BREV
Ukonvoluteret


Returneres ved varig adresseændring

Næste nummer af "MEDDELELSER" udkommer 6. april 2009.
Bidrag skal være redaktøren i hænde senest **den 27. marts kl. 12**

# MEDDELELSER

Dansk Selskab for Teoretisk Statistik

## Todages møde i DSTS 14.-15. april 2009

### CBS - Handelshøjskolen i København

Tilmelding pr. email senest torsdag d. 2 . april til Marianne Andersen, ma.mes@cbs.dk. **Bemærk den tidlige tilmeldingsfrist**, som skyldes mødets placering umiddelbart efter påske. Program, pris og betaling – se side 3-5.

**Stillingsopslag**: SAS institute side 10, Novozymes, side 11

## Kalender 2009 – se også inde i bladet

| Dato | No. | Aktivitet |
|---|---|---|
| 4/3 | 1/09 | Institut for Matematisk Fag, Københavns Universitet: Henrik Hult (Royal Institute of Technology): *A note on large deviations and importance sampling for heavy-tailed stochastic processes* |
| 5/3 | 2/09 | Thiele Centre, University of Aarhus. Rüdiger Frey (Universität Leipzig): *Option Pricing in Illiquid Markets and Nonlinear Black-Scholes Equations* |
| 16/3 | 2/09 | Biostatistisk Afdeling, Københavns Universitet. Michael W. Kattan (Cleveland Clinic): *Encouraging the Use of Statistical Prediction Models* |
| 20/3 | 2/09 | Institut for Matematisk Fag, Københavns Universitet: Sergey Foss (Heriot-Watt University): *On the asymptotics for the longest path in a class of directed graphs* |
| 26/3 | 2/09 | Thiele Centre, University of Aarhus. Sergey Foss (Heriot-Watt, Edinburgh): *Random walk in a Poisson rain* |
| 1-3/4 | 1/09 | Københavns Universitet. *100 Years of Queueing - The Erlang Centennial* |
| 3/4 | 2/09 | Institut for Matematisk Fag, Københavns Universitet: Asger Hobolth (University of Aarhus): *The coalescent process: Theory and applications* |
| 3/4 | 2/09 | Department of Biostatistics, University of Copenhagen. Thomas Lumley (University of Washington, Seattle): *Full-day course Complex sampling designs in epidemiology* |
| 14-15 /4 | 2/09 | CBS – Handelshøjskolen i København *Todages møde DSTS* |
| 10-12/6 | 1/09 | Aarhus University. Sixth International Workshop on Computational Systems Biology WCSB 2009 |

**Meddelelser** er medlemsblad for Dansk Selskab for Teoretisk Statistik (DSTS), se http://www.dsts.dk.

Selskabets formål er at fremme den statistiske videnskab og dens anvendelser.

Indmeldelse og adresseændring i DSTS gøres via http://www.dsts.dk/da/index.html.

Selskabet har en elektronisk nyhedsliste E-Meddelelser, se http://www.dsts.dk/da/index.htm.

Bidrag og stillingsopslag til Meddelelser sendes til redaktøren - red@dsts.dk. Bidrag i elektronisk form modtages helst i PDF format med indlejrede fonte i sidestørrelse A4, egnet til sort/hvid tryk i A5 format. Alternativt modtages Word, HTML eller ASCII.

Annoncering af stillingsopslag i Meddelelser koster kr. 1.000,- pr. side. Opgiv venligst faktureringsoplysninger. Indstik til udsendelse i konvolut sammen med Meddelelser koster kr. 3.000,- pr. standard A4 side for første side og 500,- kr per efterfølgende side.

Meddelelser udkommer 9 gange om året, den første mandag eller tirsdag i måneden undtagen januar, juli og august måned.

Udgivelsesplan for Meddelelser 2009

| Nr. | Bidrag senest | Udkommer |
|---|---|---|
| ~~1~~ | ~~23. januar~~ | ~~2. februar~~ |
| ~~2~~ | ~~20. februar~~ | ~~2. marts~~ |
| 3 | 27. marts | 6. april |
| 4 | 24. april | 4. maj |
| 5 | 22. maj | 1. juni |
| 6 | 21. august | 1. september |
| 7 | 25. september | 5. oktober |
| 8 | 23. oktober | 2. november |
| 9 | 20. november | 1. december |

# Todages møde i DSTS 14.–15. april 2009

**CBS — Handelshøjskolen i København**
Solbjerg Plads 3, Nordea Auditoriet (SP205)

Tilmelding pr. email

**SENEST TORSDAG D. 2. APRIL**

til Marianne Andersen, `ma.mes@cbs.dk`. Bemærk den tidlige tilmeldingsfrist, som skyldes mødets placering umiddelbart efter påske.

