Simpson's Paradox

From an AJPH article showing historical data. Prevalence of Insanity in the Foreign Born Population Relative to the Native Born Population: Massachusetts, 1854

	Insane	Not Insane
foreign-born	625	229375
native-born	2007	892669

odds ratio = 1.2 (1.11, 1.32)

This statistic was used to make claims that foreigners were more likely to be insane.

BUT

What about if we adjust for Status: i.e. is the person in the "Pauper Class" or the "Independent Class"

Simpson's Paradox

	Pauper Class		Independent	
	Insane	Not Insane	Insane	Not Insane
foreign-born	182	9090	442	220285
native-born	250	12513	1757	880156

Odds ratio within Pauper Class = 1.00
Odds ratio within Independent Class = 1.00

Thus if you adjust for the status or class of the person we see that, in fact, there is no relationship between being foreign and being insane than native born persons

Latent Class Models

- Search for underlying cause of correlation
- Conditioned on the underlying classes, the individuals response to the questions will be independent
- From our example, conditioning on status, the relationship between nativity and insanity is independent
- What about if we don't or can't observe this variable which should be conditioned on...i.e. it is latent
- Like in exploratory factor analysis we can examine a model that hypothesizes k underlying latent classes and see how well it fits the data
- Also like in exploratory factor analysis we need at least three observed variables otherwise we can't identify the model.
- We will be able to estimate something like "factor loadings", i.e. an estimate of how each variable loads on a particular class

Latent Class Model - Binary data - Two latent classes

The formal model (for p observed binary variables with 2 underlying classes)

- Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ be a response vector for one person to the p binary questions. For example if there are 4 questions there are 16 possible different $\mathbf{x}$ vectors (1,1,1,1) or (1,1,1,0) or (1,1,0,1), etc.
- So x_j represents the answer given to the jth question
- Let η represent the probability of being in the first latent class, then $1 - \eta$ is the probability of being in the second latent class
- Let π_{j1} and π_{j2} represent the probability of answering yes to the the jth question given that you are in class 1 or 2 respectively.
- Assuming conditional independence, i.e. given that we know what class an individual is in, then responses to each question are independent we have

$$p(\mathbf{x}) = \eta \Pi_{j=1}^{p} \pi_{j1}^{x_j} (1 - \pi_{j1})^{1-x_j} + (1 - \eta) \Pi_{j=1}^{p} \pi_{j0}^{x_j} (1 - \pi_{j0})^{1-x_j}$$

- Since each x_j is assumed to be independent conditional on class the model just takes the product of the Bernoulli's within in each class
- We want to estimate η, π_{j1}, and π_{j2}
- Estimation using maximum likelihood estimation also provides a goodness of fit

Latent Class Model

Let **c** represent the underlying latent class variable with K categories. Let η_k represent the probability of being in the kth latent class in the population, s.t. $\sum_{k=1}^{K} \eta_k = 1$

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}, c = k) = \sum_{k=1}^{K} p(\mathbf{x}|\, c = k)\eta_k$$

Similar to the latent trait model, by assuming conditional independence for the elements x_j in **x** given **c** we can write

$$p(\mathbf{x}|\, c = k) = \prod_{j=1}^{p} p(x_j|\, c = k)$$

where we define $\pi_{jk} = p(x_j|c = k)$. Thus the latent class model in general can be written as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \eta_k \prod_{j=1}^{p} \pi_{jk}$$

Where there a total of $K * p$ conditional probabilities π_{jk} and $K - 1$ marginal probabilities η_k to estimate.

Goodness of Fit for Latent Class Models

- χ^2 test compares observed frequencies with expected frequencies under the model of k latent classes
- For the unrestricted model with K classes and m_j categories for x_j

$$df = \prod_{j=1}^{p} m_j - \left(K + \sum_{j=1}^{p} K(m_j - 1) \right)$$

- For binary observations if $n >> 2^p$ Chi Square test should be good, but how much bigger???
  - Literature has mixed responses, some say $n = 2*2^p$ is ok, others say $n = 16*2^p$ is necessary.
- One technique of dealing with small counts for particular responses is to group the responses so that each "group" has at least 5 responses
- Also possible to use the parametric bootstrap to obtain goodness of fit value.
- To compare models with different numbers of latent classes, it is not correct to do a chi-square difference test because the models are not acutally nested. Common to use AIC or BIC to compare models with different numbers of classes. Choose the one with smallest AIC or BIC.

