Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

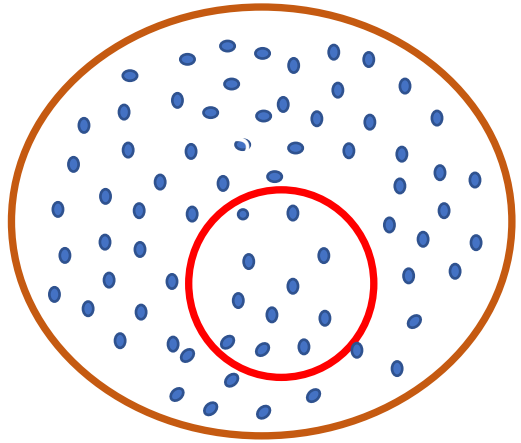
email: soumyad@cse.iitk.ac.in

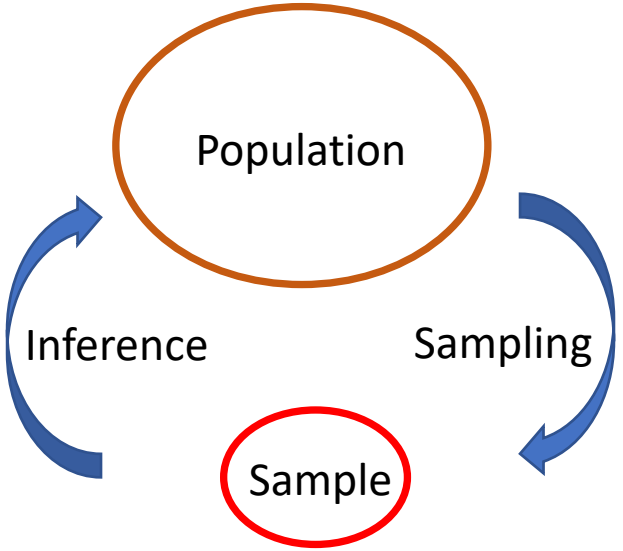
Slots to Check Graded Mid-Sem Paper

- Mar 18th 11am-1pm Tuesday
- Mar 19th 11am-1pm Wednesday
- More slots may become available later

Population, Sample, and Sampling

- "Population" is the entire set of items from which you draw data for a statistical study (sampling frame)
- A "sample" is a subset of a population
- "Sampling" is the process of selecting a subset from a population and is called sample
- Goal of sampling:
  - The primary objective of sampling is to select a subset of data from a large population, which might be impossible to handle and sometimes we cannot even have access to the entire population
    - Data reduction and representative selection for performing statistical analysis and inference about the population
    - Save time, space, and money
    - Practical approach for solving challenging problems





Real-life Motivating Applications

- Applications of Sampling and sample-based analysis is widespread
  - Sampling is everywhere since it is impossible to keep all the data

Sampling is the fundamental tool for any kind of survey

- Suppose we want to measure the average height of the males in India
- Based on our capability, we can measure heights for 10,000 males/day
- India has around 717 million male population*
- It would take 71,700 days, roughly 197 years!
- Would you do it? Do you think this is practical approach even with more resources?
- **What if tomorrow I want to know the average height of the female population in India?**

Sampling to further scientific discovery

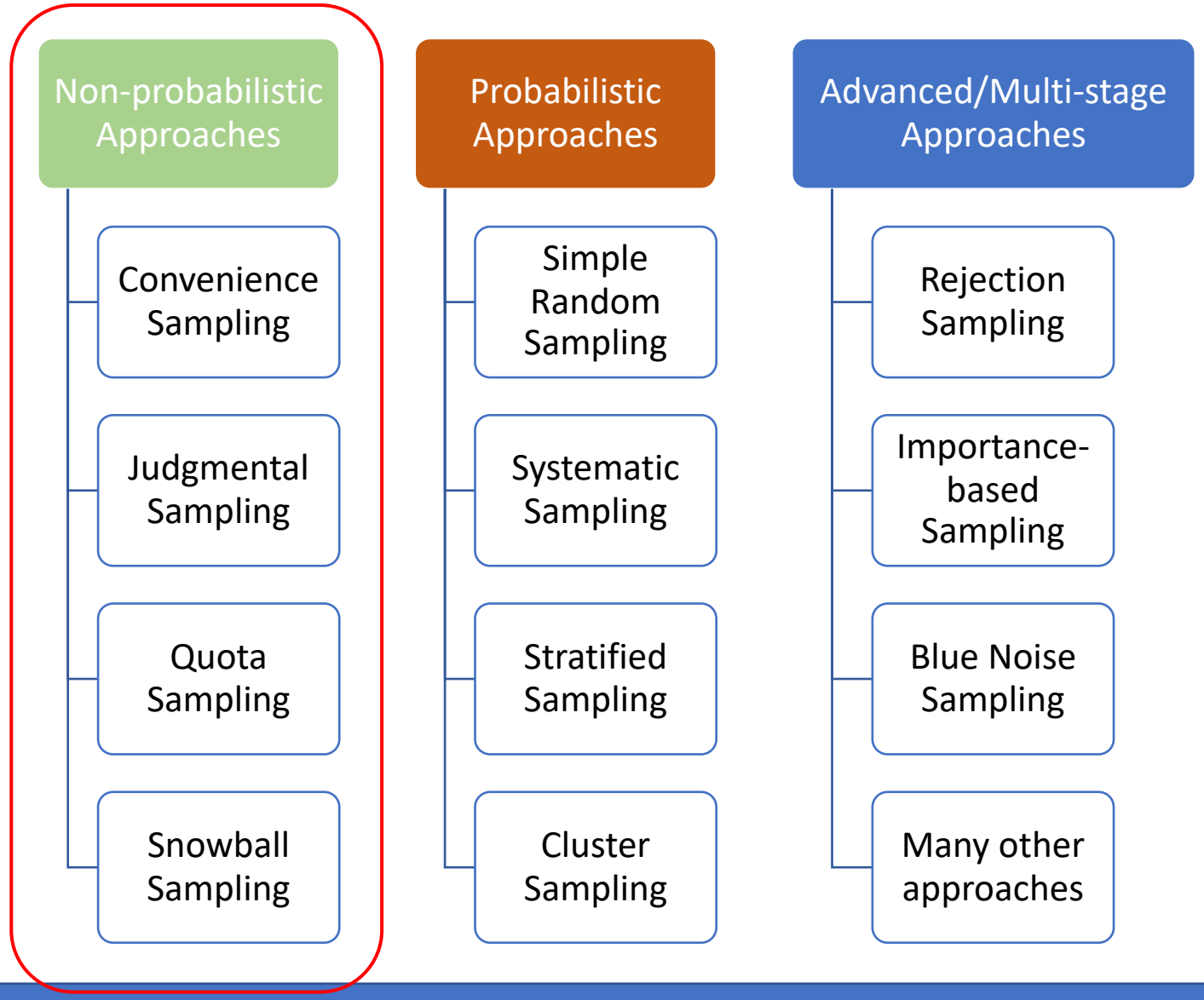
- Suppose we wish to predict climate accurately for the near future or want to understand fundamental physics or want to assess the impact of an asteroid hitting our earth!
- We use large-scale computational simulations that attempts to model these phenomena accurately
- These simulations generate petabytes (10^{15} bytes) of data, soon to reach exabyte (10^{18})
- We simply cannot keep/analyze all the data and even if we try, the cost and resource will be prohibitive

How can/should we sample big data to achieve our goals above?

