Program

Twelfth Annual Winter Workshop
Categorical Data Analysis

Department of Statistics
University of Florida
January 15-16, 2010

Contents

COVER GRAPH: These data contain 880 valid cases, each from an interview with a Scottish national after the election. Comparison of \$90%\$ credible intervals for probit regression with a Dirichlet process random effects model (black) to those from a probit regression with a normal random effects model (blue). (Kyung et al (2009, Annals of Statistics)

Sponsors

National Science Foundation, InfoTech, the College of Liberal Arts & Sciences, Graduate School, and the Department of Statistics of the University of Florida.

Organizing Committee

Alan Agresti
George Casella
Michael Daniels
Linda Young

Invited Speakers

Alan Agresti, University of Florida
Jim Albert, Bowling Green State University
Jon Forster, University of Southampton (Great Britain)
Gary Koch, University of North Carolina
Diane Lambert, Google
Joseph Lang, University of Iowa
Xihong Lin, Harvard University
Stuart Lipsitz, Brigham and Women's Hospital, Harvard University
Peter McCullagh, University of Chicago
Art Owen, Stanford University
Nancy Reid, University of Toronto (Canada)

Acknowledgements

The organizers thank the Department of Statistics staff, Robyn Crawford, Tina Greenly, Summer Rozear and Marilyn Saddler for their tremendous efforts in helping to set up this meeting and make it run smoothly.

Conference Schedule

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Thursday, January 14, 2010

7:00-9:00pm Reception Emerson Hall, Al and Judy Warrington Room

Friday, January 15, 2010

8:00-8:30am Continental Breakfast

8:30am Welcome by:
Alan Dorsey, Associate Dean, College of Liberal Arts & Sciences
Linda Young, Conference Chair

8:45-9:30am **Session 1:** ***History and Overview of Categorical Data Methods***
Chair: Linda Young, University of Florida

Speaker: Alan Agresti, University of Florida
Some Historical Highlights in the Development of Categorical Data Methodology

9:30-10:00am **Break**

10:00-11:45 **Session 2:** ***Semiparametric Methods***
Chair: Mike Daniels, University of Florida

Speakers: Xihong Lin, Harvard University
Nonparametric & Semiparametric Regression with Missing Outcomes Using Weighted Kernel & Profile Estimating Equations

Diane Lambert, Google
Robust Doubly Robust Estimates

11:45am **Conference photo on the Emerson Alumni Hall stairs to the second floor**

12:00-l:45pm **Lunch** (Gator Comer Dining Center)

l:45-3:30pm **Session 3:** ***Higher-Order Approximations and Other Methods of Improving Inference***
Chair: Alan Agresti, University of Florida

Speakers: Nancy Reid, University of Toronto
Accurate Approximation for Inference on Vector Parameters

Joseph Lang, University of Iowa
Multinomial-Poisson Homogeneous Models: A Practitioner ’s Guide

3:30-5:30pm **Poster Session:** Emerson Alumni Hall, Presidents Room B

All workshop sessions meet in the 209 Teaching Classroom, Emerson Alumni Hall

Saturday, January 16, 2010

8:30-9:00am Continental breakfast

9:00-10:45am **Session 4:** ***Wilcoxon-Type Rank Methods for Ordinal Data***

Chair: Ronald Randles, University ofFlorida

Speakers: Stuart Lipsitz, Brigham & Women's Hospital, Harvard University
Wilcoxon Rank-Sum Test for Clustered and Complex Survey

Gary Koch, University of North Carolina
Stratified Multivariate Mann-Whitney Estimators for the Comparison of Two Treatments with Randomization Based Covariance Adjustment

10:45-ll:15am **Break**

ll:15-l:00pm **Session 5:** ***Complicating Issues in Using Logistic Regression***

Chair: Brett Presneli, University ofFlorida

Speakers: Peter McCullagh, University of Chicago
Sampling Bias in Logistic Models

