B-Stat News

The Newsletter
of the

Belgian Statistical Society
Belgische Vereniging voor Statistiek
Société Belge de Statistique

Number 35 – September 2005

Editeurs responsables/Verantwoordelijke uitgevers :
Mia Hubert, Departement Wiskunde, Katholieke Universiteit Leuven,
W. de Croylaan 54, 3001 Leuven
Philippe Lambert, Institut de Statistique, Université catholique de Louvain,
voie du Roman Pays 20, 1348 Louvain-la-Neuve

THE COUNCIL OF THE SOCIETY

President

Adelin Albert, Université de Liège

Vice-President

Paul Janssen, Universiteit Hasselt

Administrators

Marc Aerts, Universiteit Hasselt (treasurer).
Luc Bijnens, Janssen Pharmaceutica
Frans Desmedt, Institut National de Statistique
Uwe Einmahl, Vrije Universiteit Brussel
Irène Gijbels, Katholieke Universiteit Leuven
Marc Hallin, Université Libre de Bruxelles
Mia Hubert, Katholieke Universiteit Leuven
Philippe Lambert, Université catholique de Louvain
Jean-Paul Rasson, Facultés Univ. Notre-Dame de la Paix, Namur
Didier Renard, Eli Lilly
Christian Ritter, Monnet Centre Intl. Lab. and Université catholique de Louvain

Secretary of the Society

Gentiane Haesbroeck, Université de Liège.

Website of the Society

www.sbs-bvs.be

Webmaster

Laurence Seidel : laurence.seidel@ulg.ac.be

TABLE OF CONTENTS

EDITORIAL

Dear member of the Belgian Statistical Society,

We have been serving the SBS-BVS as co-editors of B-Stat News since January 2002. After twelve issues of the Bulletin, we both think that fresh blood should be brought into this project that embodies the memory of our Society.

Working with two editors, one from each linguistic regimen, was thought by the Board to be a way to share the workload implied by this job while ensuring a representation for the two main Belgian Communities. Our 4-year experience confirms that this working process has given full satisfaction and our advice would be to continue that way.

Obtaining contributions from our members is a difficult task and sometimes a last minute project for the contributors. Most entries come from universities. Therefore, our Bulletin only partially reflects the large number of activities in statistics in Belgium: further encouraging and getting contributions from non-academic members is a major challenge for the future. Having such a member as (co-)Editor might be part of the solution.

We would like to take this opportunity to thank all the persons who helped us during our editorial work, in particular the contributors, the current and past Presidents of the Society (Prof. L. Simar, N. Veraverbeke and A. Albert) and, last but not least, the ULg team (Adelin Albert, Gentiane Haesbroeck, Anna Marchetta and Laurence Seidel) for its logistic support.

Let us conclude this editorial by the same sentence as in January 2002: B-Stat News *is your newsletter: it will be what you make of it.*

Mia Hubert and Philippe Lambert
Editors of B-Stat News

CALL FOR NEW EDITORS

If you are willing to serve as Editor for B-Stat News, please send a motivating letter and a short CV to the SBS-BVS Secretary

Prof. Gentiane Haesbroeck
Université de Liège
Institut de Mathématique
Grande Traverse, 12 (B37)
B-4000 LIEGE BELGIUM
Tel : 04/366.95.94 - Email : g.haesbroeck@ulg.ac.be

THE DEADLINE FOR APPLICATION IS

DECEMBER 2, 2005

The BIOSTATISTICS Section of the
Belgian Statistical Society (SBS-BVS)
organizes a **training** (16 November)
followed by a **workshop** (17 November)

TRAINING:

PRACTICAL BAYESIAN MODELLING WITH BUGS / BRUGS CODA / WBDIFF IN PHARMACEUTICAL R&D

16 November 2005, EORTC, Brussels, Belgium

COURSE OUTLINE

The Bayesian methodology has shown to be extremely flexible for dealing with relatively complex models. Thanks to the Markov Chain Monte-Carlo (MCMC) approaches, the Bayesian paradigm frees up the researchers to use unreasonable and simplified models due to technical burdens and allows them to focus only on the development of appropriate models with the main goal of summarizing, of predicting, of understanding the mechanisms generating the observed data.

