PhUSE 2011

Paper SP07

An Animated Guide:
An Introduction to Latent Class Clustering in SAS®

Russ Lavery, Contractor, Bryn Mawr, USA

ABSTRACT

This is the first in a planned series of three papers on Latent Class Analysis. Latent Clustering Analysis (LCA) is a method that uses categorical variables to discover hidden, or latent, groups and is used in market segmentation and
medical research. This paper explores PROC LCA, a free SAS add-in created by The Methodology Center at Penn State University. Using this free, and easy to install, add-in allows users of SAS to perform Latent Class Clustering using syntax with which they are already familiar.

The managerial output of the Latent Cluster Analysis, (sometimes called Latent Class Analysis) is similar to output from other clustering methods.

A researcher/statistician can use PROC LCA to produce a report that, after interpretation, contains the following information:
1) The project identified N clusters (and we gave them “cute and interpretable names” – like “Binge Drinker” or “Abstainer”)
2) The clusters have these characteristics (80% of “people in Cluster 1” answered Yes to Question 1)
3) The clusters have these relative frequencies (numbers of obs in each of the Clusters we identified”)
4) There is x% certainty that a subject belongs in the Cluster to which s/he was assigned.

The above results can be used to create models that can predict, from additional raw data, the likelihood of a new subject belonging to one of the discovered clusters.

Code and data are included in the appendix of the paper.

INTRODUCTION

PROC LCA can be downloaded for free from the Methodology Center at Penn State University. Installation is very easy and the appropriate URL is at the end of the paper.

Comments can be found on the web that assert the superiority of LCA for theoretical reasons. LCA does not assume
normality, homogeneity or a linear relationship and assertions are made that these lack of assumptions make LCA better than other methods. My web search investigating the reasons for the superiority of LCA returned many results
supporting the use of LCA, but a large faction of the supporting entries could be linked to one company – a company
that makes LCA software. The creators of PROC LCA do not make any claim that LCA is superior to other methods.

LCA is often used when:
1) The data is best collected in frequencies of observations in categories
2) The assumptions of LCA (we are investigating a process that is categorical and not continuous) match the theory being investigated
3) The results, the identified categorical classes, are easy to communicate to clients.
Use of LCA assumes local independence of variables within clusters and this could be a reason for selecting another

technique. LCA assumes that the membership in different latent classes is the only thing that causes observed variables to be correlated. If you believe that the variables being used in the correlation are related within found clusters (height and weight might be an example of correlated variables) you might want to use another technique. For pharmaceutical marketers (where TRx and NRx are correlated) might want to consider other methods as well as
LCA.

Latent cluster analysis can be used in a variety of business and academic situations. It can be used to discover groups (e.g. non-drinker, social drinker, binge drinker, addicted drinker … severely depressed, moderately depressed, not depressed) from categorical data. This makes LCA a very general tool, especially when the researcher is willing to "bucket" continuous variables into levels and thereby create discrete variables from continuous (e.g. bucket drinks per week from numbers to categories as follows: 0 drinks → never ; 1-5 drinks-→ sometimes ; 6-15 drinks → always. or 0-3 → low ; 4-7→medium ; 8-15 → high). This means this LCA can be applied to almost any data set. PROC LCA can be applied to almost any sort of cross-tabbed data – even to "found" data from a scholarly articles or found in business reports.

The basic process is to:
1) "Bucket" continuous variables into categorical variables with two or three levels.

2) Use PROC LCA to make a very complex nested frequency table (if we had 6 vars at 3 levels we get 243 cells).

3) Use PROC LCA to solve for 2,3,4,…N cluster solutions.

4) Determine which solution (2-cluster, 3-cluster etc.) best represents reality.

Task: recover, from an overall and highly nested cross table, tables for each segment.

An overly simple (just two questions) frequency table.

Overall	Rare Arrest	Occasional Arrest	Regular Arrests
Rare Complaints	420	88	43
Occasional Complaints	148	299	164
Regular Complaints	73	112	353

Recovered latent classes and their frequency tables

23%

Desk Officers	Rare Arrest	Occasional Arrest	Regular Arrests
Rare Complaints	388	1	2
Occasional Complaints	4	0	0
Regular Complaints	5	0	0

6%

Homicide Detectives	Rare Arrest	Occasional Arrest	Regular Arrests
Rare Complaints	11	0	0
Occasional Complaints	59	0	2
Regular Complaints	28	0	0

18%

Vice Squad Detectives	Rare Arrest	Occasional Arrest	Regular Arrests
Rare Complaints	0	0	0
Occasional Complaints	0	0	0
Regular Complaints	0	0	300

53%

Foot Patrol	Rare Arrest	Occasional Arrest	Regular Arrests
Rare Complaints	21	87	41
Occasional Complaints	85	299	162
Regular Complaints	40	112	53

The above data is from an imaginary study.

Figure 1

In Figure 1, we see an overall cross-tab of complaints vs. arrests broken down into four cross-tabs (with % in each cluster) and the clusters named / interpreted.

Figure 1 is an example of a source table, and LCA results, for a police-related example that will be developed later in
the paper. Figure 1 figure shows that PROC LCA was able to take a 2x2 table (the top table) and find a 4 cluster solution that, when summed, approximates the overall cross tab. PROC LCA estimated the answers/characteristics for each cluster and the cluster sizes (as percent of the total sample). We see, in Figure 1, that twenty-three percent of the subjects were assigned to the cluster that was, after analysis of the pattern of answers, named "Desk Officers". In practice the cross-tabs will involve more variables and be more complicated.

5) Create models predicting cluster membership, using variables not used in the segmentation. Statistical techniques
for categorical Y variables can be used to predict the cluster membership for new observations.

It should be noted that the recovered clusters, or recovered segments, are categorical – but often can be considered

to have some aspects of ordering. This ordering "aspect" is sometimes raised as a criticism of the use of LCA and is raised by people supporting the use of other techniques. As an exploration of this issue we can consider alcohol use habits.

There is some ordering in "frequency and quantity of drinking" as one goes from an "abstainer" to an "addicted drinker". However a social drinker and a binge drinker might consume the same amount of alcohol per unit month and the groups might not "separate", if alcohol consumption per month were plotted on an equal interval scale.

The creators of PROC LCA freely acknowledge that, in many cases, interpretable results come from modeling output
variables as ordered variables and good results will come from using techniques other than LCA. The creators of PROC LCA make no claim that Latent Cluster Analysis is TRUTH in all cases. However; some research formulations are more accurately served, and some theoretical considerations are more closely mapped, by categorical Y and categorical X variables. In these cases LCA can be the most appropriate methodology.

CONSTRUCTS OR LATENT VARIABLES:

When researchers use the word latent they use it to refer to something that is hidden (un-measurable), or underlying, and which generates observable, measurable characteristics. The factors, in factor analysis, are latent variables. Many path analytic problems focus on latent variables. In fact, most social or business research focuses on latent variables. A review of the idea of latency would be helpful and is the subject of this section.

Figure 2 is a chart that could be used to explain the constructs involved in investigating whether "Watching TV Violence" causes "Increased Aggressiveness". "Watching TV violence" and "Increased Aggressiveness" are theoretical variables that cannot be observed directly (and so are latent variables). They can also be thought of as generalizations or conceptual ideas. They are latent variables. Latent variables are often called constructs.