*https://statisticstimes.com/demographics/country/india-sex-ratio.php

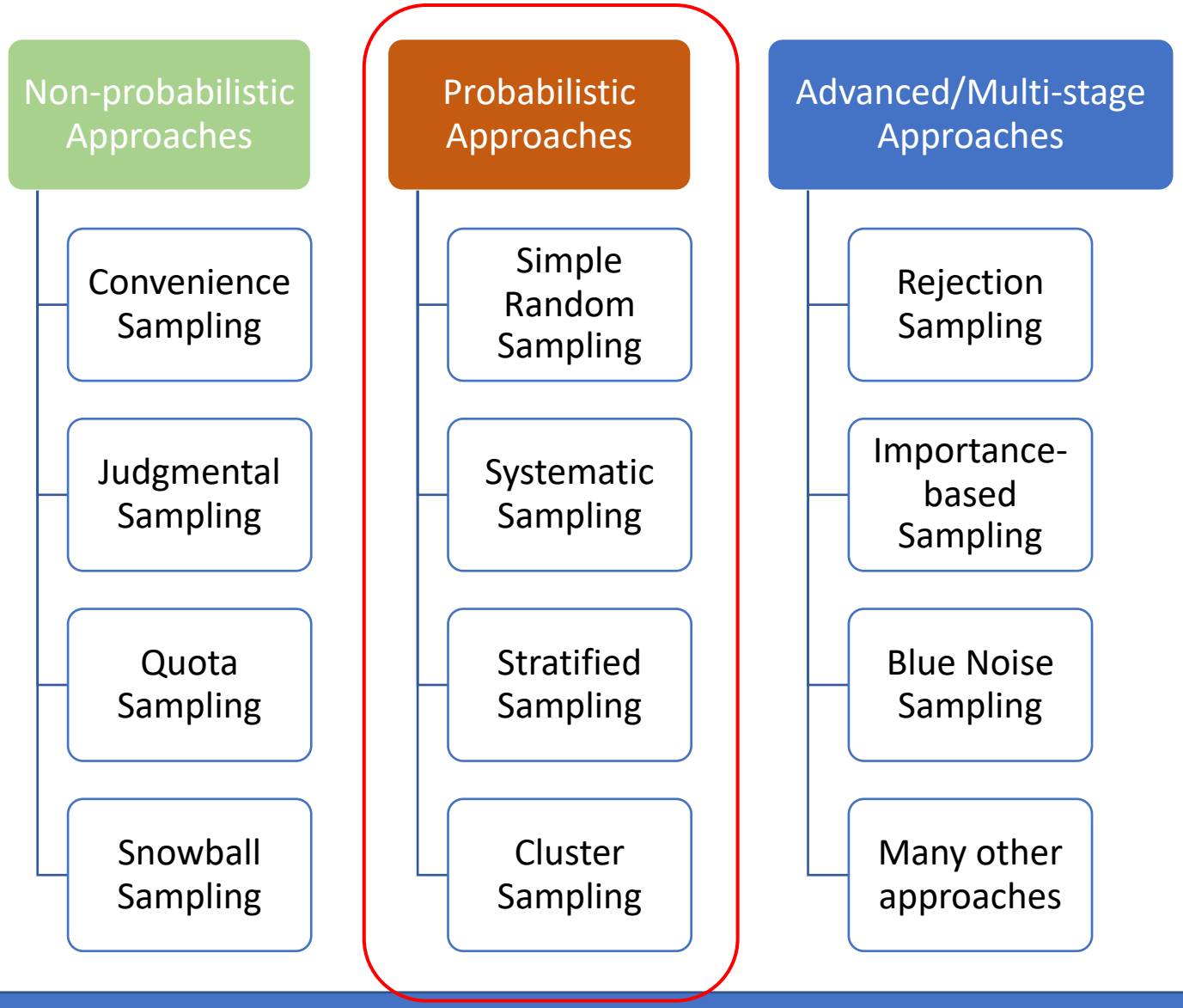
Classification of Sampling Techniques

- Non-probabilistic approaches
  - Items selected by not considering their probability of occurrence



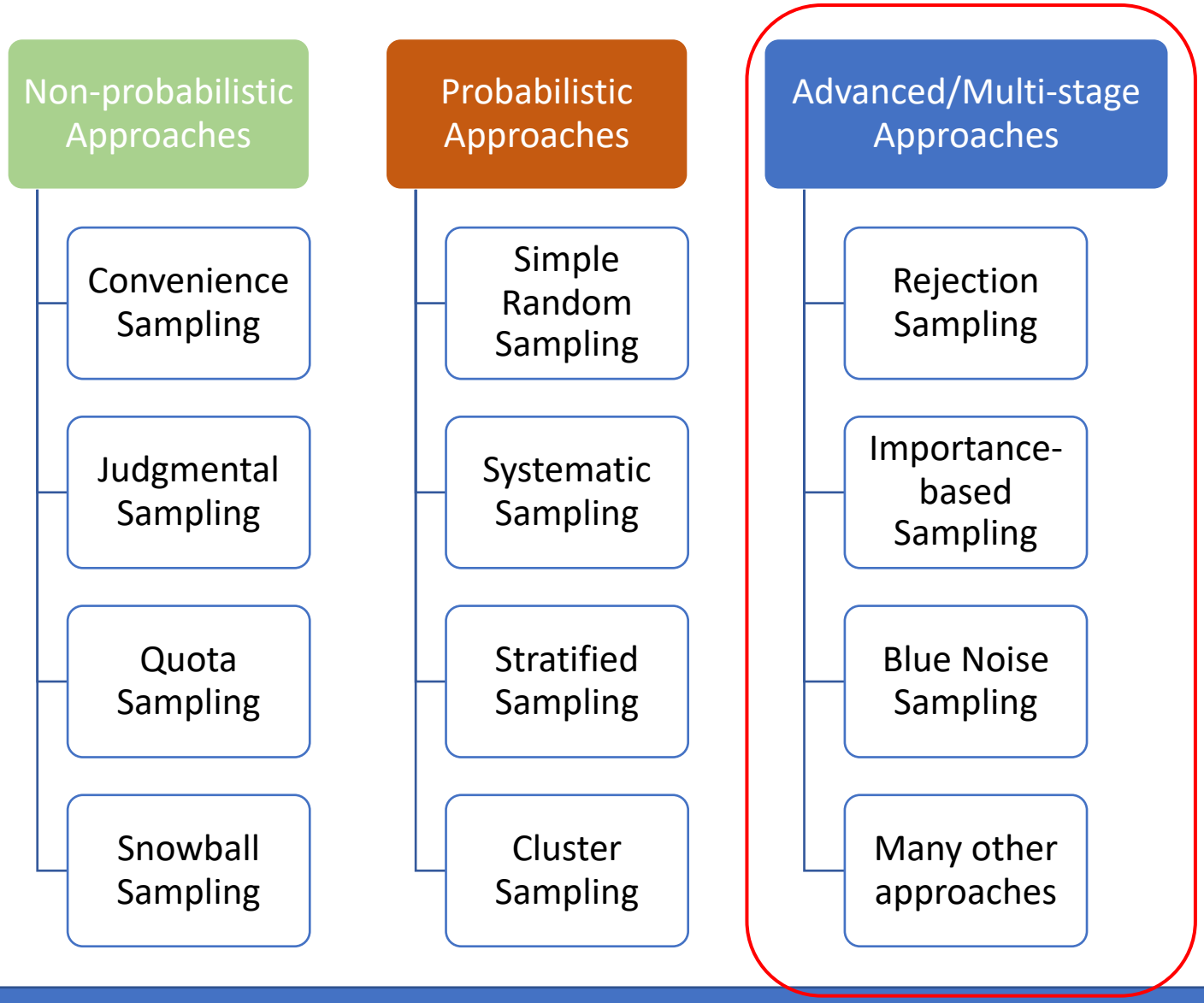
Classification of Sampling Techniques

- Non-probabilistic approaches
  - Items selected by not considering their probability of occurrence
- Probabilistic approaches
  - Items selected based on their occurrence in the population
  - Prevalent in Data Science applications
  - Gives good estimations of statistic



Classification of Sampling Techniques

- Non-probabilistic approaches
  - Items selected by not considering their probability of occurrence
- Probabilistic approaches
  - Items selected based on their occurrence in the population
  - Prevalent in Data Science applications
  - Gives good estimations of statistic
- Advanced approaches
  - Largely probabilistic
  - Often data-driven
  - Sometimes application-specific