Over the last decade, multiple researchers have worked on different pieces of software in order to give access to the MCMC technologies. The most popular package for practical applications is definitively the BUGS (Bayesian inference Using Gibbs Sampling) software, now available under an open source license. The family of BUGS software provides to practitioners a nice general framework for running complex Bayesian evaluations. The package is composed by a language for model specification and a robust sampler for simulating from the posterior distributions. The core Bugs package has been gradually reinforced by multiple satellite packages like BRugs, Coda, WBDiff. BRugs provides a direct connection between R and OpenBugs that greatly facilitates the pre and post evaluation manipulations of the data indispensable for relatively heavy real applications. Coda is a bunch of R subroutines with the main purposes of diagnosing, through different graphs and summaries, potential problems on

the convergence of the simulated chains. Finally, WBDiff, as an add-in to Bugs, allows the practitioners to include differential systems in the model description, even if those systems do not admit any explicit solutions.
The goal of the one-day training is essentially to drive the future potential users through the key concepts associated to the Bugs package and its different sister packages previously listed. The training is centered on four hands-on sessions of one-hour-and-a-half each. Each session handles a specific application selected from the pharmaceutical R&D world.

INSTRUCTOR: Dr BENOIT BECK

Dr Benoît Beck is working as a senior research scientist for Eli Lilly and Company (Mont-Saint-Guibert, Belgium) in the European Early Phase Statistic department. Before joining Eli Lilly and C., he completed his PhD in Sciences (Orientation: statistics) at Louvain-la-Neuve University, during which he also spent some times at the Imperial college as well as at the University of Ottawa. Dr Beck's principle area of expertise is pre-clinical statistics, working on In Silico, In Vitro, In Situ and In Vivo data evaluations. He is working with cross-functional multidisciplinary teams on the development of new compounds looking for tomorrow medicines. Dr. Beck is playing a leadership role on the use of the Bayesian methodology within Eli Lilly and C. with the intent to impact the Lilly decision-making processes.

WHO SHOULD ATTEND ?

Some familiarity with Bayesian ideas and with parametric distribution theory will be assumed. A quick review of basic Bayesian concepts and of decision theory will be proposed.
Although some basic knowledge about Bugs and R could be seen as a significant advantage for an active participation to the training, no specific knowledge will be assumed.

The course will start at 9.00 and finish at about 17.00. Note that **the number of participants is limited**.

More details concerning the practical aspects of this course as well the **registration form** (**deadline: November 1st**) can be found at
http://www.stat.ucl.ac.be/~lambert/BiostatWorkshop2005

The BIOSTATISTICS Section of the
Belgian Statistical Society (SBS-BVS)
organizes a **training** (16 November)
followed by a **workshop** (17 November)

WORKSHOP ON BAYESIAN INFERENCE IN BIOMEDICAL APPLICATIONS

17 November 2005, EORTC, Brussels, Belgium

After the 2-day training by Nicky Best (Imperial College, London) in early May and the November 16 short course by Dr Benoît Beck, the Biostatistics Section concludes its 2005 activities on Bayesian methods by a workshop.

PRELIMINARY PROGRAM

9.00 Welcome, registration and coffee

9.30 Adaptive MCMC,
by *Gareth Roberts*, University of Lancaster

10.30 Diazepam pharamacokinetics from preclinical to phase I using a Bayesian population pharmacokinetic physiological model with informative prior distributions in WinBUGS,
by *Iva Gueorguieva*, Eli Lilly, UK

11.15 Coffee break

11.45 On the smoothing of PET time-activity curves by Bayesian P-Splines,
by *Astrid Jullion*, Université catholique de Louvain

12.15 Lunch

13.45 The Bayesian analysis of two clinical trials in endocrinology,
by *Allun Bedding*, Eli Lilly, UK

14.30 Bayesian proportional hazards model with time varying regression coefficients,
by *Philippe Lambert*, Université catholique de Louvain

15.15 Coffee break

15.45 Complex stats for a seemingly simple problem: the analysis of fluctuating asymmetry,
by *Stefan Van Dongen*, Universiteit Antwerpen

16.30 A practical implementation of the Gibbs sampler for mixture of distributions: application to the determination of specifications in food industry,
by *Myriam Maumy*, Université Louis Pasteur, Strasbourg.

