

Abstracts for Olle Nerman symposium 2017-09-01 – 09-02

Title: The Distant Past may be Lost in Randomness

Speaker: Peter Jagers

The initial stages of growing populations are often characterised by a branching type pattern. Individuals (particles, molecules, cells, organisms, heavenly bodies or galaxies) generate new individuals and give rise to exponential growth changing or even eliminating the wild, random dynamics of the system. When we look back into history, the start can typically not be recovered from observations of the system at large size, but is hidden by a veil of randomness.

This problem arose in the very concrete context of quantitative PCR, where the molecule number can only be observed after many amplification rounds, and it is important to find out whether the process started from one or a couple of DNA segments.

Title: Regenerative multi-type Galton-Watson processes

Speaker: Serik Sagitov

The general Perron-Frobenius theorem describes the growth of powers of irreducible non-negative kernels. In the special case of kernels with an atom this result can be obtained using a regeneration method. If such a kernel is sub-stochastic, then the regeneration method can be intuitively explained in terms of the so-called split-chain.

In this talk we present a probabilistic interpretation of the regeneration method in terms of what we call regenerative Galton-Watson processes. These multi-type branching processes have an intrinsic structure of a single-type Crump-Mode-Jagers process with time inhomogeneous immigration.

Title: Fixed points of the smoothing transform

Speaker: John Biggins

In this talk I will look at some of the work on the fixed points of the smoothing transform, paying particular attention to the critical role played by Olle Nerman's work on branching processes.

Title: Modeling and estimation of overdiagnosis in prostate cancer screening

Speaker: Marianne Månsson

The risk of overdiagnosis is an issue regarding PSA-based screening for prostate cancer (PSA= prostate-specific antigen), and the major reason that prevents population-based screening programs from being introduced. The mechanisms of overdiagnosis and different approaches to estimate the amount of overdiagnosis and related concepts such as lead and sojourn times will be discussed.

Title: Scan-o-matic: High-Resolution Microbial Phenomics at a Massive Scale

Speaker: Anders Blomberg

The capacity to map traits over large cohorts of individuals—phenomics—lags far behind the explosive development in genomics. For microbes, the estimation of growth is the key phenotype because of its link to fitness. We introduce an automated microbial phenomics framework that delivers accurate, precise, and highly resolved growth phenotypes at an unprecedented scale. Advancements were achieved through the introduction of transmissive scanning hardware and software technology, frequent acquisition of exact colony population size measurements, extraction of population growth rates from growth curves, and removal of spatial bias by reference-surface normalization. Our prototype arrangement automatically records and analyzes close to 100,000 growth curves in parallel. We demonstrate the power of the approach by extending and nuancing the known salt-defense biology in baker's yeast. The introduced framework represents a major advance in microbial phenomics by providing high-quality data for extensive cohorts of individuals and generating well-populated and standardized phenomics databases.

Title: Joint modeling in Chronic Obstructive Pulmonary Disease: Understanding the association between clinical endpoints

Speaker: Alexandra Jauhainen

Chronic obstructive pulmonary disease (COPD) is a progressive obstructive lung disease characterized by persistent airflow limitation. Patients with COPD suffer from a variety of symptoms as well as severe flare-ups of the disease called exacerbations.

The prevention or modification of exacerbations is considered a key efficacy measure in clinical COPD trials, but trials and drug programs are complicated by a relatively low exacerbation event rate, significant patient heterogeneity, inadequate statistical methods and a currently poor understanding of the longitudinal association between exacerbation risk and other COPD endpoints.

Joint modelling is a powerful methodology where separate submodels are joined into a single one to infer dependence and association between the different endpoints modelled. We have used the joint modelling framework to analyze associations between clinical endpoints like cough and sputum production, lung function and occurrence of exacerbations and their connection to treatment response.

I will show some recent results indicating the value of joint modelling techniques to disentangle the association between COPD endpoints, and outline how this work will be continued in order to help design more efficient clinical trials.

Title: Survival analysis around a cross-section and unobserved heterogeneity

Speaker: Niels Keiding

Consider lifetimes originating at a series of calendar times t_1, t_2, \dots . At a certain time t_0 a cross-sectional sample is taken, generating a sample of *current durations* (backward recurrence times) of survivors until t_0 and a *prevalent cohort study* consisting of survival times left-truncated at the current durations. A Lexis diagram is helpful in visualizing this situation.

Survival analysis based on current durations and prevalent cohort studies is now well-established as long as all covariates are observed.

The general problems with *unobserved covariates* have been well understood for ordinary prospective follow-up studies, with the good help of hazard rate models incorporating frailties: as for ordinary regression models, the added noise generates attenuation in the regression parameter estimates.