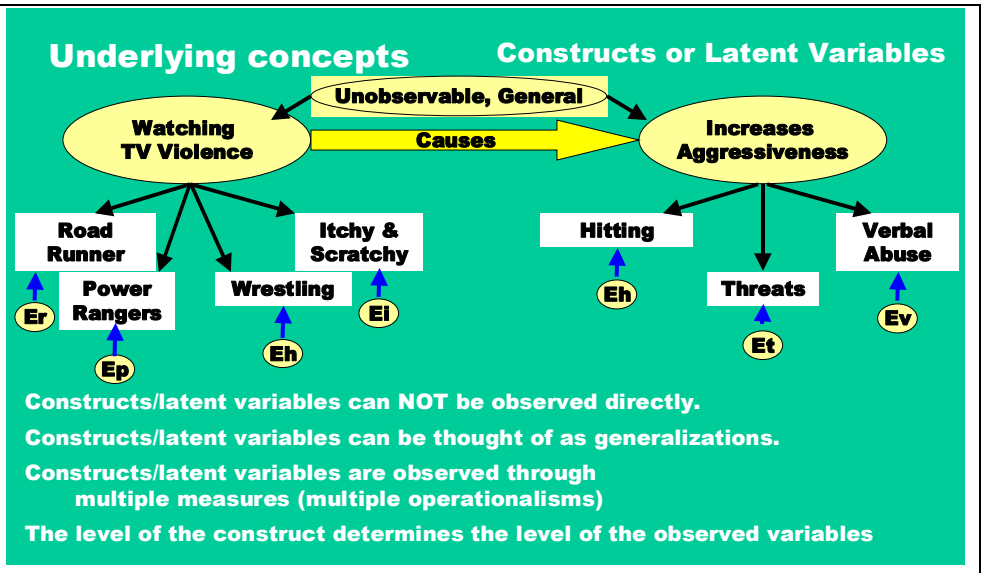


Figure 2

Because these two variables cannot be observed, they are said to be latent variables. We assume that the unobservable, latent variables determine the levels of the variables we can see. The variables we can see, like the number of hours a child spends watching roadrunner cartoons, are called operationalisms. So we have, in research models, unobservable constructs and observable operationalisms of the latent variables.

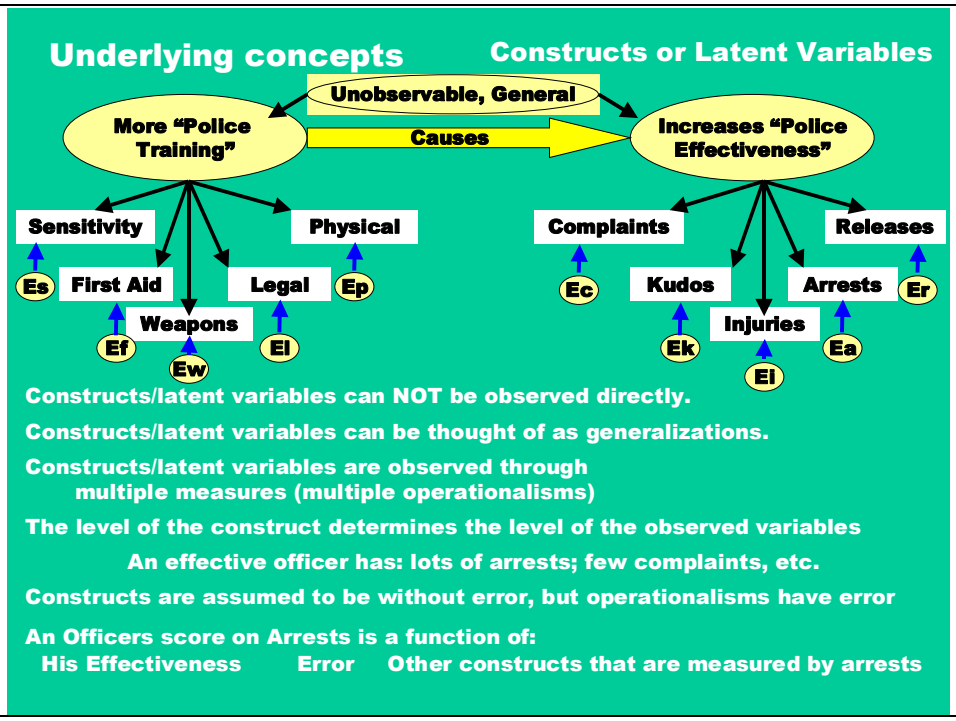
As a child watches moreTV violence, the number of hours spent watching Road Runner cartoons and Wrestling (etc.) increase. Constructs are generalizations and investigating generalizations makes research more useful than if the research were interpreted using just observable variables. On a policy level, it is more useful to say "watching TV violence" increases "aggressiveness in the nation's schools" than it is to say "watching roadrunner cartoons makes kids in kindergarten punch each other out". If one is only thinking in terms of an operational variable, say watching
roadrunner cartoons, the statements one can make are much less interesting and also, potentially, less precise and

less generalizable. If the children who are allowed to watch roadrunner cartoons exhibit greater aggressiveness, one is left wondering if allowing children to watch cartoons where the background is a desert scene (as in road runner cartoons) increases aggressiveness. Increased generalizability of results, the ability to use these results in many settings and to make statements about new and slightly different situations, are important characteristics of research done using the above construct-operationalism model.

The construct-operationalism way of approaching a research problem is taught through a course called research methodology. People trained in social research, or business research, are required to take this course and are explicitly trained to conceptualize their research projects in this way. People in the hard sciences, like statistics, physics and chemistry, usually do not get a formal exposure to this material. However; I suggest that constructoperationalism is the true model for research - even in the hard sciences.

As an example of construct-operationalism in hard science, consider the following scenario. Imagine some people are doing clinical research. They want to tell FDA that their new "active drug" cures some disease in the general population. "Active drug" is an abstraction/generalization of all of the batches of the "active drug" that were used in the study. Individual production runs of the test drug can be considered as operationalisms of the construct "active drug". The subjects that receive the test drug can be thought of as operationalisms of the construct "patients in the population with the disease". When we say "the new drug cures a disease", we want that statement to be interpreted as the drug has the ability to cure cases of the disease in the general population. We do not want to use a research process that only allows us to say: "batch number 32 of the drug cured these seven people". Even in hard science, we want to generalize results.

Figure 3 shows the layout of another research project. In this project researchers want to investigate if increased "Police Training" causes increased "Police Effectiveness".

Figure 3 shows the layout of another research project. This project investigates if increased "Police Training" causes increased "Police Effectiveness". The research hypothesis describes relationships between the generalizations / latent variables / constructs. The constructs are: 1) level of "Police Training" and 2) increased "Police Effectiveness".	
	Figure3

In this research project we might give the police officers cultural sensitivity training and then measure how much they learned. We would give them first aid training and measure how much they learned from the training.

The latent variables, or constructs, are considered to be without error. The observed "training" variables (scores on sensitivity, scores on legal issues and on physical capabilities and the operationalisms for the construct increased "Police Effectiveness") are all measured with error. Creating good operationalisms is a very difficult task and problems with operationalisms deserve, at least, the short discussion that follows.

Ideally, your operationalisms are done:

1) with great precision of measurement

and the operationalisms measure
2) all of the construct of interest and
3) only the construct of interest.

Firstly, recognize that error can enter into the measurement of operationalisms through all of the usual measurement
problems. When taking tests to measure cultural sensitivity or testing skill with weapons, the officer may be distracted or tired and score below his/her average. The measuring instrument, also, may be unreliable and introduce error.

The other two issues are subtler, and almost philosophical. The problem is that your operationalisms do not: 1) measure all of the construct of interest (they under-measure) or 2) that the operationalisms measure other constructs besides the construct of interest (they are contaminated).

Think first of the problem of under-measuring a construct. Possibly, a well-trained policeman is a good reader of body language and uses that ability to read body language to help manage stressful interactions on the street. If that
is true, the operationalisms in Figure 3 do not adequately measure a well-trained policemen - and are incomplete.

