**Week 9 Discussion Board:**

In the context of predictive modeling why do we segment (or stratify) populations?  How should we construct these segments?  Should segments be constructed from an application point of view or a purely statistical point of view?  Give examples of why we would segment and how we should segment.

In the context of predictive modeling why do we segment (or stratify) populations?

Segmenting populations is necessary for accuracy purposes when creating predictive models. I found an example of why segmentation is necessary from a medical perspective. “Predictive modeling is a set of tools used to stratify a population according to its risk of nearly any outcome…ideally, patients are risk-stratified to identify opportunities for intervention before the occurrence of adverse outcomes that result in increased medical costs (Cousins, Shickle, Bander).”  The segmentation allows for better predictions for specific targeted outcomes.

 How should we construct these segments?    
          In the context of my medical example, individuals were stratified based on ordinal risk of  general health. This stratification allowed for specific delineations that narrowed the target segment seeking to be predicted. Everitt and Dunn discuss dispersion in section 6.3, which can be a helpful tool for numerically analyzing segments. Overall, it is my tyro opinion that segmentation should be conducted based on conclusions from the EDA coupled with an outside application.  
  
Should segments be constructed from an application point of view or a purely statistical point of view?    
          The construction of segments should be an iterative process that starts with a desired application strategically but tactically is implemented from a statistical point of view. One of the weaknesses with multivariate cluster analysis is drawing misleading conclusions. To placate this weakness, statistical point of view should be heavily enforced.   
  
Give examples of why we would segment and how we should segment.

For medical reasons, segmentation allows for better treatment of diseases and the prevention of expensive care.  The segmentation described above was based on risk of illness.

           From a business standpoint, segmentation allows for a refined target demographic. The segmentation that is conducted is based on core business competencies paired with market data. It should be noted that companies can also use segmentation for illegal purposes. Recently, Wells Fargo was hit with a large lawsuit based on segmenting customers based on race and predatory lending issues.

Segmentation is very useful when looking to identify an optimal target population, whether it be for marketing/sales purposes or to understand current customer demographics.  Stratification segments should be constructed so that they create the most variance in the target characteristics of your population.     
  
At my company our group frequently leverages customer stratification to proactively assist with business operations.  One example is we stratify our customers based a combination of value and risk (using a multivariate logistic regression model) to assist with our call-center operations.  When a low-value/high-risk customer calls to cancel their insurance policy, we forgo any retention efforts and simply “let them go”.   This reduces call-center costs (time on the phone) and reduces the number of low-value/high-risk customers on our books.  On the contrary, when a high-value/low-risk customer calls, we automatically route the call to the best available representative (“results based routing”) and execute maximum retention efforts.       
  
We relay the value/risk information to the representative answering the call as a screen-pop (like an email) so they know what to do and why.  The model is updated weekly (autonomously) so the stratified customer value/risk information stays current. When used appropriately, the segmentation of populations can be very powerful.     
  
The most similar (non-proprietary) example I could find is something FICO does (or sells, rather).

Dr Bahtti: Typically we use segmentation analysis to develop a "suite of models" as opposed to a single predictive model.

Segmentation allows for the grouping within populations that have distinctiveness to formulate classes in which can be used to build a model that can predict or describe a response. Having a suite of models instead of a single predictive model allows the analyst to compare multiple models  and various groups found through clustering to understand which one would have optimal predictive accuracy. An analyst should adhere to the principle of parsimony and having a suite of models would allow for this to happen.  
  
Segmentation based on analytics could be done through the initial analysis of the data. Where classes or groups are formed based on similarities among the individual data points. The variables that have strong relationships with each other could be combined into one group. While segmentation based on statistics, like Ryan mentioned, could be based on hierarchical methods and other similar processes, in which a set criteria is used to form a class of attributes. Data availability could be as simple as the constraints of time and availability of funds. In that segmentation allows for the recognition of specific classes that are able to build a predictive model without collecting or using enormous amounts of data.

In the context of predictive modeling, segmentation of a population is used to partition individuals into groups that are similar in characteristics. This should assist in understanding of future behaviors by recognizing which individuals are closely related and have exhibited similar actions. For example in marketing, many customers are not homogenous and differ with respect to one another in their preferences, wants, and needs. To be effective, it is practical to separate the market into groups of resembling behaviors. Segments should be constructed so that individual points that have been grouped together  are consequential  to the purpose of the research and distinct from one another. Segments can be created so there is a specific context or selected from predefined criteria, but ultimately should conducive to the purpose of the analysis. Segments should be constructed from an application point of view for this reason, it would be imperative that the segments are created so that the model that will be created provides the best information to make sound decisions. Though statistical tools, like clustering, can provide insight and define previously unrecognizable groups, it still remains the duty of the analyst to make sure that the cases do fit well with the definitional goal.

An example of why populations should be segmented could be understood through marketing analysis. Market segmentation enables for effective marketing strategies by selecting groups with similar needs and responses. Targeting will allow for determination into which segments to serve or which group responds well to advertising of a product. For marketing analysis, it is important to position a product so that it will compete well in the market with similar products. For marketing, segmentation can be divided by cohort, demographics, product usage, socioeconomic status, and other characteristics. This will all depend on what the product is and which groups respond to it.

Cluster analysis falls into the category of statistical problems known as "unsupervised learning".  What is an unsupervised learning problem?  
  
When performing a cluster analysis, how should we select the number of clusters to use?  Are there any metrics to help us in this selection?  Are there any caveats to keep in mind when using these metrics?

An Unsupervised learning problem is one in which there is no training data provided with previous classification of data into one or more groups.  The data is not classified and the goal is often to explore the data to find elements which share common characteristics and thus form natural clusters. Thus clustering can result in identify new categories of data which had not been previously identified. In evaluating clusters, we examine the distance among points within a cluster and the distance among points in different clusters. Good clusters will have small inter-cluster distances and large intra-cluster distances.  One measure of performance used for clustering algorithms such as KMeans is the silhouette.  The silhouette examines how dissimilar each data point within a given cluster is from the other data points in that cluster and then examines distances among data points from different clusters.  The ratio of these two metrics yields the silhouette.

The selection to the number of clusters to be used will depend on the clustering method, which can range from being heuristic to being a more formal procedure based on statistical models. Clustering methods will usually follow a hierarchical strategy or one in which the observations are relocated among tentative clusters. Hierarchical methods proceed by stages producing a sequence of partitions, each corresponding to different number of clusters.  They can be either agglomerative, which means that the groups are merged or they can be divisive, which means that one or more groups are split at each stage. At each stage of hierarchical clustering, the splitting or merging is chosen so as to optimize some criterion. Agglomerative hierarchical methods use heuristic criteria, such as single link, complete link, or sum of squares. Most of the cluster criteria used in multivariate analysis are based of the three dispersion matrices, which include total dispersion, within-group dispersion, and between group dispersion. In model based methods, a maximum-likelihood criterion is used for merging groups. It is important to keep in mind that those that use hierarchical classification should be wary of using heuristic methods if they are not clearly necessary (Everitt & Dunn, 2001). This would mean that when clustering methods are used, the segmentations made should be relevant to the purpose of exploring the data. If cluster are formed with no meaning, the model constructed would be ineffective for its practical purpose.

Impurity measures define how well classes are separated. In general, the impurity measure should be largest when the data has been split evenly for attribute values and should be zero when all the data belong to the same class. There are various impurity measures, which include the Gini Index. The Gini Index measures the divergences between the probability  distributions of the target attribute's values.