

Latent Class Analysis

Stephanie Lanza, Ph.D.

Upcoming Seminar:
December 7-8, 2018, Philadelphia, Pennsylvania

LATENT CLASS ANALYSIS

Stephanie T. Lanza, Ph.D.
Statistical Horizons
November 11-12, 2016

INSTRUCTOR

Stephanie T. Lanza
Scientific Director,
The Methodology Center
<http://methodology.psu.edu/>
Professor,
Department of Biobehavioral Health
The Pennsylvania State University

OVERVIEW OF WORKSHOP (DAY 1)

- Introduction to latent class analysis (LCA)
- The LCA mathematical model
- Latent class homogeneity and separation
- Brief SAS tutorial
- SAS PROC LCA demo

- Model identification, selection, starting values
- Multiple-groups LCA
- Measurement invariance across groups

OVERVIEW OF WORKSHOP (DAY 2)

- Brief review of binary and multinomial logistic regression
- LCA with covariates
- LCA with a distal outcome
- LCA for repeated measures

- Introduction to latent transition analysis (LTA)
- The LTA mathematical model
- SAS PROC LTA demo
- LCA and LTA in professional writing and grant proposals
- Open discussion

Abbreviations

- Throughout course, remember that:
- **LCA** stands for latent class analysis
- **LTA** stands for latent transition analysis (a longitudinal extension of LCA)

Introduction to LCA, Model specification

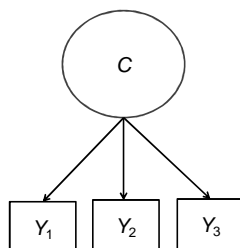
Introduction to LCA: Some basic Ideas

- Individuals can be divided into subgroups, or latent classes, based on unobserved construct
- True class membership is unknown
- Latent classes are mutually exclusive and exhaustive

Ideas underlying LCA

- Measurement of that construct typically based on several categorical indicators
- There is error associated with the measurement of the latent classes
- Like confirmatory factor analysis (specify number of classes), but latent variable is categorical

Graphical depiction of latent class variable



- C is the latent class variable
- Y_1, Y_2, Y_3 are three observed indicators of C

Parameters estimated in LCA

- Latent class membership probabilities
 - e.g. *probability of membership in HIGH DEPRESSION latent class*
- Item-response probabilities
 - e.g. *probability of reporting Felt Lonely given membership in HIGH DEPRESSION latent class*

Brief example of LCA: Depression in adolescence

- From Lanza, Flaherty, & Collins (2003)
- Eight indicators of adolescent depression ($2^8=256$ possible response patterns):

Sad

- Couldn't shake blues
- Felt depressed
- Felt lonely
- Felt sad

Disliked

- People unfriendly
- Disliked by people

Failure

- Life was failure
- Life not worth living

Brief example of LCA: Depression in adolescence

Five latent classes of depression:

Latent Class	Prevalence
No depression	41%
Sad	18%
Disliked	17%
Sad and disliked	15%
High depression	9%

**Example of LCA:
Drinking classes in 12th grade**

- Data from 2004 cohort of Monitoring the Future public release
- N=2490 high school seniors who answered at least one question on alcohol use (48% boys, 52% girls)
- Goals of study: to explore alcohol use behavior among high school seniors; to explore gender differences in measurement and in behavior

**Example of LCA:
Drinking classes in 12th grade**

Seven indicators of drinking behavior:

Item	Proportion 'Yes'
Lifetime alcohol use	82%
Past-year alcohol use	73%
Past-month alcohol use	50%
Lifetime drunkenness	57%
Past-year drunkenness	49%
Past-month drunkenness	29%
5+ drinks in past 2 weeks	26%

We will use LCA to:

- Identify and describe underlying classes of drinking behavior in 12th grade students
- Include grouping variable (gender)
 - Examine gender differences in behavior
 - Test for measurement invariance across males and females

Selecting the number of classes

Classes	G ²	df	AIC	BIC	BLRT
1	9510	120	9524	9564	NA
2	3019	112	3049	3137	.01
3	911	104	957	1091	.01
4	209	96	271	452	.01
5	4	88	81	308	.01
6	4	80	98	372	.08
7	3	72	113	434	---

(We will revisit this...)