Next think of the problem of operationalisms measuring other “things” (being contaminatied). A good operationally should also only measure the construct of interest –though this is never possible in practice. While some operationalisms are better than others, ALL operationalisms measure both the construct of interest and other constructs. Looking at Figure 3 we see that all of the pictured operationalisms will measure officer effectiveness is but they also measure the “officer type of job”. Desk officers will not get injured often – regardless of how effective they might be.

While never having been a policeman, and not even being a fan of police drama, I will let my imagination flow and describe the life of police officers. Your understanding of my assumptions about the life of police officers will further your understanding of the example in the paper. My understanding of the life of a police officer was used in creating dummy data for PROC LCA and in interpreting the PROC LCA output.

As an overview, I will state that an effective police officer makes lots of arrests and receives letters of praise (kudos –kudos are an archaic word for praise) from the grateful citizens s/he helps. An effective police officer has few complaints, few injuries, and has few arrested subjects released because of paperwork or legal issues. All of these are generally true.

Thinking of the issues of measurement contamination, let’s think of how an officer’s job affects the measures that we’ve created for police effectiveness.

In my imagination, a desk sergeant has a paperwork job and works inside the police station building. As a result, s/he sees few citizens. He will have few complaints filed, and receive few kudos from citizens. He will have few injuries, make a few arrests and have few of his “arrested suspects” released for legal reasons.

A foot patrol officer has a chance to confront violent criminals and meet the public. A good foot patrol officer has moderate Complaints, kudos, injuries, and arrests. I assume these officers have about the average number of arrested suspects released.

Homicide detectives actively work on just a few cases and they routinely talk to very upset and easily offend people. Homicide detectives have moderate complaints, kudos, injuries, and low arrests. I assume, since everyone struggles
so intensely to be released from a homicide charge, that they have a lot of their “perps” released after arrest.

People on the vice squad can make lots of arrests (because in many large cities it’s easy to pull someone off a street
corner and arrest them) but this kind of officer gets lots of complaints. In my imaginary police world, vice suspects often file complaints against the officers as a way of letting off their aggravation at being arrested and also as a form

of revenge (by making the officers spend time at his desk filling out forms in response to the complaint). A vice squad officer has high complaints, low kudos, low injuries, and high arrests. I assume these officers have about the average number of arrested suspects released.

With the above story in mind, you can easily see how our operationalisms for effective policing also measure the construct “officer type of job”. To make the situation worse, the operationalisms measure other constructs as well. In
the real world, operationalisms always measure more than one construct and are always contaminated. Good operationalisms have less “contamination” than poor operationalisms. As you can imagine, researchers can have lots of discussions over operationalism issues.

As an aside, it is never safe to operationalize a construct with just one variable. Additionally, self-reported variables are notoriously unreliable. Constructs should be measured by several different variables and using variables of different types. Good research practice requires a researcher should use a mixture of self-report, historical records, direct observation by the research team members, and even physical measurements to capture a construct.

More detail about our data follows:
A program, in the abstract of the paper, will allow you to reproduce the results shown below. Changing parameters in
the part of the SAS program that generates dummy data will allow you to experiment with how well PROC LCA can find TRUTH - in whatever messy data set you create.

I will spend a little time describing the parameters that I used to create the data used in this analysis. The parameters are used to create dummy data against which I ran PROC LCA. I think generating dummy data is an effective learning technique. The learning process is to create dummy data with known characteristics and then see how well the technique being used can recover the known characteristics designed into the data.
Please see Figure 4 for our model and the parameters. PROC LCA assumes ONLY ONE latent variable (unlike factor analysis). LCA assumes that categorical classes in the latent variable cause the values we observe in the operationalisms. It also assumes the errors in the operationalisms are uncorrelated.

In my imaginary research project I reviewed six months of information on arrests, complaints, prisoners released for some problem with paperwork or the legality of the arrest, injuries to the police officer, and kudos. Kudos is an archaic word for praise, and I needed a word that did not start with P because I’d used a variable beginning with p in
a previous slide. On the slide below you can see the average counts in six months for each of these variables and by
officer type. I used base SAS, and a loop, to generate hundreds of observations with these characteristics. I also added an error term to each variable for each observation so the officers in a cluster did not have identical characteristics. Feel free to change these parameters to play with the model and PROC LCA.

The variables started out as being continuous and I used PROC Rank to group them into three categories that I labeled: rare, occasional and regular.

In the SAS program in the appendix, there are several sections that produce cross tabs of variables. These cross tabs are not required for a research project but is useful to investigate how the magnitude of the designed in error term has affected the observed results.

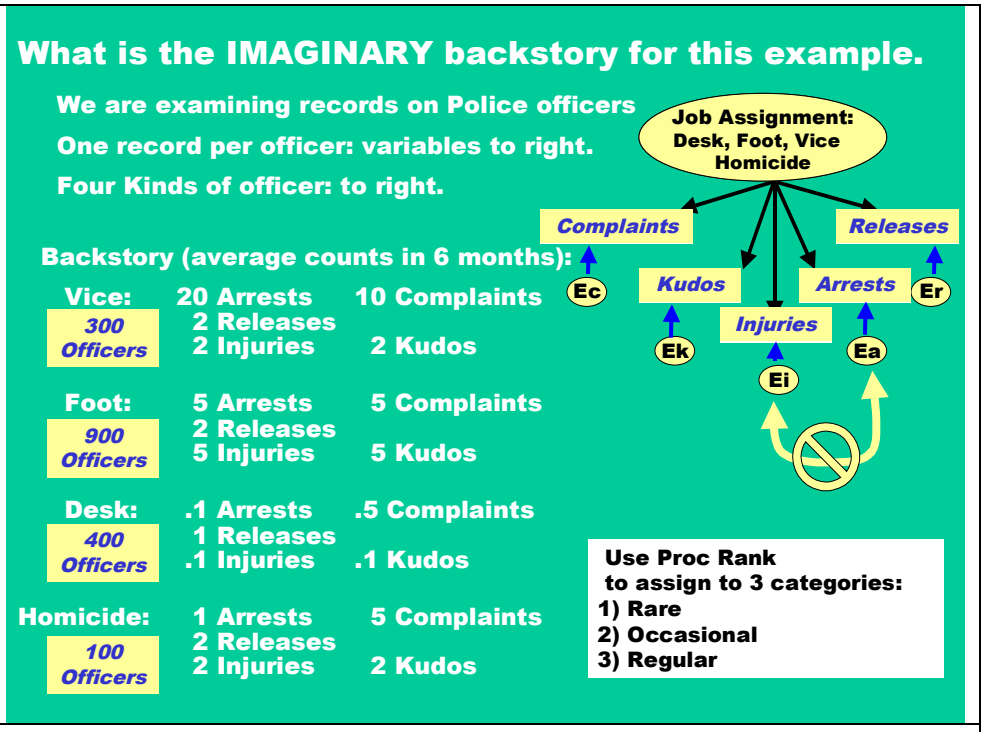


Figure 4

Please look back at Figure 1 and notice that two desk officers made regular arrests. This is the result of the error parameter in the program introducing error to the measured variable.

