SMURF Workshop

Survey Methods and their Use in Related Fields

Neuchâtel, 20-22 August, 2018

SMURF
Survey Methods and their Use in Related Fields

Venue information

The conference will take place in the Faculty of Science, University of Neuchâtel:

Faculty of Science, University of Neuchâtel
11 Emile Argand Street, 2000 Neuchâtel
Auditoire Louis-Guillaume, level 2, wing F (opposite to the lake side)

For those coming by car, there are parking places available close to the Faculty. For those coming from the Hotel "Alpes et Lac", the meeting point is in the hall of the hotel (in front of the reception desk) 30 minutes before the first talk. The Faculty can be reached by foot from there.

Trail at the Areuse gorges

The planned social activity is a hike in the Areuse Gorges, which are close to Neuchâtel. We will go there by train, from the railway station of Neuchâtel. The train leaves Neuchâtel at 14:41 on platform 1 (Sector D). The travel is free for holders of the Neuchâtel tourist card (anyone staying at an hotel can freely request it).

The train will take us until the station "Champ-du-Moulin" where we will start walking towards Boudry. The path is easy, but good shoes are recommended for the hike.

If the weather is not good enough (the hike is not recommended if the trail is wet), we suggest to visit the Latenium, which is an archeological museum dedicated to the "La Tène Culture", a pre-celtic civilization. More informations are available on the conference website.

Social dinner at Le Silex

The social dinner will take place at the restaurant Le Silex, on Tuesday at 19:15 (7:15 PM). From the city center, take the bus 101 towards Marin and leave the bus at the stop "Port" (Hauterive port). The restaurant is located at the port. From the railway station of Neuchâtel, take the funicular in the railway station (Fun'ambule) and then take the bus 101. The buses are included in the Neuchâtel Tourist Card.

Schedule

	Speaker	Chair
Monday		
8:30 - 9:10	Registration	
9:15 - 9:30	Opening	Matthieu Wilhelm
9:30 - 10:20	Jean Opsomer	F. Jay Breidt
10:20 - 10:50	Coffee Break	
10:50 - 11:40	Simon Barthelmé	Matthieu Wilhelm
11:40 - 12:30	Vincent Loonis	Matthieu Wilhelm
12:30 - 14:00	Lunch break	
14:00 - 14:50	Lorenzo Fattorini	Yves Tillé
14:50 - 15:40	Guillaume Chauvet	Yves Tillé
15:40 - 16:10	Coffee Break	
Tuesday		
9:00 - 9:50	Yves Tillé	Yves Berger
9:50 - 10:40	Yann Busnel	Yves Berger
10:40 - 11:10	Coffee Break	
11:10 - 12:00	David Haziza	Guillaume Chauvet
12:00- 12:50	Yves Berger	Guillaume Chauvet
12:50 - 14:00	Lunch Break	
14:00 - 18:00	Social Event	
19:30 - 21:00	Social Dinner at Le Silex	
Wednesday		
9:15 - 10:05	Mathieu Gerber	Matthieu Wilhelm
10:05 - 10:35	Coffee Break	
10:35 - 11:25	Hendrik P. Lopuhaä	Jean Opsomer
11:25 - 12:15	Takumi Saegusa	Jean Opsomer
12:15 - 14:00	Lunch Break	
14:00 - 14:50	Patrice Bertail	David Haziza
14:50 - 15:40	F. Jay Breidt	David Haziza
15:40 - 16:00	Concluding remarks	Yves Tillé

Abstracts of the presentations

Data sub-sampling with Determinantal Point Processes
Simon Barthelmé, CNRS – Université Grenoble-Alpes
Monday 20 August, 10:50

Determinantal Point Processes (DPPs) are a class of point processes that exhibit "repulsion". This property can be leveraged to obtain high-diversity subsets, meaning that DPPs can be used to subsample various objects (surfaces, datasets, graphs, etc.) with relatively high fidelity. This idea has been suggested by several authors and holds tremendous theoretical appeal. However, many difficulties crop up in the implementation, and our goal has been to lift some of them. One aspect is that DPPs come in two variants: fixed sample size (so-called k-DPPs) and varying sample size. DPPs with varying sample sizes are more tractable, since their inclusion probabilities admit a closed form. k-DPPs make more sense in many applications, but are less tractable, since inclusion probabilities are much harder to compute. We show that k-DPPs and DPPs are asymptotically equivalent, which leads to tractable formulas for inclusion probabilities.
Joint work with Pierre Olivier Amblard and Nicolas Tremblay.

