# WELCOME TO THE 2021 NDACAN SUMMER TRAINING SERIES!

- The session will begin at 12pm EST.
- Please submit questions to the Q&A box.
- This session is being recorded.

# NDACAN SUMMER TRAINING SERIES

National Data Archive on Child Abuse and Neglect

Cornell University & Duke University

Children's Bureau

An Office of the Administration for Children & Families


# DATA STRATEGIES FOR THE STUDY OF CHILD WELFARE

# NDACAN SUMMER TRAINING SERIES SCHEDULE

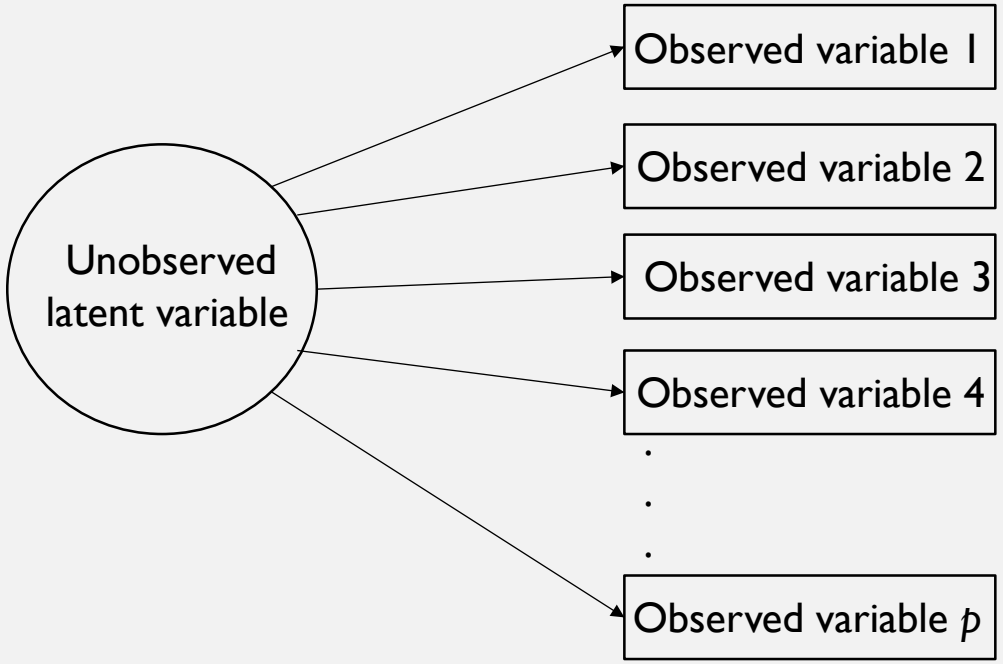
- July 7, 2021 - Introduction to NDACAN
- July 14, 2021 - Survey Based Data
- July 21, 2021 - Administrative Data and Linking
- July 28, 2021 - VCIS Data and Special Populations
- August 4, 2021 - Multilevel Modeling Workshop
- **August 11, 2021 - Latent Class Analysis Workshop**