If you want to play with the program, and create a custom data set, look in the SAS program for the lines below

```
/*TrueSegments: the names of the classes/segments.
    In a research project, this will not be known */
/*ClassPrev   :number of respondents in each class.     In a research project, this will not be known */
/*ComplntsAvg :average number of complaints for the class in 6 months
   - from police records  */
/*KudosAvg :average number of complimentary letters for the class in 6 months
   - from police records*/
/*InjuryAvg   :average number of injuries for the class in 6 months
   - from police records*/
/*ArrestAvg   :average number of arrests made for the class in 6 months
   - from police records*/
/*ReleaseAvg  :average number of suspects released for procedural reasons
   for the class in 6 months - from police records*/
%let TrueSegments  =  Desk  Foot Hom  Vice  ; /*Name of segments*/
%let SgmntCount     = 400    900  100   300 ; /*officers/segment*/
%let ComplntsAvg    =    .5    5    5    10 ; /*avg compaints/segment*/
%let KudosAvg       =    .1    5    2     2 ; /*avg Kudos/segment*/
%let InjuryAvg      =    .1    5    2     2 ; /*avg Injuries/segment*/
%let ArrestAvg      =    .1    5    1    20 ; /*avg arrests/segment*/
%let ReleaseAvg     =    .1    2    2     2 ; /*avg released/segment*/
```

You can change the parameters to produce new datasets and see how well Proc LCA performs against your data.
Issues with Latent Cluster Analysis

ISSUES WITH LATENT CLUSTER ANALYSIS

Latent cluster analysis can have algorithmic problems with local minimum and maximum. It is a good idea to make several runs of PROC LCA using several different “starting points” to assure that the output is stable. If the output from your several runs is unstable, the creators of PROC LCA suggest you create a loop to runs several hundred LCAs with different starting points. The creators of PROC LCA suggest you compile the output of the hundred LCA runs and see where the modal points for output parameters might be and they describe this process in their book.

Missing values are often present in data and do not cause Proc LCA undue problems. It is assumed that missing values are missing completely at random and if this is not the case, additional effort and thought must be allocated to the data itself.

LCA assumes that, after a valid solution is found, variables are uncorrelated within clusters and this might not be the case.

PROC LCA output is harder to read than the examples in the book that was written by the authors of PROC LCA.

The book format is used in Figure 5. I should warn the reader that Figure 5 shows new data – data not collected with the police example.

I apologize for using a new data set but it was very difficult to construct one data set that was interesting enough, small enough and well behaved enough to allow it to be used to illustrate all of the points in this paper.

While I prefer to stick with one data set through a whole paper. It just was not possible.

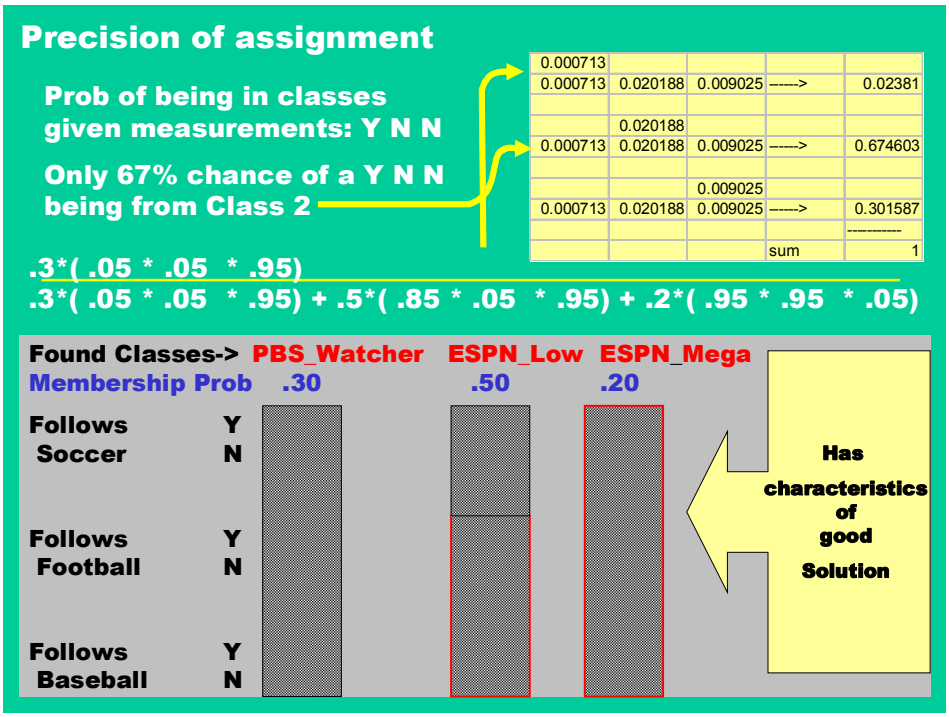


Figure 5

In Figure 5 you see the results from PROC LCA in an easy-to-read format. In the grayish box in the bottom of the slide you see the results from running a PROC LCA on a new, and imaginary, data set.

What you don't get from PROC LCA, are names of the "Found Classes" (e.g. "PBS_Watcher"). The "Found Classes" are the results of the researcher interpreting data and matching patterns on the PROC LCA output with his/her understanding of reality.

We see, in Figure 5, that the researcher is interpreting a three LCA solution. The membership probabilities, or the relative size of the groups, are 30%, 50% and 20%.

The researcher had asked three questions of the subjects.
The questions were:

Do you follow soccer?
Do you follow football?
Do you follow baseball?
The answers were a categorical yes or no.

We assume that the researcher has also run a two-cluster LCA, a three cluster LCA, a four-cluster LCA, and likely even a five-cluster LCA. Figure 5 only shows the results the three-cluster LCA and also how the researcher has mapped the numbers in the output listing to his understanding of external reality. What we see, in the output listing, are the percentage of people in the three-cluster solution that have answered a certain way. In cluster one, few people follow any of the sports. In cluster two (named ESPN_low) the people follow soccer. Maybe this cluster is made up of young people who played soccer in school or people exposed to soccer outside of the US. Members of the third cluster seem to follow all sports. 95% of the people in the third cluster (ESPN_mega) said yes to all of the three questions.

Figure 5 has the characteristics of a good solution – at least after a first level of examination. The sizes of the clusters and the pattern of responses to the questions map well to our understanding of the external reality. The patterns are distinct and interpretable. The sizes of the segments are "reasonable".

The second level of examination is to investigate the stability of assignments. This is done over each pattern of responses and, with three two-level responses, there are eight possible patterns of yes and no responses. One test

for stability is shown in the formula at the top of the Figure 5. In this formula, we are investigating the chance of responding yes, no, and no to the three questions ….. *given that you are from a particular class*.

You can see, the probability we are calculating is: the percent of the total number of subjects in a "Found Class" times the probability of a yes on question one times the probability of a yes on question two times the probability of a yes on question three ….then that whole thing divided by… that same calculation done for all groups.

We see is that the most common response pattern in the group we've named ESPN_low (the response pattern YNN)
only has a 67% chance of coming from somebody in ESPN_low. Almost 2/3 of the subjects with this response pattern were assigned to other clusters. This should decrease our faith in this solution.