Art Owen, Stanford University
Infinitely Imbalanced Logistic Regression

1:00-2:45pm **Lunch** - free time

2:45-4:30pm **Session 6:** ***Bayesian Inference for Categorical Data***

Chair: Malay Ghosh, University ofFlorida

Speakers: Jim Albert, Bowling Green State University
Good Smoothing

Jonathan J. Forster, University of Southampton
Bayesian Model Averaging for Categorical Data

5:30-8:30pm. Barbecue at the home of Alan Agresti and Jacki Levine
1632 NW 24th Street, Gainesville, FL 32605

Invited Talks

Some Historical Highlights in the Development of Categorical Data Methodology

Alan Agresti
University of Florida

This talk surveys the development of the basic methods of categorical data analysis, starting with Karl Pearson's introduction of the chi-squared test in 1900. It focuses on the contributions of statisticians who introduced methods such as inferences about associations in contingency tables, logistic regression and probit models, loglinear models, and Bayesian approaches. The last part of the talk will highlight some more recent developments, focusing on contributions made by the invited speakers at this workshop.

Good Smoothing

Jim Albert
Bowling Green State University

One of the most influential Bayesians in the analysis of categorical data was I. J. Good. One problem addressed by Good in many papers was the treatment of tables with sparse counts. Good’s famous 1967 JR.SS paper on testing for equiprobability is reviewed with a focus on the smoothing problem. Following Good’s strategy, Bayesian procedures for smoothing proportion and two-way contingency table data are described. These ideas are generalized to simultaneously estimate proportions and odds ratios in a data mining application.

Bayesian Model Averaging for Categorical Data

Jonathan J. Forster
University of Southampton

It is common for multivariate categorical data (which may be represented as a contingency table) to be unbalanced or sparse, particularly when the dimensionality is large. Then, estimating cell probabilities, or predicting the unobserved population in a finite population sampling analysis, typically relies on some kind of modelling to provide smoothed estimates. In this talk I will investigate Bayesian model averaging as a estimation method for multivariate categorical data which allows multiple models to be entertained. I will discuss default choices of model class, and of prior distributions on model parameters, across a range of applications.

Stratified Multivariate Mann-Whitney Estimators for the Comparison of Two Treatments with Randomization Based Covariance Adjustment*

Gary Koch
University of North Carolina

Methodology for the comparison of two randomly assigned treatments for strictly ordinal response variables has discussion in terms of multivariate Mann-Whitney estimators with stratification adjustment. Although such estimators can have direct computation by two-way analysis of variance for within stratum ridits (as ranks/(sample size) within each stratum), determination of a consistent estimator for their covariance matrix is through methods for multivariate U-statistics. The scope for these methods includes ways of managing randomly missing data and for invoking randomization based covariance adjustment for no differences between treatments for background or baseline covariables. The assessment of treatment differences can be through confidence intervals or statistical tests for the adjusted Mann-Whitney estimators in their own right or for their counterparts from linear logistic models. Three examples have illustrative results presented for the methods in this paper. The first example is a randomized clinical trial with four strata and a univariate ordinal response variable. The second example is a randomized clinical trial with four strata, two covariables (as age and baseline) and four ordinal response variables. The third example is a randomized two period crossover clinical trial with four strata, three covariables (as age, screening, first baseline) and three response variables (as first period response, second baseline, second period response). For each of these examples, the results from analyses with adjusted Mann-Whitney estimators are interpretable in terms of the probability of better outcomes for test treatment than the control treatment. When the distributions of each of the response variables for the two treatments have compatibility with the proportional hazards assumption, the interpretation of adjusted Mann-Whitney estimators can be through their hazard ratio counterparts as well.

* Atsushi Kawaguchi 1, Gary G. Koch2, and Xiaofei Wang3
IBiostatistics Center, Kurume University, 67 Asahi-Machi Kureme-City Fukuoka 830-0003, Japan. 2Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599-7420, U.S.A. 3Department of Biostatistics and Bioinformatics, Duke University Medical Center, Box 2721 2424 Erwin Road, Suite 1102 Durham, NC 27710, U.S.A.