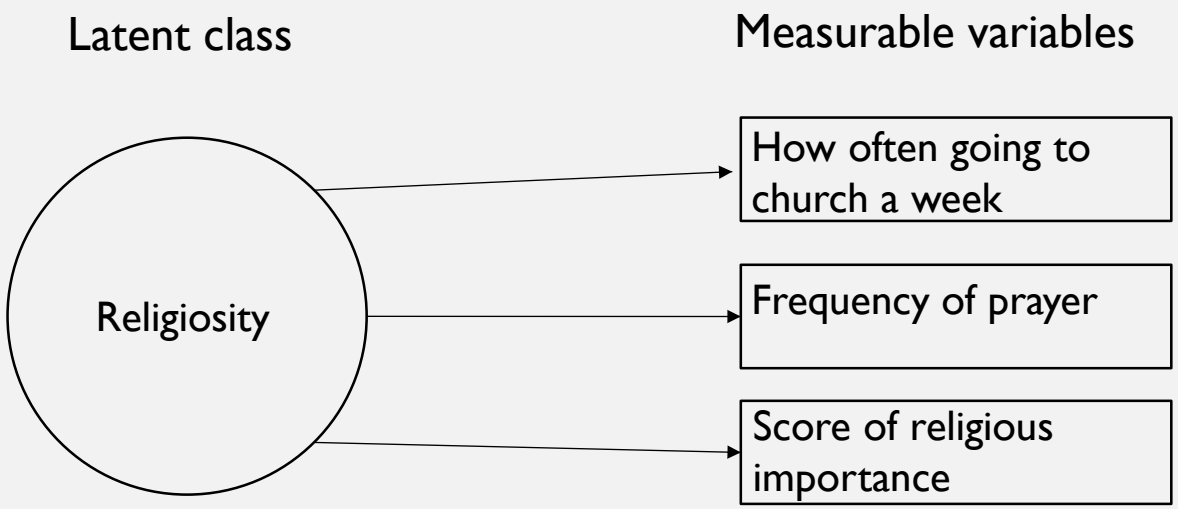
# SESSION AGENDA

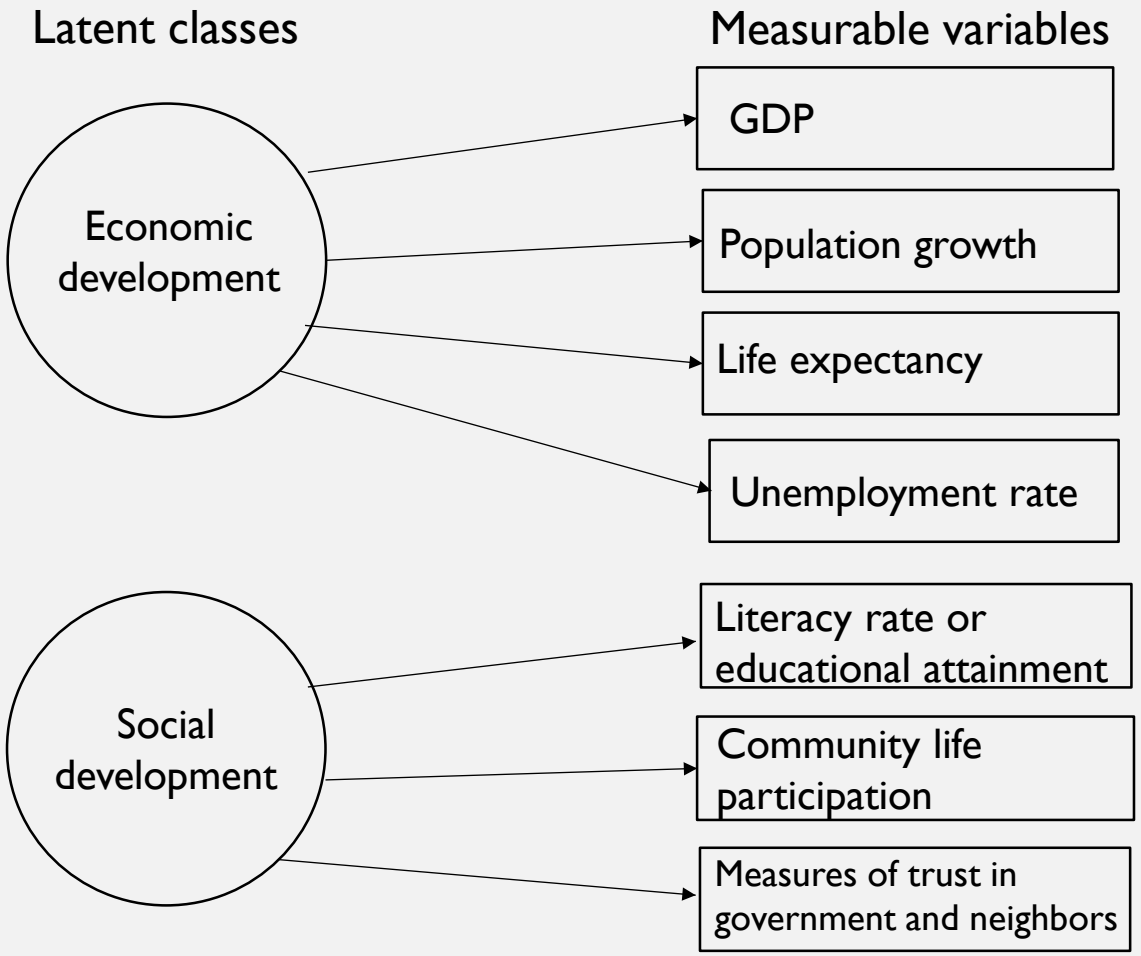
- What are latent variables and what is latent class analysis?
- The math and algorithms behind the scenes
- Example of applying LCA

# LATENT CLASS ANALYSIS (LCA)

- A latent variable is a variable that can't directly be observed but explains much of the relationship and covariation between observed variables
- Could have more than one latent variable that affects certain groups of observed variables




Latent class
Measurable variables
Religiosity
How often going to church a week
Frequency of prayer
Score of religious importance


Latent classes
Measurable variables
Economic development
GDP
Population growth
Life expectancy
Unemployment rate
Social development
Literacy rate or educational attainment
Community life participation
Measures of trust in government and neighbors

# WHAT IS LCA?

- Latent class analysis/modeling is the statistical technique used to try to identify the unobserved/latent subgroups within data
- Analogous to factor analysis (assumes continuous variables) but with categorical variables instead
- Subset of structural equation modeling

| | | Latent variables | |
|---|---|---|---|
| | | Continuous | Categorical |
| Observed variables | Continuous | Factor analysis | Latent profile analysis |
| | Categorical | Item response theory/latent trait analysis | Latent class analysis |

Reproduced from Table 1.1, Collins & Lanza 2009

# WHEN WOULD YOU WANT TO USE IT

- Used in exploratory analysis to uncover (hidden or suspected) latent structures or as confirmatory analysis to test hypotheses about latent structures
- Used for dimension reduction, clustering, and/or regression techniques