Pris: 500 kr. (250 for studerende, dog ikke ph.d. studerende). Betaling til DSTS, Jyske Bank, reg. nr. 7853, konto 1117188

Handelshøjskolens bygning på Solbjerg Plads ligger lige bag ved Frederiksbergcentret på Falkoner Alle, og over for den vestlige trappe fra Frederiksberg Metrostation. Man kommer lettest til SP205 ved at gå op til 2. sal ad balkontrappen ca. midt i vandrehallen.

Hvis der sker programændringer vil det blive meddelt på

`www.mes.cbs.dk/~sttt/2dages/`

Eventuelle spørgsmål til `tuetjur@cbs.dk`.

**Tirsdag:**

14.00-14.05 Velkommen.

14.05-14.55 **Søren Johansen**, Institut for Matematiske Fag, KU, and **Bent Nielsen**, Nuffield College, Oxford: *An analysis of the indicator saturation estimator as a robust regression estimator.*

An algorithm suggested by Hendry (1999) for estimation in a regression with more regressors than observations, is analyzed with the purpose of finding an estimator that is robust to outliers and structural breaks. This estimator is an example of a one–step M–estimator based on Huber's skip function. The asymptotic theory is derived in the situation where there are no outliers or structural breaks using empirical process techniques. Stationary processes, trend stationary autoregressions and unit root processes are considered.

14.55-15.20 Kaffe/te.

15.20-16.10 **Matthias Buch–Kromann**, Computational Linguistics, ISV, CBS: *What is the probability of a word sequence? N–gram smoothing in statistical machine translation.*

Statistics plays a crucial role in state–of–the–art machine translation systems such as translate.google.com. The systems are based on two submodels, a translation model that models how short word sequences are translated from one language to another, and an n–gram model that assigns probabilities to word sequences in the target language. The n–gram model presents particular statistical challenges because of the data sparseness and the categorical nature of the data. In the talk,

I will describe the Kneser Ney algorithm, which is the current state–of–the–art algorithm for n–gram modelling, and propose a more general and principled alternative, called Hierarchy–based Partition Models (HPMs). HPMs can be loosely viewed as semi–parametric correction estimators for categorical data, based on variable–bandwidth kernel smoothing where the topology of the sample space is given by a polyhierarchy rather than some distance metric.

16.10-17.00 **Ib Skovgaard**, Department of Basic Sciences and Environment, Faculty of Life Sciences, KU: *Count models for amplified samples.*

Joint work with Anders Tolver Jensen (same place) and the private biotechnological company Vipergen.

Models are introduced for counts of sequences from amplified material, reflecting the major steps of the process: chemical selection, amplification by PCR (Polymerase Chain Reaction) and sequence sampling. This results in a complex of count models, the basic element of which is a scaled Poisson mixture of zero–truncated Poisson distributions. A key ingredient is the so–called Stirling distribution of the second kind. Even with few parameters, the multivariate distribution of several such counts provides a flexible class for modelling overdispersion compared to the multinomial distribution, given the total count. In a semi parametric model with unknown proportions of initial material between the replicates, estimation is delicate because maximum likelihood breaks down due to singularity, while semiparametric estimating equations lead to good estimates. In general, estimation is a challenge, but the models are in many ways surprisingly tractable. The models have been developed for counts of molecules from a combinatorially assembled molecular library and some results will be shown.

17.00-18.00 Pause — øl og vand kan købes.

18.00-22.00 Middag i CBS's kantine.

**Onsdag:**

9.30-10.20 **Tue Tjur**, Center for Statistik, CBS: *R–square statistics for logistic regression models.*

Many different analogues to the coefficient of determination in ordinary regression models have been proposed in the context of logistic regression. Our starting point is a study of three definitions related to quadratic measures of variation. We propose a new quantity, "the coefficient of discrimination", and recommend its use as a standard measure of explanatory power. In its interpretation, this quantity has no immediate relation to the classical versions of R–square, but it turns out to be related to these by two exact relations, which imply that all these statistics are asymptotically equivalent.