Robust Doubly Robust Estimates

Diane Lambert
Google

Search engines that place ads on the web are dynamic systems. New tools that help advertisers fine-tune and enhance their campaigns are always being developed and offered to advertisers, and advertisers are always changing their ads and target audience. In an ideal world, the effectiveness of both tools for advertisers and campaigns by advertisers would be evaluated with statistical experiments, but in the real world such experiments are rare. Worse still, the outcome is often a rare binary event, analysis is often automated, and results are evaluated without the benefit of a statistician. Causal models and doubly robust estimation are obvious candidates for estimation in such circumstances, but they may also behave poorly if some of the controls had little chance to be treated, even if the goal is only to estimate the effect on the treated. This talk describes a score test and alternative estimate that can be used to detect the presence of untreatable controls and further robustify estimated treatment effects.

Multinomial-Poisson Homogeneous Models: A Practitioner's Guide

*Joseph B. Lang**
University of Iowa

Multinomial-Poisson homogeneous (MPH) models make up a broadly-applicable class of contingency table models, a class that includes both log-linear and non log-linear models. For example, any model that places smooth constraints on the table probabilities can be formulated as an MPH model. Other examples of MPH models include loglinear and generalized loglinear models, and the multinomial linear predictor models of Grizzle et al. (1969). The impetus behind the introduction of MPH models was to free researchers to think outside, or inside, the loglinear model box when faced with contingency table data. Specifically, with MPH models, researchers can more directly formulate hypotheses of interest and carry out the corresponding likelihood-based inferences.

Whereas previous research focused on the derivation of theoretical results, this presentation is more of a practitioner's guide to MPH modeling. We give a simplified description of the models and explain the utility of using homogeneous constraints. In addition, we explain how to use the R program mph.fit to carry out maximum likelihood estimation of MPH models and the special sub-class of Homogeneous Linear Predictor (HLP) models. Several examples illustrate the MPH/HLP implementation. Along the way, the examples will touch on other topics including improved interval estimation, estimability of contingency table parameters, and graphical displays of model goodness of fit.

*Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA, joseph-lang@uiowa.edu, www.stat.uiowa.edu/jblang.

Nonparametric and Semiparametric Regression with Missing Outcomes Using Weighted Kernel and Profile Estimating Equations

XihongLin
Harvard University

We consider nonparametric and semi-parametric regression when an outcome is missing at random (MAR). We first consider nonparametric regression of a scalar outcome on a covariate under MAR. We show that nonparametric kernel regression estimation based only on complete cases is generally inconsistent. We propose inverse probability weighted (IPW) kernel estimating equations and a class of augmented inverse probability weighted (AIPW) kernel estimating equations for nonparametric regression under MAR. Both approaches do not require specification of a parametric model for the error distribution and the estimators are consistent when the probability that a sampling unit is observed, i.e., the selection probability or the response probability, is known by design or is estimated using a correctly specified model. We show that the AIPW kernel estimator is double-robust in that it yields valid inference if either the model for the response probability is correctly specified or a model for the conditional mean of the outcome given covariates and auxiliary variables is correctly specified. In addition, we argue that adequate choices of the augmented term in the AIPW kernel estimating equation help increase the efficiency the estimator of the nonparametric regression function. We study the asymptotic properties of the proposed IPW and AIPW kernel estimators. We extend the results to semi-parametric regression under MAR where one covariate effect is modeled non-parametrically and some covariate effects are modeled parametrically. IPW and AIPW profile-kernel estimating equations are proposed to estimate the parametric component. Asymptotic semi-parametric efficiency is studied. We perform simulations to evaluate their finite sample performance, and apply the proposed methods to the analysis of the AIDS Costs and Services Utilization Survey data.

