Statistical and Machine Learning Methods

Stratified Sampling, Isolation Forest, and DBSCAN

GOD

December 23, 2025

Table of Contents

Stratified Sampling

Introduction to Sampling

Why Sampling?

- Population data often too large or expensive to collect
- Need representative subset for analysis
- Inference from sample to population

Sampling Methods:

- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling $\leftarrow$ Our Focus
- Cluster Sampling

What is Stratified Sampling?

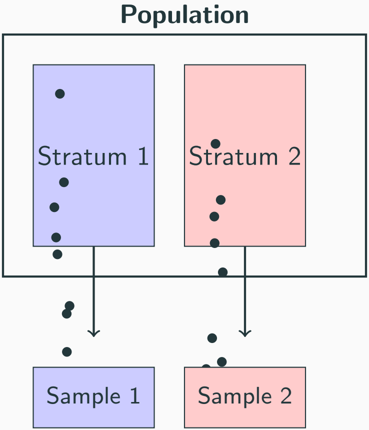
Definition: A probability sampling technique where the population is divided into homogeneous subgroups (strata) and samples are drawn from each stratum.

Key Characteristics:

- Population divided into mutually exclusive groups
- Each group shares similar characteristics
- Random sampling within each stratum
- Combines samples from all strata

Stratification Variables: Age, gender, income, education, geographic location, etc.

Stratified Sampling: Visual Representation



Types of Stratified Sampling

1. **Proportionate Stratified Sampling**
   - Sample size from each stratum proportional to stratum size
   - Formula: $n_i = n \times \frac{N_i}{N}$
   - Where n_i = sample from stratum i, N_i = stratum size
2. **Disproportionate Stratified Sampling**
   - Sample sizes not proportional to stratum sizes
   - Used when strata have different variances
   - Optimal allocation: $n_i = n \times \frac{N_i \sigma_i}{\sum N_i \sigma_i}$

Mathematical Framework

Stratified Sample Mean:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$$

where $W_h = \frac{N_h}{N}$ (stratum weight), $\bar{y}_h$ = sample mean in stratum h

Variance of Stratified Sample Mean:

$$Var(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \frac{\sigma_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right)$$

Standard Error:

$$SE(\bar{y}_{st}) = \sqrt{Var(\bar{y}_{st})}$$

Advantages and Disadvantages

Advantages:

- More precise estimates than simple random sampling
- Ensures representation of all subgroups
- Allows separate analysis of strata
- Reduced sampling error

Disadvantages:

- Requires prior knowledge of population
- More complex to implement
- Stratification variables must be known for entire population

Practical Example

Survey of University Students

Year	Population	Proportion	Sample (n=400)
Freshman	5000	0.25	100
Sophomore	4000	0.20	80
Junior	6000	0.30	120
Senior	5000	0.25	100
Total	20000	1.00	400

Each year level forms a stratum, ensuring representation across all years.

Isolation Forest for Anomaly Detection

Introduction to Anomaly Detection

What are Anomalies?

- Data points that differ significantly from the majority
- Also called outliers, novelties, or exceptions
- Can indicate errors, fraud, or rare events

Applications: Fraud detection, network intrusion, manufacturing defects.

Why Isolation Forest?

Traditional Approaches: Statistical (Z-score), Distance-based (k-NN), Density-based (LOF).

Key Insight: Anomalies are few and different, therefore easier to isolate than normal points.

- Anomalies require fewer random partitions to be isolated
- Path length to isolate a point indicates anomaly score

Algorithm: Building Isolation Forest

Algorithm 1 iForest(X, t, ψ)

```
Input: X - data, t - number of trees, ψ - subsample size
Initialize Forest = {}
for i = 1 to t do
    X′ ← sample(X, ψ)
    Forest ← Forest ∪ iTree(X′, 0, l)
end for
return Forest
```

Parameters:

- t: Number of trees (typically 100)
- ψ: Subsample size (typically 256)

Anomaly Score Calculation

Interpretation:

- $s \approx 1$: Clear anomaly (short path)
- $s \approx 0.5$: Normal point
- $s < 0.5$: Likely normal (deep path)

DBSCAN for Clustering

What is DBSCAN?

Density-Based Spatial Clustering of Applications with Noise

Core Advantages:

- Discovers clusters of arbitrary shape
- Robust to outliers (identifies noise)
- No need to specify number of clusters

Core Concepts

Two Parameters:

- ε (eps): Radius of the neighborhood
- MinPts: Min points to form a dense region

Point Classifications:

- Core Point: $\geq$ MinPts within ε
- Border Point: Within ε of a core point but low density
- Noise Point: Neither core nor border point

DBSCAN Algorithm Overview

Algorithm 2 DBSCAN Simplified

```
for each unvisited point P do
    Mark P as visited
    Find Neighbors in ε
    if count < MinPts then
        Mark P as Noise
    else
        Create new Cluster; ExpandCluster(P)
    end if
end for
```

Conclusion

Summary of Methods

- **Stratified Sampling:** Essential for ensuring subgroup representation in heterogeneous populations. It minimizes sampling error compared to simple random sampling.
- **Isolation Forest:** A highly efficient, non-parametric approach to anomaly detection that excels by "isolating" outliers rather than modeling normal points.
- **DBSCAN:** A powerful density-based clustering tool that identifies non-linear patterns and noise, making it superior to K-Means for complex spatial datasets.

Comparison Table

Feature	Strat. Sampling	iForest	DBSCAN
Primary Use	Data Collection	Anomaly Detection	Clustering
Metric	Proportion/Weights	Path Length	Density/Distance
Complexity	Low (Manual/Stat)	$O(n \log n)$	$O(n \log n)$
Key Benefit	Representativeness	High Dimensions	Arbitrary Shapes

Final Thoughts

Thank You!

Questions?