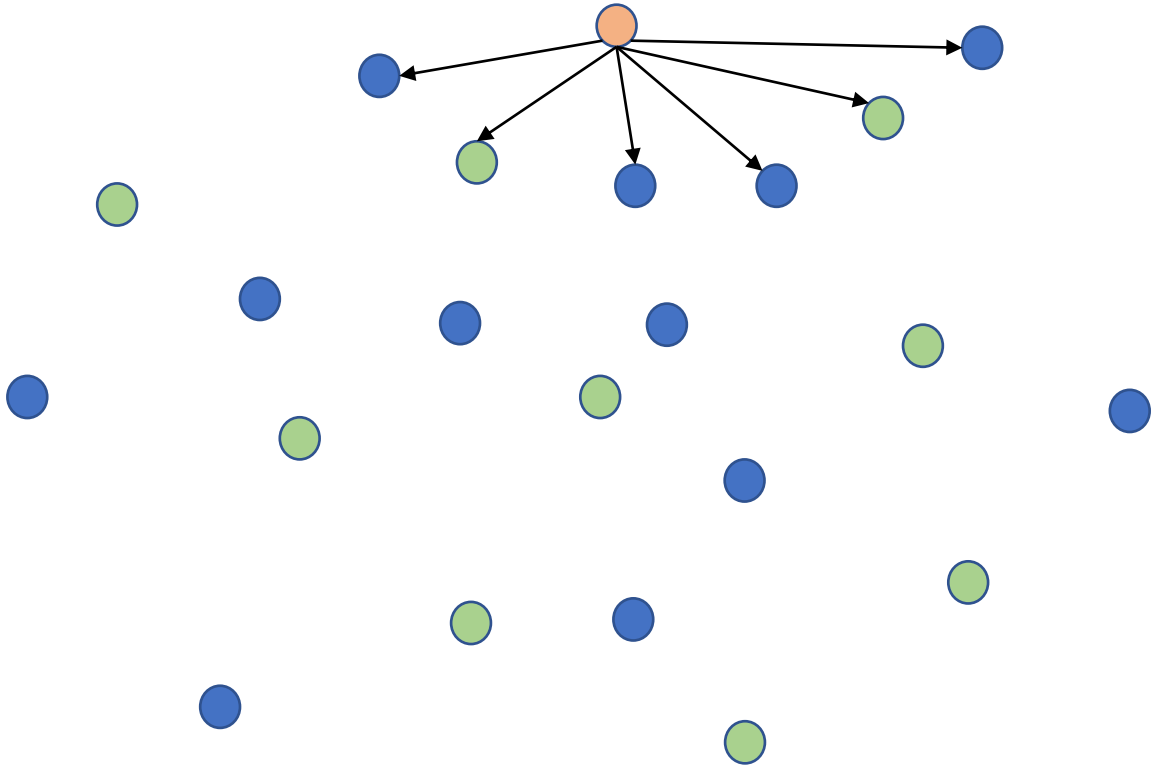


Non-Probabilistic Sampling Approaches

Convenience sampling

- Convenience sampling
  - It is one of the easiest and common form of sampling
  - Sample observations are selected based on ease of accessibility and convenience
  - Sample is not a true representation of the population
  - Generalization and statistical inference using samples generated by convenience sampling may not be accurate
  - Can be used for initial or informal pilot study
  - Also known as **grab sampling** or **accidental sampling**

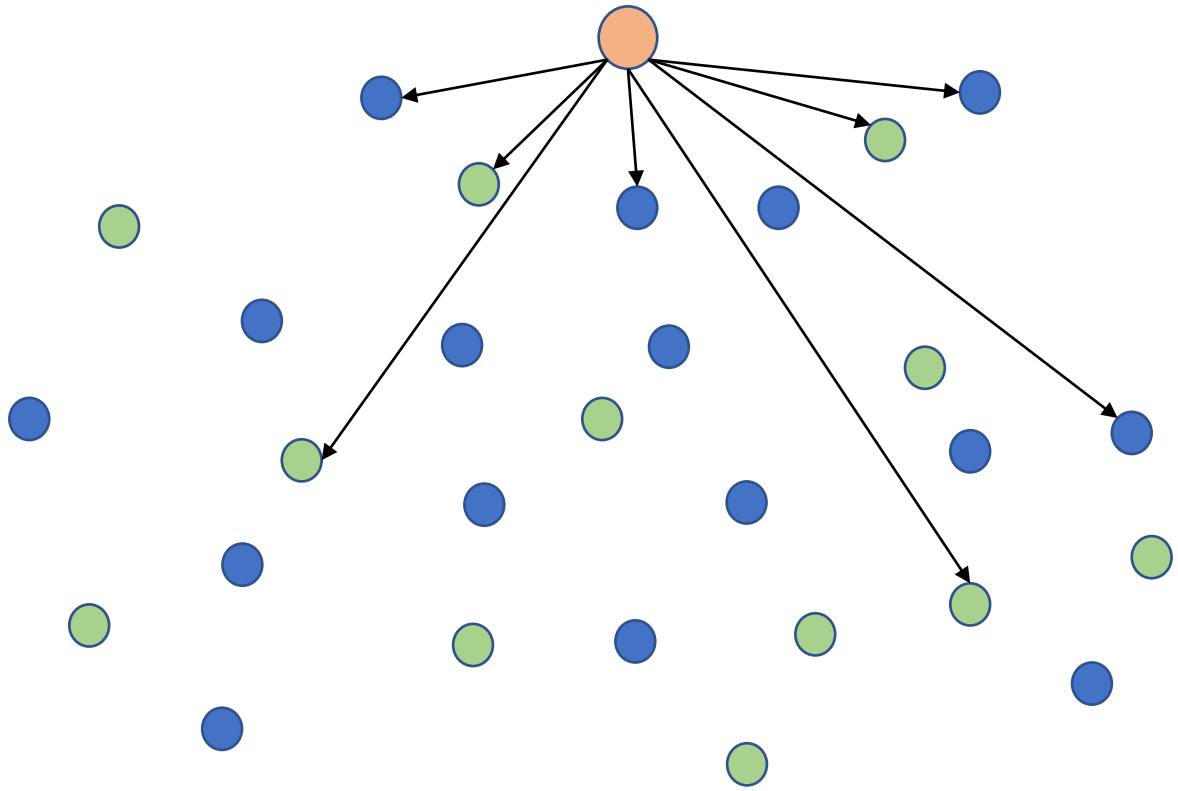
Convenience sampling



Judgmental sampling

- Judgmental sampling
  - It is a non-probabilistic method where existing knowledge is used to select sample observations from the population
  - Sample is not a true representation of the population
  - Generalization and statistical inference using samples generated by convenience sampling may not be accurate
  - Judgmental sampling could be computationally less expensive than others and gives the sample set where the user is interested in