Wilcoxon Rank-Sum Test for Clustered and Complex Survey

Stuart Lipsitz
Brigham and Women's Hospital, Harvard University

The Wilcoxon rank-sum test is one of the most frequently used statistical tests for comparing an ordered categorical outcome between two groups subjects. Even in cases where more complicated analyses are subsequently performed, initial summaries in terms of bivariate analyses are regularly reported. With a study with clustered data, such might arise in a cluster randomized study of two treatments for LDL cholesterol control (<3.4,3.4-4.1, 4.1-4.9,>4.9), one would be interested in testing for equal treatment effects on LDL cholesterol control using a Wilcoxon rank-sum type test interested. However, since the subjects within a cluster are not independent, the assumptions needed for application of the Wilcoxon rank-sum test do not hold. With independent subjects, the Wilcoxon rank-sum test can be shown to be a score test from a proportional-odds cumulative logistic regression model, in which the "continuous outcome" is treated as the ordinal outcome, and the group is a dichotomous covariate; here the Wilcoxon rank-sum test is the score test for no effect of the dichotomous covariate. With clustered and complex sample survey data, we propose formulating a similar proportional-odds cumulative logistic regression model, in which the 'continuous outcome' is treated as the ordinal outcome, and the group is a dichotomous covariate. Our extension of the Wilcoxon rank-sum test to clustered and complex survey data will be a generalized estimating equations score test for no group effect under a working independence assumption for this cumulative logistic regression model. The test can easily be obtained in common statistical software such as SAS Proc Genmod or SAS Proc Survey Logistic.

Sampling Bias in Logistic Models

Peter McCullagh
University of Chicago

It is common for multivariate categorical data (which may be represented as a contingency table) to be unbalanced or sparse, particularly when the dimensionality is large. Then, estimating cell probabilities, or predicting the unobserved population in a finite population sampling analysis, typically relies on some kind of modelling to provide smoothed estimates. In this talk I will investigate Bayesian model averaging as a estimation method for multivariate categorical data which allows multiple models to be entertained. I will discuss default choices of model class, and of prior distributions on model parameters, across a range of applications.

Infinitely Imbalanced Logistic Regression

Art B. Owen
Stanford University

In binary classification problems it is common for the data sets to be very imbalanced: one class is very rare compared to the other. In this work we consider the infinitely imbalanced case where the rare class has fixed finite sample size n, while the common class has sample size N -»-oo. For logistic regression, the infinitely imbalanced case often has a useful solution. The logistic regression intercept typically diverges to — oo as expected. But under mild conditions, the rest of the coefficient vector approaches a non trivial, interpretable and useful limit. Perhaps surprisingly, the limiting parameter vector depends on the n points from the rare class only through their sample mean.

Accurate Approximation for Inference on Vector Parameters

Nancy Reid
University of Toronto

The theory of higher order approximation is most easily developed for scalar parameters of interest, and for continuous responses. We consider how the derivations can be extended to discrete responses and how this can be combined with directional tests. The goal is to provide higher order approximations for inference in multi-dimensional problems for categorical data. This work is joint with Don Fraser, Anthony Davison and Nicola Sartori.

Poster Abstracts

Changing Approaches of Prosecutors Towards Juvenile Repeated Sex-Offenders: A Bayesian Evaluation

*Dipankar Bandyopadhyay**
Medical University of South Carolina

Existing state-wide data bases on prosecutors' decisions about juvenile offenders are important, yet often unexplored resources for understanding changes in patterns of judicial decisions over time. We investigate the extent and nature of change injudicial behavior towards youths following the enactment of a new set of mandatory registration policies between 1992 and 1996 via analyzing the data on prosecutors' decisions of moving forward for youths repeatedly charged with sexual violence in South Carolina. We use a novel extension of random effects logistic regression model for longitudinal binary data via incorporating an unknown change-point year. For convenient physical interpretation, our models allow the proportional odds interpretation of effects of the explanatory variables and the change-point year with and without conditioning on the youth-specific random effects. As a consequence, the effects of the unknown change-point year and other factors can be interpreted as changes in both within youth and population averaged odds of moving forward. Using a Bayesian perspective, we consider various prior opinions about the unknown year of the change in the pattern of prosecutors' decision. Based on the available data, we make posteriori conclusions about whether a change-point has occurred between 1992 and 1996 (inclusive), evaluate the degree of confidence about the year of change-point, estimate the magnitude of the effects of the change-point and other factors, and investigate other provocative questions about patterns of prosecutors' decisions over time.

