# Introduction to Latent Class Analysis

## Video 1 transcript

https://www.ncrm.ac.uk/resources/online/all/?id=20806

Oliver Perra: So, before I start, I would like to summarise some of the things I've said in the first presentation. I highlighted that Latent Class Analysis is a person-centred approach. It assumes that a sample is made up by a mixture of individuals who belong to a few groups or classes. It is also a measurement model, so, the underlying classes explain persons’ patterns of behaviour that we observe.

 And the latent classes are therefore types, categories of people that share the same propensity to display certain behaviour. And the categorical nature of these underlying classes also means that groups of individuals may differ in more than one dimension. They may differ in categorical ways each from the other.

 So, in the rest of this second presentation, I will provide formal definitions of the Latent Class Analysis model because I think this is important to better understand the assumptions of the models and how it works.

 So, I will start with example I used before of four observed categorical variables, for example, different symptoms of depression, which I called A, B, C, D here. Each of these symptoms can be observed and graded in three ordered categories of response. One, if symptoms are never observed during the time of study. Two, if they are sometimes. Three, if they are observed most of the times.

 The latent variable is called X and these are supposed to have two categories, one and two.

 Once again, I emphasise that the examples I'm using include categorical observed variables, but Latent Class Analysis works with any type of observed variables from nominal to continuous variables.

 So, here I'm presenting a formula where P indicates probability, the lowercase subscript indicates values of variables and the vertical bar symbol indicates conditional probabilities. For example, here the probability of A given latent class X.

 So, this formula basically indicates that the probability of observing a combination of categories A, B, C, D together with category X of latent class is equal to the probability of latent class X, this one here, multiplied by the conditional probability of a given latent class X multiplied by the conditional probability of B given latent class X and so on.

 So, what this means is basically that the probability of observing a combination of answers to the four indicators of a latent class given together with a latent class is given by the product of the probability of the latent class and their conditional probabilities.

 This formula represents the conditional probability assumption, the probability of observing different responses in the indicators in the observed variables are independent of each other given latent class affiliation.

 The constraint imposed to this probability is that the sum of the probabilities of being in one class or another sums up to one. Which means that the latent classes sum to one or as there is a latent class for each respondent. Latent classes are exhaustive, as I highlighted in the first presentation.

 And similarly, the sum of the conditional probabilities also add up to one.

 So, here I'm interested in writing the probability of observing a pattern of answers where individuals report the symptoms, most of the time Level 3, in all the four symptoms when they are in Category 1 one of latent class X. The probability of observing this pattern of answers is given by the product of the probability of being in Latent Class 1 multiplied by the conditional probabilities of reporting each item most of the time given membership to Latent Class 1.

[0:05:17]

 So, to illustrate how this works, I will start with an example of fictional data. So, the data of a similar study would look like this where rows represent participants and then we have columns that represent the frequencies of different symptoms from one to three.

 So, using the formula provided, I'm going to show what type of parameters the latent class output will show. And the first one will be the class membership probabilities. So, these are the probabilities of being Latent Class 1 of Latent Variable X and Latent Class 2. And these probabilities are constrained to sum up to one.

 So, the output of Latent Class Analysis will look something like this where we have the probability of being in Class 1 of Latent Class X and the probability of being in Latent Class 2. And as you can see, these add up to(?) to one, what they indicate is basically if we take a random individual from our study there is a probability of 12% roughly that an individual being Class 1 and the probability of approximately 88% of random individual to be in Class 2. Note that the output doesn't tell us what these classes are or how we should interpret these classes.

 We will get information about how we should interpret the classes looking at the second parameter that Latent Class Analysis will provide. These are the conditional item response probabilities. So, these are the probability of different type of answers given affiliation to one latent class. So, for example, the probability of never displaying low mood if someone is in Latent Class 1, the probability of displaying low mood sometimes if someone is in Latent Class 1, the probability of displaying low mood most of the time if someone is in Latent Class 1 and so on.

 And again, remember these probabilities are constrained to sum up to one, so, the probability of the three type of answers or frequencies of the variable low mood within Latent Class 1 all sum up to one.

 So, here we can see the conditional probabilities by classes and here, for example we see the probability of never displaying low mood is someone is in Latent Class 1.

 Whereas here we see the other conditional probabilities. For example, the probability of reporting low mood sometimes if someone is in Latent Class 1.

 These conditional probabilities all sum up to one, so the probability of different types of responses to item low mood for people in Latent Class 1 all add up to one as you can see here, and similarly for the other indicators and the other class.

