Data Engineering
Lecture 10: Data Integration

Nada Sharaf

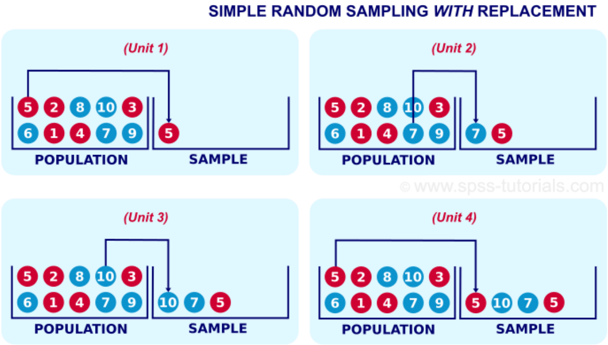
The German International University

Flipped Classroom

In sampling, what is

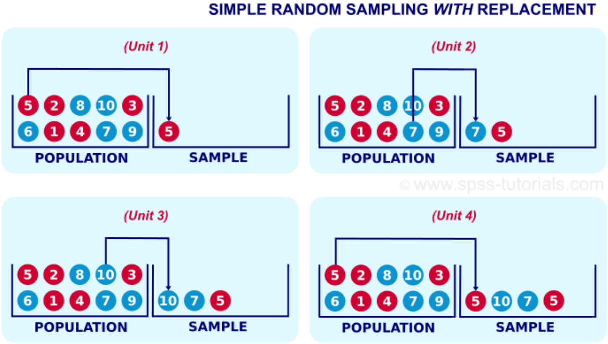
- Simple Random Sample Without Replacement of sizes
- Simple Random Sample with Replacement of sizes
- Cluster Sample
- Stratified Sample

Simple Random Sample with Replacement



https://www.spss-tutorials.com/simple-random-sampling-what-is-it/

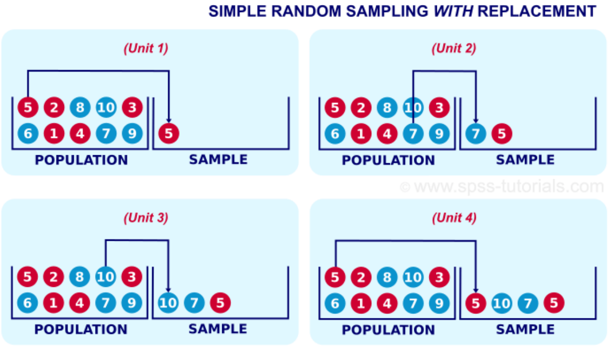
Simple Random Sample with Replacement



https://www.spss-tutorials.com/simple-random-sampling-what-is-it/

- We record the tuple/ some of its properties
- Then we place it back

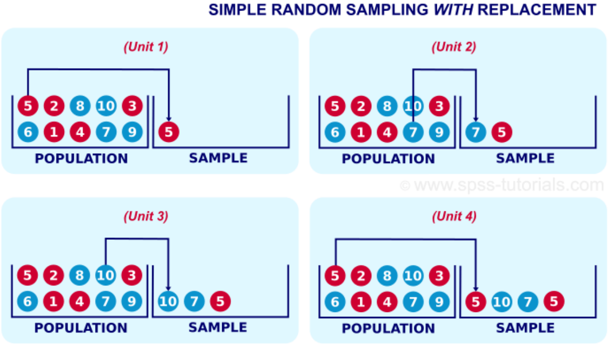
Simple Random Sample with Replacement



https://www.spss-tutorials.com/simple-random-sampling-what-is-it/

- We record the tuple/ some of its properties
- Then we place it back
- All records have a chance of 0.1 in the above example