PRACTICAL ASPECTS

More details (including the **registration form**) concerning the November 16 short course and the November 17 workshop can be found at
http://www.stat.ucl.ac.be/~lambert/BiostatWorkshop2005

The **deadline** for registration is **November 1st**.

Organizers: Jan Bogaerts (EORTC), Bruno Boulanger (Eli Lilly), Cécile Dubois (UCB) and Philippe Lambert (UCL).

ANNOUNCEMENTS

2ND INTERNATIONAL MEETING METHODOLOGICAL ISSUES IN ORAL HEALTH RESEARCH: ASSESSING AND IMPROVING DATA QUALITY,

Ghent, 19-21 April 2006

As for the first edition, this international meeting aims to bring together oral health researchers and statisticians interested in the analysis of dental data. The ultimate goal is to stimulate collaboration between the two parties eventually leading to the improvement of the methodological quality of oral health papers and the stimulation of new statistical research useful for the analysis of dental data. The topic of this meeting is "Assessing and improving data quality" and therefore a state of the art will be given on: design issues in oral health studies and on methods that assess, improve and take the quality of oral health studies into account both from a dental as well as a statistical perspective. The invited speakers will give a presentation suitable and attractive for oral health researchers as well as for statisticians. Plenary and discussion sessions will provide oral health researchers and statisticians the opportunity to confront their views. Ample time will be available for oral health researchers and statisticians to meet informally.

Participants are invited to present their work either as an oral presentation or as a poster. Abstracts should be submitted via e-mail before **January 1, 2006** to jeannine.rongy@med.kuleuven.be following the guidelines available on the Website:
http:/med.kuleuven.be/biostat/conferences/Dental2006/

The organisers will notify acceptance or rejection of papers before **February 28, 2006**. Two researchers (preferably from developing countries) will be supported up to an amount of 1250€ each to participate in the meeting. Candidates for this grant are invited to submit a two-pages abstract (see instructions on Website) no later than November 1, 2005. Winners will be notified before November 15, 2005.

The meeting will be organized in Het Pand (Ghent), a medieval abbey (see http://www.gent.be/). Hotel accommodation can be found at
http:/med.kuleuven.be/biostat/conferences/Dental2006/accom

3rd Young Researchers Day (YRD)

INSTITUTE OF STATISTICS (UCL)

December 2nd , 2005, Louvain-la-Neuve

YRD

The young researchers of the Institute of Statistics (UCL) are pleased to announce their third Young Researchers Day (YRD). This meeting consists in a half-day workshop stimulating discussions between young researchers. It is supported by the graduate School of the Institute of Statistics (UCL) and the F.N.R.S.

The Third Young Researchers Day (YRD) will take place on **December 2nd, 2005** in Louvain-la-Neuve. It will follow the same spirit as the 2nd YRD which gathered around 100 attendees last year.

The topic of the 3rd YRD is

On the use of Bayesian Statistics: Some applications in various fields

In other words, we will concentrate on the practical use of different Bayesian methods. Participants are assumed to be familiar with the theoritical background of Bayesian Statistics. A good introduction has already been given by the Biostatistics Section of the Belgian Statistical Society ("Bayesian Inference in Biomedical Applications using WinBUGS", N.Best, 3-4 May 2005). They also propose a workshop (November 17) where the participants will have the possibility to practically approach Bayesian Statistics in Biomedical applications. The objective of this YRD is to offer a larger overview of various applications of Bayesian Statistics showing also their uses in hydrology, actuarial sciences and time series.

www.fnrs.be

This half-day is fully organized by the Ph.D. students of the Institut de statistique (Université catholique de Louvain). The idea is to listen to five young speakers on their own research in that field. Each speaker will have a long time to explain their research. A special attention will be paid on discussions between the participants and the speakers of the workshop. Participants are young researchers, as well as Professors and people working in other companies.