Checking, of each response pattern probability and also each subject's classification probability (in another calculation that was not shown) is essential if these "found classes" are to be used as dependent variables in a later steps in a modeling process. Said in another way, good practice requires checking two (question pattern and subject) types of classification probability if a researcher wants to use cluster assignments as Y variables in further modeling,

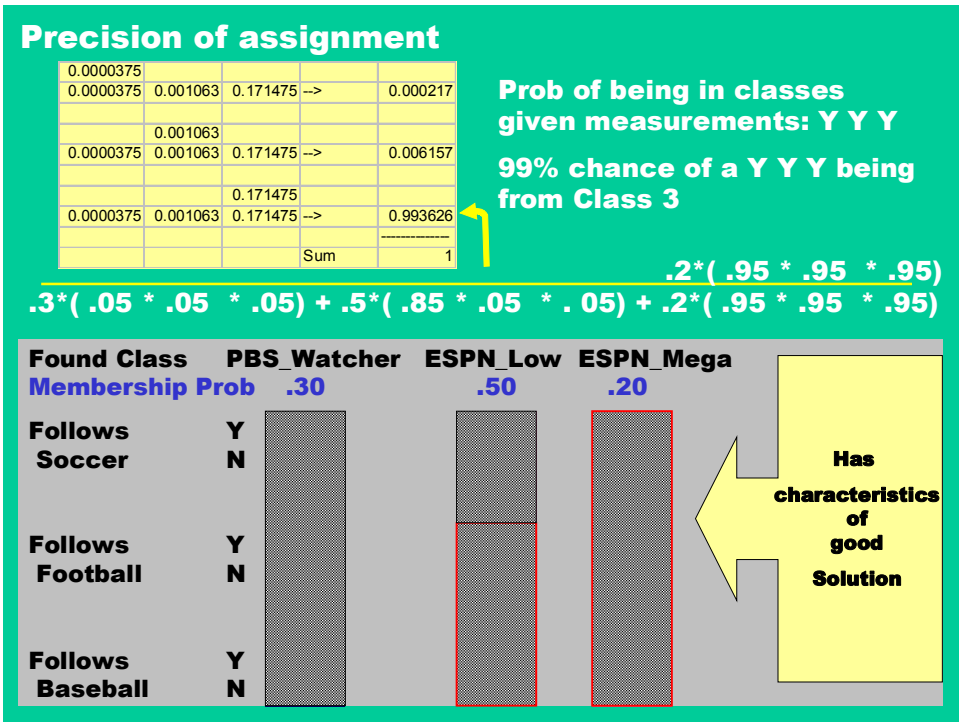
Inside a particular solution the stability of assignment can vary greatly among classes. In the figure to the right you can see that the probability of a response pattern YYY is very strongly associated with the group we've named ESPN_mega. A high (.99) likelihood of assignment is a characteristic of a good solution. Compare this the the classification probabilities for a YNN that is shown in Figure 5. Let's discuss some more characteristics of a good solution. Unfortunately, to do this I will bring in yet another new data set	
	Figure 6

Figure 7 uses a new data set to illustrate facilitate discussion of how good questions exhibit dependence across "Found Classes".

For question one, all "found classes" had the same response Probability. This question does not help us find a solution.

Theory, or our understanding of reality, should have directed us away from asking this question.

The third and fourth questions have great differences in their response patterns across the "Found Classes". These questions will be of use to the researcher.

Criteria 1: distribution varies across classes 50

Independence: Knowing something about X does not tell you anything about Y.

If the prob. of a level of response is the same across all classes, the variable is not of much use in the model or the number of classes is incorrect.

Found Class		Non_drinker	Social	Addict	Binge
		☹ INDEPENDENT and of little use ☹			
Drank Beer in last year	Y	.99	.99	.99	.99
	N	.01	.01	.01	.01
Drank this week	Y	.02	.30	.95	.50
	N	.98	.70	.05	.50
5 drinks-1 day in last year	Y	.01	.02	.50	.95
	N	.99	.98	.50	.05
		☺ DEPENDENT and of use ☺			
Drank Alone in last year	Y	.01	.05	.95	.25
	N	.99	.95	.02	.75
Was Sick while drinking in last year	Y	.01	.03	.50	.80
	N	.99	.97	.50	.20

Figure 7

Figure 8 illustrates another characteristic of good questions. Good questions, in LCA, have responses with either high or low probabilities.

This part of the analysis is very parallel to factor analysis, where questions with high factor loadings help identify the factors.

The nondrinkers are very different in their patterns from other people.

Additionally, because They have probabilities close to zero and one we can consider this to be a homogeneous group – a real group that likely exists in the population as a whole.

Criteria 2: Item-response probs are close to 1 or 0

High or low probabilities are associated with good solutions.

If the prob. of a level of response is the same across all classes, the variable is not of much use in the model or the number of classes is incorrect.

Found Class		Non_drinker	Social	Addict	Binge
Drank this week	Y	.02	.30	.95	.50
	N	.98	.70	.05	.50
		Real Group !	Real Group ?		
5 drinks-1 day in last year	Y	.01	.35	.50	.95
	N	.99	.65	.50	.05
Drank Alone in last year	Y	.01	.40	.95	.25
	N	.99	.60	.02	.75
		Homogeneous	Heterogeneous		

	Figure 8

The “Found Class” that was named “social drinkers” is much less homogeneous than the “non-drinker” “Found Class”. The fact that their response percentages are not close to zero and one could be caused by a natural heterogeneity in a group both truly exists and that is truly heterogeneous. It could also be caused by this being a poor solution. It could be that the “Found Class” named “social drinker” is not one true class. It could be that that this
group is a mixture of two homogeneous groups resulting in the heterogeneous group we see in Figure 8. It could be that the “Found Class” named “Social Drinker”, when we ask PROC LCA for a five group solution, separates into two
groups who have their probabilities close to zero and one.

MEASURES OF HOMOGENEITY AND SEPARATION:

Because of the individual differences in classification error people often look at the classification probabilities over all
the subjects using a formula that has the name of “Entropy”. The formula is:

$$E = 1 - \frac{\sum_{i=1}^{n} \sum_{C=1}^{C} -p_{ic} \log p_{ic}}{N \log C}$$

MEASURES OF GOODNESS OF FIT:

LCA has a few measures of goodness of fit and they are all based on a statistic called G^2.

Unfortunately, to have the example fit easily on a slide I must introduce still another new data set.

This new data set has two questions and you can see its nested frequency table on the left-hand side of Fig. 9.

PROC LCA looks at the observed nested frequency table and tries to create a nested frequency table that closely approximates the observed table.

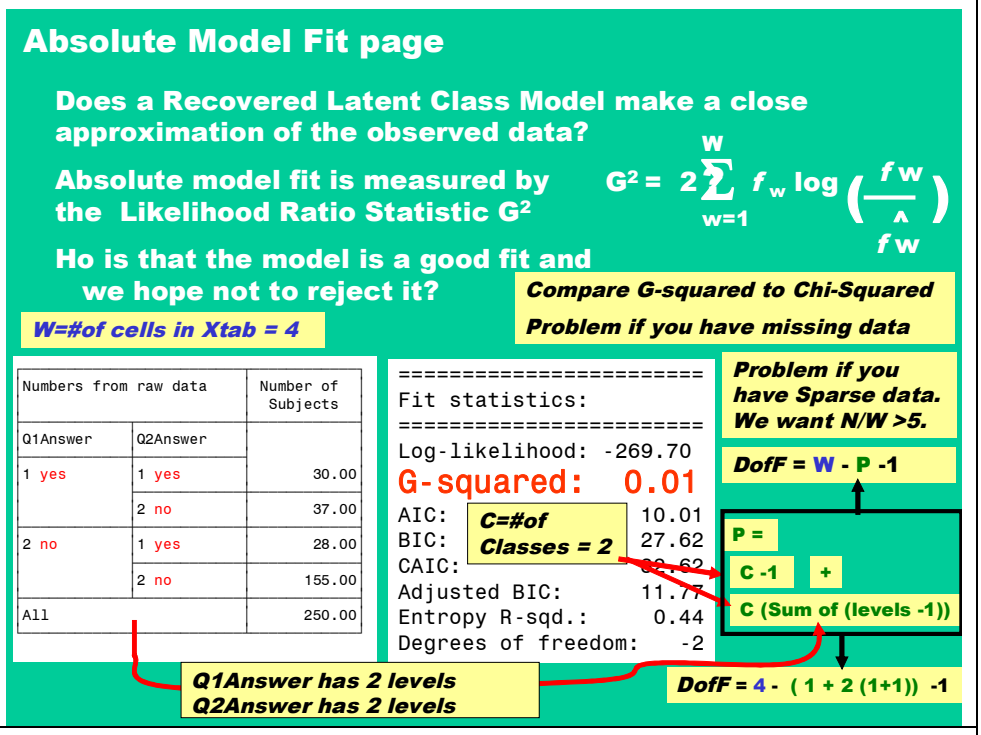


	Figure 9

I pasted the fit statistics from the PROC LCA into the slide in Figure 9. The PROC tells us that G2 is .01. G2 is then compared to a chi-squared and the degrees of freedom for G2 is the messy calculation shown in Figure 9. LCA

does
this calculation for us, but let us duplicate its work as a bit of practice.