Predicting the latent class

How to calculate which class is more likely given a particular response...use Bayes rule

$$
\begin{aligned}
p(c = k|\,\mathbf{x}) &= \frac{p(c = k, \mathbf{x})}{p(\mathbf{x})} \\
&= \frac{p(c = k)p(\mathbf{x}|\,c = k)}{p(\mathbf{x})} \\
&= \frac{\eta_k \prod_{j=1}^{p} \pi_{jk}}{\sum_{k=1}^{K} \eta_k \prod_{j=1}^{p} \pi_{jk}}
\end{aligned}
$$

This provides an estimate of the probability that an individual is in class k given their response vector $\mathbf{x}$. The predicted class $\hat{c}$ is then taken to be class corresponding to the largest probability.

Macready and Dayton (1977) example

Macready and Dayton data

This data set arises from educational testing where one wishes to study the learning process in children. Macready and Dayton (1977) used a two-class model on a set of four tests to identify two distinct classes; that of "masters" and "non-masters". The frequency distribution of the response patterns together with the expected frequencies for each response pattern under the two-class model and the probability of belonging to the "master" class are given in Table 9.10. The estimated posterior probabilities show that most of the individuals can be allocated to the "master" class or the "non-master" class with high confidence. For example, individuals who have responded to at least two items correctly have probabilities greater than 0.9 of being allocated into the "master" class. On the other hand, individuals who got all items wrong (0000) are allocated to the "non-master" class with probability 0.98. There are, however, three response patterns (0100, 0001, 0010) that have probabilities close to a half. These include only 11 individuals out of 142. Therefore, the latent class model has been able to classify with high confidence most of the response patterns in the two classes.

Table 9.10 *Observed and predicted frequencies and estimated class probabilities for the two-class model, Macready and Dayton data*

Observed frequency	Expected frequency	$\hat{Pr}$(master | **x**)	Class	Response pattern
15	14.96	1.00	2	1111
23	19.72	1.00	2	1101
7	6.19	1.00	2	1110
4	4.90	1.00	2	0111
1	4.22	1.00	2	1011
7	8.92	0.91	2	1100
6	6.13	0.90	2	1001
5	6.61	0.98	2	0101
3	1.93	0.90	2	1010
2	2.08	0.97	2	0110
4	1.42	0.97	2	0011
13	12.91	0.18	1	1000
6	5.62	0.47	1	0100
4	4.04	0.45	1	0001
1	1.31	0.44	1	0010
41	41.04	0.02	1	0000

Macready and Dayton (1977) example

The $X^2 = 9.5$ and the $G^2 = 9.0$ on six degrees of freedom indicate a near perfect fit to the data. The percentage of G^2 explained is 91%. The parameter estimates and standard errors of the model are given in Table 9.11.

Table 9.11 *Estimated conditional probabilities, $\hat{\pi}_{ij}$, and prior probabilities, $\hat{\eta}_j$, with standard errors in brackets for the two-class model, Macready and Dayton data*

Item (i)	$\hat{\pi}_{i1}$	$\hat{\pi}_{i2}$
1	0.21 (0.06)	0.75 (0.06)
2	0.07 (0.06)	0.78 (0.06)
3	0.02 (0.03)	0.43 (0.06)
4	0.05 (0.05)	0.71 (0.06)
$\hat{\eta}_j$	0.41 (0.06)	0.59 (0.06)

Members of the first class have small estimated probabilities of answering items correctly. This class is clearly the "non-master" one. Members in the second class have for all items much higher probabilities of answering correctly. This class is the "master" class.

You could re-analyze the data using a latent trait model and compare the expected frequencies obtained under the latent class model and the latent trait model.