Confirmed speakers:

- Katrien Antonio (K.U.L., Belgium)
- Arnost Komarek (K.U.L., Belgium)
- Simon Woodhead (University of Bristol, England)
- Theodore Kypraios (Lancaster University, England)
- Mathias Hofmann (Ludwig-Maximilians-Universität München,Germany)

For practical informations:

www.stat.ucl.ac.be/YRD

For any further questions or suggestions, send an e-mail to *yrd@stat.ucl.ac.be.*

The organisers:
Carlos Almeida, Angélique Baclin, Hilmar Böhm, Céline Bugli, Alexandra Daskovska, Anouar El Ghouch, Nancy François, Gery Geenens, Cédric Heuchenne, Julien Hunt, Astrid Jullion, Maria Key Prato, Thomas Laloux, Alexandre Lambert, Céline Le Bailly de Tilleghem, Giovanni Motta, Réjane Rousseau, Bianca Teodorescu.

Institut de statistique
Voie du Roman Pays, 20
1348 Louvain-la-Neuve
Belgique
yrd@stat.ucl.ac.be

ww.fnrs.be

COURSE ON INTERVAL CENSORING WITH MEDICAL APPLICATIONS

November 15, 2005

10.00 Introduction, likelihood and identifiabilty problems

12.00 Sandwich lunch

13.00 Frequentist methods for the response: Estimation

15.00 Coffee break

15.30-17.30 Frequentist methods for the response: Hypothesis testing

November 16, 2005

10.00 Interval censoring in covariates (frequentist approach)

12.00 Sandwich lunch

13.00 A Bayesian approach to interval censoring: response and covariates

15.00 Coffee break

15.30-17.00 Splus possibilities for interval censoring

Course Instructors: Guadalupe Gómez[1], M. Luz Calle[2], Ramon Oller[2] and Klaus Langohr[1] ([1]Technical University of Catalonia, Barcelona, [2] Universitat de Vic, Barcelona)

Lecture Room: “Opleidingslokaal”, 2nd floor at UZ St Rafaël, Kapucijnenvoer 35, Leuven

Costs and registration: Free of charge for IAP members, for non-IAP members 100€. For all participants registration is mandatory and need to contact Jeannine Rongy (*jeannine.rongy@med.kuleuven.be*).

LECTURES ON INTERVAL CENSORING WITH MEDICAL APPLICATIONS

November 17, 2005

10.00 Model characterizations and identifiability aspects for the validity of the simplified likelihood for interval censoring data,
by *Ramon Oller* (Universitat de Barcelona)

11.00 Coffee Break

11.30 Bivariate problems involving interval censoring,
by *G*eurt Jongbloed (Free University of Amsterdam)

12.30 Sandwich lunch

13.30 Bayesian accelerated failure time models for clustered doubly-interval- censored data and with flexible distributional assumptions,
by *Arnošt Komárek* (KULeuven)

14.00 Measures of association for bivariate interval censored data,
by *Kris Bogaerts* (KULeuven)

14.30 Likelihood maximization using web based optimization tools for interval censoring patterns,
by *Klaus Langohr* (Technical University of Catalonia, Barcelona)

15.30 Closing of the meeting with coffee

Lecture Room: "Opleidingslokaal", 2nd floor at UZ St Rafaël, Kapucijnenvoer 35, Leuven

Costs: Free of charge. Further information can be obtained from Jeannine Rongy (jeannine.rongy@med.kuleuven.be)

ANNOUNCEMENTS FROM THE IAP-NETWORK

`Statistical techniques and modeling for complex substantive questions with complex data’

The IAP (Interuniversity Attraction Pole, Phase V, 2002-2006) network is a research network in statistics between the universities of KULeuven, UHasselt, ULB, UCL (main coordinator), UJF Grenoble (France) and RWTH Aachen (Germany). The research project focuses on statistical techniques for dealing with complex problems and/or complex data structures. The project is organized around six work packages: Functional estimation, Time series, Survival analysis, Mixed models, Classification and mixture models, and Incompleteness and latent variables.

Forthcoming activities of the network :

- September 30, 2005: the fourth annual workshop will take place at KULeuven on the theme `*how to deal with heterogeneity’*.
- October 2005: short course by Yoav Benjamini (Tel Aviv University, Israel) at UCL.
- November 15-17, 2005 : Course and Lectures on `*Interval Censoring with Medical Applications’* organized by KUL, UCL and UHasselt at KUL.