The degrees of freedom calculation is W - P –1. Calculating W is easy but calculating P takes several steps.

W is the number of cells in the crosstab and here it is four. P is the result of a sub-calculation.

To calculate P, we start from the most nested question (Q2 is shown nested within Q1) and as we work "outwards", We do the following:

For each question we take a look at the number of levels of that question and subtract one.
(here we see Q1: 2-1=1 and Q2: 2-1=1)
We sum up the result of that subtraction over all the questions. 1+1

We multiply that sum by the number of Found Classes. We had asked PROC LCA for a two-class solution. 2*(1+1)

We add the result of this multiplication to a number we calculate by subtracting one from the number of classes requested from PROC LCA and we have P. Thankfully, PROC LCA does this calculation for us.

The calculation of G^2 itself is more complicated. The formula basically compares the frequency found in a cell (f_W) with the expected value of f_W-hat calculated by PROC LCA. If the numbers are close, then we have a good fit, so a small value of G^2 indicates a good model. We assume that the model is good and reject the model when G^2 gets large.

We are going to construct, from response probabilities predicted by LCA, the probability table from the raw data that
can be seen in Figure 10

Figure 10 shows more calculations for G^2.

Note that the words class and group are used interchangeably.

The white box in the top left shows output from PROC LCA.

The probabilities of being in the two classes are: .3478 and .6522.

I know, because I dummied up the data, that there are 250 observations in the data set. That lets us predict the number of people in each class to be: 86.95 (.3478*250) and 163.05 (250 *.6522) subjects.

For Class 1 the probability of a Yes-Yes is .6402 * .5468

Absolute Model Fit

$$G^2 = 2\sum_{w=1}^{W} f_w \log\left(\frac{f_w}{\hat{f}_w}\right)$$

```
Class:                    1       2
                     0.3478  0.6522
Rho estimates(item response probs):
Response category  1: yes
Class:                    1       2
  Q1Answer     :      0.6042  0.0882
  Q2Answer     :      0.5468  0.0637
Response category  2: no
Class:                    1       2
  Q1Answer     :      0.3958  0.9118
  Q2Answer     :      0.4532  0.9363
```

Fit statistics:
G-squared: 0.01

Class:	1	2
% of N in class	0.3478	0.6522
N (# of obs)	X 250	X 250
(pred)n in class	86.95	163.05

			Class1 n	Classn	N
			86.95	163.05	250

			Obs. Cell Count	Calculated n	Q1 Percent	Q2 Percent	Predicted N-for cell
Class1	Q1=1 Y	Q2=1 Y	30	86.95	0.6042	0.5468	28.72624189
		Q2=2 N	37	86.95	0.6042	0.4532	23.80894811
	Q1=2 N	Q2=1 Y	28	86.95	0.3958	0.5468	18.81801811
		Q2=2 N	155	86.95	0.3958	0.4532	15.59679189
Class2	Q1=1 Y	Q2=1 Y	30	163.05	0.0882	0.0637	0.916070337
		Q2=2 N	37	163.05	0.0882	0.9363	13.46493966
	Q1=2 N	Q2=1 Y	28	163.05	0.9118	0.0637	9.470214663
		Q2=2 N	155	163.05	0.9118	0.9363	139.1987753
						Total	250

Figure 10

PROC LCA thinks that there are 86.95 people in Class 1 and, for Class 1, the probability of a Yes-Yes is .6402 * .5468. So we expect class 1 to contribute 28.72 (86.95 * .6402 * .5468.) "counts" to the Yes-Yes

cell in the cross-tab PROC LCA created from the original data.

PROC LCA thinks that there are 163.05 people in Class 2 and, for Class 2, the probability of a Yes-Yes is .0882 * .0637. So we expect Class 2 to contribute .916 (163.05 *.0882 * .0637) "counts" to the Yes-Yes cell in the cross-tab that PROC LCA created from the original data.

The expected number of subjects responding Yes-Yes is: 28.72 + .916 = 29.64 You can look at Figure 9 to see the table from the raw data had a count of 30.0. The fit is close for this cell.

PROC LCA thinks that there are 86.95 people in Class 1 and, for Class 1, the probability of a No-Yes is .3958 * .5468. So we expect class 1 to contribute 18.82 "counts" to the No-Yes cell in the cross-tab PROC LCA created from the original data.

PROC LCA thinks that there are 163.05 people in Class 2 and, for Class 2, the probability of a No-Yes is .0882 * .0637. So we expect Class 2 to contribute 9.47 "counts" to the Yes-Yes cell in the cross-tab that PROC LCA created from the original data. The expected number of subjects responding No-Yes is: 18.82 + 9.47 = 28.29 and you can look at Figure 9 to see the table from the raw data had a count of 28.0.

We see that calculated counts are fitting the data pretty well and we should expect that our test statistics to be small and that we will "be allowed" to continue to assume Ho.

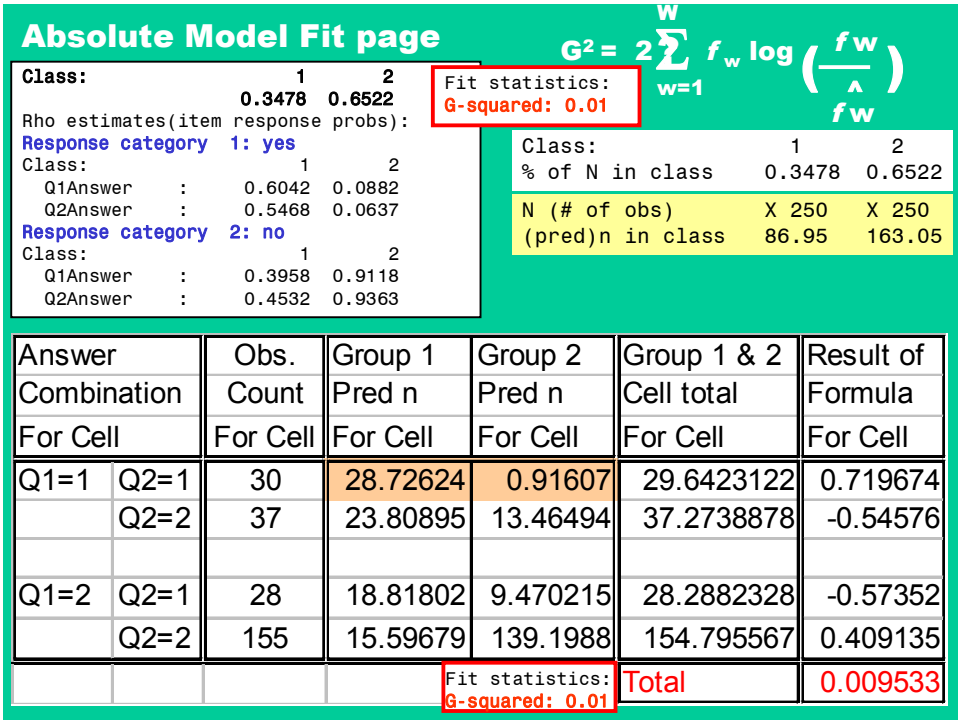
Figure 11 shows the end of the calculations.

Here, for each of the 4 cells, we calculated the predicted cell count.