*This is a joint work with Debajyoti Sinha, Florida State University, Stuart Lipsitz, Harvard Medical School, Boston and Elizabeth Letourneau, Medical University of South Carolina. This work was funded by a CDC grant.

Generating Correlated Binary Variables Using Pair-Copulas

*Lianfu Chen**
Texas A&M University

Sampling from correlated binary variables with specified mean vector and correlation matrix but without fully specified joint distribution is a challenging problem. Erhardt and Czado (2008) proposed a method for approximately sampling from high dimensional count variables with specified mean and correlation matrix based on Gaussian paircopula. We develop an analogue of this method using the student t pair-copula to simulate correlated binary variables. An optimization routine to determine the parameters in the copula sequentially using the quasi-Newton method is introduced. We then compare our method to the methods of Erhardt & Czado (2008), Emrich & Piedmonte (1992) and Qaqish (2003). Instead of using the maximal relative bias and the average number of acceptance as in Erhardt & Czado (2008), we suggest using the mean absolute element wise difference and the Kullback-Leibler loss as the performance measure when making comparisons.

*This is a joint work with Mohsen Pourahmadi, Texas A&M University.

Bootstrapping Possibly Misspecified Models

Mihai C. Giurcanu
University of Louisiana at Lafayette

In this paper we analyze the asymptotic properties of various bootstrap procedures for possibly misspecified moment condition models. We prove that the usual uniform bootstrap, the centered-bootstrap, and the biased-bootstrap yield consistent estimators of the null distributions of the t-test, Wald test, and QIF test statistics if the model is correctly specified. Moreover, we show that the uniform bootstrap estimator of the null distribution of the J-test is inconsistent and converges in distribution to some random non-central chi-square distribution. We also prove that, under misspecification, the ordinary uniform bootstrap is consistent and that the centered-bootstrap fails. An empirical study shows the finite sample behavior of various resampling procedures for a possibly misspecified panel data model.

Analysis of Zero-Inflated Clustered Count Data Using Marginalized Model Approach

Keunbaik Lee *
Louisiana State University Health Sciences Center

Min and Agresti (2005) proposed random effect hurdle models for zero-inflated clustered count data with two-part random effects for a binary component and a truncated count component. In this paper, we propose new marginalized models for zero-inflated clustered count data using random effects. The marginalized models are similar to Dobbie and Welsh's (2001) model in which generalized estimating equations were exploited to find estimates. However, our proposed models are based on likelihood-based approach. Quasi-Newton algorithm is developed for estimation. We use these methods to carefully analyze two real datasets.

*This is a joint work with Yongsuung Joo, Dongguk University, Korea, Joon Jin Song University of Arkansas, Fayetteville.

New Algorithm for Iteratively Weighted Partial Least Squares for High Dimensional Data

Yanzhu Lin *
Purdue University

In order to deal with the generalized linear model (GLM) with a large number of predictors (p) for a small number of subjects (n), we incorporate partial least squares (PLS) to GLM framework. Different from previous methods based on the extension of PLS to categorical data, we sequentially construct the components by treating the working response generated by GLM as the input of PLS. To avoid the long-standing convergence issue of maximum likelihood (ML) estimation of GLM, bias correction to the likelihood, which was originally presented by Firth, has been applied. After generating all of the components, we fit the GLM by treating the components as the predictors. We also consider incorporating penalized framework to our method to generate the sparse loadings for each component, which will result in a GLM with a large number of original predictors having zero coefficients.