 And here you can see that these conditional probabilities in helping type the meaning of the different classes. In particular Class 1 includes people that have a high probability of reporting the different symptoms most of the time here and here, whereas Latent Class 2 includes people that have a low probability of reporting the symptoms most of the time. And in fact, they have a higher probability of never displaying these symptoms.

 So, we can then interpret Latent Class 1 as the latent class of people that are more likely to be depressed, whereas Latent Class 2 is the class of people that are less likely to be depressed. Those conditional probabilities therefore work in a way like the factors loading in factor analysis. They help us understand the relationship between the latent classes, the latent categories and the observed variables, and help us interpret and provide meaning to the latent classes.

 We can visualise these conditional probabilities for ease of interpretation, and one way is to report to the probability of one category of response across all the items and by class. In this example I reported that conditional probability of observing symptoms most of the time by latent classes and this shows how the individuals in Latent Class 2 differ in respect to the probability of reporting symptoms most of the times compared to individuals in Latent Class 1.

[0:10:59]

 Another possibility of reporting the conditional probabilities is to report the cumulative probability of two categories of response as long as these categories are ordered. And for example, here I displayed the cumulative probability of reporting symptoms sometimes or most of the time. Once again, this shows how individuals in their two classes differ in their propensity to report symptoms frequently or not.

 But so far the example I've have shown shows latent classes that differ along dimension of being likely to display frequent symptoms of depression or not. But sometimes the latent classes can represent more complex patterns. The example I'm showing here show that individuals in Latent Class 1 appear more progressive in their attitudes. They are more likely to endorse abortion rights, gender equality, and so on. Those in Latent Class 2 appear more conservative in their attitudes. They are less likely to endorse all these items. But also, Latent Class 3 do not easily fit along a dimension of being progressive or conservative. In fact, they appear more liberal on gender related issues, but more conservative on issues like taxes and environment. And so, they represent differences along another dimension than just being conservative or progressive. And that's an example of how Latent Class Analysis can highlight qualitative differences in our samples.

 But I also wanted to mention in this stage that it's possible to constrain these conditional probabilities, and this is useful when we want to test hypotheses about the associations between the observed variables and the latent classes.

 For example, here I constrain the conditional probability observing symptoms for people in the depression class so that the conditional probability was the same across all symptoms. Because the conditional probability sum up to one, it is only necessary to constrain two of these probabilities.

 But a model like these would indicate that no symptom is a better indicator of the underlying depression status and it's possible to test these hypotheses by comparing the model with the constrained conditional probabilities with a model without these constraints using the likelihood ratio test.

 The final parameter that latent class will provide is the posterior probability. So, this represents each individual's probability of being in one latent class or another based on the pattern of responses. And these probabilities also sum up to one.

 So, the output will look something like this where we have the conditional probabilities of being Latent Class 1 and Latent Class 2 and you can see, for example, that ID 102 has a high probability of being Latent Class 1 based on the pattern of symptoms.

 Conversely, ID 104 has a more uncertain date and class affiliation. The probability of this individual of being Latent Class 1 is 42%. And this individual has a 58% probability of being Latent Class 2.

 So, this example also illustrates how the latent class membership is uncertain as individuals are assigned to the different latent classes with some probability and some uncertainty. So, we cannot consider latent status assigned to an individual as a certainty and when we want to use the latent class membership as an outcome or predictor, we also need to control for this uncertainty in assigned participants to one class or the other. And a measure of the precision of this classification is provided by entropy, and this is based on the individual's posterior probabilities of being in the latent classes. So, if every person had the same probability of being in each of the latent classes, estimated entropy will be equal to zero. And conversely, if everyone had a posterior probability of being related in one latent class equal to one and zero probabilities of being in the other classes, entropy will be equal to one. So, lower values of entropy indicate classifications. They are less precise.

[0:16:23]

 So, to summarise, I have described how Latent Class Analysis is a probability based model. The associations between the latent classes and the indicators are probabilistic and later class affiliation is also estimated with error and ascent uncertainty.

 I've also mentioned that it's possible to test some assumptions in the models, some hypotheses, for example, by constraining the associations between indicators and the latent variables we can test different type of hypotheses about these associations.

National Centre for Research Methods (NCRM)
Social Sciences
Murray Building (Bldg 58)
University of Southampton
Southampton SO17 1BJ
United Kingdom

**Web** www.ncrm.ac.uk
**Email** info@ncrm.ac.uk
**Tel** +44 23 8059 4539
**Twitter** @NCRMUK