Latent Class Analysis
This is just a very brief introduction to the general concepts of latent class analysis. The topic deserves much more space than this, but this handout will give you a general idea of the purpose of the analysis and some suggestions for further readings.

Latent Variables
Latent class analysis is a kind of measurement model which estimates an unobserved construct, or latent variable, defined by a set of observed variables. The idea is much like a traditional factor analysis model in which a set of observed variables define an underlying continuous construct. A traditional factor analysis, for example, might involve a set of questions about various political attitudes to try to define an underlying construct of political conservatism. The advantage of such an approach is that the shared variance among the questions can be extracted to create a more reliable measure of political conservatism, removing measurement error and variance that is unique to any of the observed variables. That latent variable can then be used in regression model to improve the estimates of the associates with other variables by correcting for the attenuation that occurs with measurement error. In contrast to the factor analysis model, the latent class model groups individuals in order to identify types of voters, perhaps obtaining patterns that reflected groups such as moderate liberal, radical progressive, apolitical, and libertarian. The classes are assumed to represent nominal categories of voters and do not represent a continuum. As with factor models, the observed variables, or indicators, can be continuous or binary measures (ordinal indicators are also possible). The term latent profile analysis is used for the special case in which indicators are continuous, but latent class analysis is used more generally to refer to models whether binary or continuous indicators are involved.

Latent Class Analysis
The latent class measurement model (without the regression portion) seeks to find some set of mutually exclusive and exhaustive categories that group cases based on a set of observed variables. Either exploratory or confirmatory approaches to latent class models are possible, analogous to the distinction between exploratory and confirmatory factor analysis. For exploratory models, the number of latent classes is not specified, usually because no clear hypothesis exists about the number of latent classes. In the confirmatory form, which I will mostly focus on here, the number of classes is specified and the software provides an estimate of the fit of the data to the hypothesized number of classes. Though the true number of classes is still unknown in the confirmatory model, a number ultimately must be specified in the application. The meaning of each latent class must be inferred from the data or theory, and it is up to the researcher to name and interpret them. Contrast the example of unknown political class described above from an example of a known political class, such as when the respondent can be assigned as a Democrat, Independent, or Republican based on self-declaration or registration documents, for instance. Latent class membership of any one individual is estimated in a probabilistic fashion.

The figure below follows confirmatory factor analysis and structural equation modeling conventions to depict a latent class model. The ellipse is used to represent the latent variable, called $\eta$ (the Greek eta) with a superscript $C$ for some number of classes, and the square boxes represent the measured variables, $y_j$. 

![Latent Class Model Diagram]
Each arrow represents a type of regression, where the observed variable $y_j$ is predicted by the latent variable. The $v$ (the Greek nu) is the intercept for that regression. If the indicators were binary, we would use $\tau$, the Greek letter tau, for a threshold, as with generalized linear modeling thresholds. Latent class models do not have traditional factor loadings (the regression slope), and this is why the arrows are represented with dotted lines here. Instead, the pattern of measurement intercepts (for continuous indicators) or response probabilities (for binary indicators) across classes indicate the strength of the relationship between factor and item. The epsilons, $\varepsilon$, represent measurement residuals that include unaccounted for (unique) variance including measurement error.

The model provides $C - 1$ latent class factor mean estimates, $\alpha_c$. These values are like the intercepts from a logistic regression. Because there are no predictors of this latent class variable, we have an intercept only model and no regression coefficient. If there are only two classes, then then there will be one intercept value. Just like in logistic regression, the estimates of $\alpha_c$ are in logit form and we would need to use the logistic transformation to obtain an estimated probability of membership in a given class, exactly as we would with a logistic model.

$$\hat{\pi}_c = P(\eta^c = c) = \frac{e^{\alpha_c}}{1 + e^{\alpha_c}}$$

The class membership probability is the estimate of the proportion of the sample that belongs to a certain class. In other words, for a two class model, $\hat{\pi}_1$ would tell us the proportion of cases we expect to be members of the first class, where $c = 1$. If there are three classes, there will be two intercept values. The program may choose the last category as the referent class by default, but the referent group is the same as the use of $Y = 0$ as the referent group in binary or multinomial logistic regression. When there are more than two classes, we extend the transformation, using the multinomial logistic transformation.

$$\hat{\pi}_c = \frac{1}{1 + \sum_{c=1}^{C} e^{\alpha_c}}$$

The probabilities for all of the classes must sum to 1.0.

$$\sum_{c=1}^{C} \hat{\pi}_c = 1$$

Mixture Models
A focus on multiple indicators in a more traditional latent class analysis has served to provide an introduction to some of the underlying principles of categorical latent variables. A more general framework, suggested by Muthén and colleagues (Clark & Muthén, 2009; Lubke & Muthén, 2005; Muthén, 2001; 2002; Muthén & Muthén, 2000), combines categorical and continuous latent variables in the same model. These structural equation mixture models open a variety of possible avenues for investigating hypotheses involving unknown groups in two important ways—(1) associations among categorical latent variables, continuous latent variables, continuous observed variables, or categorical observed variables can be examined, and (2) categorical latent variables can be used in a flexible manner that allows classification of observed or latent variables.