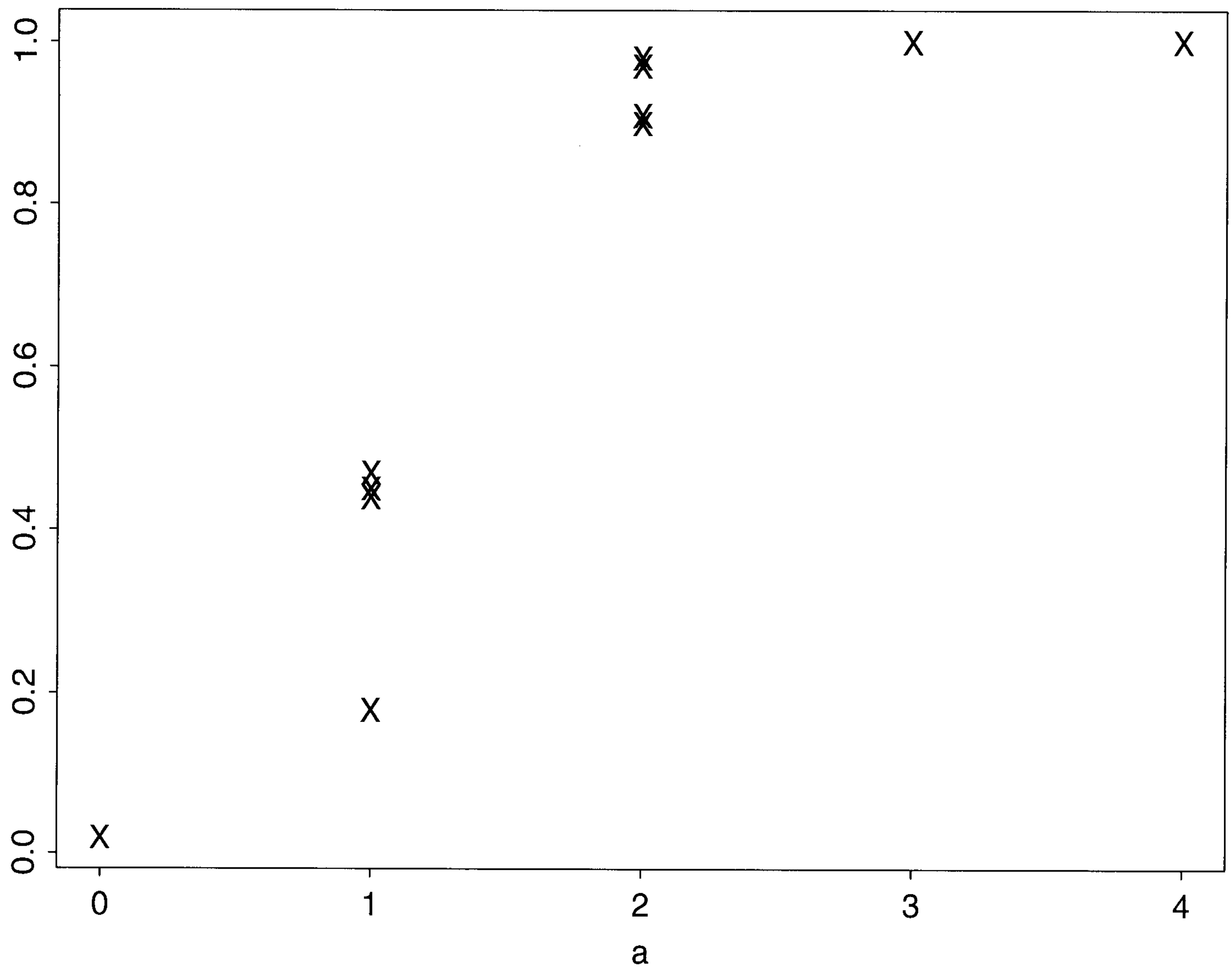
Macready and Dayton (1977) example

Given that a person had the first two tests correct and the second two tests wrong, their probability of being in the master class is

$$p(c = Master|\, \mathbf{x} = (1100)) = .91$$

$$\begin{aligned} &= \frac{p(c = 1 \text{ } and \text{ } \mathbf{x} = (1100))}{p(\mathbf{x} = (1100))} \\ &= \frac{.59 * (.75 * .78 * (1 - .43) * (1 - .71))}{.59 * (.75 * .78 * (1 - .43) * (1 - .71)) + .41 * (.21 * .07 * (1 - .02) * (1 - .05))} \\ &= .91 \end{aligned}$$

Macready and Dayton (1977) example



Macready and Dayton (1977) example in Mplus 3

```
Mplus VERSION 3.1 DEMO MUTHEN & MUTHEN

INPUT INSTRUCTIONS
  TITLE: mix1
  DATA: FILE IS bart.dat;
  VARIABLE: NAMES ARE u1-u4;
            USEV ARE  u1-u4;
            CATEGORICAL = u1 - u4;
            CLASSES = c(2);
  ANALYSIS: TYPE=MIXTURE;
  savedata: file is C:\Documents and Settings\melanie\Desktop\lcaestimates;
           save = cprobabilities;
```

Mplus 3 and onwards can implement random starting values

A common problem with using the maximum likelihood method for mixture models, in particular latent class models, is that the likelihood function may have several local maxima. This is a problem because numerical maximization techniques are notoriously fooled by local maxima because they are only, in fact, looking for local maxima rather than global maxima. Of course the maximum likelihood solution should be the global maxima.

