Workshop Proposal

1. Title: Foundations and Frontiers in Statistics
(A Workshop in Celebrating Mary Thompson's Contributions to the Statistical Sciences)

2. Location: University of Waterloo

3. Dates: October 28-29, 2011

4. Program Committee: Richard Cook, Paul Marriott, David Matthews, Changbao Wu and Grace Yi

5. Confirmed Speakers:

Hugh Chipman	Acadia University, Canada
Geoffrey Fong	University of Waterloo, Canada
Mark Handcock	University of California at Los Angeles, United States
Xihong Lin	Harvard University, United States
Bruce Lindsay	Pennsylvania State University, United States
Sharon Lohr	Arizona State University, United States
Erica Moodie	McGill University, Canada
Nancy Reid	University of Toronto, Canada
Chris Skinner	University of Southampton, United Kingdom
Dylan Small	University of Pennsylvania, United States
Rob Tibshirani	Stanford University, United States
Jane-Ling Wang	University of California at Davis, United States

6. Scientific Themes

The purpose of this workshop is to bring together leading researchers to discuss recent advances in statistical sciences. Themes include survey sampling, biostatistics and statistical genetics, statistical methodology for social sciences, causal inference, likelihood-based inferences and statistical learning.

(1) Survey Sampling (Sharon Lohr and Chris Skinner)

Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata: Government statistical agencies often apply statistical disclosure limitation techniques to survey microdata to protect the confidentiality of respondents. There is a need for valid and practical ways to assess the protection provided. This paper develops some simple methods for disclosure limitation techniques which perturb the values of categorical identifying variables. The methods are applied in numerical experiments based upon census data from the United Kingdom which are subject to two perturbation techniques: data swapping (random and targeted) and the post randomization method. Some simplifying approximations to the measure of risk are found to work well in capturing the impacts of these techniques. These approximations provide simple extensions of existing risk assessment methods based upon Poisson log-linear models. A numerical experiment is also undertaken to assess the impact of multivariate misclassification with an increasing number of identifying variables. It is found that the misclassification dominates the usual monotone increasing relationship between this number and risk so that the risk eventually declines, implying less sensitivity of risk to choice of identifying variables. The methods developed in this paper may also be used to obtain more realistic assessments of risk which take account of the kinds of measurement and other nonsampling errors commonly arising in surveys.

(2) Statistical Methods for Social Sciences (Geoff Fong and Mark Handcock)

Respondent-driven sampling: an assessment of current methodology: Respondent-driven sampling (RDS) employs a variant of a link-tracing network sampling strategy to collect data from hard-to-reach populations. By tracing the links in the underlying social network, the process exploits the social structure to expand the sample and reduce its dependence on the initial (convenience) sample. The current estimators of population averages make strong assumptions in order to treat the data as a probability sample. We evaluate three critical sensitivities of the estimators: (i) to bias induced by the initial sample, (ii) to uncontrollable features of respondent behavior, and (iii) to the without-replacement structure of sampling. Our analysis indicates: (i) that the convenience sample of seeds can induce bias, and the number of sample waves typically used in RDS is likely insufficient for the type of nodal mixing required to obtain the reputed asymptotic unbiasedness; (ii) that preferential referral behavior by respondents leads to bias; (iii) that when a substantial fraction of the target population is sampled the current estimators can have substantial bias. This paper sounds

a cautionary note for the users of RDS. While current RDS methodology is powerful and clever, the favorable statistical properties claimed for the current estimates are shown to be heavily dependent on often unrealistic assumptions. We recommend ways to improve the methodology.

(3) Biostatistics and Statistical Genetics (Xihong Lin and Jane-Ling Wang)

Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies: GWAS have emerged as popular tools for identifying genetic variants that are associated with disease risk. Standard analysis of a case-control GWAS involves assessing the association between each individual genotyped SNP and disease risk. However, this approach suffers from limited reproducibility and difficulties in detecting multi-SNP and epistatic effects. As an alternative analytical strategy, we propose grouping SNPs together into SNP sets on the basis of proximity to genomic features such as genes or haplotype blocks, then testing the joint effect of each SNP set. Testing of each SNP set proceeds via the logistic kernel-machine-based test, which is based on a statistical framework that allows for flexible modeling of epistatic and nonlinear SNP effects. This flexibility and the ability to naturally adjust for covariate effects are important features of our test that make it appealing in comparison to individual SNP tests and existing multimarker tests. Using simulated data based on the International HapMap Project, we show that SNP-set testing can have improved power over standard individual-SNP analysis under a wide range of settings. In particular, we find that our approach has higher power than individual-SNP analysis when the median correlation between the disease-susceptibility variant and the genotyped SNPs is moderate to high. When the correlation is low, both individual-SNP analysis and the SNP-set analysis tend to have low power. We apply SNP-set analysis to analyze the Cancer Genetic Markers of Susceptibility (CGEMS) breast cancer GWAS discovery-phase data.

