**Sampling Techniques**

Sampling techniques are methods used to select a subset of individuals or observations from a larger population to make inferences about the population. Here are some common sampling techniques, their definitions, use cases, and examples with real-world datasets used in machine learning (ML) or artificial intelligence (AI):

1. Simple Random Sampling

**Definition**: Every member of the population has an equal chance of being selected.

**Use Case**: When the population is homogeneous, and you want to ensure each member has an equal opportunity to be included.

**Example**: Selecting 1000 customers from a database of 100,000 customers for a satisfaction survey.

**Real-World Dataset**: In ML, this could be used when selecting a subset of images from a large image dataset like ImageNet for initial model training.

2. Stratified Sampling

**Definition**: The population is divided into strata, or groups, based on a characteristic, and random samples are taken from each stratum.

**Use Case**: When you want to ensure representation from different subgroups within the population.

**Example**: Dividing a population into age groups and then randomly sampling from each age group to ensure age representation.

**Real-World Dataset**: In ML, this could be used when creating a balanced dataset from an imbalanced dataset, such as ensuring equal representation of different classes in a fraud detection dataset.

Here are five examples of stratified sampling:

1. **Surveying College Students:** Divide students into strata based on their major (e.g., Science, Arts, Business) and sample proportionally from each group.
2. **Customer Satisfaction Survey:** Divide customers into groups based on age ranges (e.g., 18-25, 26-35, 36-45) and sample from each age group.
3. **Medical Research:** Stratify patients by gender (male, female) and sample equally from each group.
4. **Employee Feedback:** Group employees by department (e.g., HR, Sales, IT) and sample a fixed number of employees from each department.
5. **Political Polling:** Stratify voters by region (e.g., North, South, East, West) and sample proportionally from each region.

3. Systematic Sampling

**Definition**: Every kth member of the population is selected.

**Use Case**: When a random sample is needed but a list of the population is available.

**Example**: Selecting every 10th person from a list of attendees at a conference.

**Real-World Dataset**: In ML, this could be used when creating a dataset from a log file, selecting every nth entry to create a sample dataset.

Here are five additional examples of systematic sampling:

1. **Store Inventory Check:** Check every 15th item in the inventory list for stock levels.
2. **Traffic Study:** Record the speed of every 20th car passing a certain point on a highway.
3. **Restaurant Feedback:** Select every 10th customer from the reservation list to provide feedback.
4. **Water Quality Testing:** Take a water sample from every 7th well in a region.
5. **Fitness Survey:** Measure the fitness levels of every 5th member at a gym.

4. Cluster Sampling

**Definition**: The population is divided into clusters, some of which are randomly selected, and all members of the selected clusters are sampled.

**Use Case**: When the population is large and widely dispersed geographically.

**Example**: Dividing a city into neighborhoods and then randomly selecting some neighborhoods to survey all households within them.

**Real-World Dataset**: In ML, this could be used when dealing with geographically distributed data, such as selecting specific regions from a national dataset for detailed analysis.

Here are five examples of cluster sampling:

1. **School Research:** Randomly select a few schools from a city and survey all the students within those selected schools.
2. **Agricultural Study:** Randomly choose several farms in a region and collect soil samples from all fields within those farms.
3. **Employee Survey:** Randomly select several branches of a company and survey all employees in the chosen branches.
4. **City Traffic Analysis:** Select a few neighborhoods within a city randomly and study traffic patterns on all streets within those neighborhoods.
5. **Healthcare Study:** Randomly choose a few hospitals and examine all patients admitted within a certain timeframe in those hospitals.

5. Snowball Sampling

**Definition**: Existing study subjects recruit future subjects from among their acquaintances.

**Use Case**: When the population is hard to reach or identify.

**Example**: Studying a hidden population like drug users where subjects refer others.

**Real-World Dataset**: In ML, this could be used in network analysis, such as mapping social networks where participants help recruit other participants.

Here are five examples of snowball sampling:

1. **Social Network Research:** Start by interviewing a few key individuals in a specific social network, and then ask them to refer others in the network who meet the study criteria.
2. **Drug Use Study:** Begin with a few known drug users and ask them to refer other users they know for the study.
3. **Homeless Population Survey:** Start with a few known homeless individuals and ask them to introduce other homeless people they know.
4. **Undocumented Immigrants Study:** Interview a few undocumented immigrants and request they refer others in similar situations.
5. **Rare Disease Research:** Identify a few patients with a rare disease and ask them to refer other patients they know with the same condition.