So one idea for being more confident that we have found the global maxima is to give many different starting values for the numerical procedure and take the solution to be the one with the largest likelihood value.

Mplus 3 and onwards implements this technique for latent class analysis by essentially considering many different starting values for the π_{jk} parameters. So, while it was necessary in earlier versions of Mplus, to provide starting values, in Mplus 3 it is not. Which makes our job as users of the software easier.

For more description of the problem of local maximum in latent class models see e.g.,
http://ourworld.compuserve.com/homepages/jsuebersax/local.htm

Macready and Dayton (1977) example in Mplus

```
SUMMARY OF ANALYSIS
Number of groups                                               1
Number of observations                                       142
Number of y-variables                                          0
Number of x-variables                                          0
Number of latent class indicators (u)                          4
Number of structural continuous latent variables               0
Number of mixture continuous latent variables                  0
Observed variables in the analysis
   U1          U2          U3          U4
  Categorical variables
   U1          U2          U3          U4
Categorical latent variable in the analysis
   C
Estimator                                                    MLR
Maximum number of iterations                                1000
Convergence criterion                                  0.100D-05
Maximum number of iterations for mixture model               100
Convergence criteria for mixture model
  Loglikelihood change                                 0.100D-06
  Derivative                                           0.100D-05
Latent class regression model part
  Number of M step iterations                                  1
  M step convergence criterion                         0.100D-05
  Basis for M step termination                         ITERATION
Latent class indicator model part
  Number of M step iterations                                  1
  M step convergence criterion                         0.100D-05
  Basis for M step termination                         ITERATION
  Maximum value for logit thresholds                          15
  Minimum value for logit thresholds                         -15
  Minimum expected cell size for chi-square            0.100D-01
Optimization algorithm                                       EMA
Input data file(s)
  bart.dat
Input data format  FREE
```

Macready and Dayton (1977) example in Mplus

```
THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Loglikelihood

          H0 Value                        -331.764

Information Criteria

          Number of Free Parameters              9
          Akaike (AIC)                     681.527
          Bayesian (BIC)                   708.130
          Sample-Size Adjusted BIC         679.653
            (n* = (n + 2) / 24)
          Entropy                            0.754

Chi-Square Test of Model Fit for the Latent Class Indicator Model Part

          Pearson Chi-Square

          Value                              9.459
          Degrees of Freedom                     6
          P-Value                           0.1493

          Likelihood Ratio Chi-Square

          Value                              8.966
          Degrees of Freedom                     6
          P-Value                           0.1755
```

Macready and Dayton (1977) example in Mplus

```
FINAL CLASS COUNTS AND PROPORTIONS OF TOTAL SAMPLE SIZE
BASED ON ESTIMATED POSTERIOR PROBABILITIES

  Class 1         58.70857          0.41344
  Class 2         83.29143          0.58656


CLASSIFICATION OF INDIVIDUALS BASED ON THEIR MOST LIKELY CLASS MEMBERSHIP

Class Counts and Proportions

  Class 1               65          0.45775
  Class 2               77          0.54225


Average Class Probabilities by Class

                 1        2

  Class 1     0.875    0.125
  Class 2     0.024    0.976
```

Macready and Dayton (1977) example in Mplus

```
MODEL RESULTS

                   Estimates     S.E.  Est./S.E.

CLASS 1

CLASS 2

LATENT CLASS INDICATOR MODEL PART

 Class 1

 Thresholds
    U1$1               1.333    0.402      3.313
    U2$1               2.613    0.883      2.960
    U3$1               4.004    2.120      1.888
    U4$1               2.898    1.143      2.535

 Class 2

 Thresholds
    U1$1              -1.117    0.323     -3.455
    U2$1              -1.267    0.402     -3.155
    U3$1               0.275    0.238      1.158
    U4$1              -0.883    0.306     -2.886

LATENT CLASS REGRESSION MODEL PART

 Means
    C#1               -0.350    0.271     -1.290
```