An updated information can be found at

www.stat.ucl.ac.be/IAP

L'UER « OPÉRATION » DE HEC-ECOLE DE GESTION DE L'ULG, LES SERVICES DE STATISTIQUE DE L'ULG ET L'IREM DE LIÈGE

organisent, le mercredi 26 octobre 2005, une après-midi consacrée à

L'enseignement de la statistique et des probabilités

en hommage au professeur

Léopold BRAGARD

Horaire :

- De 14 H à 14 H 45 : hommages au professeur BRAGARD
- De 14 H 45 à 15 H 30 : conférence de Pierre DAGNELIE
- De 15 H 30 à 16 H : pause – café
- De 16 H à 16 H 45 : conférence de Jean-Jacques DROESBEKE

Participation gratuite - Invitation cordiale à tous

Lieu :

Université de Liège au Sart Tilman, Bâtiment B31 (Faculté de Droit),
Salle du Conseil (3ème étage)

Contact : J.Bair@ulg.ac.be, V.Henry@ulg.ac.be

CONTINUING EDUCATION

Course on Statistics

To meet the needs of users of statistical methods, the Center for Statistics of Ghent University, in co-operation with the Institute for Continuing Education in Science (ICES), organizes a series of courses on statistics each year. Our goal is to provide insight in the basics of statistical research. Practical sessions on PC allow participants to obtain this through hands-on experience. Our courses are aimed at professionals and students with an academic training, who wish to refresh their knowledge, keep it up to date or discover new areas of research. The program is designed to offer very specific knowledge and skills through separate modules.

The 2005-2006 program consists of the following 7 modules:

M1: Survey Analysis	(5 evenings, Sep.-Oct. 2005)
M2: Introduction to SPSS	(4 evenings, October 2005)
M3: Introductory Statistics. Basics of Statistical Inference	(7 evenings, Nov.-Dec. 2005)
M4: Analysis of Variance	(7 evenings, Jan.-Feb. 2006)
M5: Linear Regression	(6 afternoons, Feb.-Mar. 2006)
M6: Applied Longitudinal Analysis	(3 days, April 2006)
M7: Structural Equation Modeling	(5 evenings, Apr.-May 2006)

Please watch our websites www.ipvw-ices.UGent.be and www.cvstat.UGent.be for other short courses or seminars on specialised topics that will be announced during the course of the year.

Students and personnel employed in the government, the non-profit and social-profit sector can participate at a considerably reduced registration fee. The Flemish Community regards continuing training as an important aspect in its economic policy. Again this year professionals are granted financial support through the government's introduction of training cheques. More information about this stimulating initiative can be found on the ICES-website, and on the website of the Flemish Community, http://www.opleidingscheques.be, and the VDAB, http://www.vdab.be/opleidingscheques.

In addition it is possible, for several of these modules, to obtain a certificate of Ghent University upon succeeding in the exam. These modules can therefore be incorporated as a course in a Ph.D. training.

Detailed information about this and other courses, is available at our website. If you would like to be kept informed personally about this or other courses please fill out the 'Information Request Form' at our website. Brochures can be sent to you upon request.

Ghent University – ICES, Krijgslaan 281 S3, 9000 Ghent, Belgium
Tel. +32 (0)9 264 44 26   *Fax* +32 (0)9 264 49 83
E-mail: Heidi.DeDobbelaere@UGent.be
Website : www.ipvw-ices.UGent.be.

FORTHCOMING STATISTICAL EVENTS IN BELGIUM

September 12-16, 2005 – La Roche en Ardenne – European courses in advanced statistics *Regression quantiles and applications*. Contact person: Catherine Vermandele.
Website: http://www.ulb.ac.be/soco/lmtd/ecas2005

September 30, 2005 – Leuven (KULeuven) – IAP-network workshop on the theme *How to deal with heterogeneity*.
Website: http://www.stat.ucl.ac.be/IAP

October 14-15, 2005 – Corsendonk – 13th Annual Meeting of the BVS-SBS, organized by the University of Ghent. Contact person: Olivier Thas.
Website: http://www.bss2005.be/

October 26, 2005 – Liège (ULg) – Après-midi consacrée à *L'enseignement de la statistique et des probabilités* en hommage au Prof. L. Bragard. Organisateurs : ULg et IREM de Liège. *Details available in this issue.*