The far right column in Figure 11 shows the Contribution, for each cell, to the total G^2.

Unfortunately the formula can not be shown inside the cells in this PowerPoint but the formula is repeated in the top right-hand corner of Figure 11 and the result is in the bottom right-hand corner.

We have reproduced the G^2 statistic (.009533 ~= .01).



Absolute Model Fit page

$$G^2 = 2\sum_{w=1}^{W} f_w \log\left(\frac{f_w}{\hat{f}_w}\right)$$

```
Class:                    1       2
                     0.3478  0.6522
Rho estimates(item response probs):
Response category  1: yes
Class:                    1       2
  Q1Answer     :      0.6042  0.0882
  Q2Answer     :      0.5468  0.0637
Response category  2: no
Class:                    1       2
  Q1Answer     :      0.3958  0.9118
  Q2Answer     :      0.4532  0.9363
```

Fit statistics: G-squared: 0.01

Class:	1	2
% of N in class	0.3478	0.6522
N (# of obs)	X 250	X 250
(pred)n in class	86.95	163.05

Answer Combination For Cell		Obs. Count For Cell	Group 1 Pred n For Cell	Group 2 Pred n For Cell	Group 1 & 2 Cell total For Cell	Result of Formula For Cell
Q1=1	Q2=1	30	28.72624	0.91607	29.6423122	0.719674
	Q2=2	37	23.80895	13.46494	37.2738878	-0.54576
Q1=2	Q2=1	28	18.81802	9.470215	28.2882328	-0.57352
	Q2=2	155	15.59679	139.1988	154.795567	0.409135
				Fit statistics: G-squared: 0.01	Total	0.009533

Figure 11

The G^2 statistic is an absolute measure of model fit.

Our usual process is to try to create a model that explains well and that is parsimonious (few clusters and few parameters to estimate).
To do this, we also need
relative measures of fit that will allow us to compare different models as we strive for parsimony.

There are two approaches to evaluating a parsimonious model. They are: 1) the likelihood ratio test of G^2 s and 2) comparing information criteria.

Relative Model Fit

You can test relative model fit as you develop a more parsimonious mode.

Two Approaches:

- **Likelihood ratio test of G^2**
- **Compare Information Criteria**

Do repeated analysis and Plot G2, AIC and BIC as a function of the number of classes recovered.

Likelihood ratio test of G^2

Models must be nested

The more parsimonious model has more parameter "restrictions"

Do the added "restrictions" cause a significant drop in G2

$G^2_{change} = G^2_{restricted} - G^2_{full}$

$DofF_{change} = DofF_{restricted} - DofF_{full}$

Compare Information Criteria

$AIC = G^2 + 2P$ and $BIC = G^2 + [\ln(N)]P$

Smaller is better

Figure 12

AIC and BIC penalize more complex models and the values for AIC and BIC can be seen in Figure 9.

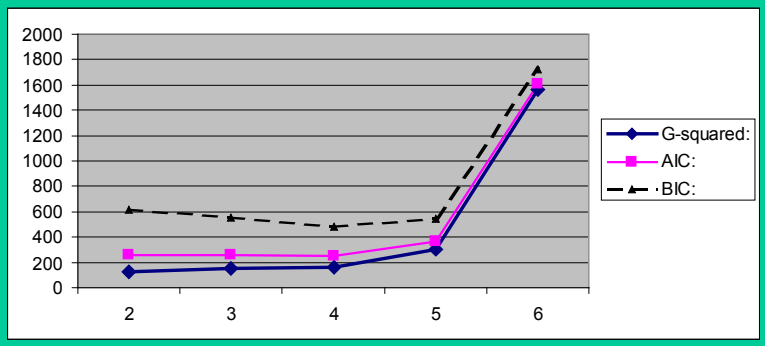
RESULTS FROM THE POLICE EXAMPLE: THE CODE INCLUDED IN APPENDIX

Please refer back to Figures 3 & 4 for information about the police example.
In the example, I created a dummy data set with 4 different types of police officer.

Imagine I checked officer records as to the number of: complaints, kudos, injuries, arrests, and prisoners later released.

In the program, I looped to create an observation for each officer, added error, and bucketed the numbers ("rarely", "occasionally" and "regularly").

Worked example with code



# of Clusters Recovered	2 Clusters	3 Clusters	4 Clusters	5 Clusters	6 Clusters
G-squared:	129.46	150.74	161.5	306.46	1566.23
AIC:	259.46	258.74	247.5	370.46	1608.23
BIC:	612.96	552.41	481.35	544.49	1722.43

Figure 13

As I started the example, I looked at the raw frequencies and saw variability on the data set level. This gave some hope that patterns of responses might exist in the 5-variable cross-tab that PROC LCA would create. If there is no variability in the data set, all responses are uniform and there is one pattern of responses (one cluster). If there is variability in the data set, there might be clusters that are uniform in their responses or there might be one cluster of people with a lot of variability within the cluster.

I instructed PROC LCA to run analyses for two clusters up to six clusters. The fit statistics, in Figure 13, suggest that
two, three or four clusters might be reasonable solutions. I looked at the PROC LCA output for three and four clusters and was most comfortable interpreting the four-cluster solution against my understanding of reality. Output from the four-cluster solution is included below (Figures 14 to 16).

I added some text boxes of information onto the figures so that we wouldn't have to remember as much about the problem. While questions are hidden by my added boxes of typing/information, the questions are in the order in which we always have presented them and can be discovered by looking back in the paper.

Remember we had three levels of response, created by the use of PROC Rank. We assigned subjects to "rarely", "occasionally" and "regularly" (coded rare, occasional and regular to save space).

Unfortunately, the output from PROC LCA is grouped by level of response over all questions. This is different from the format used in the Proc LCA book and what I have used in previous slides. This way Proc LCA presents results is awkward, at least for me, to interpret and I have avoided it until now.

Using the format from PROC LCA:
First, we will first see an analysis of the lowest response level ("RARE") in Figure 14.
Then we will then see an analysis of the middle response level ("OCCASIONAL") in Figure 15.
Then we will then see an analysis of the highest response level ("REGULAR") in Figure 16.
As mentioned before, the format of the output in the book is easier to use than the output from PROC LCA (shown below).

RARE:
First, look at the sizes of the clusters. PROC LCA gives us the percentage of the original data that was assigned to a particular cluster.

The returned percentages closely match what we know about the environment.

What we called Homicide is the smallest class. 76.18% of the homicide officers "Rarely get kudos". 74% "Rarely are injured". 99% "Rarely make arrests". These are characteristics of people who work on a few cases and who do not interact with the general public. 99.23% of the desk officers

```
 Parameter Estimates

Class membership probabilities: Gamma estimates (standard errors)
Class:                    1          2          3          4
                     0.1767     0.2409     0.0500     0.5323
                   (0.0100)   (0.0111)   (0.0075)   (0.0123)
                  Vice 300   Desk 400   Hom 100    Foot 900
Item response probabilities: Rho estimates (standard errors)

Response category  1: Rare

Class:                    1          2          3          4
  GrpComplaint:      0.0004     0.9791     0.0097     0.1647
                    (0.0013)   (0.0160)   (0.0328)   (0.0124)

%let TrueSegments =Desk Foot Hom Vice;
%let SgmntCount   = 400  900 100  300 ;      0.7618     0.0016
%let CmpIntsAvg   =  .5    5   5   10 ;     (0.0520)   (0.0016)
%let KudosAvg     =  .1    5   2    2 ;
%let InjuryAvg    =  .1    5   2    2 ;      0.7438     0.0015
%let ArrestAvg    =  .1    5   1   20 ;     (0.0531)   (0.0015)
%let ReleaseAvg   =  .1    2   2    2 ;
  GrpArrests   :     0.0059     0.9923     0.9970     0.1634
                    (0.0190)   (0.0043)   (0.0079)   (0.0124)

  GrpReleases  :     0.1386     0.8694     0.1466     0.1854
                    (0.0203)   (0.0175)   (0.0655)   (0.0130)
```

"rarely make an arrest".	
	Figure 14

OCCASIONAL:
In Figure 15 we see the percentages of subjects classified as "Occasional".