Macready and Dayton (1977) example in Mplus

```
LATENT CLASS INDICATOR MODEL PART IN PROBABILITY SCALE

 Class 1

 U1
    Category 1         0.791    0.066     11.911
    Category 2         0.209    0.066      3.139
 U2
    Category 1         0.932    0.056     16.584
    Category 2         0.068    0.056      1.216
 U3
    Category 1         0.982    0.037     26.325
    Category 2         0.018    0.037      0.480
 U4
    Category 1         0.948    0.057     16.741
    Category 2         0.052    0.057      0.923

 Class 2

 U1
    Category 1         0.247    0.060      4.105
    Category 2         0.753    0.060     12.543
 U2
    Category 1         0.220    0.069      3.191
    Category 2         0.780    0.069     11.332
 U3
    Category 1         0.568    0.058      9.748
    Category 2         0.432    0.058      7.402
 U4
    Category 1         0.292    0.063      4.617
    Category 2         0.708    0.063     11.169

QUALITY OF NUMERICAL RESULTS
     Condition Number for the Information Matrix              0.842E-02
       (ratio of smallest to largest eigenvalue)
```

Deciphering Mplus

When you tell Mplus that a variable is categorical it orders the observed responses numerically and then recodes them as 1,2,3, etc.

data	Mplus
0	1
1	2

In my notation, when the x_j are binary (0,1), then $\pi_{jk} = p(x_j = 1|c = k)$. Mplus and other software which do latent class analysis, perform estimation of parameters on the logit scale rather than on the original probability scale. This is because it is easier to do numerical computation on unconstrained values between $-\infty$ to ∞ than on numbers constrained to be between 0 and 1 (as are probabilities). Note a large negative logit corresponds to a probability close to 0, a large positive logit corresponds to a probability close to 1, and a logit of 0 corresponds to a probability of 0.50.

PROC LCA add on in SAS

There is a Proc that can be downloaded from the Penn State Methodology Center website at http://methcenter.psu.edu/index.php/home

Latent Class Analysis
Measuring categorical latent variables - Project EAT example

EXAMPLE: Measuring unhealthy weight control behavior. Hypothetically a categorical latent variable.

Have you done the following in the last year in order to lose weight or maintain your weight: (yes, no)

	marginal	2-class		3-class			4-class			
To control weight	1	1	2	1	2	3	1	2	3	4
fasted	17.9	38.8	2.8	58.5	32.6	2.6	24.9	71.4	29.2	2.6
ate little	44.1	92.5	9.0	94.2	89.9	7.9	74.1	100.0	87.5	6.9
diet pills	6.3	13.6	1.1	40.4	6.5	1.2	49.4	31.0	6.1	0.8
vomit	6.3	15.0	0.1	45.0	7.1	0.1	33.2	43.6	5.6	0.1
laxatives	1.6	3.5	0.2	17.1	0.1	0.2	20.7	11.7	0.0	0.2
diuretics	1.4	3.3	0.1	16.1	0.1	0.1	29.4	7.9	0.0	0.1
food substitutes	9.3	19.2	2.1	41.6	13.1	2.1	54.4	34.5	12.0	1.7
skipped meals	44.4	89.7	11.1	85.7	89.7	9.5	43.5	100.0	87.4	8.5
smoked more cigs	9.3	18.6	2.5	39.1	13.1	2.5	9.8	47.6	11.2	2.4
% in each class	100	42.0	58.0	8.4	35.2	56.4	2.6	8.0	34.7	54.7

Estimated π_{jk} (probability of saying yes to the variable j given that the individual is in latent class k) under latent class models with different K

Longitudinal profile Latent class example

American Journal of Epidemiology

Vol. 157, No. 9
Printed in U.S.A.
DOI: 10.1093/aje/kwg049

Developmental Typology of Trajectories to Nighttime Bladder Control: Epidemiologic Application of Longitudinal Latent Class Analysis

Tim J. Croudace[1], Marjo-Riitta Jarvelin[2], Michael E. J. Wadsworth[3], and Peter B. Jones[1]

[1] Department of Psychiatry, University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 2QQ, United Kingdom.
[2] Department of Epidemiology and Public Health, Faculty of Medicine, Imperial College of Science, Technology, and Medicine, London SW7 2AZ, United Kingdom.
[3] MRC National Survey of Health and Development, University College London, London WC1E 6BT, United Kingdom.

Received for publication November 27, 2001; accepted for publication November 14, 2002.