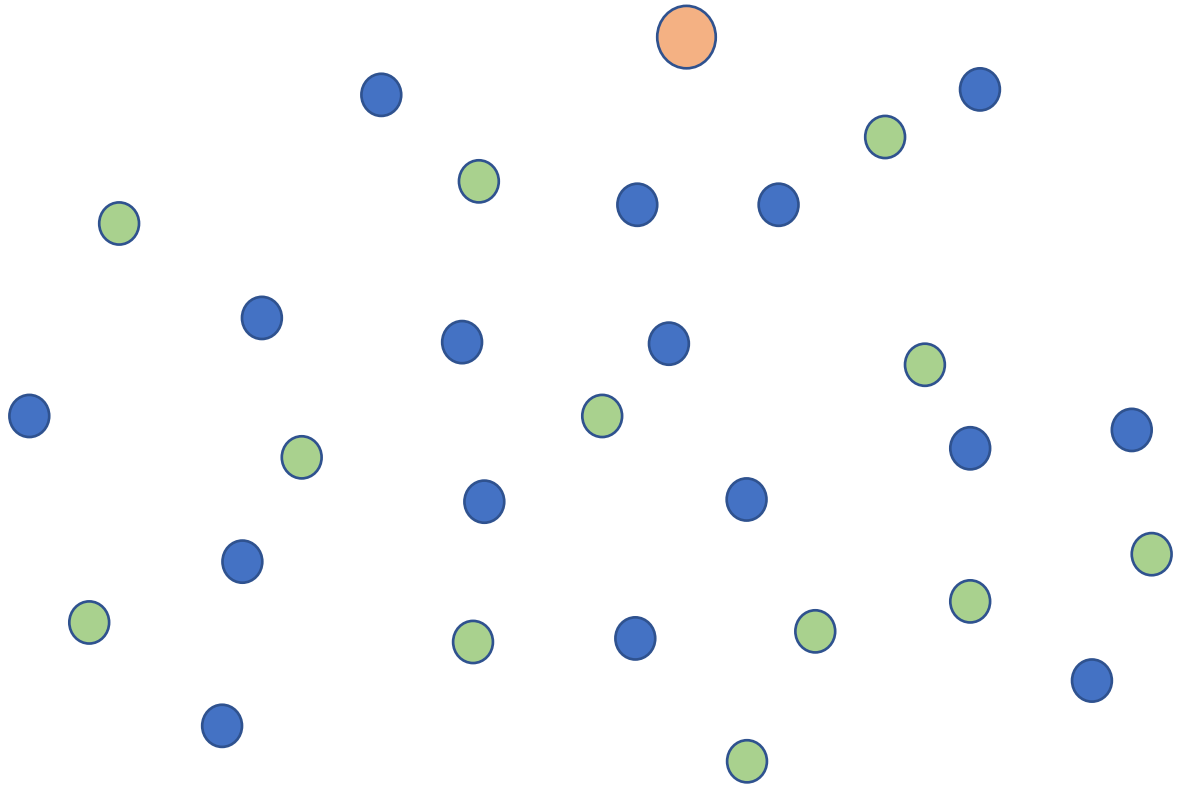
Judgmental sampling



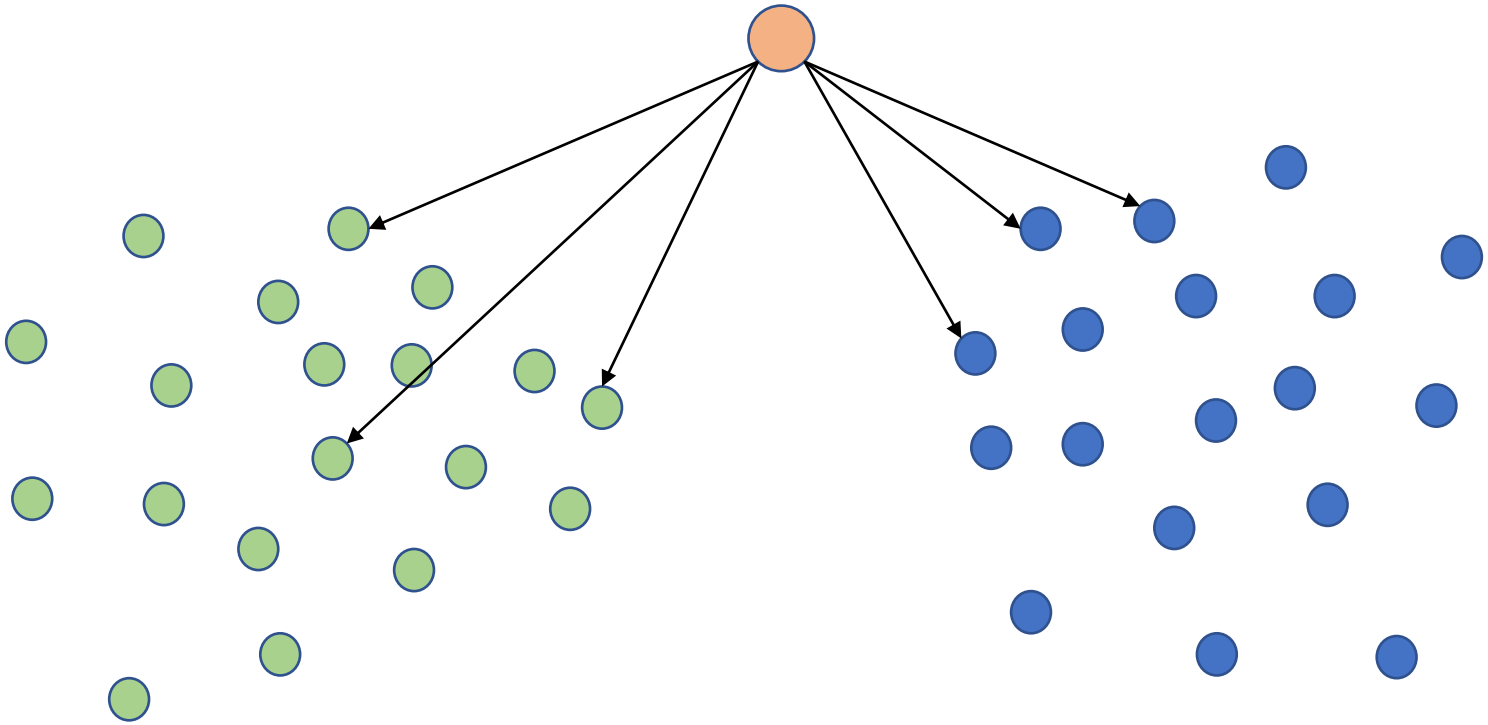
Quota sampling

- Quota sampling
  - It is a non-probabilistic method where sample observations are selected based on some pre-defined ‘quota’
  - First the population is divided into mutually exclusive groups based on certain characteristics and traits
  - Then judgmental sampling is performed inside each group to select observations to satisfy a pre-defined criterion
  - May have bias in selected sample
  - This is a non-probabilistic version of **stratified sampling**

Quota sampling



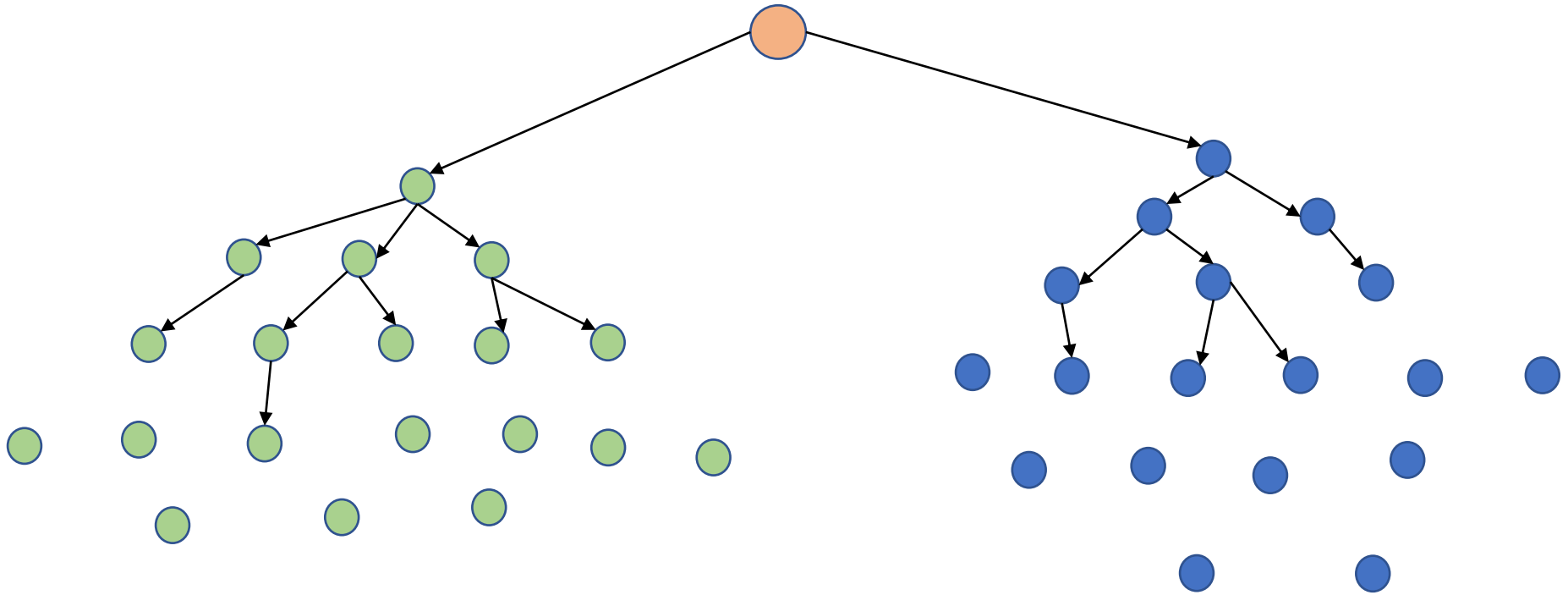
Quota sampling



Snowball sampling

- Snowball sampling
  - It is a non-probabilistic sampling method where the current selected observations dictate how subsequent observations will be selected
  - This is also known as **chain sampling** or **referral sampling**
  - The sampled observations grow like a **rolling snowball**, hence the name
  - The sampling process starts with a small pool of observations and then the selection process propagates via nominations of the initial observations
- This method is heavily used in social computing, graph sampling applications
- Produces a biased estimate of the population but can often reveal hidden patterns

