What is Latent Class Analysis?

Statistical Concepts for clinical investigators
Michael P. Anderson
Oklahoma Shared Clinical and Translational Resource
July 8, 2025

OSCTR Oklahoma Shared Clinical and Translational Resources
NIGMS U54 GM104938

Introduction

Latent Class Analysis (LCA) is a powerful statistical technique used to identify unobserved (latent) subgroups within a population based on patterns of responses across a set of observed variables. Rather than assuming that individuals belong to one observed group, LCA uses probabilistic modeling to infer hidden groupings that explain how individuals differ across key characteristics. It is particularly useful in health and behavioral sciences where underlying group structures are suspected but not directly measured.

LCA is conceptually similar to cluster analysis but differs in that it assigns individuals to classes with probabilities rather than fixed groupings. It also shares theoretical similarities with factor analysis but emphasizes person-centered rather than variable-centered modeling. LCA models require researchers to select meaningful indicator variables, estimate the optimal number of latent classes, and assess the accuracy of class assignment using measures such as entropy and average posterior probability (AvePP). The method has expanded to include covariates and distal outcomes using a structured three-step approach to improve interpretability and reduce bias.

Best Practices for Latent Class Analysis

- Select strong indicator variables that are theoretically linked to the latent construct.
- Ensure sufficient sample size, typically 300+ total cases and at least 50 per expected class.
- Use fit indices (e.g., BIC, AIC, entropy) in combination with interpretability when deciding on the number of classes.
- Avoid including covariates or distal outcomes in the initial LCA model; use the 3-step approach instead.
- Inspect profile plots to check for parallel patterns (Salsa effect), which may indicate a lack of meaningful latent structure.
- Report class assignment uncertainty metrics (e.g., AvePP > 0.8) and validate models in external samples if possible.

Strengths and Limitations of LCA

Strengths

- Identifies unobserved subgroups in heterogeneous populations.
- Provides probabilistic rather than deterministic class assignments.
- Flexible modeling with categorical, continuous, or mixed indicators.
- Incorporates class uncertainty into downstream analyses.

Limitations

- Requires large sample sizes for reliable estimation.
- Can overextract classes with large data or weak indicators.
- Subjectivity in interpreting class labels and profiles.
- Sensitive to initial values and estimation settings in software.
- Salsa effect can lead to spurious latent classes when data reflect severity along a continuum.

Brief Example

A pediatric research team used LCA to examine the complexity of congenital heart disease (CHD) in 1,482 children. They selected seven medical history indicators, including stroke and cardiac surgeries, to identify patterns of CHD complexity. LCA revealed four distinct subgroups: Mild, Moderate 1, Moderate 2, and Severe. Each child was assigned a probability of membership to each group, and the classifications were used in subsequent analyses to examine quality of life differences across CHD complexity levels.

References

Anderson MP, Bard D (2024) A Gentle Introduction to Latent Class Analysis for Researchers in Pediatrics. J Pediatr. 2024; 271, 114069
https://www.sciencedirect.com/science/article/pii/S0022347624001720?via%3Dihub

O'Connor A, et al(2023) Differences in quality of life in children across the spectrum of congenital hear disease. J Pediatr. 2023; 263, 113701
https://www.jpeds.com/article/S0022-3476(23)00564-4/fulltext