*This is a joint work with Dabao Zhang, Min Zhang, Department of Statistics Purdue University

Generalized Additive Partial Linear Models — Polynomial Spline Smoothing Estimation and Variable Selection Procedures

*Xiang Liu**
University of Rochester

We study generalized additive partial linear models, proposing the use of polynomial spline smoothing for estimation of nonparametric functions, and deriving quasi-likelihood based estimators for the linear parameters. We establish asymptotic normality for the estimators of the parametric components. The procedure avoids iterative backfitting and thus results in gains in computational simplicity. We further develop a class of variable selection procedures for the linear parameters by employing a non-concave penalized likelihood, which is shown to have an oracle property. Monte Carlo simulations and an empirical example are presented for illustration.

*This is a joint work with Li Wang, Hua Liang and Raymond J. Carroll.

What Can Go Wrong When Ignoring Correlation Bounds in the use of Generalized Estimating Equations

Roy T. Sabo
Virginia Commonwealth University

Abstract: The analysis of repeated measure or clustered data is often complicated by the presence of correlation. Further complications arise for discrete responses, where the marginal probability-dependent Fréchet bounds impose feasibility limits on the correlation that are often more restrictive than the positive definite range. Some popular statistical methods, such as Generalized Estimating equations (GEE), ignore these bounds, and as such can generate erroneous estimates and lead to incorrect inferential results. In this paper we give examples of the repercussions of incorrectly using GEE in the presence of correlated binary responses. We also provide two alternative strategies: (i) using GEE to select a data-driven correlation value within the Fréchet bounds, and (ii) the use of likelihood based latent variable modeling (such as multivariate probit) to get around the problem all together.