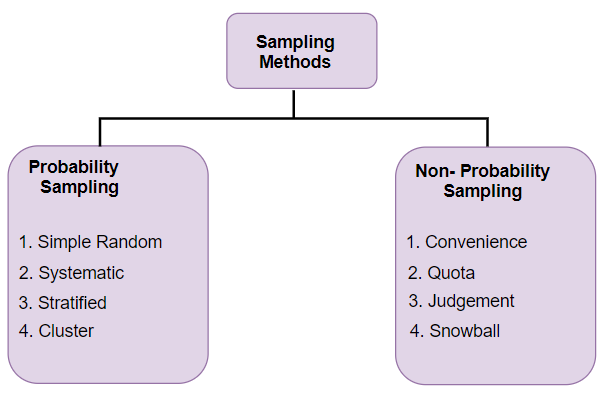
Example Use Cases with Datasets

* **Simple Random Sampling**: In an ML project predicting customer churn, a simple random sample of customer data could be used for training the model.
* **Stratified Sampling**: In a healthcare dataset, ensuring equal representation of different age groups to train a model predicting disease outcomes.
* **Systematic Sampling**: Analyzing time-series data by selecting every nth data point to create a manageable dataset for training a predictive model.
* **Cluster Sampling**: Using satellite images divided by geographical regions to train a model for land use classification.
* **Snowball Sampling**: Building a dataset for social network analysis by using initial participants to recruit others.

Each sampling method has its strengths and weaknesses, and the choice of technique depends on the research question, the nature of the population, and the resources available.

<https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/>

### Simple Random Sampling

1. **Student Survey:** A university randomly selects 400 out of 20,000 students to survey about online course satisfaction. 
2. **Health Survey:** A health organization randomly selects 1,000 individuals from a city population of 100,000 to study dietary habits.
3. **Employee Feedback:** A company randomly selects 50 employees from a staff of 2,000 to participate in a job satisfaction survey.
4. **Customer Satisfaction:** An e-commerce company randomly selects 1,000 customers from their database of 50,000 to survey about their shopping experience.
5. **Environmental Study:** Researchers randomly select 200 out of 10,000 trees in a forest to measure the impact of pollution.

### Systematic Sampling

1. **Factory Quality Control:** A factory inspects every 50th product from a daily production of 10,000 units.
2. **Library Book Check:** A library checks every 10th book in their collection of 5,000 books for damage.
3. **Customer Survey:** A retail store surveys every 25th customer that enters the store over a week.
4. **Highway Traffic Study:** Traffic analysts record the speed of every 20th car passing a checkpoint.
5. **Warehouse Inventory:** A warehouse manager checks every 100th item in a stock of 20,000 items for accuracy.

### Stratified Sampling

1. **Political Survey:** A political analyst divides a population into age groups (18-29, 30-44, 45-59, 60+) and randomly selects individuals from each group to ensure representation.
2. **Employee Survey:** A company divides employees into departments (HR, IT, Sales, Marketing) and randomly selects individuals from each department for a satisfaction survey.
3. **Medical Research:** A study divides participants by gender and age groups to ensure a representative sample of the population.
4. **Education Study:** Researchers divide students into grade levels (freshman, sophomore, junior, senior) and randomly select students from each level for a survey on study habits.
5. **Market Research:** A company divides consumers into income brackets and randomly selects individuals from each bracket to study purchasing behavior.

### Cluster Sampling

1. **School Study:** A researcher divides a city into clusters of schools and randomly selects 10 schools. All students in these schools are surveyed about their dietary habits.
2. **Urban Development Survey:** A city is divided into neighborhoods (clusters), and five neighborhoods are randomly selected for a survey on community services.
3. **Agricultural Study:** A region is divided into clusters of farms, and a few clusters are randomly selected. All farms within these clusters are studied for crop yield.
4. **Healthcare Survey:** A country is divided into clusters based on healthcare districts, and several districts are randomly selected to survey patient satisfaction.
5. **Corporate Training Evaluation:** A large corporation divides its offices into clusters by location, randomly selects a few offices, and surveys all employees in these offices about training effectiveness.

### Convenience Sampling

1. **App Testing:** A startup company asks its employees to participate in the beta testing phase of a new app.
2. **Customer Feedback:** A café owner surveys customers who visit the café over a weekend to gather feedback on a new menu item.
3. **Event Feedback:** Organizers of a conference ask attendees to fill out a feedback form immediately after the event.
4. **Product Testing:** A cosmetics company gives free samples to people in a shopping mall and asks for their opinions.
5. **Classroom Survey:** A teacher surveys students in their own class to get quick feedback on a new teaching method.