November 15-17, 2005 – Leuven (KULeuven) – Course and lectures on *Interval censoring with medical* applications as a joint organization of KULeuven, Universiteit Hasselt and UCL (through the IAP network).
Details available in this issue and at http://www.stat.ucl.ac.be/IAP

November 16-17, 2005 – Woluwé-St-Lambert – EORTC – **Short course** (November 16) by Dr. Benoît Beck on *Practical Bayesian Modelling with Bugs/Brugs/Coda/WBDiff in Pharmaceutical R&D* and **workshop** (November 17) on *Bayesian inference with biomedical applications* organized by the Biostatistics Section of the SBS-BVS.
Details available in this issue and at
http://www.stat.ucl.ac.be/~lambert/BiostatWorkshop2005

December 2, 2005 – Louvain-la-Neuve (Institut de statistique, UCL) – 3rd Young Researchers Day. *Details available in this issue and at*
http://www.stat.ucl.ac.be/YRD

April 19-21, 2006 – Ghent – Second International Meeting on Methodological Issues in Oral Health Research: Assessing and Improving Data Quality. *Details available in this issue and at*
http:/med.kuleuven.be/biostat/conferences/Dental2006/

RECENT PhD THESES

Faculté universitaire des Sciences agronomiques de Gembloux

Romain Glèlé Kakaï. *Comparaison empirique des règles de classement et des estimateurs des taux d'erreur en analyse discriminante* (May 12, 2005 – Promotor: R. Palm)

Monte Carlo study is achieved to compare, for two groups discriminant analysis, three classification rules and twenty estimators of error rates associated to the rules, in 480 situations related to the type of distribution, the overlap of the populations, the number of variables, the samples size and the heteroscedasticity degree of the model, which is measured by a parameter Γ defined in the study. The results of this study show that the quadratic rule is the best only for severe heteroscedastic normal distributions with high overlap. The linear rule is the best for homoscedastic normal or moderate non normal models with low overlap. The logistic rule is the best for severe non normal models except when homoscedasticity occurs. In the other situations, linear and logistic rules have almost the same performance. By considering the parameters computed on the data samples, the linear rule is the best when normality occurs whereas logistic rule performs well with non-normality. As far as the error rate estimators are concerned, the results obtained indicate the best performance of e632 estimator, for the computation of the actual error rate associated to the discriminant rules used. The parametric estimators eOS, eO, eM and eL can also be advised for the computation of actual error rate but their performance depends on the normality and the heteroscedasticity degree of the populations. Otherwise, the effect of groups number on the performance of classification rules and error rate estimators, achieved in the case of normal populations, allows to notice the invariability of rules and estimators performance, excepted logistic rule, whose performance decreases with the increasing of groups number

Yves Brostaux. *Etude du classement par forêts aléatoires d'échantillons perturbés à forte structure d'interaction* (July 4, 2005 – Promotor: J.J. Claustriaux)

Amongst classification methods, forests of decision trees (*Random Forests*, Breiman, 2001) are highly versatile concerning descriptive attributes' or target variable's nature and shape of the concept to estimate.

Their diffusion in agronomical sciences is slowed by a lack of information about their ability to learn models with high interaction structures using learning samples with few examples and affected by random noise and irrelevant attributes. This research aim to fill this gap by a systematic exploration of those factors' effects and of the parameters of the Random Forests method, which is done by computer simulations, taking as a reference the classification trees generated by Breiman's *CART* method (1984). Results show that generating random forests with a partially deterministic attributes selection and a forest size of at least 100 or 500 trees give the best prediction accuracy. Those random forests show a significant increase in prediction accuracy on *CART* trees, even for low learning sample size (50 examples). This advantage reduce with the global perturbation level (noise and irrelevant attributes) but increase with the learning sample size, as random forests aren't affected by the asymptotic limitation of the learning curve showed by *CART* method.