Snowball sampling



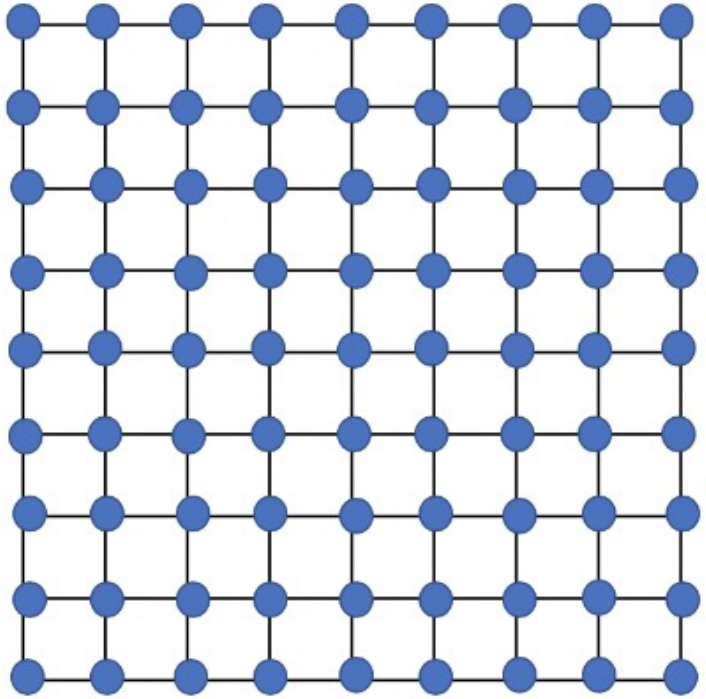
Probabilistic Sampling Approaches

Systematic Sampling

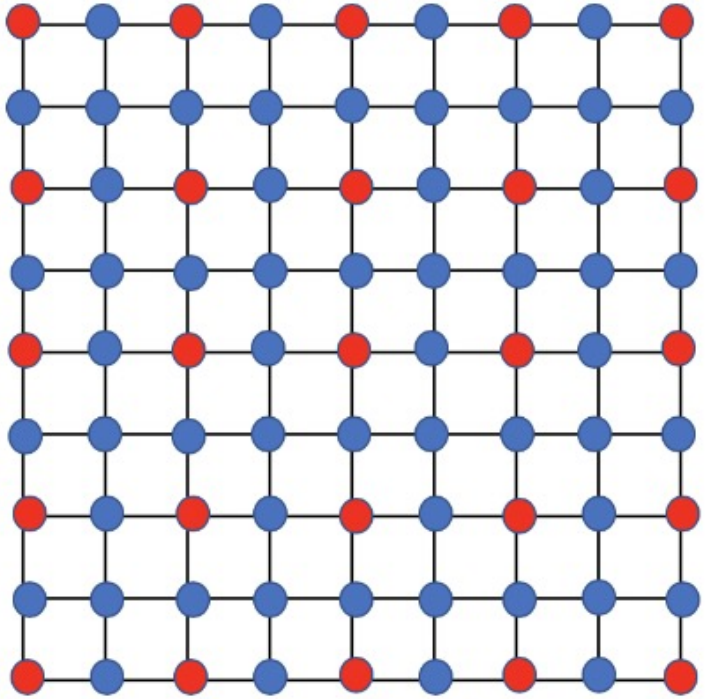
- Observations or data points are selected at regular interval from the population
- Steps:
  - Calculate the sampling interval (I = N/n)
  - Draw a random number (<=I) for the starting data point
  - Draw every I^{th} data point from the starting point
- Ensures a good representativeness of the population in the selected sample

Systematic Sampling

- Observations or data points are selected at regular interval from the population



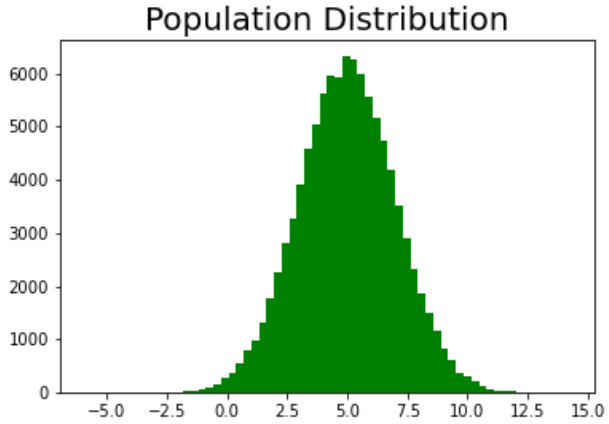
9X9 = 81 data points before sampling



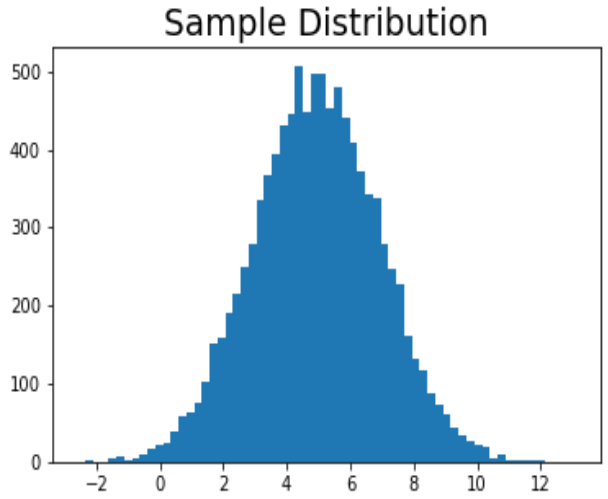
25 data points selected after sampling

Simple Random Sampling (SRS)

- The most basic sampling technique, widely used with favorable properties
  - Provides theoretical basis for the more complicated methods
- Idea: Every item/point in the population has equal probability of being selected
  - If we have N points and we wish to select a sample of n points (n <= N) then each point has initially probability 1/N to get selected
  - In practice, we can generate random numbers using the indices of points to select the desired number of points
- Random sampling gives unbiased estimations about the population
  - Statistic estimated on sample faithfully reflects the statistic about population
    - Mean, variance, higher order moments etc.