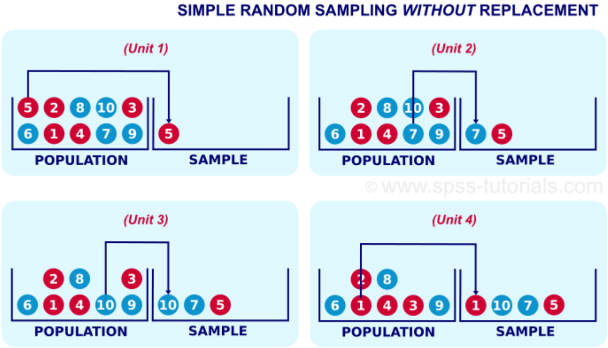
Simple Random Sample with Replacement



https://www.spss-tutorials.com/simple-random-sampling-what-is-it/

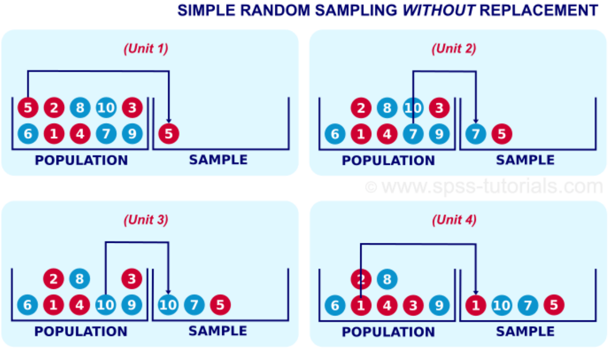
- We record the tuple/ some of its properties
- Then we place it back
- All records have a chance of 0.1 in the above example
- Independant items

Simple Random Sample without Replacement



https://www.spss-tutorials.com/simple-random-sampling-what-is-it/

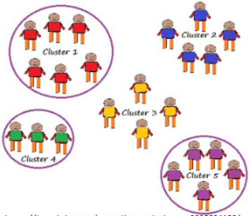
Simple Random Sample without Replacement



https://www.spss-tutorials.com/simple-random-sampling-what-is-it/

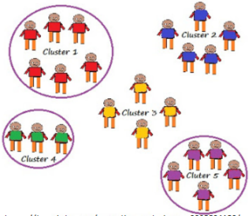
- Dependant items
- Random Sample

Cluster Sample



https://laptrinhx.com/sampling-techniques-3995824180/

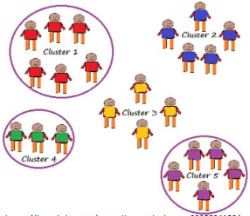
Cluster Sample



https://laptrinhx.com/sampling-techniques-3995824180/

- Equal Chance of being selected
- Cluster then choose some of clusters

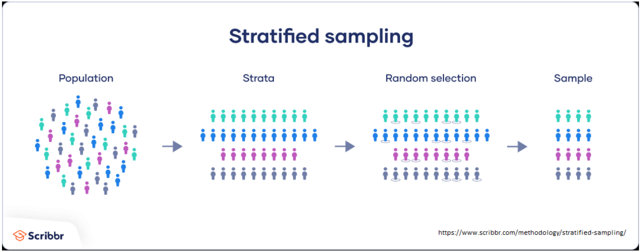
Cluster Sample



https://laptrinhx.com/sampling-techniques-3995824180/

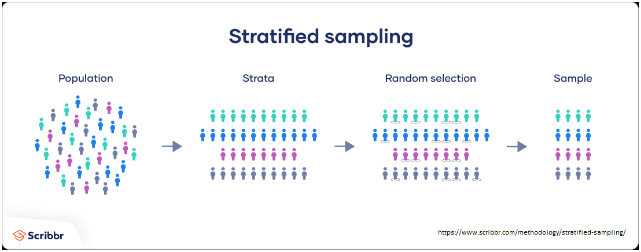
- Equal Chance of being selected
- Cluster then choose some of clusters
- One-Stage Sampling: All elements of the selected clusters are part of the sample
- Two-Stage Sampling: Select elements from the selected clusters
- Multi-Stage Sampling: Select elements from the selected clusters through different levels and stages