Sample Size Re-calculation in Sequential Diagnostic Trials

Liansheng Larry Tang
George Mason University

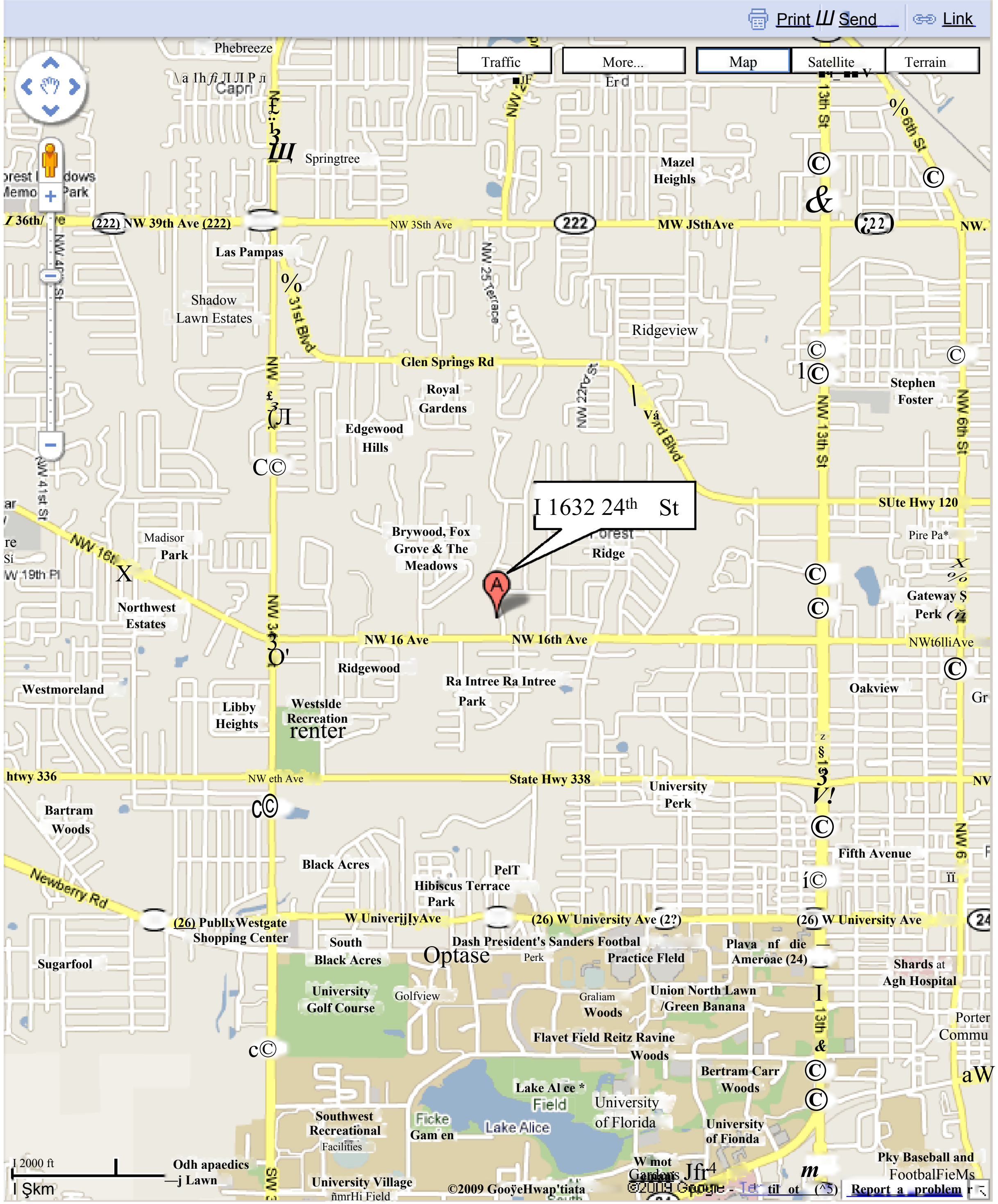
Before a comparative diagnostic trial is carried out, maximum sample sizes for the diseased group and the non-diseased group need to be obtained to achieve a nominal power to detect a meaningful difference in diagnostic accuracy. Sample size calculation depends on the variance of the statistic of interest, which is the difference between receiver operating characteristic (ROC) summary measures of two medical diagnostic tests. To obtain an appropriate value for the variance, one often has to assume an arbitrary parametric model and the associated parameter values for the two groups of subjects under two tests to be compared. It becomes more tedious to do so when the same subject undergoes two different tests, because the correlation is then involved in modeling the test outcomes. The calculated variance based on incorrectly specified parametric models may be smaller than the true one, which will subsequently result in smaller maximum sample sizes, leaving the study underpowered. In this paper we develop a nonparametric adaptive method for comparative diagnostic trials to update the sample sizes using interim data while allowing early stopping during interim analyses. We show that the proposed method maintains the nominal power and type I error rate through theoretical proofs and simulation studies.

Goodness of Fit Tests in Generalized Linear Mixed Models

Min Tang
University of Maryland

Generalized linear mixed models (GLMMs) are widely used in many fields. The random effects allow one to accommodate correlations due to repeated measure on the same individual, such as in longitudinal studies or correlations that are seen in family studies in genetic epidemiology. We propose omnibus chi-squared goodness of fit tests for GLMMs that jointly assess the appropriateness of the measured covariates and the latent random effect. The test statistic is a quadratic form in the observed and expected differences in cells defined by available covariates. The expected numbers are computed under the proposed model with MLEs in place of true parameters. We show that under mild conditions, this test statistic has an asymptotic chi-square distribution. We explicitly derive the test statistic for linear, logistic and count data and illustrate its performance on several data sets.

MAP to Alan Agresti's home



Directions: From Emerson Alumni Hall (O'Connell parking lot) travel West on University Avenue; turn right (North) on NW 22nd Street; turn left (West) on 16th Avenue; turn right (North) on NW 24th Street; 1632 is the first house on the left-hand side of the street.