10.20-10.50 Kaffe/te.

10.50-11.40 **Klaus Kaae Andersen**, Informatics, Section for Statistics, DTU: *Real–time quality control by exponential smoothing.*

The rapid development of information technology has increased the number of applications where real–time quality control is both feasible and useful. In this talk we present an exponential weighted moving average control chart well suited for real–time surveillance of adverse events. The procedure is based on inter–events counts for failures recorded when the failures occur. The method is easy to implement and allows for real–time monitoring when integrated with a web–based interface. The method and its implementation is presented, and two examples of applications are given. The first example concerns routine surveillance of medical procedures at a surgical unit, while the second example concerns surveillance of adverse events in on–line games.

11.40-12.30 **Gorm Gabrielsen**, Center for Statistik, CBS: *Paired comparisons and designed experiments.*

Within the social sciences and especially within the theory of consumer behavior the concepts of preferences and choice behavior are common. Preferences may be assigned to individuals, different subgroups or to a population. The method of paired comparisons with continuous responses seems to be a very suitable method to measure and analyze preferences. However, beyond the paper of Scheffe (1952), very little work seems to have been done in this area. In the talk I will present the method of paired comparisons and discuss advantages and disadvantages of the method. The model can be considered as a general linear model and therefore most estimation and test problems are easily solved. Furthermore, I will show that it is possible to establish an orthogonal decomposition of subspaces and thereby to generalize the "usual" concepts of e.g. orthogonal two–way–layout, block design, confounding etc.

12.30-13.00 Sandwich og vand — kan indtages før afgang eller medtages.

DEPARTMENT OF MATHEMATICAL SCIENCES
UNIVERSITY OF COPENHAGEN

## SEMINAR IN APPLIED MATHEMATICS AND STATISTICS

FEBRUARY 20, 2009

UNIVERSITETSPARKEN 5
DK-2100 COPENHAGEN Ø
DENMARK

http://www.math.ku.dk/

**Wednesday, March 4, 2009, 15:15, aud. 10, H.C. Ørsted Institute**
**Speaker: Henrik Hult, Royal Institute of Technology**
**Title: A note on large deviations and importance sampling for heavy-tailed stochastic processes**
Abstract: In the light-tailed case it is well known that results on large deviations provide insights into the design of efficient rare event simulation algorithms. I will illustrate some aspects of this connection for heavy-tailed processes and present some current work on the design of efficient importance sampling algorithms in the heavy-tailed setting.

**Friday, March 20, 2009, 14:15, aud. 10, H.C. Ørsted Institute**
**Speaker: Sergey Foss, Heriot-Watt University**
**Title: On the asymptotics for the longest path in a class of directed graphs**
Abstract: We start with a finite set $\{1,2,\ldots,n\}$. Two elements $i<j$ are connected by a direct link (from $i$ to $j$) with a probability $p(i,j)$ which may depend on their distance only. A path is a collection of consecutive links and its length is the number of links. Let $L(n)$ be the maximal path length. We study its asymptotic properties, as $n$ goes to infinity. Then we discuss various multidimensional analogies of the model. This work may be considered as a continuation and generalisation of results from F&Konstantopoulos (2003).

**Friday, April 3, 2009, 14:15, aud. 10, H.C. Ørsted Institute**
**Speaker: Asger Hobolth, University of Aarhus**
**Title: The coalescent process: Theory and applications**
Abstract: The coalescent process is becoming a standard tool for understanding genetic variation within a species. During this talk, I plan to describe some recent and ongoing projects where the coalescent process is applied. The projects include: 1) A discussion of some special cases where exact calculations of basic properties can be derived, 2) Understanding human evolutionary history from whole-genome analysis of human-chimpanzee-orangutan-macaque alignments, and 3) Allowing for varying population size.