An empirical likelihood approach for modelling survey data
Yves Berger, University of Southampton
Tuesday 21 August, 12:00

Data are often selected with unequal probabilities from a clustered and stratified population. We propose a design-based empirical likelihood approach for regression parameters of generalised linear models. It differs from the mainstream (pseudo-)empirical likelihood approach, because it takes into account of the selection process. It provides asymptotically valid inference for the finite population parameters; that is, it gives consistent maximum empirical likelihood estimators and pivotal empirical likelihood ratio statistics. Hence, this approach can be used for point estimation, hypothesis testing and confidence intervals, without the need of variance estimates or linearisation. We will show how the approach can be extended for hierarchical models. We will also show how nonresponse can be taken into account. We will show that standard estimators such as Horvitz-Thompson, regression and calibration estimators are particular cases of the general approach proposed. We will use the 2006 PISA survey data as an illustrative example.

Survey sampling for big data
Patrice Bertail, Université Paris X - Nanterre
Wednesday 22 August, 14:00

Subsampling methods as well as general sampling methods appear as a natural tools to handle very large database (big data in the indivual dimension) when traditional statistical methods or statistical learning algorithms fail to be implemented on too large datasets. The choice of the weights of the survey sampling scheme may reduce the loss implied by the choice of a much more smaller sampling size (according to the problem of interest).

I will first recall some asymptotic results for general survey sampling based empirical processes, indexed by class of functions (see Bertail and Clémençon, 2016, Scandinavian Journal of Statistics), for Poisson type and conditional Poisson (rejective) survey samplings. These results may be extended to a large class of survey sampling plans via the notion of negative association of most survey sampling plans (Bertail, Rebecq, 2017).

However when one is interested in controlling generalization capability of statistical learning algorithms based on survey sampling techniques, asymptotic results are not sufficient. Hoeffding or Bennett type inequalities, classical deviation inequalities in the i.i.d. setting can be applied directly to Poisson or associated survey sampling plan leading however to suboptimal bounds . We show here how to overcome this difficulty for rejective sampling or conditional sampling plans. In particular, the Bennet/Bernstein type bounds established highlight the effect of the asymptotic variance of the (properly standardized) sample weighted sum and are shown to be much more accurate than those based on the negative association property.

On testing for informative selection in survey sampling
F. Jay Breidt, Colorado State University
Wednesday 22 August, 14:50

Consider sampling from a finite population that is modeled as a realization from a stochastic generating mechanism, or superpopulation. The goal is to make inference about a distributional property of the superpopulation based on data from the sample. Informative selection occurs if the conditional distribution of a response in the sample, given that it was selected for observation, is not the same as the distribution of a response in the finite population. Inference must be modified to account for this informative selection. Methods of testing for informative selection, both parametric and nonparametric, are reviewed and extensions are discussed. Methods are compared analytically and via simulation.

How to Generate Uniform Samples on Large-scale Data Streams
Yann Busnel, IMT-Atlantique, Rennes
Tuesday 21 August, 9:50

Within a network or distributed computer system, an ideal random sampling service should return a pointer to a node (which may be a computer, a process, a service, etc.), corresponding to an independent sample without bias to the group under consideration. This so-called uniform sampling service offers a simple primitive as a base brick for many applications in very large scale systems, such as information dissemination, metrology (via counting operations and statistical metrics), logic clock synchronization, etc. Unfortunately, the inevitable presence of malicious agents in these open systems hinders the construction of these sampling services.

Here we propose a solution to the problem of uniform sampling in large-scale computer systems in the presence of Byzantine behaviour. The latter reflect the non-compliance of the results of a system that does not meet its specifications. The Byzantine faults that are the most difficult to apprehend, come mainly from deliberate attacks aimed at making the system fail (sabotage, viruses, denial of service, etc.). We propose a first algorithm allowing to uniform on the fly a data (or items) stream of unbounded size, under the assumption that the exact probabilities of occurrence of the items are known. We model the behaviour of our algorithm using a Markov model and provide the results of the stationary and transient regime study. Our second algorithm releases the strong hypothesis of knowledge of item's probability of occurrence in the initial stream. These probabilities are then estimated on the fly using an aggregate data structure with a memory space proportional to the log of the size of the stream. We then evaluate the resilience of this algorithm to targeted and flooding attacks. In addition, we quantify the effort that the opponent must provide (i.e., the number of items to inject into the initial stream) to violate the uniformity property.

Properties of pivotal sampling : applications to spatial sampling and to sampling in dataflows
Guillaume Chauvet, ENSAI, Rennes
Monday 20 August, 14:50

A large number of sampling algorithms with unequal probabilities have been proposed in the literature, see for example Tillé (2011) for a review. The choice of a sampling algorithm is based on both statistical and practical considerations. On one hand, statistical properties are required, such as the weak consistency and the asymptotic normality of estimators. On the other hand, it may be necessary to introduce constraints in the selection of units. For example, when sampling in a dataflow, it is attractive to have a sequential sampling procedure under which the decision of selecting or not a unit is done as soon as it enters the dataflow.