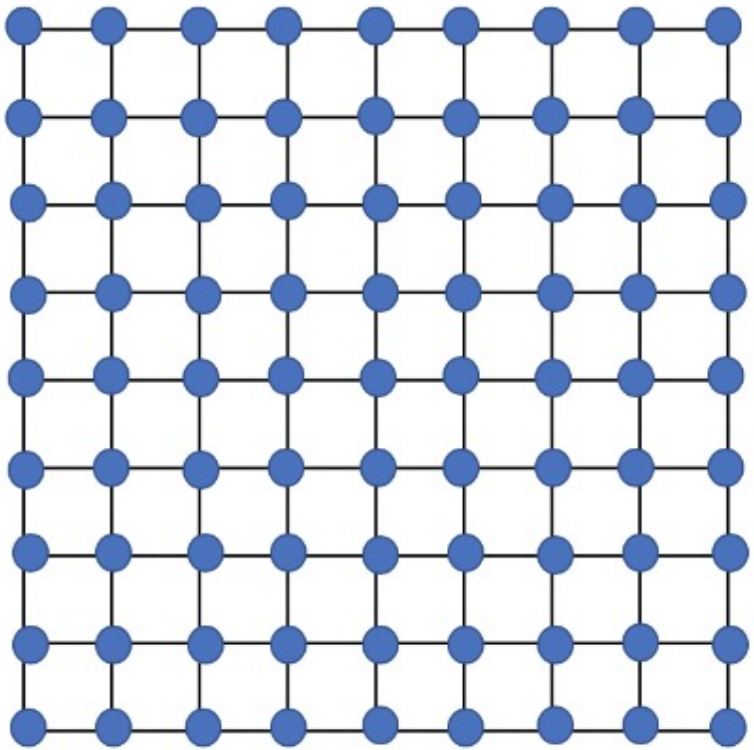
Number of points = 100000
Average (mean): 5.0
Standard Deviation = 2.0



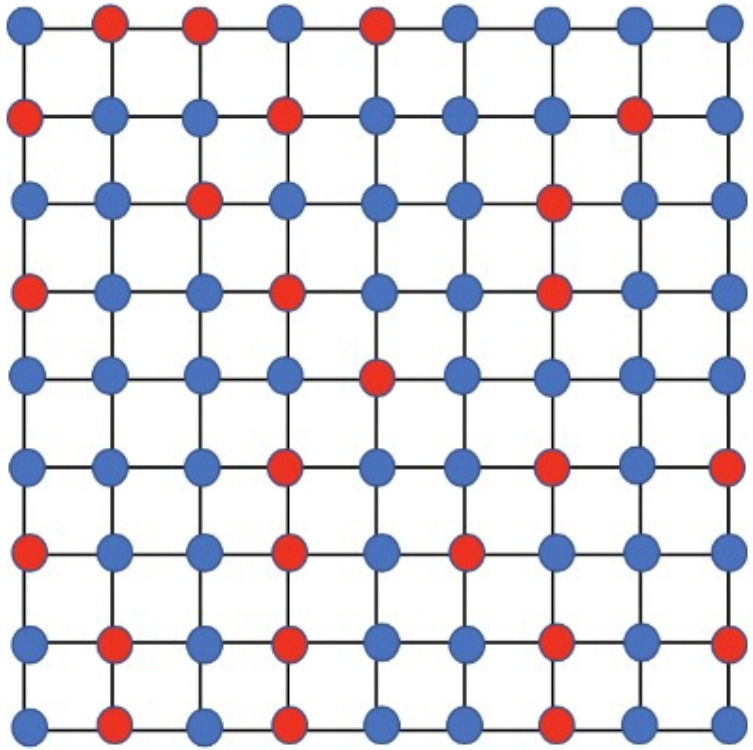
10% sample
Average (mean): 4.974
Standard Deviation = 2.008

Simple Random Sampling (SRS)

- Randomly select points from population



9X9 = 81 data points before sampling



25 data points selected after sampling

Randomization Theory for Simple Random Sampling

- Simple Random Sampling (SRS) gives unbiased estimator about the population
- Let us show that $\bar{y}$ (sample mean) is an unbiased estimator of $\bar{y}_U$ (population mean)
- We are selecting n items from a population of size N, s is the set of items selected
- Let Z_i be an indicator random variable and is defined as follows

$$Z_i = \begin{cases} 1 \; if \; item \; i \; is \; in \; the \; sample \\ 0 \; otherwise \end{cases}$$

Then we have $$\bar{y} = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{y_i}{n}$$

When we select n items out of the N items in the population, and $\{Z_1, Z_2, \ldots, Z_N\}$ are identically distributed Bernoulli random variables, the probability of this event is:

$$P(Z_i = 1) = P(select \; item \; i \; in \; sample) = n/N$$

Randomization Theory for Simple Random Sampling

$$Z_i = \begin{cases} 1 \; if \; item \; i \; is \; in \; the \; sample \\ 0 \; otherwise \end{cases}$$

$$P(Z_i = 1) = P(select \; item \; i \; in \; sample) = n/N$$

Then we have $$\bar{y} = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{y_i}{n}$$

We also see that since Z_i is a Bernoulli random variable,

$$E[Z_i] = \mathrm{P}(Z_i = 1) = \mathrm{n/N}$$

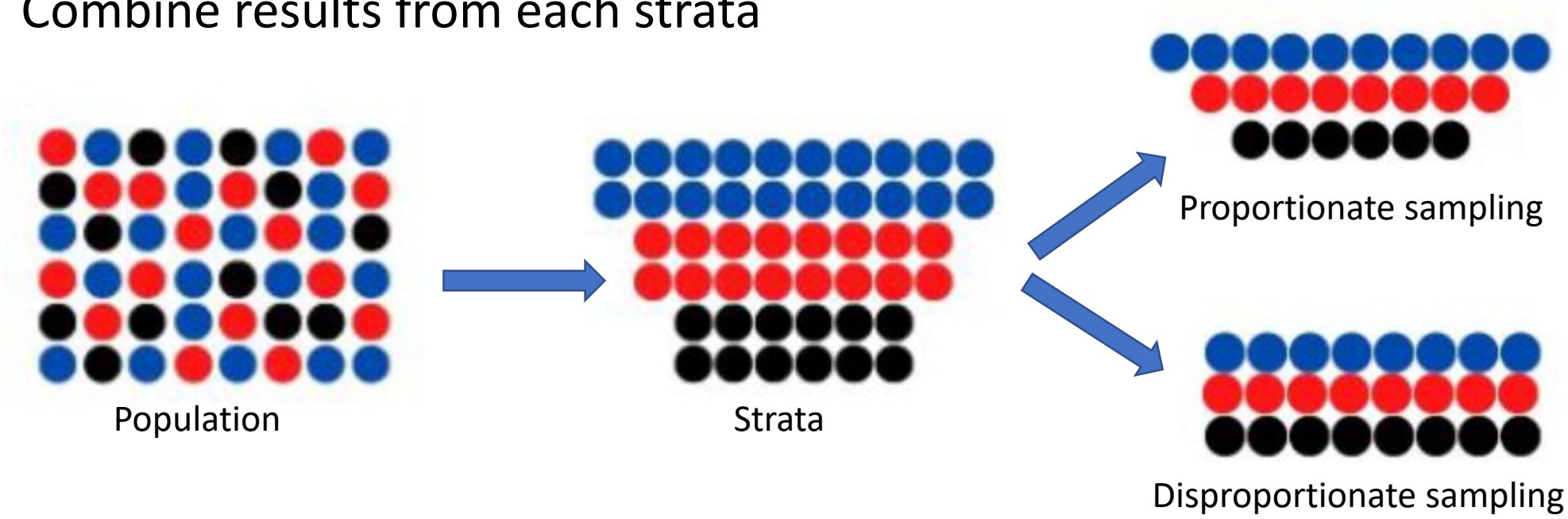
and finally, we have $$E[\bar{y}] = E\left[\sum_{i=1}^{N} Z_i \frac{y_i}{n}\right] = \sum_{i=1}^{N} E[Zi] \frac{y_i}{n} = \sum_{i=1}^{N} \frac{n}{N} \frac{y_i}{n} = \sum_{i=1}^{N} \frac{y_i}{N} = \bar{y}_U$$

Stratified Sampling

- Classify the population into several homogeneous strata
  - This process is called stratification
- Determine the sample size
- Randomly sample points from each strata
  - Disproportionate sampling
  - Proportionate sampling
- Combine sampled results from each strata