TABLE 1. Nighttime bed-wetting outcomes (n = 3,272) in the Medical Research Council 1946 National Survey of Health and Development*

Boys (n = 1,719) No.	Boys %	Pattern	Girls (n = 1,553) No.	Girls %
1,362	79.2	000000	1,311	84.4
103	6.0	100000	83	5.3
39	2.3	010000	33	2.1
45	2.6	110000	18	1.2
12	0.7	001000	9	0.6
2	0.1	101000	4	0.3
8	0.5	011000	4	0.3
3	0.2	111000	9	0.6
5	0.3	000100	6	0.4
2	0.1	100100	2	0.1
2	0.1	010100	2	0.1
2	0.1	110100	2	0.1
2	0.1	001100	1	0.1
5	0.3	011100	1	0.1
8	0.5	111100	5	0.3
11	0.6	000010	9	0.6
1	0.1	100010	1	0.1
5	0.3	010010		
3	0.2	110010		
5	0.3	001010	3	0.2
2	0.1	011010	2	0.1
3	0.2	111010	3	0.2
6	0.3	000110	2	0.1
2	0.1	100110		
1	0.1	010110	1	0.1
3	0.2	000001	4	0.3
1	0.1	100001		
2	0.1	010001	1	0.1
		100101	29	1.9
		001101	1	0.1
2	0.1	110001		
1	0.1	010101	1	0.1
		011101	1	0.1
1	0.1	110110		
2	0.1	001110	6	0.4
3	0.2	101110	2	0.1
2	0.1	011110	4	0.3
26	1.5	111110	10	0.6
1	0.1	000011	1	0.1
1	0.1	010011		
2	0.1	011011		
1	0.1	010111		
1	0.1	001111	1	0.1
7	0.4	011111	2	0.1
24	1.4	111111	7	0.5

* Outcome patterns for a binary repeated measure (1 = wetting at least occasionally during the past month, 0 = not wetting in the past month) at ages 4, 6, 8, 9, 11, and 15 years, ordered in terms of increasing prevalence of persistent wetting.

TABLE 2. Results of unrestricted longitudinal latent class analysis in the Medical Research Council 1946 National Survey of Health and Development (pooled sexes, n = 3,272)

	Three classes (LLCA*-3)	Four classes (LLCA-4)	Five classes (LLCA-5)
Sequential model comparisons (T + 1 classes vs. T classes)	3 vs. 2	4 vs. 3	5 vs. 4
Log-likelihood value for model with T + 1 classes	–3,243.605	–3,211.173	–3,201.380
Log-likelihood value for model with T classes	–3,344.440	–3,243.605	–3,211.173
–2 difference in log-likelihood	201.669	64.863	19.587
Difference in no. of parameters (T + 1 classes vs. T classes)	7	8	8
Lo-Mendell-Rubin adjusted LRT* value	198.171	63.877	19.289
Lo-Mendell-Rubin adjusted LRT p value	<0.0001	<0.0001	0.0322
Bootstrap LRT p value	<0.01	<0.01	>0.50
Chi-square goodness-of-fit tests			
Degrees of freedom	43	36	29
LRT χ^2	123.588	58.725	39.138
p value	<0.0001	0.0098	0.0990
Bootstrap p value†	<0.01	0.02	0.11
Pearson χ^2	132.431	49.416	35.966
p value	<0.0001	0.0674	0.1746
Bootstrap p value†	<0.01	0.10	0.40
Information criterion‡			
Akaike's Information Criterion	6,527.210	6,476.347	*6,470.760*
Bayesian Information Criterion	6,649.073	*6,640.862*	6,677.927
Sample-size-adjusted Bayesian Information Criterion	6,585.524	*6,555.071*	6,569.894
Entropy	0.856	0.913	0.897
Condition number§	$0.120E^{-03}$	$0.783E^{-03}$	$0.379E^{-03}$

* LLCA, longitudinal latent class analysis; LRT, likelihood ratio test.

† Bootstrap p values were based on 200 resamples.

‡ Minimum values are shown in italic type.

§ Condition number = ratio of the largest eigenvalue to the smallest eigenvalue for the Fisher information matrix. Small values less than $10E^{-09}$ indicate problems with model identification.

TABLE 4. Parameter estimates from the four-class longitudinal latent class analysis model (*n* = 3,272; 60.5% of the cohort) in the Medical Research Council 1946 National Survey of Health and Development

