DATA7001
INTRODUCTION TO DATA SCIENCE

Module 2 Getting the Data I Need

Module Topics

- Types of Data

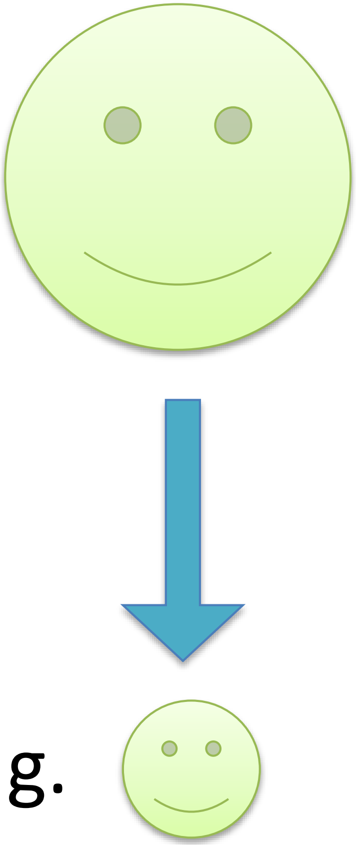
- Data Ingestion

- Managing Data Privacy

- **Sampling Big Data**

(Structured) Data Sampling – Why?

- Reduction of data
  - Volume of data – storage, accessibility
  - Convenience – laptop vs. cluster
  - Smaller dataset with same data structure
  - Generally applicable



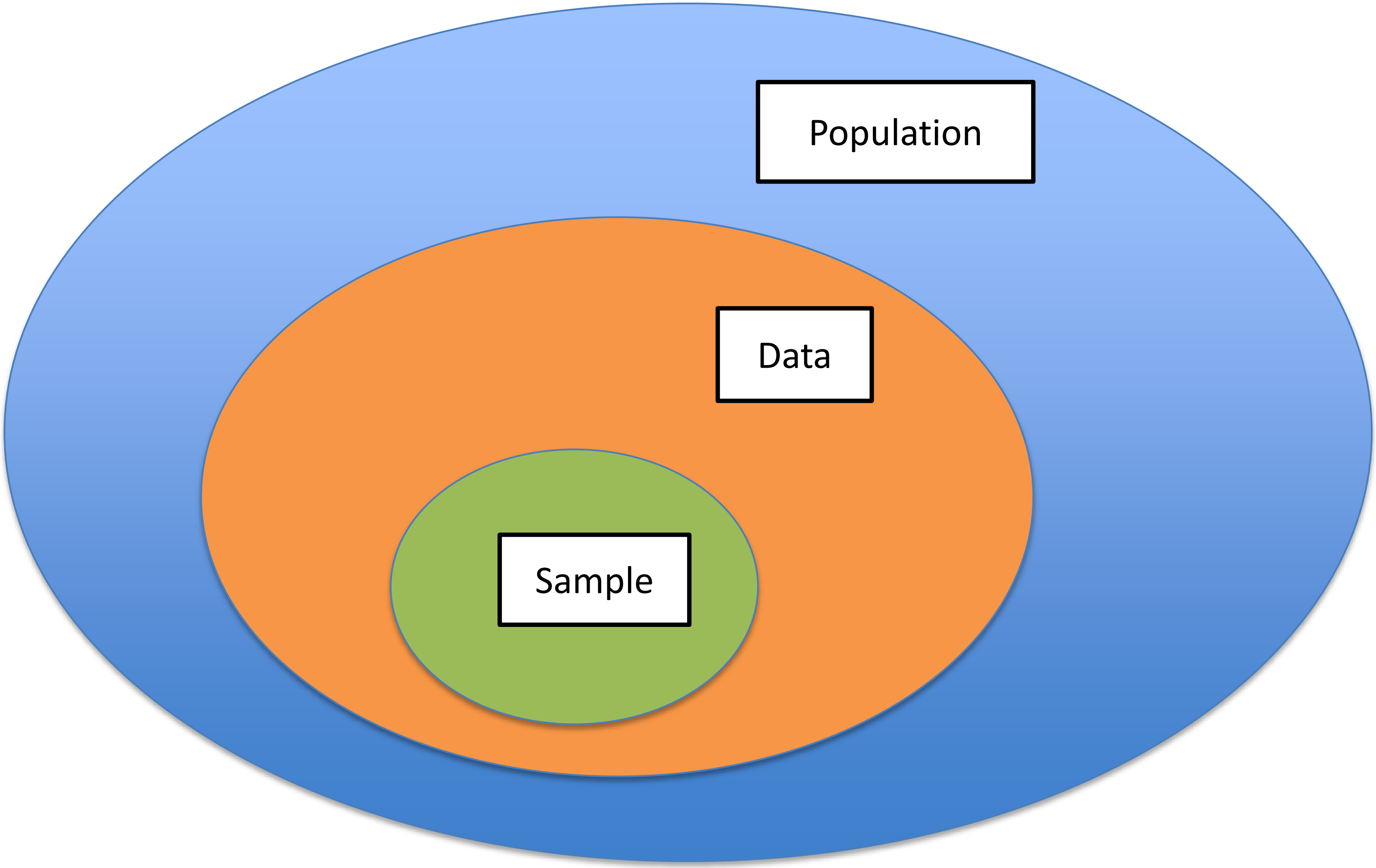
- Other data reduction methods exist (e.g. Summarization, PCA)

Data Sampling – What?