In this work, we consider the use of the pivotal method (Deville and Tillé, 1998), also known as Srinivasan sampling design (Srinivasan, 2001), which has a number of interesting properties. It is based on a principle of duels between units, and leads to a variance reduction if the order of the units in the population is informative. This is a sequential method, so that it can be particularly helpful when sampling in a dataflow. Also, it enables to avoid selecting neighbouring units which makes it useful in the context of spatial sampling, when we wish to select spatially balanced samples.

In this talk, we will describe the principles of the method, and prove that it guarantees good statistical properties for a Horvitz-Thompson estimator (weak consistency, central limit-theorem, exponential

inequality) under mild assumptions. We will present two applications of the method. The first one is a modification of the Generalized Random Tesselation Sampling method, which is commonly used for spatial sampling. This is joint work with Ronan Le Gleut. The second work is related to sampling in a dataflow, with estimation on a sliding window. This is joint work with Emmanuelle Anceaume, Yann Busnel and Nicolo Rivetti.

A challenge for statisticians: the design-based spatial interpolation
Lorenzo Fattorini, Università degli Studi di Siena
Monday 20 August, 14:00

Accurate and updated wall-to wall maps depicting the spatial pattern of ecological and economic attributes throughout the study area represents a crucial information for evaluations, decision making and planning. Traditionally maps, as well as most of the issues of spatial statistics, are approached in a model-based framework (e.g. Cressie 1993). Recently we have attempted to construct maps in a complete design-based framework simply exploiting the inverse distance weighting interpolator and deriving the properties from the characteristics of the sampling scheme adopted. We first approached the problem of making maps for finite population of spatial units, when the survey variable is the amount of an attribute within units (Fattorini et al. 2018a). Subsequently, we considered the problem of making maps for continuous populations when the survey variable is, at least in principle, defined at each point of the continuum representing the study area (Fattorini et al. 2018b). Finally, we have faced the problem of constructing maps for finite populations of marked points. Our design-based approach to spatial mapping avoids the massive modelling involved in model-based approaches, i.e. the use of spatial models on lattices required for finite populations of spatial units (e.g. Cressie, Chapter 6), the use of second-order stationary spatial processes required for continuous populations (e.g. Cressie, Chapter 3) and the marked point processes in the plane required for finite populations of marked points (e.g. Cressie, Chapter 8). Design-based asymptotic unbiasedness and consistency of the resulting maps are achieved exploiting different asymptotic scenarios for the three cases, at the cost of supposing i) some forms of smoothness of the survey variables throughout the study area; ii) some sort of regularities that are necessary in the case of finite populations such as regularities in the shape of spatial units or regularities in the enlargements of the point populations; iii) asymptotically balanced spatial sampling schemes; iv) the use of distance functions sharing some mathematical properties. It is worth noting that iii) is satisfied by the more common sampling schemes adopted in spatial surveys and iv) does not constitute an assumption because it can be readily ensured by the user.

Cressie, N. (1993) *Statistics for Spatial Data*. New York: Wiley.
Fattorini, L., Marcheselli M, Pratelli L (2018a) Design-based maps for finite populations of spatial units. *Journal of the American Statistical Association*, to appear.
Fattorini, L., Marcheselli, M., Pisani, C., Pratelli, L. (2018b) Design-based maps for continuous spatial populations. *Biometrika*, to appear.

Survey Methods and their use in Monte Carlo algorithms
Mathieu Gerber, University of Bristol
Wednesday 22 August, 9:15

The problem of sampling in finite population arises in various Monte Carlo procedures and is notably at the heart of sequential Monte Carlo (SMC) algorithms. In this talk I will first explain the importance for this class of algorithms of what is called 'resampling' schemes in the Monte Carlo literature, that can be viewed as a certain type of unequal probability sampling methods. I will then outline the desirable properties that resampling algorithms should have in the context of SMC before presenting some recent results on that topic as well as some open questions. In the second part of the talk I will discuss some other uses of unequal probability sampling algorithms within Monte Carlo methods, with applications ranging from the statistical analysis of Big data sets to the construction of low discrepancy point sets.

Estimation in the presence of influential units in finite population sampling: an overview
David Haziza, Université de Montréal
Tuesday 21 August, 11:10

Influential units are those which make classical estimators very unstable. The problem of influential units is particularly important in business surveys, which collect economic variables, whose distribution are highly skewed (heavy right tail). There are two main inferential frameworks in survey sampling: the design-based framework and the model-based framework. To measure the influence of a unit under either framework, we use the concept of conditional bias of a unit and argue that it is an appropriate measure of influence measure. Using the conditional bias of a unit, we will show how to construct robust estimators/predictors of population totals. The robust estimators involves a psi-function, which depends on a tuning constant. The choice of the tuning constant will be discussed. Finally, the problem of internal and external consistency in the context of robust estimators will be presented.