	Trajectory class							
	Normal		Persistent		Chronic		Onset	
	Estimate	SE*	Estimate	SE	Estimate	SE	Estimate	SE
Prevalence								
Crude	0.876	0.013	0.065	0.015	0.032	0.002	0.025	0.006
General population†	0.881		0.064		0.038		0.017	
Class-specific conditional probability	0.057	0.006	0.557	0.104	0.832	0.058	0.112	0.082
	0.007	0.006	0.732	0.116	0.961	0.035	0.324	0.105
	0.005	0.002	0.161	0.040	0.994	0.027	0.481	0.107
	0.002	0.001	0.069	0.029	0.937	0.041	0.516	0.109
	0.004	0.002	0.036	0.026	0.856	0.046	0.674	0.128
	0.002	0.001	0.033	0.017	0.396	0.056	0.144	0.059
Average latent class assignment probability conditional on assignment by maximum probability rule‡								
Assigned latent trajectory class								
Normal	0.976§		0.017		0.000		0.006	
Persistent	0.098		0.859¶		0.002		0.038	
Chronic	0.000		0.021		0.916§		0.062	
Onset	0.008		0.103		0.068		0.819¶	

* SE, standard error. All standard errors are bootstrap standard errors based on 3,000 resamples.

† General population estimates are weighted estimates based on the prevalence of the assigned classes. Weights were derived from the inverse of the sampling fractions.

‡ Children were assigned to the latent trajectory classes for which the posterior probability of latent class membership was highest. High values on the leading diagonal are indicative of good separation of the latent classes and reflect the quality of the empirical classification.

§ The mean "Normal" trajectory class assignment probability for the 87.6% of the sample in the first latent trajectory class was very high (0.976); this was also the case (0.916) for the "Chronic" class.

¶ The classes defined by trajectories of increasing (Onset) or decreasing (Persistent) probability of night wetting were less well discriminated (0.819 and 0.859, respectively). This pattern indicates that those classified in trajectory class 2, Persistent, also had nonzero probabilities for membership in trajectory class 1, Normal. Similarly, those classified in trajectory class 4, Onset, also had nonzero probabilities for membership in trajectory classes 2, Persistent, and 3, Chronic. In general, all of the off-diagonal elements were very low, indicating well-separated classes (entropy statistic = 0.913).