- Select data subset, usually according to probability rules
  - Simple Random Sampling
    - Each item has an equal chance of appearing in the sample
  - Weighted Random Sampling
    - Each item has a weight
    - Appears in sample proportional to weight
  - Stratified Sampling
    - Distinct groups (strata) present in data
    - Maintain representation of all groups in the sample

- Many other approaches (e.g. systematic sampling)

Data Sampling – What?



Data Sampling – What?

- Population
  - Set of items of interest (e.g. individuals, households)
- Data
  - Information pertaining to (usually part) of the population of interest
  - **NB: Often, we only have data on a sample of the population!**
- Sample
  - Subset of data, (random) representative of whole dataset

Data Sampling – How?

- Sampling Without Replacement (WOR)
  - Each time we add an item to the sample, it is excluded from being added again
  - No item is duplicated in the sample
  - Sampled items are DEPENDENT

- Sampling With Replacement (WR)
  - Each time we add an item to the sample, it is NOT excluded from being added again
  - Items could be duplicated in the sample
  - Sampled items are INDEPENENT

- **NB: We will ONLY consider WR!**

Data Sampling – How?

- Simple Random Sampling
  - Given *n* items in the dataset, want to select *m* items for the sample, WR (where $m \ll n$)
  - For each of the *m* items in the sample, choose item *i* in the dataset with probability $p_i = 1/n$

Data Sampling – How?

- Simple Random Sampling

	DATA ITEM	CATEGORY1	VALUE1
	1	F	27
	2	F	21
	3	F	18
	4	F	35
X 2	5	F	31
	6	F	22
	7	M	37
	8	F	21
	9	F	37
	10	M	55

SAMPLE ITEM	CATEGORY1	VALUE1
1	F	21
2	F	31
3	F	31
4	F	21

NB: *Sampling Error* with SRS can lead to loss of data features

Data Sampling – How?

- Weighted Random Sampling
  - Given *n* items in the dataset, each with a (positive) weight w_i, want to select *m* items for the sample, WR (where $m<<n$)
  - For each of the *m* items in the sample, choose item *i* in the dataset with probability p_i proportional to w_i
  - **NB: The weights should be designed to capture data features of particular interest**

Data Sampling – How?

- Weighted Random Sampling (e.g. PPS)

	DATA ITEM	CATEGORY1	VALUE1
	1	F	27
	2	F	21
	3	F	18
	4	F	35
	5	F	31
	6	F	22
	7	M	37
	8	F	21
	9	F	37
X 2	10	M	55

SAMPLE ITEM	CATEGORY1	VALUE1
1	M	55
2	F	35
3	F	37
4	M	55

PPS: Probability Proportional to Size

Data Sampling – How?

- Stratified Random Sampling
  - Given *n* items in the dataset, each belonging to one of *s* strata, want to select *k* items from each stratum giving *m*=*sk* items for the sample, WR (where *m*<<*n*)
  - For each of the *s* strata, choose each of the *k* samples for that stratum uniformly at random (i.e. according to SRS within the stratum)
  - **NB: Strata can be created artificially by selecting ranges of a numerical variable (e.g. income bands)**

Data Sampling – How?

- Stratified Random Sampling

DATA ITEM	CATEGORY1	VALUE1
1	F	27
2	F	21
3	F	18
4	F	35
5	F	31
6	F	22
7	M	37
8	F	21
9	F	37
10	M	55

X 2

SAMPLE ITEM	CATEGORY1	VALUE1
1	F	21
2	F	21
3	M	37
4	M	37

NB: Two samples taken uniformly at random from each category `F’ and `M’

Data Sampling – How?

- In general, stratified random sampling may not sample the same number of items from each stratum.
- Instead, the idea is to make sure that the “right” number of items are sampled from each stratum.
  - E.g. we may want to to preserve the proportion of the strata in some study
  - E.g. we may want to oversample a rare strata in order to perform meaningful statistical analysis on these rare strata.

Data Sampling – When?

- Sampling can occur during or after data collection
  - Here, we focus on the latter case

- Sampling methods (particularly SRS) are also used for analytic purposes (e.g. cross-validation of statistical models)

- Simple Random Sampling is easy; however, can lose data features (e.g. unusual items)

- Weighted Random Sampling or Stratified Sampling can be used to address this problem

Task and Discussion

For each of the three sampling methods, give an example of a dataset for which the method is appropriate.

POLL QUESTIONS - SAMPLING