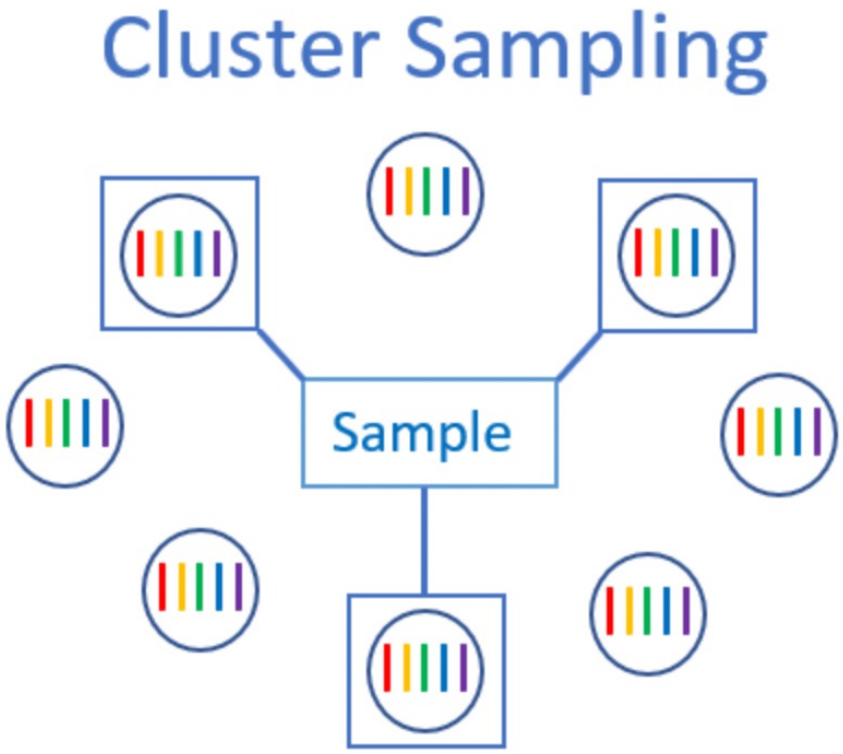
Stratified Sampling

- Classify the population into several homogeneous strata
- Determine the sample size
- Randomly sample points from each strata
  - Disproportionate sampling
  - Proportionate sampling
- Combine results from each strata



Cluster Sampling

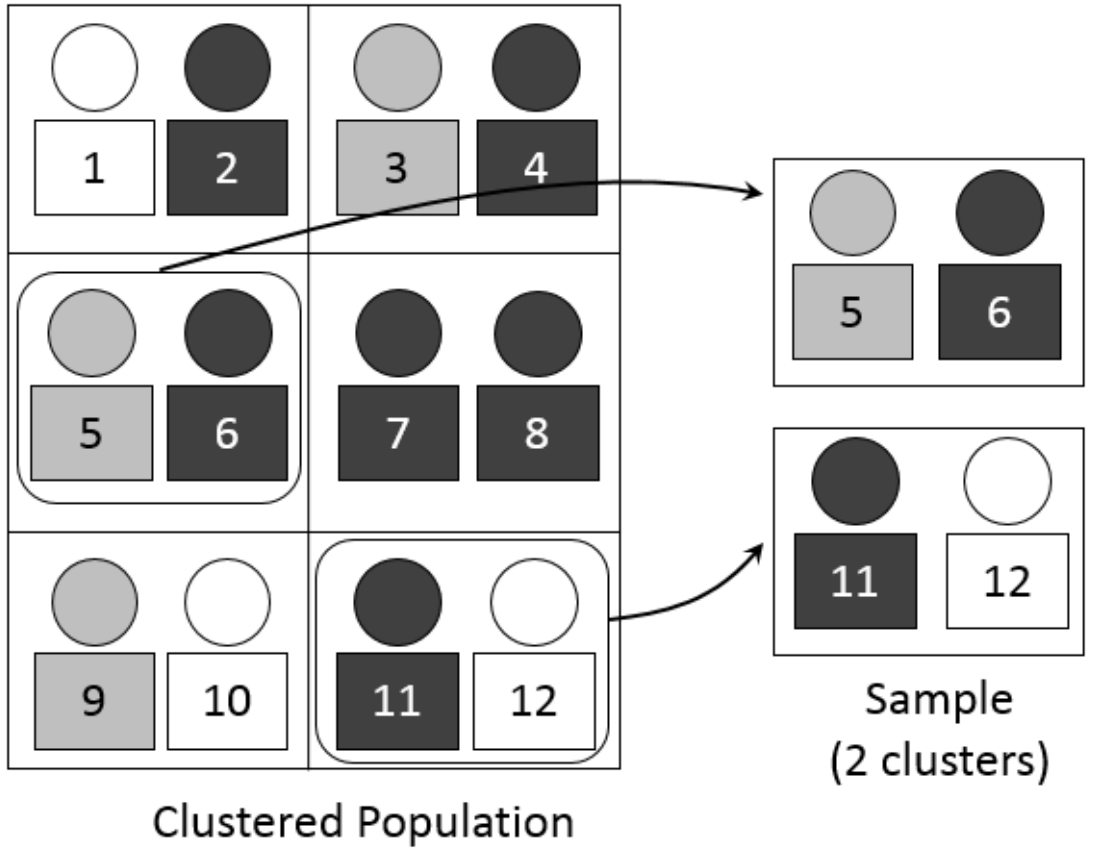
- The population is first clustered into mutually exclusive heterogeneous groups
- The clustering is done based on some global criteria
- Each cluster must represent the population as best as possible
- Clusters are internally heterogeneous but externally homogeneous
- Then sample is selected from a randomly selected single or multiple group of clusters



https://statisticsbyjim.com/basics/cluster-sampling/

Single-stage Cluster Sampling

- After the clusters are formed, either a single or a set of clusters are selected at random
- All the data points in such clusters are combined for analysis
- This is suitable when the data set is not too large, and a subset of clusters can be handled efficiently



Two-stage Cluster Sampling

- After the clusters are formed, either a single or a set of clusters are selected at random
- Simple random sampling is performed inside each cluster to select subset of data points from each selected cluster
- All the selected points are are combined for analysis
- This is more of a practical approach that can handle relatively large data sets as two steps of filtering is applied

Cluster Sampling: Advantages

- If the cluster generation process produces clusters that are very similar to the entire population and each cluster can represent it well, then using cluster sampling method reliable results can be produced
- Often suitable for large scale data sets
- Applicable when the entire population is impossible to access

Advanced Sampling Approaches

Inverse Transform Sampling

- What happens behind the scene when we generate sample points from a specific type of distribution
- Set up:
  - We have numbers between $U \sim [0,1]$ coming from a Uniform distribution
  - We want points that follow $\text{Exponential}(\lambda)$ distribution

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Pdf of exponential distribution

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Cdf of exponential distribution

Inverse Transform Sampling

- We have uniform numbers between $U \sim [0,1]$ coming from a Uniform distribution
- We want points that follow Exponential(λ) distribution

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

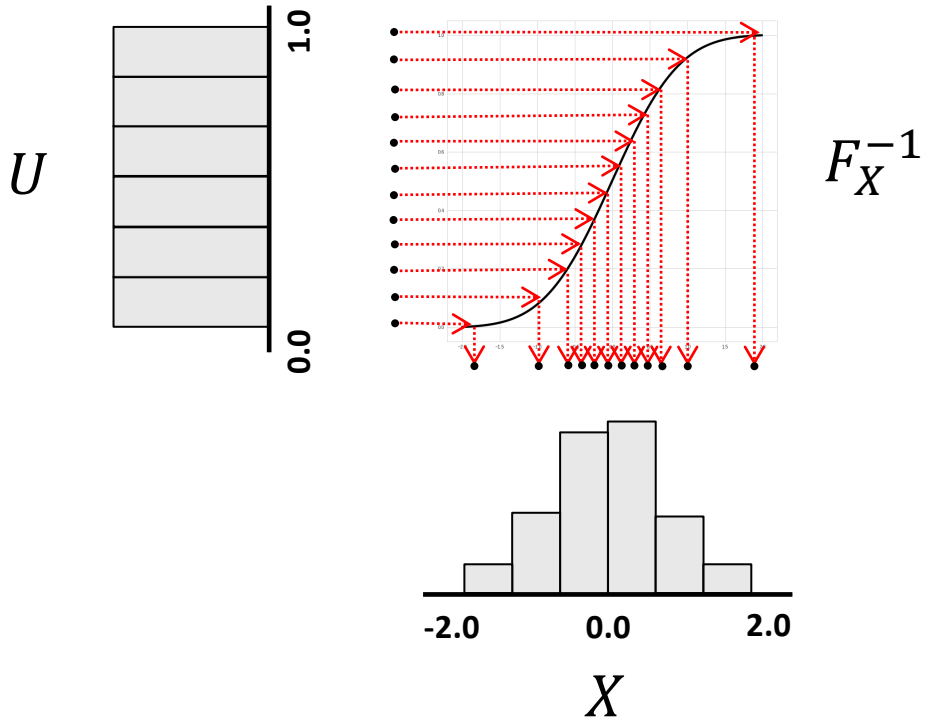
- We want $T(U) = X$ so that we transform uniform numbers to Exponential distribution

$$F(X) = P(X \leq x) = P(T(U) \leq x) = P\left(U \leq T^{-1}(x)\right) = T^{-1}(x)$$

