The field of statistics has innumerable applications in the world around us. Depending on the area of application, both the data collected and – more importantly – the specific research questions may suggest particular statistical techniques that are most appropriate. For example, in the behavioral sciences, questions often have to do with behavioral change over time; this suggests a need for methods to analyze longitudinal data. In addition, the behaviors under study often are too complex to adequately measure them over time using a single, error-free measure; latent variable models provide an opportunity to measure an underlying construct using multiple available measures.

One important area of research in the behavioral sciences is sexual risk behavior. This behavior can lead to unplanned pregnancies or the acquisition of sexually transmitted infections including HIV. It is the intersection of multiple aspects of behavior (e.g., sexual intercourse, the number of sexual partners, inconsistent condom use) that places individuals at highest risk, although few studies have attempted to characterize behavior according to all of these dimensions. Latent transition analysis (LTA) can be used to model transitions over time in discrete latent variables and investigate predictors of initial behavior status and transitions in behavior.

This approach will be demonstrated in an empirical study using a sample of 2937 adolescents from the NLSY97. Dating and sexual risk behavior is modeled as a discrete latent variable indicated by number of dating partners, past-year sex, number of sexual partners, and possible STD exposure. Development across three time points in the behavior is examined and compared across genders, and use of various substances is explored as predictors of sexual risk behavior. Use of a new SAS procedure developed at Penn State, PROC LTA, will be described.

Results suggest five latent classes of dating and sexual behavior: Non-daters (18.6% at Time 1), Daters (28.9%), Monogamous (11.7%), Multi-partner safe (23.1%), and Multi-partner exposed (17.7%). Non-daters who change status are most likely to transition to the Daters status, and Daters who change status tend to move to the Monogamous and the Multi-partner safe statuses. Interestingly, the highest probability of transitioning to the Multi-partner exposed status is among individuals in monogamous relationships. Significant gender differences were identified (p<.0001), with females more likely to belong to the Monogamous class and males the Multi-partner safe class. Drunkenness and marijuana use are stronger predictors of high-risk sexual behavior than cigarette use.
We propose a model-robust exchange algorithm which produces exact experimental designs that maximize the minimum $\mathcal{D}$-efficiency with respect to a set of user-specified models. The method relaxes the optimal design assumption that the model form is known completely in advance, and produces designs for which each possible model in the set will be well estimated. Further, we present a generalization of this method which allows the user to express varying levels of interest in each potential model; the resulting design will be suggestive of these differences. Some asymptotic properties of the maximin criterion are also explored, including a condition which guarantees a design with the same generalized $\mathcal{D}$-efficiencies for each possible model. The new algorithm provides experimenters with more control of these efficiencies than comparable procedures in the literature and constitutes an effective alternative to usual design strategies under model uncertainty.

The primary goal of this talk is the analysis of long sequence data generated in biology, such as SNP data. Suppose we have observed $n$ current descendant sequences of length $L$, one interesting question is that how to estimate the unknown ancestral distribution from the observed descendants, considering realistic biology complexities such as mutation and recombination. We have developed a statistical model by extending the ancestor mixture model (Chen and Lindsay (2006)) with both mutation and recombination to estimate the ancestral distribution. However, though we can write out the full likelihood for ancestral distribution explicitly, there is an enormous computation challenge when applying it on data due to an enormous number of recombination possibilities, which grows exponentially in sequence length. Therefore, we apply composite likelihood as an approximation to solve the problem. In this talk, we first introduce our developed statistical model and composite likelihood method. Then, some simulation results are shown to investigate the performance of composite likelihood in long sequence data.
People debate whether or not a covariate-adjusted approach should be used as the primary analysis. As expected, omitting predictive covariates often leads to misspecified models in which the parameter of interest is difficult to interpret, particularly when omitted covariates interact with the main predictive variable. Under a generalized linear model framework, we derive the analytical relationship between the parameters of interest in the potentially misspecified model and the true model. Meanwhile, we show that for a broad class of generalized linear models, the estimates obtained from a covariate-adjusted model have greater variances compared to those from an unadjusted model. These theoretical results are illustrated and validated through two examples and a simulation study. We allow models to include treatment/covariates interactions, and hence in terms of accuracy, our results include analogue conclusions in Gail et al (1984) as special cases. In terms of precision, we make substantial extensions of results in Robinson and Jewell (1991) to a broad class of generalized linear models including the most frequently used ones.

We present some new methods based on kernel density estimation and a modal expectation-maximization (MEM) method for clustering DNA haplotype sequences. For the simple mission of clustering binary sequences, we define a kernel density estimator for the sequences and use it to define a weight function for each sequence. Then we start from each data sequence and examine all other sequences along with the weight function to find the nearest mode of the density. We then cluster the sequences that share the same mode. For the haplotype problem, we construct a likelihood function for the genotype-haplotype problem that depends on the unknown haplotype type density and then use likelihood EM to create an updated density that partially maximizes the likelihood. The performance of the method regarding haplotype inference is tested on large datasets with the comparison to the available methods such as Phase. It shows that the new method yield comparable results while requiring less computational time. In a similar fashion, we can define a density estimator for the binary sequences (haplotypes) in the presence of recombination and mutation. One direct outcome is the resulting tree structure converges to a single ancestor faster than the ones that are based on a model of mutation alone.
SESSION 2
1:30 – 3:30
Chaired by Dr. Naomi Altman

1:30-2:30
Andreas Artemiou (advisor: Dr. Bing Li)
“Support Vector Machine for Sufficient Dimension Reduction”

We propose a new method for supervised dimension reduction using support vector machines. This is based on previous inverse slicing methods but support vector machines, instead of sample moments, are applied between slices to extract effective predictors. This method has two advantages; it is robust against outliers and it can extract nonlinear features.

2:30 – 2:50
Debashis Ghosh
“Causal Inference Methods in Medical Research”

In this talk, I will give a brief introduction to the field of causal inference, whose techniques are being increasingly used in scientific and medical research. I then will describe a causal inference working group that has begun this semester at Penn State, along with some recent research directions in the area.

3:00 – 3:30
Student Poster Session

Jialin Xu (advisor: Dr. Yu Zhang)
“A Generalized Linear Model for Peak Calling in ChIP-Seq Data”

Qing Wang (advisor: Dr. Bruce G. Lindsay)
“The Unbiased Estimator of the Variance of a U-Statistic and Its Resampling Realization”

3:15-4:00
Coffee/Cookie Break
330 Thomas Building
4:00-5:00

**Keynote Address**
201 Thomas Building

**James Farber**
United States Census Bureau

“Statistical Research and Application at the Census Bureau”

The U.S. Census Bureau has been referred to as the nation’s preeminent statistical agency. The decennial census is unquestionably the marquee program of the bureau. However, many smaller programs play vital roles in government, industry, and academia. Sound statistical methodology is fundamental to the effectiveness of Census Bureau programs. Many areas in the bureau maintain active programs of applied statistical research to evaluate and improve censuses and surveys. I will talk about my experiences as a mathematical statistician at the Census Bureau, highlight interesting projects in several areas of the bureau, and offer thoughts about career opportunities.

5:30-7:30

**Dinner Reception**
330 Thomas Building

Thank you for attending the 2010 Department of Statistics Alumni Workshop. A special thanks to all SAC and Admission Committee Members for their hard work and dedication.

We look forward to seeing you again at next year’s event!!!