Stratified Sample



https://www.scribbr.com/methodology/stratified-sampling/

Stratified Sample



https://www.scribbr.com/methodology/stratified-sampling/

- Define characteristics
- Define strata (group with the same characteristics)

Data Integration

Data Integration

- When you search the web, different documents and data sources are being queried
- Data with different formats are somehow shared and used

Data Integration: Why

- Data from different sources need to be merged in one place
- Data might have different formats

Data Integration: Why

- Data from different sources need to be merged in one place
- Data might have different formats
- Data can be from within the same organization or external data

Data Integration: Why

- Data from different sources need to be merged in one place
- Data might have different formats
- Data can be from within the same organization or external data
- Integration should reduce redundancies and inconsistencies

Data Integration: Other Challenges

- Semantic heterogeneity
  - Different scales
  - Different representations of data

Data Integration: Other Challenges

- Semantic heterogeneity
  - Different scales
  - Different representations of data
  - e.g. databases, files, html

Data Integration: Other Challenges

- Semantic heterogeneity
  - Different scales
  - Different representations of data
  - e.g. databases, files, html
- Redundancies
- Entity Specification (Keys, ...)

Data Integration: Other Challenges

- Semantic heterogeneity
  - Different scales
  - Different representations of data
  - e.g. databases, files, html
- Redundancies
- Entity Specification (Keys, ...)
- Number of sources

Heterogeneity Sources: Schema

- One Employee table vs. multiple Employee tables: Schema mismatch
- first name vs first and last name: Domain mismatch
- Constraint mismatch e.g. gpa

Heterogeneity Sources: Instance

- Identifying entities: same student in two different sources without identification
- Format conflict e.g. date of birth

Heterogeneity Sources: Semantic Heterogeneity

- Different units: total price vs. number of units
- Different encodings
- Different scales, ... etc

Structure of Data Sources

- Data formats
  - Vendor-specific formats
  - XML, JSON are usually accepted but not used by all systems

Structure of Data Sources

- Data formats
  - Vendor-specific formats
  - XML, JSON are usually accepted but not used by all systems
- Prioritizing constraints

Structure of Data Sources

- Data formats
  - Vendor-specific formats
  - XML, JSON are usually accepted but not used by all systems
- Prioritizing constraints
- Some data types are more challenging e.g. images, audio, video where there are no specific attributes

Schema Integration

- Merging multiple schemas into one schema

Schema Integration

- Merging multiple schemas into one schema
    1. Schema Transformation

Schema Integration

- Merging multiple schemas into one schema
  1. Schema Transformation
     - ★ Have homogenous formats
     - ★ Extract the model from data

Schema Integration

- Merging multiple schemas into one schema
  1. Schema Transformation
     - ★ Have homogenous formats
     - ★ Extract the model from data
  2. Schema matching
     - ★ Identify items that are semantically related
     - ★ e.g. student, graduate student

Schema Integration

- Merging multiple schemas into one schema
  1. Schema Transformation
     - ★ Have homogenous formats
     - ★ Extract the model from data
  2. Schema matching
     - ★ Identify items that are semantically related
     - ★ e.g. student, graduate student
  3. Schema integration

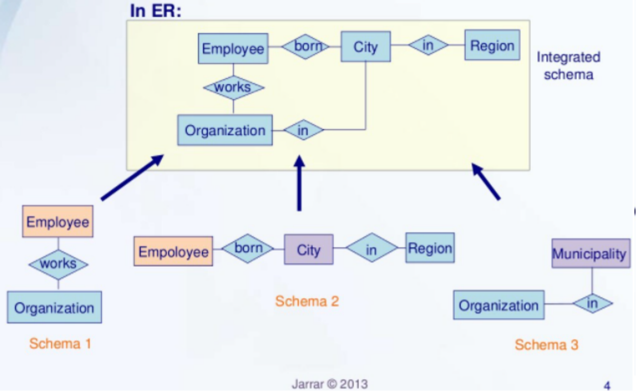
Schema Integration

- Merging multiple schemas into one schema
  1. Schema Transformation
     - ⋆ Have homogenous formats
     - ⋆ Extract the model from data
  2. Schema matching
     - ⋆ Identify items that are semantically related
     - ⋆ e.g. student, graduate student
  3. Schema integration
  4.
     - ⋆ Global-as-View GAV
     - ⋆ Local-as-View LAV