So, we have, $F(X) = T^{-1}(x)$ *hence,* $T(x) = F^{-1}(X)$

Remember: Distribution Transformation Property

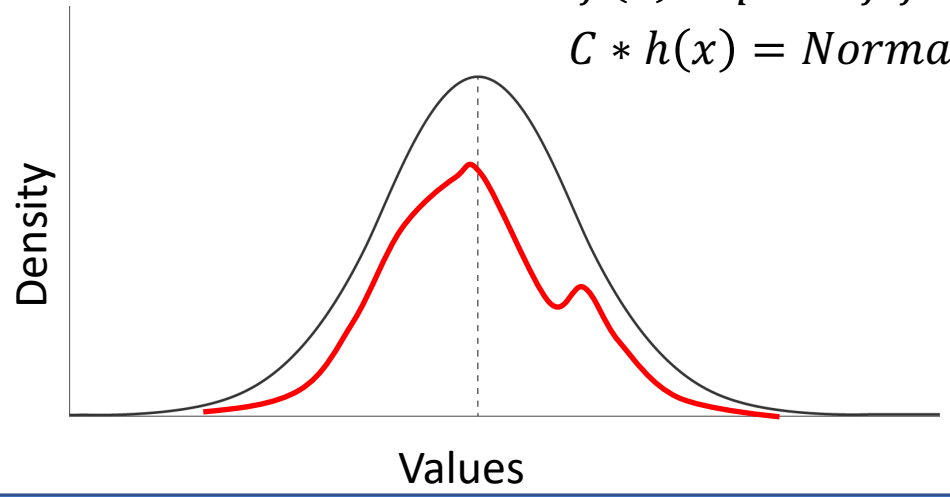
- If U is a uniform random variable (i.e., $U \sim Unif[0,1]$) and F_X is a CDF of random variable X, then its inverse function $F_X^{-1}(U)$ corresponds to the random variable X (i.e., $F_X^{-1}(U) \sim X$)



Rejection Sampling

- Rejection sampling is a method for generating samples from a function $f(x)$ with distribution with density $P(x)$ by drawing sample points from another distribution $h(x)$ that is easier to sample
  - When we do not have $P(x)$ or its CDF is difficult to compute
- Steps:
  - Generate a sample point from $h(x)$
  - Accept the sample point with acceptance prob: $\frac{f(x)}{C * h(x)}$, C is a constant to ensure $C * h(x) \geq f(x)$

$$f(x) = plot\ of\ function\ (Red)$$
$$C * h(x) = Normal\ Distribution\ (Black)$$



Intuition: High acceptance probability for a specific sample point drawn from $h(x)$ indicates that the sample is highly likely for distribution $P(x)$ and so accept it.

- C is a normalizing constant to ensure $C * h(x) \geq f(x)$ as the ratio $\frac{f(x)}{C*h(x)}$ is interpreted as probability value

Why Rejection Sampling Works?

- We have $f(x)$ and the PDF of x is $\mathrm{P}(x)$. We do not know $\mathrm{P}(x)$!
- General form of $\mathrm{P}(x) = \frac{f(x)}{NC}$. *NC* is normalizing constant
- General form of $\mathrm{NC} = \int_{-\infty}^{\infty} f(x)dx$
- From Bayes' theorem:

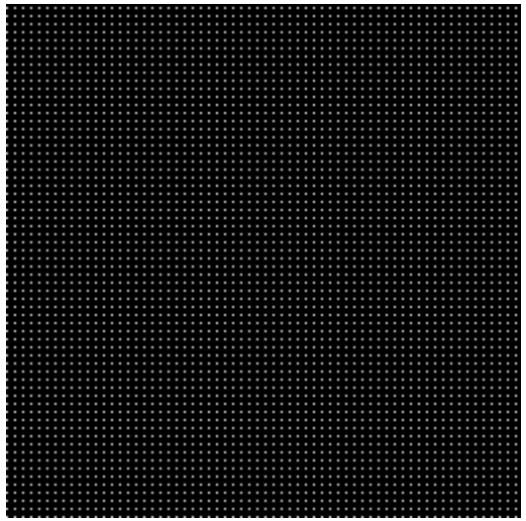
$$D(x|A) = \frac{P(A|x)D(x)}{P(A)} = \frac{\frac{f(x)}{C*h(x)}*h(x)}{P(A)}$$

$$P(A) = \int_{-\infty}^{\infty} h(x) * \frac{f(x)}{C*h(x)}dx = \frac{1}{C}\int_{-\infty}^{\infty} f(x)dx = \frac{NC}{C}$$

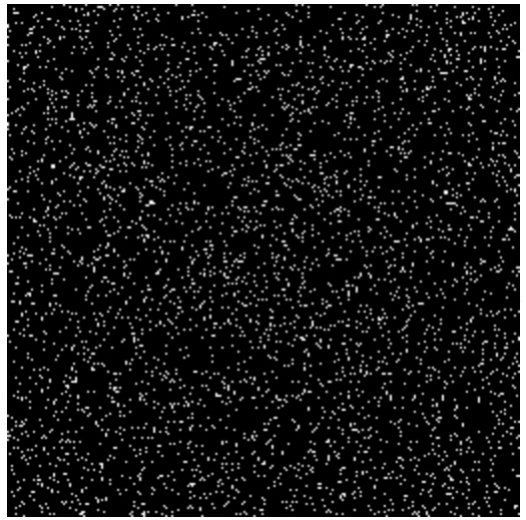
So, $D(x|A) = \frac{\frac{f(x)}{C}}{\frac{NC}{C}} = \frac{f(x)}{NC} = P(x)$

Blue Noise Sampling

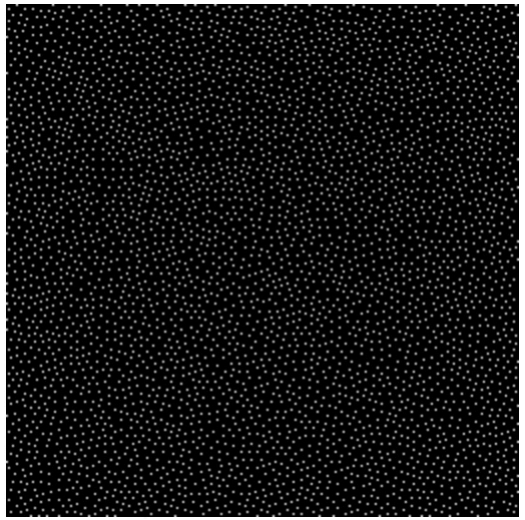
- Blue noise: Characterized by a power spectral density that decreases logarithmically with frequency
  - Blue noise has **more energy at higher frequencies** and less at lower frequencies
  - Blue noise sampling: Blue noise sampling aims to create a distribution of points that appears random but avoids clustering or patterns, resulting in a more uniform and visually appealing distribution
    - emphasize higher frequencies while suppressing lower frequencies



Systematic



Random



Blue Noise

Blue Noise Sampling

- Poisson disk / Dart Throwing for Blue noise sampling
- No two samples withing a radius *r* are allowed
- Sample points are picked from a uniform distribution, and the sample points that obey the minimum distance property with respect to the sample points currently in the set are kept, while the others are discarded.