The five-class model: ρ, γ

Item	Probability of 'Yes' response				
	Class 1 (18%)	Class 2 (22%)	Class 3 (9%)	Class 4 (17%)	Class 5 (34%)
Lifetime alcohol use	.00	1.00	1.00	1.00	1.00
Past-year alcohol	.00	.61	1.00	1.00	1.00
Past-month alcohol	.00	.00	1.00	.39	1.00
Lifetime drunk	.00	.24	.29	1.00	1.00
Past-year drunk	.00	.00	.00	1.00	1.00
Past-month drunk	.00	.00	.00	.00	.92
5+ drinks past 2 wk	.00	.00	.16	.00	.73

What would you name these five classes?

Labeling the classes

Item	High probability of 'Yes' response				
	Non-Drinkers (18%)	Experi-menters (22%)	Drinkers (9%)	Bingers (17%)	Heavy Drinkers (34%)
Lifetime alcohol use		√	√	√	√
Past-year alcohol		√	√	√	√
Past-month alcohol			√		√
Lifetime drunk				√	√
Past-year drunk				√	√
Past-month drunk					√
5+ drinks past 2 wk					√

Latent class notation

- Let C represent the number of latent classes, which in our example will be five, e.g. NONDRINKERS, BINGERS
 - $c = 1, \dots, C$
- 7 dichotomous items measuring the latent classes (yes/no)

Latent class notation

- Let Y refer to the vector of response patterns
- Let y represent a particular response pattern
- Example: $y = (Y, Y, N, N, N, N, N)$

The traditional latent class model parameters

- γ_c = probability of membership in latent class c
(**class membership probabilities**)
- p_{ic} = probability of response i to Item 1
conditional on membership in latent class c (**item-response probabilities**)

The traditional latent class model

$$P(Y = y) = \sum_{c=1}^C \gamma_c \prod \rho$$

– Assumption of local independence: Items are independent within each class

$$P(Y = y) = \sum_{c=1}^C \gamma_c \prod \rho$$

γ_c = probability of membership in Latent Class c
(e.g. *probability of membership in BINGERS latent class*)

ρ_{ic} = probability of response i to Item 1, conditional on membership in Latent Class c , etc.
(e.g. *probability of reporting 'Yes' to Item 1, conditional on membership in the BINGERS latent class*)

More about ρ parameters

- The ρ parameters express the relation between:
 - the discrete latent variable in a latent class analysis and
 - the observed variable indicators
- Similar conceptually to factor loadings
 - basis for interpretation of latent classes
- ρ 's are probabilities (between 0 and 1)

A note on missing data

- Most LCA and LTA software can handle missing data
- Missing data mechanisms
 - **MAR (missing at random)**
 - *Missingness is completely random, or related to observed items*
 - **MNAR (missing not at random)**
 - *Missingness related to unobserved items; more difficult to adjust for this type of missingness*
- Software assumes data are MAR

Homogeneity and Latent Class Separation

Rho parameter interpretation

- Rho parameters analogous to factor loadings; both
 - Express relation between manifest and latent variables
 - Form basis for interpreting latent structure
- But
 - Factor loadings are *b*-weights
 - Rho parameters are PROBABILITIES

How would you interpret these latent classes?

Probability of correctly performing practical task	Latent Class 1	Latent Class 2
Task 1	Low	High
Task 2	Low	High
Task 3	Low	High
Task 4	Low	High
Task 5	Low	High

How would you interpret these latent classes?

Probability of Responding 'Yes'	Latent Class 1	Latent Class 2
Tried Alcohol	Low	High
Been Drunk	Low	High
Tried Tobacco	Low	High
Tried Marijuana	Low	High
Tried Meth	Low	High

How would you interpret these latent classes?

Probability of Responding 'Yes'	Latent Class 1	Latent Class 2	Latent Class 3
Tried Alcohol	Low	High	High
Been Drunk	Low	High	High
Tried Tobacco	Low	High	High
Tried Marijuana	Low	Low	High
Tried Meth	Low	Low	High

How would you interpret these latent classes?