65.66% of the homicide detectives occasionally get complaints.

.26% of the officers in the group that we have named "Desk Officers", make occasional arrests.

.09% of the officers in the group that we have named homicide make occasional arrests.

55.02% of the officers in
the group that we have named foot, make occasional arrests.

```
 Parameter Estimates

Class membership probabilities: Gamma estimates (standard errors)
Class:                      1          2          3          4
                       0.1767     0.2409     0.0500     0.5323
                     (0.0100)   (0.0111)   (0.0075)   (0.0123)
                    Vice 300   Desk 400   Hom 100    Foot 900
Item response probabilities: Rho estimates (standard errors)

Response category  2: Occasional

Class:                      1          2          3          4
  GrpComplaint:        0.0006     0.0160     0.6566     0.6061
                     (0.0019)   (0.0123)   (0.0729)   (0.0164)

%let TrueSegments  =Desk  Foot  Hom   Vice;
%let SgmntCount    = 400   900   100   300 ;
%let CompIntsAvg   =   .5    5     5    10 ;
%let KudosAvg      =   .1    5     2     2 ;
%let InjuryAvg     =   .1    5     2     2 ;
%let ArrestAvg     =   .1    5     1    20 ;
%let ReleaseAvg    =   .1    2     2     2 ;
                                          18     0.2209     0.1180
                                        38)   (0.0496)   (0.0109)
                                          04     0.2144     0.1473
                                        71)   (0.0488)   (0.0121)
  GrpArrests   :       0.0003     0.0026     0.0009     0.5502
                     (0.0010)   (0.0025)   (0.0034)   (0.0167)

  GrpReleases  :       0.5439     0.0935     0.6146     0.5570
                     (0.0292)   (0.0151)   (0.0659)   (0.0166)
```

	Figure 15

REGULARLY:
Finally, look at the patterns where the classification is regularly or high.

99.91% of the vice squad officers make a high number of arrests. They also have 31.75 % of the officers classified as "high % for suspects later released".

The .9991 on the GrpComplaint row tells us that almost all vice squad officers are classified as getting "regular" complaints.

The recovered sizes of the four groups and the characteristics of the groups agree with the parameters that were used in creating the dummy data for PROC LCA.

```
 Parameter Estimates

Class membership probabilities: Gamma estimates (standard errors)
Class:                      1          2          3          4
                       0.1767     0.2409     0.0500     0.5323
                     (0.0100)   (0.0111)   (0.0075)   (0.0123)
                    Vice 300   Desk 400   Hom 100    Foot 900
Item response probabilities: Rho estimates (standard errors)

Response category  3: Regular / High

Class:                      1          2          3          4
  GrpComplaint:        0.9991     0.0048     0.3337     0.2292
                     (0.0023)   (0.0070)   (0.0712)   (0.0143)

%let TrueSegments  =Desk  Foot  Hom   Vice;
%let SgmntCount    = 400   900   100   300 ;
%let CompIntsAvg   =   .5    5     5    10 ;
%let KudosAvg      =   .1    5     2     2 ;
%let InjuryAvg     =   .1    5     2     2 ;
%let ArrestAvg     =   .1    5     1    20 ;
%let ReleaseAvg    =   .1    2     2     2 ;
                                          51     0.0173     0.8804
                                        37)   (0.0187)   (0.0110)
                                          74     0.0418     0.8512
                                        44)   (0.0256)   (0.0122)
  GrpArrests   :       0.9938     0.0051     0.0021     0.2864
                     (0.0190)   (0.0035)   (0.0072)   (0.0153)

  GrpReleases  :       0.3175     0.0371     0.2388     0.2576
                     (0.0273)   (0.0095)   (0.0501)   (0.0146)
```

	Figure 16

The managerial output from PROC LCA is similar to the output from other clustering methods. You get characteristics of the clusters that an analyst can use, by comparing the output to his/her understanding of the environment. The managerial output is clusters with useful name, sizes of the clusters and estimates of classification
accuracy.

CONCLUSION:

PROC LCA is a free add-in to SAS that you can download from the Methodology Center at Penn State University. Itis easy to install and use. The authors of the program have also written a very readable companion book: “Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences”.

We saw in the paper above that PROC LCA was very successful in recovering, from a dummy data set, the structure
that was designed into the data set - even after error was added to the data.

There is discussion among researchers as to the superiority of LCA over traditional clustering methods. Regardless of the position one takes on this issue, LCA is good to have this tool on your tool-belt.

The code for the example used in this paper is included in the appendix.

A summary of the steps in a PROC LCA analysis is:

1) Review the literature and talk with your subject matter expert. While there are statistics and programming skills needed in a Latent Cluster Analysis, interpreting the results requires knowledge of the underlying issues.

2) Check out the overall univariate frequencies. If there is little variability in the overall frequencies, there isn’t much of a hint that these variables will support a successful investigation. If the variables have three levels, and the marginal levels are all .33, you might be able to find interesting segments but the data does not suggest that. If the variables have three levels, and the marginal levels are .1 .4 and .6, there is some evidence that people inside the large sample are not replying uniformly and it gives stronger hope for the idea that there are patterns of responses inn the data that are coming out of different latent classes.

3) Consider creating a validation holdout sample, or even two holdouts (improvement and validation). As in any model results are biased towards the sample on which the model was developed. It is good practice to hold some percentage of your original data out of the modeling process and use that holdout to see if the model generalizes to new data.

4) Run PROC LCA and ask PROC LCA for solutions involving different numbers of latent classes. Run PROC LCA with different starting points to investigate the existence of local minimums/maximums.

5) Plot the goodness of fit measures for each of the N latent class solutions that you ask PROC LCA to produce. This should give hints as to the true number of latent classes in the data.

6) Interpret solutions with good test statistics and select a solution.

REFERENCES

Thanks to the group that created the free Proc LCA program :
The Methodology Center http://methodology.psu.edu/index.php
The Methodology Center is an interdisciplinary center that comprises faculty, research associates, post-docs, and students from several academic disciplines, including human development, psychology, statistics, and public health. Their work is funded by the National Institutes of Health and by the National Science Foundation.
Linda M. Collins, Ph.D. &Director, The Methodology Center
The Methodology Center
The Pennsylvania State University 204 E. Calder Way, Suite 400 State College, PA 16801

Collins & Lanza “Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences” (2010) Wiley Book web site is: http://methodology.psu.edu/latentclassbook/

Magidson & Vermunt "Latent class models for clustering: A comparison with K-means" , Canadian Journal of Marketing Research, Volume 20, 2002 http:www.statisticalinnovations.com

Thompson, David M. " Performing Latent Class Analysis Using the CATMOD Procedure" Proceeding of the SAS user group International (31) http://www2.sas.com/proceedings/sugi31/201-31.pdf

Magidson, J. a. "Latent Class (Finite Mixture) Segments: How to find them and what to do with them. (2010)" Presented at 2010 Sensometrics Meeting: Rotterdam, The Netherlands http://www.statisticalinnovations.com/articles/Magidson.Sensometrics2010.pdfin

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Russ Lavery Independent Contractor Bryn Mawr. Pennsylvania, U.S.A.
Email: Russ.Lavery@verizon.net
Web: Russ-Lavery.com