Determinantal sampling design
Vincent Loonis, INSEE, Paris
Monday 20 August, 11:40

In this presentation, recent results about point processes are used in sampling theory. Precisely, we define and study a new class of sampling designs: determinantal sampling designs. The law of such designs is known, and there exists a simple selection algorithm. We compute exactly the variance of linear estimators constructed upon these designs by using the first and second order inclusion probabilities. Moreover, we obtain asymptotic and finite sample theorems. We construct explicitly fixed size determinantal sampling designs with given first order inclusion probabilities. We also address the search of optimal determinantal sampling designs.
Joint work with Xavier Mary, Université Paris X - Nanterre

Functional central limit theorems for single-stage sampling designs
Rik Lopuhäa, TU Delft
Wednesday 22 August, 10:35

In survey sampling one samples n individuals from a fixed population of size N according to some sampling design, and for each individual in the sample one observes the value of a particular quantity. On the basis of the observed sample, one is interested in estimating population features of this quantity, such as the population total or the population average. Well known estimators for population features that take into account the inclusion probabilities corresponding to the specific sampling design are the Horvitz-Thompson estimator and Hájèk's estimator. Distribution theory for these estimators is somewhat limited, partly due to the dependence that is inherent to several sampling designs and partly due to the more complex nature of particular population features. In this talk I will present a number of functional central limit theorems for different types of empirical processes obtained from suitably centering the Horvitz-Thompson and Hájèk empirical distribution functions. Basically, these results are obtained merely under conditions on higher order inclusion probabilities corresponding to the sampling design at hand. This makes the results generally applicable and allows more complex sampling designs that go beyond the classical simple random sampling or Poisson sampling. As an application I will use the results in combination with the functional delta method to establish the limit distribution of estimators for certain economic indicators, such as the poverty rate and the Gini index.
This is joint work with Hélène Boistard and Anne Ruiz-Gazen.

Revisiting design-based inference
Jean Opsomer, Colorado State University & Westat
Monday 20 August, 9:30

Design-based inference is being challenged due to declining response rates and rising costs, and the increasing availability of large non-probability samples. In the first part of this talk, I aim to present arguments for the continued relevance of this traditional survey paradigm, but also propose ways to make it more appropriate to today's data collection environment. In the second part, I present some recent results on nonresponse adjusted estimators using constraints. The results will illustrate how traditional approaches continue to be competitive with more sophisticated ones.

Large sample theory for merged data from multiple sources
Takumi Saegusa, University of Maryland
Wednesday 22 August, 11:25

We study infinite-dimensional M-estimation for merged data from multiple data sources in a super population framework. A setting we consider is characterized by (1) duplication of the same units in multiple samples, (2) unidentified duplication across samples, (3) dependence due to finite population sampling. Applications include data synthesis of clinical trials, epidemiological studies, disease registries and health surveys. Our estimator is adopted from a well-known technique from analysis of multiple-frame surveys in sampling theory. The main statistical issue is a theoretical gap between sampling theory and theory of infinite-dimensional M-estimation. Sampling theory handles dependence well but it is often the case that relatively simple estimators are treated in the finite-population framework where randomness from a probabilistic model is considered fixed. Theory of infinite-dimensional M-estimation, on the other hand, deals with complex statistical models but it relies on results from empirical process theory most of which assume i.i.d. data. To fill this gap, we extend empirical process theory to biased and dependent samples with duplication. Specifically we develop the uniform law of large numbers and uniform central limit theorem with applications to general theorems for consistency, rates of convergence and asymptotic normality in the context of data integration. Our results are illustrated with simulation studies and a real data example using the Cox proportional hazards model.

Fast Procedures for Selecting Unequal Probability Samples From a Stream
Yves Tillé, Université de Neuchâtel
Tuesday 21 August, 9:00

Probability sampling methods were developed in the framework of survey statistics. Recently sampling methods are the subject of a renewed interest for the reduction of the size of large data sets. A particular application is sampling from a data stream. The stream is supposed to be so important that it cannot be stored. When a new unit appears, the decision to conserve it or not must be taken directly without examining all the units that already appeared in the stream. We will examine the existing possible methods for sampling with unequal probabilities from a stream. Next, we propose a general result about subsampling from a balanced sample that enables us to propose several new solutions for sampling and subsampling from a stream. Several new applications of this general result will be presented.