Isabelle Carletti. *Etude de méthodes de transformation en rangs et en rangs alignés comme alternatives non paramétriques à l'analyse de la variance pour les plans factoriels équilibrés à deux effets fixes* (July 4, 2005 – Promotor: J.J. Claustriaux)

This study examined the type I error and power properties of usual ANOVA, rank transform and aligned rank transform methods in the context of a balanced 2x4 fixed-effects layout. Different versions of F-test and Kruskal-Wallis test are used with four transformations (ranks, Blom normal scores, Van der Waerden normal scores, Savage scores) and three techniques of alignment (least-square, Lehmann, medians). In order to compare the robustness and the power of tests, simulations were employed for a variety of situations (absence or presence of nuisance parameters, small or large sample size, normality or non normality, equal or unequal variances). This study also examined these properties after a normal pre-test (moment test on standardized residuals) and a variance pre-test (Bartlett test or Brown-Forsythe test). The results showed that ANOVA is the more robust test among studied methods for all situations. Even if ANOVA has a moderate inflation in the type I error rate for some situations of unequal variances, non parametric alternatives of that study are not recommended for two-way fixed-effects layouts.

Université catholique de Louvain

Cédric Heuchenne. *Mean preservation in censored regression using preliminary nonparametric smoothing* (August 18, 2005 – Promotor: I. Van Keilegom)

In this thesis, we consider the problem of estimating the regression function in a general heteroscedastic regression model. This model assumes unknown location function (e.g. conditional mean, median, truncated mean,...), unknown scale function and error independent of X. The response Y is subject to random right censoring, and the covariate X is completely observed.

In the first part of the thesis, the location function is assumed to be a polynomial conditional mean. A new estimation procedure for the unknown parameters of this polynomial is proposed. It extends the classical least squares procedure to censored data. The proposed method is inspired by the method of Buckley and James (1979), but is, unlike the latter method, a non-iterative procedure due to nonparametric preliminary estimation. The asymptotic normality of the estimators is established. Simulations are carried out for both methods and they show that the proposed estimators have usually smaller variance and smaller mean squared error than the Buckley-James estimators.

For the second part, suppose that location belongs to some parametric class of regression functions. A new estimation procedure for the true, unknown parameters, is proposed, that extends the classical least squares procedure for nonlinear regression to the case where the response is subject to censoring. The proposed technique uses new `synthetic' data points that are constructed by using a nonparametric relation between Y and X. The consistency and asymptotic normality of the proposed estimators are established, and the estimators are compared via simulations with estimators proposed by Stute in 1999.

In the third part, we study the nonparametric estimation of general location (or scale) functions. It is well known that the completely nonparametric estimator of the conditional distribution of Y given X suffers from inconsistency problems in the right tail (Beran, 1981), and hence the location function cannot be estimated consistently in a completely nonparametric way, whenever it involves the right tail of this distribution function (like e.g. for the conditional mean). We propose two alternative estimators for this location that do not share the above inconsistency problems. The idea is to make use of the assumed general heteroscedastic regression model, in order to improve the estimation of the conditional distribution, especially in the right tail. We obtain the asymptotic properties of the two proposed estimators. Simulations show that the proposed estimators outperform the completely nonparametric estimator in many cases.

Oana Purcaru. *Modelling dependence in actuarial science, with emphasis on credibility theory and copulas* (August 19, 2005 - Promotor: M. Denuit)

One basic problem in statistical sciences is to understand the relationships among multivariate outcomes. Although it remains an important tool and is widely applicable, the regression analysis is limited by the basic setup that requires to identify one dimension of the outcomes as the primary measure of interest (the "dependent" variable) and other dimensions as supporting this variable (the "explanatory" variables). There are situations where this relationship is not of primary interest. For example, in actuarial sciences, one might be interested to see the dependence between annual claim numbers of a policyholder and its impact on the premium or the dependence between the claim amounts and the expenses related to them. In such cases the normality hypothesis fails, thus Pearson's correlation or concepts based on linearity are no longer the best ones to be used. Therefore, in order to quantify the dependence between non-normal outcomes one needs different statistical tools, such as, for example, the dependence concepts and the copulas.

This thesis is devoted to modelling dependence with applications in actuarial sciences and is divided in two parts: the first one concerns dependence in frequency credibility models and the second one dependence between continuous outcomes. In each part of the thesis we resort to different tools, the stochastic orderings (which arise from the dependence concepts), and copulas, respectively.

