complete list of the population;
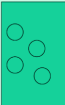 - □ A bad arrangement of the units may produce a very inefficient sample.
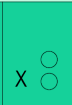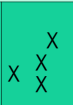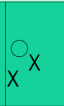
Outline
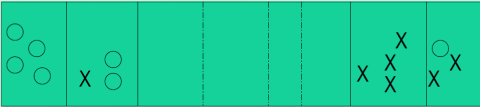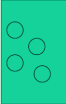
Stratified sampling

Stratified sampling

Stratum no.	1	2	h	L
Stratum size	N_1	N_2	N_h	N_H
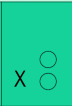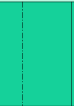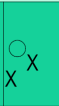
- Divides the population into a number of distinct groups (strata) based on auxiliary information;

Stratified sampling
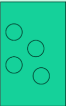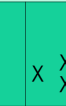
Stratum no.	1	2	h		L
Stratum size	N_1	N_2	N_h		N_H

- Divides the population into a number of distinct groups (strata) based on auxiliary information;

- Each stratum is composed of units that satisfy the condition set by the values of the stratifying variable;

Stratified sampling

Stratum no.	1	2	h	L
Stratum size	N_1	N_2	N_h	N_H

- Divides the population into a number of distinct groups (strata) based on auxiliary information;

- Each stratum is composed of units that satisfy the condition set by the values of the stratifying variable;

  □ Reduce the standard error of the estimates;

Stratified sampling

Stratum no.	1	2	h			L
Stratum size	N_1	N_2	N_h			N_H

- Divides the population into a number of distinct groups (strata) based on auxiliary information;

- Each stratum is composed of units that satisfy the condition set by the values of the stratifying variable;

  □ Reduce the standard error of the estimates;
  □ Provide separate estimates required for each sub-division of the population - "domain" estimates;

Stratified sampling

Stratum no.	1	2	h	L
Stratum size	N_1	N_2	N_h	N_H

- Divides the population into a number of distinct groups (strata) based on auxiliary information;

- Each stratum is composed of units that satisfy the condition set by the values of the stratifying variable;

  □ Reduce the standard error of the estimates;
  □ Provide separate estimates required for each sub-division of the population - "domain" estimates;
  □ Using different sampling procedures for different sub-population to increase efficiency of the estimates.

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:
  - Like SRS or systematic sampling within each stratum;

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:
  - Like SRS or systematic sampling within each stratum;
  - Independent of the other strata;

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:
  - Like SRS or systematic sampling within each stratum;
  - Independent of the other strata;
- Sampling in each stratum is carried out independently;

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:
  - □ Like SRS or systematic sampling within each stratum;
  - □ Independent of the other strata;
- Sampling in each stratum is carried out independently;
  - □ Sampling fractions may differ;

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:
  - □ Like SRS or systematic sampling within each stratum;
  - □ Independent of the other strata;
- Sampling in each stratum is carried out independently;
  - □ Sampling fractions may differ;
  - □ Selection procedures may also be different;

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:
  - □ Like SRS or systematic sampling within each stratum;
  - □ Independent of the other strata;
- Sampling in each stratum is carried out independently;
  - □ Sampling fractions may differ;
  - □ Selection procedures may also be different;
- The total sample size is distributed over all the strata - allocation

Stratified sampling Cont'd

- Division or stratification of the population into homogeneous (similar) groups called strata;
- Selecting the sample using a selection procedure:
  - □ Like SRS or systematic sampling within each stratum;
  - □ Independent of the other strata;
- Sampling in each stratum is carried out independently;
  - □ Sampling fractions may differ;
  - □ Selection procedures may also be different;
- The total sample size is distributed over all the strata - allocation
- At the end of the survey, the stratum results are combined to provide an estimate for entire population

Clustering and stratification

Defining Strata

Clustering and stratification

Defining Strata

- Choice of stratification variables (location, degree, etc):

Clustering and stratification

Defining Strata
- Choice of stratification variables (location, degree, etc):
  - Homogeneous within strata; Heterogeneous across strata;

Clustering and stratification

Defining Strata

- Choice of stratification variables (location, degree, etc):
  - □ Homogeneous within strata; Heterogeneous across strata;
  - □ Highly correlated with study variables;

Clustering and stratification

Defining Strata

- Choice of stratification variables (location, degree, etc):
  - □ Homogeneous within strata; Heterogeneous across strata;
  - □ Highly correlated with study variables;
- Number of strata;

Clustering and stratification

Defining Strata
- Choice of stratification variables (location, degree, etc):
  - □ Homogeneous within strata; Heterogeneous across strata;
  - □ Highly correlated with study variables;
- Number of strata;
  - □ Depends on availability of stratifying information in sampling frame: less information, fewer strata;

Clustering and stratification

Defining Strata

- Choice of stratification variables (location, degree, etc):
  - Homogeneous within strata; Heterogeneous across strata;
  - Highly correlated with study variables;
- Number of strata;
  - Depends on availability of stratifying information in sampling frame: less information, fewer strata;
  - At least two sampling units per stratum to be able to compute sampling error;

Clustering and stratification

Defining Strata

- Choice of stratification variables (location, degree, etc):
  - □ Homogeneous within strata; Heterogeneous across strata;
  - □ Highly correlated with study variables;
- Number of strata;
  - □ Depends on availability of stratifying information in sampling frame: less information, fewer strata;
  - □ At least two sampling units per stratum to be able to compute sampling error;
- Sample allocation to strata