# THE LATENT CLASS MODEL

# ASSUMPTIONS

- Assume we have observed variable *A* that has *J* classes/levels (*j* = 1,2,3,…, J) and observed variable *B* that has *K* classes/levels (k = 1,2,3,…,K)
- *A* and *B* need not be independent, in fact their nonindependence is usually what drives the desire to do LCA
  - Want to understand if their relationship is spurious or if it can be explained by some other (latent) variable
- Let *X* be the unobserved (latent) variable with *T* classes (t = 1,2,3,…,T)
- We do assume that variables *A* and *B* are *conditionally independent*, given the class level of *X*

# THE MATH BEHIND THE SCENES

- Let $\pi_{jt}^{A|X}$ denote the *conditional* probability of an observation being in class *j* of *A*, given that it is in class *t* of *X*. Similarly for $\pi_{kt}^{B|X}$
- Let $\pi_{t}^{X}$ be the probability of an observation being in class *t* of *X*.
- Let $\pi_{jkt}^{ABX}$ be the joint probability that an observation is in class *k* of *A*, class *j* of *B*, *and* class *t* of *X*
- The latent class model is expressed as:

$$\pi_{jkt}^{ABX} = \pi_{t}^{X} \pi_{jt}^{A|X} \pi_{kt}^{B|X}$$

- The model demonstrates the conditional independence of *A* and *B*:

$$\pi_{jkt}^{ABX} / \pi_{t}^{X} = \pi_{jt}^{A|X} \pi_{kt}^{B|X} = \pi_{jkt}^{AB|X}$$

# THE ALGORITHM BEHIND THE SCENES

- The latent class model is expressed as:

$$\pi_{jkt}^{ABX} = \pi_{t}^{X} \pi_{jt}^{A|X} \pi_{kt}^{B|X}$$

- Use Maximum likelihood estimation to estimate *class membership probabilities* (probability of an observation belonging to the latent class) and the *item response probabilities* (conditional probability that an observation provided a certain response given that they've been classified to a latent class)
- Use an iterative algorithm – Expectation Maximization (EM) – to solve for each estimate of probability on the right hand side above
  - Relies on initial values to start
  - Can iterate indefinitely unless put a cap on the number of iterations
- EM starts with random split of classes then reclassifies observations based on some 'improvement criterion' – e.g. AIC/BIC, $X^2$
  - Usually the algorithm tries a range of number of classes and one can choose the optimal based on a "scree" (aka elbow) plot
  - Should also consider interpretability and subject matter expertise to help guide how many classes seems appropriate

# HOW TO IMPLEMENT LCA

# HOW TO IMPLEMENT IN VARIOUS PROGRAMMING LANGUAGES

- In R programming language: poLCA package with same function name
- In SAS: proc lca
- In Stata: gsem function (generalized structural equation models)
- SPSS: does not seem to be supported at this time

# EXAMPLE IN STATA OF CARCINOMA DATA

# CARCINOMA DATA

- Responses given by 7 pathologists who classified 118 slides of whether carcinoma was present in uterine cervix
- Grouping variables/pathologists will group like minded diagnoses

# EXAMPLE IN STATA USING NYTD DATA

# NATIONAL YOUTH IN TRANSITION (NYTD) OUTCOMES SURVEY

- NYTD surveys youth who age out of foster care services
- The Outcomes component of this survey contains the results of youth to examine well-being, financial, and educational outcomes as they get older
- The survey is administered every 3 years and has been given to 3 cohorts by now
- This example will use a random sample of 200 youth from the Wave 3 of the survey given to youth aged 17 in FY2014
  - Data have been masked of identifying variables, and some variables have been replaced by randomly generated values to preserve anonymity

# NYTD VARIABLES

- There are 19 variables in this example data set
- 3 can be used as covariates: sex (sex), race (raceth), highest education certification (highEdCert)
- 16 variables will be used for latent class analysis, includes but not limited to: current part-time employment (currPTE), employment related skills (EmplySkills), public housing assistance (PubHousAs), current enrollment (CurrEnroll), connection to adult (CnctAdult), substance abuse (subAbuse)

# REFERENCES


- Collins, Linda M., and Stephanie T. Lanza. *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. Vol. 718. John Wiley & Sons, 2009.
- Hagenaars, Jacques A., and Allan L. McCutcheon, eds. *Applied latent class analysis*. Cambridge University Press, 2002.
- McCutcheon, Allan L. *Latent class analysis*. No. 64. Sage, 1987.
- Porcu, Mariano, and Francesca Giambona. "Introduction to latent class analysis with applications." *The Journal of Early Adolescence* 37.1 (2017): 129-158.

# QUESTIONS?

SARAH SERNAKER
STATISTICIAN, DUKE
SARAH.SERNAKER@DUKE.EDU

# THAT WRAPS UP OUR SUMMER WORKSHOP SERIES!

**Join us next summer for more topics!**