(4) Causal Inference (Erica Moodie and Dylan Small)

Causal inference in statistics and the quantitative sciences: Causal inference attempts to uncover the structure of the data and eliminate all non-causative explanations for an observed association. The goal of most, if not all, statistical inference is to uncover causal relationships. However it is not in general possible to conclude causality from a standard statistical inference procedure, it is merely possible to conclude that the observed association between two variables is not due to chance. Statistical inference procedures do not provide any information about which variable causes the other,

or whether the apparent relationship between the two variables is due to another, confounding variable. An explicit introduction of the philosophy of and approaches to causation was first brought into the statistical sciences in 1986 by Paul Holland, although references to causal approaches exist in the literature up to 60 years prior. Since then, there has been an explosion of research into the area in a variety of disciplines including statistics (particularly biostatistics), computer science, and economics. Causal inference in statistics is a broad area of research. We will provide an overview of research activities in this area, including (i) Inference and asymptotic theory; (ii) Balancing scores and inverse weighting: advances in biostatistics; (iii) Instrumental variables and structural equation models: connecting statistics and econometrics; (iv) Adaptive treatment regimes; and (v) Bayesian causal inference.

(5) Likelihood-based Inference (Bruce Lindsay and Nancy Reid)

Issues and strategies in the selection of composite likelihoods: The composite likelihood method has been proposed and systematically discussed by Besag (1974), Lindsay (1988), and Cox and Reid (2004). This method has received increasing interest in both theoretical and applied aspects. Compared to the traditional likelihood method, the composite likelihood method may be less statistically efficient, but it can be designed so as to be significantly faster to compute and it can be more robust to model misspecification. Although there are a number of ways to formulate a composite likelihood to balance the trade-off between the efficiency and computational price, there does not seem to exist a universal rule for constructing a combination of composite likelihoods that is both computationally convenient and statistically appealing. In this article we present some thoughts on the composite likelihood, drawing on basic knowledge about likelihood and estimating functions. A new efficiency result based on the Hoeffding decomposition of -statistics is given. A recommendation is given to consider the construction of surrogate density functions as a way to better bridge the gap between likelihood methods and composite likelihood methods.

(6) Statistical Learning (Hugh Chipman and Rob Tibshirani)

A Framework for Feature Selection in Clustering: We consider the problem of clustering observations using a potentially large set of features. One might expect that the true underlying clusters present in the data differ only with respect to a small fraction of the features, and will be missed if one clusters the observations using the full set of features.We propose a novel framework for sparse clustering, in which one clusters the observations using an adaptively chosen subset of the features. The

method uses a lasso-type penalty to select the features.We use this framework to develop simple methods for sparse K-means and sparse hierarchical clustering. A single criterion governs both the selection of the features and the resulting clusters. These approaches are demonstrated on simulated and genomic data

(7) Celebration of Mary Thompson's Contributions to Statistical Sciences

This workshop will also provide an opportunity to celebrate the outstanding achievements of Professor Mary Thompson who officially retired this past summer. Professor Thompson has made a deep and lasting impact in the Department of Statistics and Actuarial Science at the University of Waterloo and in the Canadian and international statistical science communities. Throughout her remarkably productive career, important advances have been made in a wide range of areas reflected by the themes of this workshop.

7. Detailed Budget:

Estimated Expenses (12 Speakers, 100 Registrants)	
Invited Speakers	16,200.00
(Travel, Accommodations, Meals: Details Attached)	
Coffee Breaks ($500 per break, 4 breaks)	2,000.00
Banquet ($50 per person, 100 attendees)	5,000.00
Opening Reception	1,500.00
Program and Associated Materials	800.00
	25,500.00

Estimated Income	
Faculty Registrants ($150 per person, 50 registrants)	7,500.00
Graduate Student Registrants ($80 per person, 50 registrants)	4,000.00
Fields Institute†	7,500.00
University of Waterloo Faculty of Mathematics	1,000.00
University of Waterloo Dept. of Statistics and Actuarial Science	5,500.00
	25,500.00

† Fields Institute funding to be used exclusively for travel expenses of invited speakers

Travel Budget by Speaker:

	Travel	Accommodation*	Meals**	Total
Hugh Chipman Wolfville, Nova Scotia	800	600	150	1,550
Geoffrey Fong Waterloo, Ontario	0	0	150	150
Mark Handcock Los Angeles, California	800	600	150	1,550
Xihong Lin Cambridge, Massachusetts	600	600	150	1,350
Bruce Lindsay University Park, Pennsylvania	600	600	150	1,350
Sharon Lohr Tempe, Arizona	800	600	150	1,550
Erica Moodie Montreal, Quebec	600	600	150	1,350
Nancy Reid Toronto, Ontario	200	600	150	950
Chris Skinner Southampton, U.K.	1,200	600	150	1,950
Dylan Small Philadelphia, Pennsylvania	600	600	150	1,350
Rob Tibshirani Stanford, California	800	600	150	1,550
Jane-Ling Wang David, California	800	600	150	1,550
(Total)	7,800	6,600	1,800	16,200

* Accommodations for invited speakers is set at $200/day for three days (October 27-29, 2011)

** Meals for invited speakers is budgeted as $50 per day for three days (October 27-29, 2011)