Estimation, Model Identification, and Fit
The most common estimator for latent class models is maximum likelihood using an expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). In the EM steps of the ML process,
conditional expectations and the posterior class membership probabilities are computed in the expectation step and parameter estimates are updated. The fit is then maximized through iterations in the maximization step. This process alternates between the two steps until an optimization criterion is reached. Estimation can be sensitive to start values and it is wise to retest any model with different start values to be certain that convergence was reached at a global solution not a local solution (Hipp & Bauer, 2006), a testing process that may be automated within the software program. A log-likelihood value obtained upon convergence is used to compute fit indices.

For identification, the number of classes must be less than the number of indicators unless there are additional constraints (e.g., equal variances for the indicators across classes). Thus, with three indicators, only two classes can be specified without special restrictions. For latent class models that have binary indicators and do not include covariates or latent variables, a likelihood ratio chi-squared \( G^2 \) and Pearson chi-squared are computed to assess fit. Degrees of freedom for these tests can be calculated by \( df = \left[ 2^J - 1 \right] - q \), where \( q \) is equal to the number of latent class means estimated \((C - 1)\) plus the number of class-specific intercepts. The likelihood ratio chi-squared may be problematic for sparse data, where the Pearson chi-squared is sometimes substituted. The likelihood ratio chi-squared is not valid for comparing different number of classes (Lanza, Bray, & Collins, 2013). For example, with binary 4 indicators and 2 classes, there are 6 degrees of freedom. With three indicators and two classes, the model is just identified.

No chi-squared model fit is available for latent class models with continuous indicators. Instead assessment of fit must rely on likelihood-based fit indices, such as the Akaike Information Criteria and the Bayesian Information Criteria, which are commonly used for evaluation of fit relative to comparison models. The sample size adjusted BIC (aBIC; Sclove, 1987) seems to perform better than other information criteria (e.g., Nylund, Asparouhov, & Muthén, 2007).

\[
aBIC = -2LL + q \cdot \ln \left( \frac{N + 2}{24} \right)
\]

The quantity \(-2LL\) is the -2 log likelihood value for the \( H_0 \) model, \( q \) is the number of free parameters, and \( N \) is the sample size. The number of free parameters \( q \) is \( 2JC - (J-1)(C-1) \). Though the model may be theoretically identified, it may not be empirically identified and restrictions may be necessary for convergence. Models with a large number of binary indicators, fewer cases, large number of classes, low membership proportions in one or more classes, or sparse data (i.e., low frequency in the contingency table) may be more susceptible to convergence issues (Lubke & Muthén, 2005; 2007).

### Determining the Number of Classes

Part of the process of latent class analysis involves deciding on the correct number of classes, sometimes called class enumeration. Although the researcher may have a priori hypotheses about the number of classes, comparisons are usually made among models with different numbers of classes to provide evidence that the number of classes is correct. The difference in the log likelihoods for two models with a different number of classes is not distributed as a chi-squared, so an exact test to compare models does not exist. The adjusted BIC is commonly used for this purpose (lower values indicating better fit) and performs fairly well (Tofghi & Enders, 2006), but a number of simulation studies suggest that more precise methods may be preferable. These methods are designed to compare two models that differ by only one latent class. Among the several proposed alternatives are a bootstrapped likelihood ratio test (Nylund et al., 2007), the Lo-Mendell-Rubin adjusted likelihood ratio test, and the Vuong-Lo-Mendell-Rubin likelihood ratio test (Lo, Mendell, Rubin, 2007; Vuong, 1989).

An additional complication is that for every latent class model there is an equivalent continuous factor model that can account for the same covariance matrix, where the continuous factor model has one fewer factors than classes in the latent class model (e.g., Bauer & Curran, 2004; Bartholomew, 1987). The equivalence of these models has sparked considerable discussion among statisticians that is likely
to be continued, but, for the applied researcher, there is currently no simple way to distinguish among the
two types of models on an empirical basis. There is no choice for the researcher but to decide which type
of model is most appropriate based on theoretical considerations in the context of the questions most of
interest.

Software
Most software programs use a maximum likelihood estimation via the expectation maximization (EM)
algorithm, but a Bayesian process is also possible. There are a variety of software programs that
estimate latent class models, including PROC LCA, a free macro for SAS (Lanza, Collins, Lemmon, &
Schafer, 2007) and Latent Gold (Vermunt & Magidson, 2005), but currently only Mplus (Muthén &
Muthén, 1998–2012) and Mx (Boker et al., 2012) integrate latent class variables within the larger
structural equation modeling framework (the so-called mixture modeling approach). Mplus uses the
maximum likelihood-EM approach with a robust standard error adjustment as the default.

References and Further Reading
Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling:
in latent variable research* (pp. 317–341). Greenwich, CT: Information Age.