Probability of Responding 'Yes'	Latent Class 1	Latent Class 2	Latent Class 3	Latent Class 4
Tried Alcohol	Low	High	High	High
Been Drunk	Low	Low	High	High
Tried Tobacco	Low	High	High	High
Tried Marijuana	Low	Low	Low	High
Tried Meth	Low	Low	Low	High

Rho parameter interpretation

- $0 \leq \rho \leq 1$
- When latent variable completely predicts manifest variable, $\rho = 0$ OR $\rho = 1$
- When latent variable does not completely predict manifest variable, $0 < \rho < 1$
- When latent variable does not at all predict manifest variable, $\rho =$ marginal probability for item
- So if we are trying to measure a latent variable, what kind of rho's do we like?

Characteristics of patterns of rho parameters

- **Homogeneity:** degree to which rho parameters for a particular latent class are close to 0 and 1
- Individuals in a class are similar to each other

Characteristics of patterns of rho parameters

- High homogeneity

Probability of correctly performing practical task	Latent Class 1	Latent Class 2
Task 1	.10	.85
Task 2	.15	.90
Task 3	.05	.89
Task 4	.10	.95
Task 5	.12	.90

Characteristics of patterns of rho parameters

- Low homogeneity

Probability of correctly performing practical task	Latent Class 1	Latent Class 2
Task 1	.45	.50
Task 2	.48	.55
Task 3	.50	.51
Task 4	.49	.56
Task 5	.46	.58

Characteristics of patterns of rho parameters

- High homogeneity

Probability of Responding 'Yes'	Latent Class 1	Latent Class 2	Latent Class 3	Latent Class 4
Tried Alcohol	.95	.90	.95	.99
Been Drunk	.99	.10	.92	.97
Tried Tobacco	.96	.93	.94	.89
Tried Marijuana	.89	.05	.10	.88
Tried Meth	.91	.10	.07	.93

Characteristics of patterns of rho parameters

- Low homogeneity

Probability of Responding 'Yes'	Latent Class 1	Latent Class 2	Latent Class 3	Latent Class 4
Tried Alcohol	.45	.55	.52	.55
Been Drunk	.48	.45	.56	.50
Tried Tobacco	.40	.56	.50	.54
Tried Marijuana	.45	.48	.43	.56
Tried Meth	.47	.43	.46	.52

Characteristics of patterns of rho parameters

- **Latent class separation:** degree to which latent classes can clearly be distinguished from each other

Characteristics of patterns of rho parameters

- High homogeneity and high latent class separation

Probability of correctly performing practical task	Latent Class 1	Latent Class 2
Task 1	.10	.91
Task 2	.15	.90
Task 3	.05	.89
Task 4	.10	.95
Task 5	.12	.90

Characteristics of patterns of rho parameters

- High homogeneity and low latent class separation

Probability of correctly performing practical task	Latent Class 1	Latent Class 2
Task 1	.80	.91
Task 2	.82	.90
Task 3	.81	.89
Task 4	.80	.95
Task 5	.84	.90

Characteristics of patterns of rho parameters

- Homogeneity analogous to concept of saturation in factor analysis
- Latent class separation analogous to concept of simple structure in factor analysis
- You can have
 - High homogeneity, high latent class separation
 - High homogeneity, low latent class separation
 - Low homogeneity, low latent class separation
 - But not low homogeneity, high latent class separation

So what kind of pattern do we want?

- Depends on your perspective
 - You might be using LCA with a “discovery” perspective
 - OR you might be using LCA with a “measurement” perspective
- Neither is better, but they are subtly different

“Measurement” perspective

- Purpose: to establish the latent variable and a good way to measure it
- May take steps to improve homogeneity and latent class separation, e.g.
 - May discard some variables that contribute to heterogeneity
 - May impose parameter restrictions to simplify model
 - May choose number of latent classes partly on this basis

“Discovery” perspective

- Purpose: to make sense out of a large contingency table
- Interpretation something like: “people in Latent Class 1 had a 40 percent probability of correctly performing Task 1”
- Homogeneity and latent class separation somewhat less important from this perspective

SAS tutorial,
PROC LCA demo,
Exercise1