Clustering and stratification

Defining Strata

- Choice of stratification variables (location, degree, etc):
  - Homogeneous within strata; Heterogeneous across strata;
  - Highly correlated with study variables;
- Number of strata;
  - Depends on availability of stratifying information in sampling frame: less information, fewer strata;
  - At least two sampling units per stratum to be able to compute sampling error;
- Sample allocation to strata
  - Proportionate allocation: an uniform sampling fraction is applied to each strata; that is, the sample size selected from each stratum is made proportionate to the population size of the stratum;

Clustering and stratification

Defining Strata

- Choice of stratification variables (location, degree, etc):
  - ☐ Homogeneous within strata; Heterogeneous across strata;
  - ☐ Highly correlated with study variables;
- Number of strata;
  - ☐ Depends on availability of stratifying information in sampling frame: less information, fewer strata;
  - ☐ At least two sampling units per stratum to be able to compute sampling error;
- Sample allocation to strata
  - ☐ Proportionate allocation: an uniform sampling fraction is applied to each strata; that is, the sample size selected from each stratum is made proportionate to the population size of the stratum;
  - ☐ Disproportionate allocation: different sampling rates are used deliberately in different strata.

Proportionate allocation

In proportionate stratification, $\frac{n_h}{N_h}$, is specified to be the same for each stratum.

Proportionate allocation

In proportionate stratification, $\frac{n_h}{N_h}$, is specified to be the same for each stratum.

- Choice of stratification variables (location, degree, etc):

Proportionate allocation

In proportionate stratification, $\frac{n_h}{N_h}$, is specified to be the same for each stratum.

- Choice of stratification variables (location, degree, etc):
  - □ This implies that the overall sampling fraction is

$$\frac{n_h}{N_h} = \frac{n}{N};$$

Proportionate allocation

In proportionate stratification, $\frac{n_h}{N_h}$, is specified to be the same for each stratum.

- Choice of stratification variables (location, degree, etc):
  - □ This implies that the overall sampling fraction is

  $$\frac{n_h}{N_h} = \frac{n}{N};$$

  - □ The number of elements taken from the h-th stratum is

  $$n_h = N_h \cdot \frac{n}{N}.$$

Proportionate allocation

In proportionate stratification, $\frac{n_h}{N_h}$, is specified to be the same for each stratum.

- Choice of stratification variables (location, degree, etc):
  - □ This implies that the overall sampling fraction is

$$\frac{n_h}{N_h} = \frac{n}{N};$$

  - □ The number of elements taken from the h-th stratum is

$$n_h = N_h \cdot \frac{n}{N}.$$

- For a given total variability in the population, the gain is greater if:

Proportionate allocation

In proportionate stratification, $\frac{n_h}{N_h}$, is specified to be the same for each stratum.

- Choice of stratification variables (location, degree, etc):
  - □ This implies that the overall sampling fraction is

  $$\frac{n_h}{N_h} = \frac{n}{N};$$

  - □ The number of elements taken from the h-th stratum is

  $$n_h = N_h \cdot \frac{n}{N}.$$

- For a given total variability in the population, the gain is greater if:
  - □ The strata mean are more heterogeneous (more unequal strata mean);

Proportionate allocation

In proportionate stratification, $\frac{n_h}{N_h}$, is specified to be the same for each stratum.

- Choice of stratification variables (location, degree, etc):
  - This implies that the overall sampling fraction is

  $$\frac{n_h}{N_h} = \frac{n}{N};$$

  - The number of elements taken from the h-th stratum is

  $$n_h = N_h \cdot \frac{n}{N}.$$

- For a given total variability in the population, the gain is greater if:
  - The strata mean are more heterogeneous (more unequal strata mean);
  - The element values within the strata are more homogeneous.

Optimum allocation

Uses widely different sampling rates for the various strata.

Optimum allocation

Uses widely different sampling rates for the various strata.

- Objective: to achieve the least variance for the overall mean for the given sample size;

Optimum allocation

Uses widely different sampling rates for the various strata.

- Objective: to achieve the least variance for the overall mean for the given sample size;
- Without cost consideration, the allocation is

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum N_h \sigma_h}.$$

Optimum allocation

Uses widely different sampling rates for the various strata.

- Objective: to achieve the least variance for the overall mean for the given sample size;

- Without cost consideration, the allocation is

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum N_h \sigma_h}.$$

- This gives better efficiency as compared to proportionate allocation:

$$V_{SRS} \geq V_{prop} \geq V_{opt}.$$

Optimum allocation

Uses widely different sampling rates for the various strata.

- Objective: to achieve the least variance for the overall mean for the given sample size;
- Without cost consideration, the allocation is

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum N_h \sigma_h}.$$

- This gives better efficiency as compared to proportionate allocation:

$$V_{SRS} \geq V_{prop} \geq V_{opt}.$$

- In practice, σ_h is unknown.

Optimum allocation

Uses widely different sampling rates for the various strata.

- Objective: to achieve the least variance for the overall mean for the given sample size;
- Without cost consideration, the allocation is

$$n_h = n \cdot \frac{N_h \sigma_h}{\sum N_h \sigma_h}.$$

- This gives better efficiency as compared to proportionate allocation:

$$V_{SRS} \geq V_{prop} \geq V_{opt}.$$

- In practice, σ_h is unknown.
- This refers to a systematic sampling with the units arranged in a certain order

Take-home messages

- Motivation of sampling
- Sampling
  - □ Simple Random Sampling
  - □ Systematic Sampling
  - □ Stratified Sampling