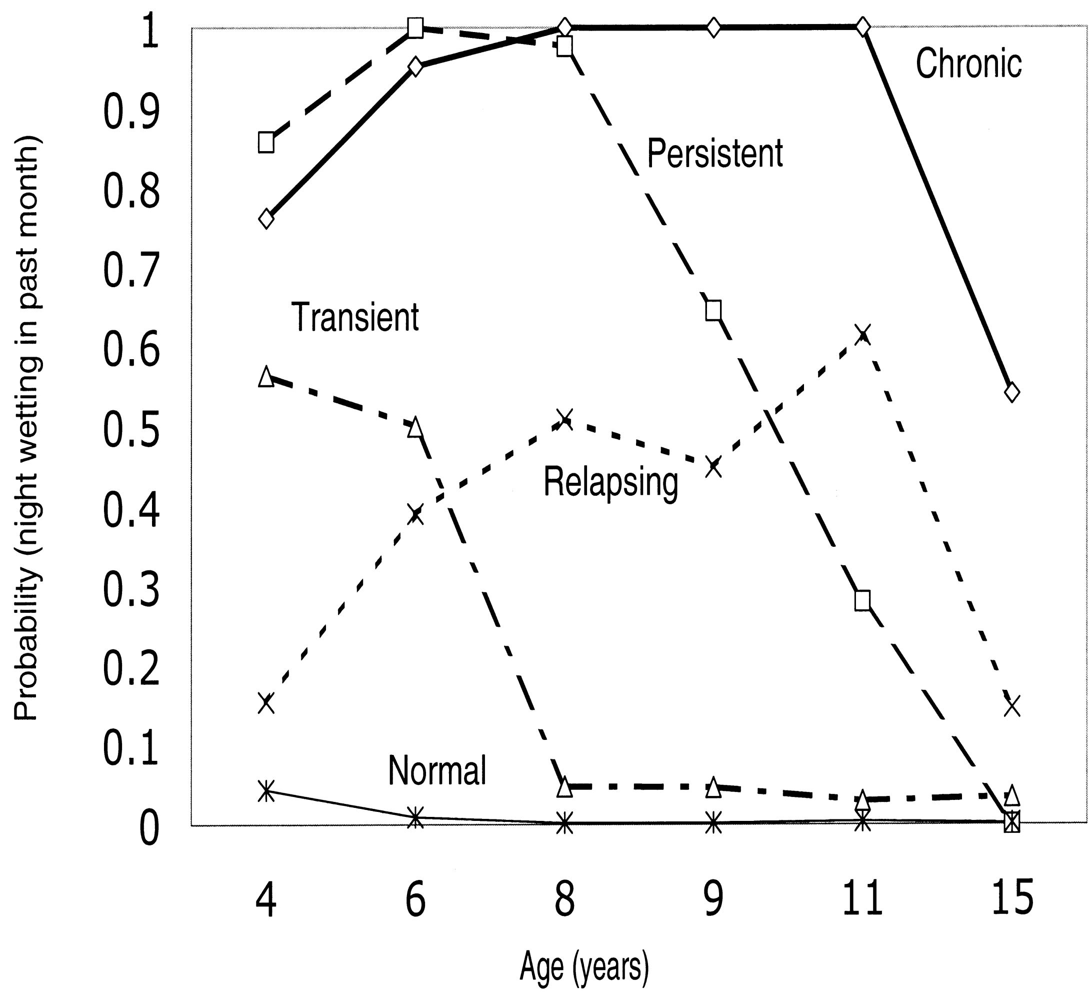


FIGURE 1. Five-class unrestricted longitudinal latent class analysis model† for repeated, binary nighttime bed-wetting outcomes at six ages (ages 4, 6, 8, 9, 11, and 15 years) in the Medical Research Council 1946 National Survey of Health and Development, including partially incomplete data‡ under the "missing at random" assumption ($n = 4{,}755$; 90.9% of the cohort§). (†The model included sex as a seventh indicator of the latent classes. ††This analysis is valid under a "missing at random" assumption; sex was included in the model to increase the likely validity of this assumption. Chi-square tests (Pearson and likelihood ratio) for the more restrictive "missing completely at random" assumption returned nonsignificant p values ($p > 0.90$) only when sex was included. §Prevalence estimates apply to an unselected, general population (inverse of sampling fractions included as weights in the model).)

Latent Class Modeling for Sensitivity and Specificity with no Gold Standard

STATISTICS IN MEDICINE, VOL. 5, 21–27 (1986)

THE VALUE OF LATENT CLASS ANALYSIS IN MEDICAL DIAGNOSIS

DAVID RINDSKOPF*

Educational Psychology, CUNY Graduate Center, 33 West 42nd Street, New York, NY 10036, U.S.A.

AND

WALLACE RINDSKOPF

SUMMARY

Assessment of the value of diagnostic indicators such as symptoms and laboratory tests results from calculation of the sensitivity and specificity of the indicators. Knowledge of the rate of occurrence of the disease allows for additional calculations of the error rates in using an indicator. These calculations are accurate only when the data on which they are based are reliable. If the diagnosis, which is used as the criterion for computing the sensitivity and specificity, is not accurate, then the resulting calculations will be in error. We show how a statistical method, latent class analysis, allows for the estimation of the characteristics of indicators even when an accurate diagnosis is unavailable. In addition, the method deals with several indicators at once, and provides a way to combine the information from all the indicators to make a diagnosis.

KEY WORDS Diagnosis Latent class analysis

Table I. Data, expected values and assignment to latent classes based on two-class model

Indicator				Frequencies			
Q-wave	History	LDH	CPK	Observed	Expected	Class	Probability of correct classification
yes	yes	yes	yes	24	21·62	2	1·000
no	yes	yes	yes	5	6·63	2	0·992
yes	no	yes	yes	4	5·70	2	1·000
no	no	yes	yes	3	1·95	2	0·889
yes	yes	no	yes	3	4·50	2	1·000
no	yes	no	yes	5	3·26	1	0·580
yes	no	no	yes	2	1·19	2	1·000
no	no	no	yes	7	8·16	1	0·956
yes	yes	yes	no	0	0	—	—
no	yes	yes	no	0	0·22	1	1·000
yes	no	yes	no	0	—	—	—
no	no	yes	no	1	0·89	1	1·000
yes	yes	no	no	0	0	—	—
no	yes	no	no	7	7·78	1	1·000
yes	no	no	no	0	0	—	—
no	no	no	no	33	32·12	1	1·000

The probabilities of correct classification cannot be computed for cells with expected values of zero, and are left blank. Also, the class to which these people should be assigned is undetermined.
The observed frequencies are from Galen and Gambino.[1]

Estimates of Sensitivity and Specificity

Table II. Parameter estimates for two-class model

	Class 1	Class 2
Unconditional class probabilities	0·542	0·458
Conditional probabilities of indicators, given latent class		
Positive Q-wave	0·000	0·767
Classic history	0·195	0·791
Flipped LDH	0·027	0·828
High CPK	0·196	1·000

Extensions to Latent Class Models

- It is possible to have multiple dimensions each having k classes
- It is possible to do confirmatory latent class modeling
- It is possible to weaken the assumption of conditional independence