Challenges

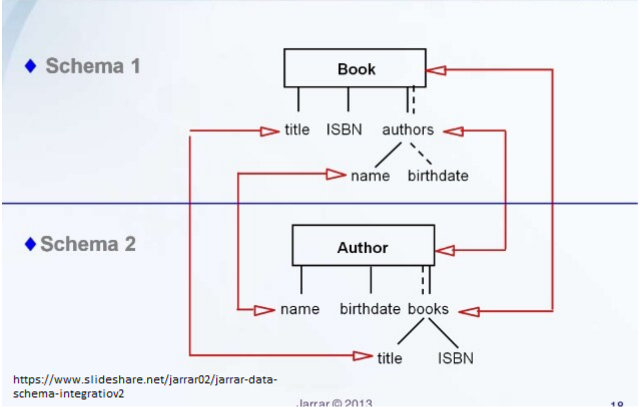
- We need to identify what is the same
- We need to identify conflicts

Examples



https://www.slideshare.net/jarrar02/jarrar-data-schema-integratiov2

Examples



- Schema 1
- Schema 2

https://www.slideshare.net/jarrar02/jarrar-data-schema-integratiov2

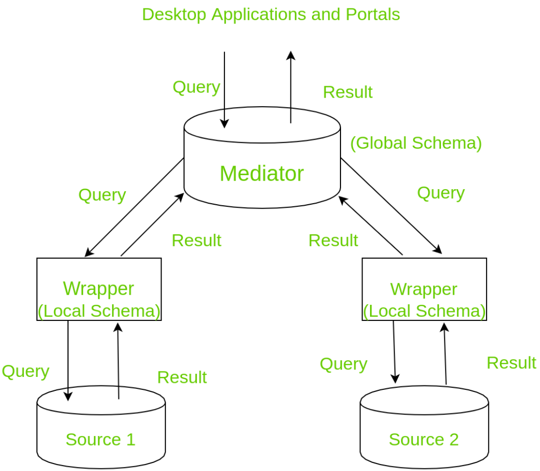
Jarrar © 2013

18

Global-as-View (GAV)

- Mediated schema (MS) is a set of views over the data sources
- The mediator converts a query to the mediator source specific queries

Global-as-View (GAV)



https://www.geeksforgeeks.org/what-is-gav-global-as-view/

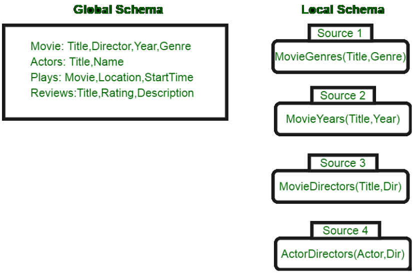
Local-as-View (LAV)

- Each data source is described as precisely as possible
- Each local schema is described as function over global schema

Local-as-View (LAV)

- Each data source is described as precisely as possible
- Each local schema is described as function over global schema
- Describe which data is available in local schema

Local-as-View (LAV)



Global Schema

Movie: Title,Director,Year,Genre
Actors: Title,Name
Plays: Movie,Location,StartTime
Reviews:Title,Rating,Description

Local Schema

Source 1
MovieGenres(Title,Genre)

Source 2
MovieYears(Title,Year)

Source 3
MovieDirectors(Title,Dir)

Source 4
ActorDirectors(Actor,Dir)

https://www.geeksforgeeks.org/local-as-view-lav/

Thank you :)