THIELE CENTRE
FOR APPLIED MATHEMATICS IN NATURAL SCIENCE

Department of Mathematical Sciences
University of Aarhus

## Activities at the Thiele Centre

**Seminar:**

**Thursday, 5 March, 2009, at 13:15–14:00, in building 1540, room K26**

Rüdiger Frey (Universität Leipzig)
Option Pricing in Illiquid Markets and Nonlinear Black-Scholes Equations

Abstract/Description :
We study properties of option prices in illiquid markets. As a starting point, we introduce various models for illiquid markets such as the model of Cetin, Jarrow and Protter (2004) or the model of Frey (2000) and show that the problem of hedging European options leads to closely related fully nonlinear versions of the standard Black-Scholes PDE.
The main part is devoted to a study of analytical properties of this PDE. In particular, we show that under natural conditions the option prices provided by the model (the solutions of the nonlinear PDE) have the properties of a convex (but non-coherent) risk measure. Moreover, we study asymptotic properties of solutions as the market depth tends to zero. The necessary technical tools are comparison and stability theorems for viscosity solutions.

**Thursday, 26 March, 2009, at 13:15–14:15, in Koll. G3 (1532.218)**

Sergey Foss (Heriot-Watt, Edinburgh)
Random walk in a Poisson rain

Abstract/Description :
Consider a person walking in one direction along the line and stopping to clean it from items falling in a Poisson rain (or "serve" or "eat" "manna"...)
We study the asymptotics of the person's position when time grows. Assume more precisely that the person either walks with a constant speed or stops for cleaning, that $n$–th cleaning time is $s_n$ and that $(\sum_1^n s_k)/n^\alpha \to c \in (0,\infty)$. What could we say about the person's location $X(t)$ at time $t$? Here the power $\alpha$ is any positive number. There are 5 zones for $\alpha$, within each zone there is a specific type of asymptotic behaviour of $X(t)$.

**Biostatistisk Afdeling**
**Københavns Universitet**

16. februar 2009
J.nr. 241-0002/07

## Seminar i anvendt statistik

Seminaret afholdes på det gamle Kommunehospital, Øster Farimagsgade 5. Der serveres te i Biostatistisk Afdelings bibliotek (opgang B, 2. sal) en halv time før.

**Mandag d. 16. marts 2009, kl. 14.00, lokale 2.1.02.** **BEMÆRK tidspunkt!**

**Michael W. Kattan**

**Quantitative Health Sciences, Cleveland Clinic**

**Encouraging the Use of Statistical Prediction Models**

Everyone would be better off if physicians were able to use statistical prediction models more often. Several barriers are involved and will be discussed. These include inadequate data collection, lack of statistician involvement, programming expense, physician reluctance, and patient unawareness. Many of these barriers are modifiable, and various solutions will be presented.

Thomas Gerds

# Full-day course: Complex sampling designs in epidemiology

### *Teacher*

Thomas Lumley, Department of Biostatistics, University of Washington, Seattle

### *Time*

April 3, 2009, 9:15-16:00

### *Location*

Center for Health and Society (Kommunehospitalet), Øster Farimagsgade 5

### *Course abstract*

Complex sampling designs with unequal sampling probabilities and non-independence of sampling occur in epidemiology when analyzing national survey data and when constructing two-phase subsamples from existing cohort or case-control samples. The same techniques used to analyze complex samples are also used in causal inference, treating observed data as if sampled from a population of potential outcomes. I will describe how to use sampling weights for estimation and inference, how to use known population information to improve efficiency, and how to do the computations in R using the 'survey' package. I will also discuss how much efficiency is lost by using sampling weights rather than maximum likelihood in two-phase designs.

1. Multistage Survey samples
   - How to incorporate strata, weights, clustering in estimation
   - Describing survey designs to R
   - Large data sets
   - Graphics and exploratory analysis
   - Regression models
   - Calibration and the 'estimated weights' paradox
2. Sampling from a cohort: Two-phase samples
   - Motivation
   - Inverse-probability weights
   - Using the whole cohort: calibration on influence functions
3. Efficiency of weighted-likelihood methods
   - Design-based vs model-based analysis of case-control design
   - Case-only design for interaction: efficiency and robustness
   - More general incomplete data designs: efficiency and robustness

### *Audience*

Primarily statisticians and epidemiologists with some prior programming experience. The R language and environment will be used throughout the course.

### *Price*

Free, but the number of seats is limited to 45. Preference is given to members of the Statistics Network at University of Copenhagen. Applications should be sent to the course director by email before March 25.

### *Course director*

Associate professor Peter Dalgaard, Department of Biostatistics, Institute of Public Health;
p.dalgaard@biostat.ku.dk phone 35 32 79 18.