For current durations and prevalent cohort studies this attenuation remains, but in addition one needs to take account of the differential selection of the survivors from initiation t_i to cross-sectional sampling at t_0 .

This talk intends to survey the recent development of these matters and the consequences for routine use of hazard rate models or accelerated failure time models in the many cases where unobserved heterogeneity may be an issue.

The work was motivated by our work on designs for estimating *time-to-pregnancy* (TTP) and will be illustrated by a simulation study in that framework.

Title: Statistical challenges when analysing emerging epidemic outbreaks

Speaker: Tom Britton

New infectious disease outbreaks have great impact on communities over the world, as most recently manifested by the Ebola outbreak. An important statistical task then is to predict the future scenario with and with out preventive measures. In the current talk we will investigate such analyses and see how it can be improved. The main catch is that in the exponentially growing phase early on in an outbreak, several biases can occur if not taken into account: events with short delays will be over-represented. We will give some examples from the Ebola outbreak and see how the biases can be removed or at least reduced. (Joint work with Gianpaolo Scalia Tomba)

Title: Integration of polygenic risk estimates into personalized risk prediction algorithms – some statistical challenges

Speaker: Krista Fischer

We will discuss the process of development and validation of algorithms for personalized prediction of complex disease risk. One aims to capture the genetic component of the risk in a Genetic (polygenic) Risk Score (GRS). Usually the GRS is defined as a linear combination of effect allele counts of several Single Nucleotide Polymorphisms (SNPs), whereas the SNPs and their corresponding weights are based on results of a large-scale meta-analysis of Genome-Wide Association Study (GWAS). As the heritable component of most common complex diseases (such as Type 2 Diabetes and Coronary Artery Disease) is highly polygenic, the

efficient GRS should account for the effects of thousands of SNPs. The efficiency of the genetic predictor will further depend on selection of the optimal weights for the SNPs (Läll et al 2016). One should also be aware of the need for non-overlapping cohorts in the process of SNP discovery (GWAS meta-analysis cohorts), GRS development and validation.

Next we demonstrate how the effects of phenotypic risk factors and GRS are combined to an overall risk score and illustrate how the absolute risks can be calculated in the case of Type 2 Diabetes risk estimation. We discuss statistical challenges that are related to specific features of the population based biobank data (left-truncation for some outcomes, mix of retrospective and prospective data for some others, etc.).

Title: Revisiting the polygenic additive liability model for multifactorial diseases

Speaker: Françoise Clerget-Darpoux

During the last decade, genome wide association studies have been very successful in detecting associations between multifactorial diseases and numerous Single Nucleotide Polymorphisms (SNPs). These associated SNPs are often used for estimating missing heritability and individual risk scores under the assumption of a polygenic additive liability. We will highlight the pitfalls of relying on such a model when there are reasons to suspect etiological heterogeneity and/or departure from the hypotheses on the environmental factor effects.

The main purpose of the genetic study of multifactorial diseases is to understand the etiology of the disease, with the hope that this understanding will contribute to better disease control and prevention. Such an objective is shunted around when assuming that all complex and heterogeneous pathological processes can be summarized by a single simplistic model.

We believe that time has come to shift from the polygenic additive model paradigm and to give room to more biologically-driven models of diseases supported by the physiopathology and genetic knowledge specific to diseases.

Title: Modeling rater agreement via exchangeability, with clinical applications of the kappa statistic.

Speaker: Mauro Gasparini

A very natural model of rater agreement is based on infinite sequences of exchangeable ratings and independent subjects. The result is a simple hierarchical exchangeable model (not necessarily Bayesian) which has not been sufficiently explored in the rater agreement literature. For binary ratings, the population kappa parameter can then be defined properly as the (Pearson) correlation coefficient between the ratings of the same subject by any two raters. For qualitative ratings into more than two classes, kappa itself can be taken as a measure of qualitative correlation between ratings. The sampling counterparts are the kappa statistics as usually defined. First and second order asymptotic distributions can be obtained and used to solve inferential problems. An application to the clinical evaluation of endoscopic findings will be presented.

Title: **Simulating the component counts of combinatorial structures**
Speaker: **Simon Tavaré**

We describe methods for simulating the component counts of random logarithmic combinatorial structures such as permutations and mappings. We exploit the Feller coupling for simulating permutations to provide a very fast method for simulating logarithmic assemblies more generally. For logarithmic multisets and selections, this approach is replaced by an acceptance/rejection method based on a particular conditioning relationship that represents the distribution of the combinatorial structure as that of independent random variables conditioned on a weighted sum. We show how to improve its acceptance rate. We illustrate by estimating the probability that a random mapping has no repeated component sizes, and establish the asymptotic distribution of the difference between the number of components and the number of distinct component sizes for a very general class of logarithmic structures. This is joint work with Richard Arratia, Andrew Barbour and Warren Ewens.