During the last decade of the 20th century, the world of insurance was confronted with important developments of the *a posteriori* tarification, especially in the field of credibility. This was dued to the easing of insurance markets in the European Union, which gave rise to an advanced segmentation. The first important contribution is due to Dionne & Vanasse (1989), who proposed a credibility model which integrates *a priori* and *a posteriori* information on an individual basis. These authors introduced a regression component in the Poisson counting model in order to use all available information in the estimation of accident frequency. The unexplained heterogeneity was then modeled by the introduction of a latent variable representing the influence of hidden policy characteristics. The vast majority of the papers appeared in the actuarial literature considered time-independent (or static) heterogeneous models. Noticeable exceptions include the pioneering papers by Gerber & Jones (1975), Sundt (1988) and Pinquet, Guillén & Bolancé (2001, 2003). The allowance for an unknown underlying random parameter that develops over time is justified since unobservable factors influencing the driving abilities are not constant. One might consider either shocks (induced by events like divorces or nervous breakdown, for instance) or continuous modifications (e.g. due to learning effect). In the first part we study the recently introduced models in the frequency credibility theory, which can be seen as models of time series for count data, adapted to actuarial problems. More precisely we will examine the kind of dependence induced among annual claim numbers by the introduction of random effects taking unexplained heterogeneity, when these random effects are static and time-dependent. We will also make precise the effect of reporting claims on the *a posteriori* distribution of the random effect. This will be done by establishing some stochastic monotonicity property of the *a posteriori* distribution with respect to the claims history. We end this part by considering different models for the random effects and computing the *a posteriori* corrections of the premiums on basis of a real data set from a Spanish insurance company.

Whereas dependence concepts are very useful to describe the relationship between multivariate outcomes, in practice (think for instance to the computation of reinsurance premiums) one need some statistical tool easy to implement, which incorporates the structure of the data. Such tool is the copula, which allows the construction of multivariate distributions for given marginals. Because copulas characterize the dependence structure of random vectors once the effect of the marginals has been factored out, identifying and fitting a copula to data is not an easy task. In practice, it is often preferable to restrict the search of an appropriate copula to some reasonable family, like the archimedean one. Then, it is extremely useful to have simple graphical procedures to select the best fitting model among some competing alternatives for the data at hand. In the second part of the thesis we propose a new nonparametric estimator for the generator, that takes into account the particularity of the data, namely censoring and truncation. This nonparametric estimation then serves as a benchmark to select an appropriate parametric archimedean copula. This selection procedure will be illustrated on a real data set.

Alexandre Lambert. *Nonparametric estimation of discontinuous curves and surfaces* (August 26, 2005 – Promotor: I. Gijbels).

In this thesis, we discuss the problem of nonparametric estimation of curves or surfaces from noisy data points where the underlying function to estimate may show some discontinuous behaviour. Most existing methods in this area can be divided into two groups: the first one focuses on the estimation of the position of the jump points (for curve) or on the detection of the edges (for surfaces), the other group tries to estimate directly the functions without preliminary estimation of the jump location. Methods using this last approach should be able to preserve some possible jump/edges while giving a smooth estimation in continuous regions of the true function.

In this work, we propose an estimator (which belongs to the second group) based on local linear kernel estimation which compromises between jump preserving and smoothing. This estimator is analysed theoretically and via simulations/comparisons with alternative procedures. The advantage of the estimator lies in its simple formulation and its non-iterative feature; consequently it is easy to use even in the random design and heteroscedastic cases. Moreover, the parameters involved in the estimator can be selected in a simple data-driven way. The proposed procedure can be used in the one-dimensional case (for curve estimation) as well in the bidimensional case (for surfaces estimation or image denoising). The problem of corner preserving in regression surface is also discussed. This work contains several cross-fertilizations between statistical and engineering methods, and builds a few bridges between two different literatures.

We further also focuses on the problem of edge detection itself by using Difference Kernel Estimators (DKE). In the statistical framework, we discuss the problem of measuring the performance of an edge detector. We study the consistency of the edge detector based on DKE using two different measures of performance. Since the choice of the smoothing parameters highly influences the final results, we study bootstrap procedures to select these in practice. In this bootstrap step we rely on results from the previously developed direct estimation method. Simulations and real data analysis are provided.

Universiteit Hasselt

Niels Hens. *Non- and semi-parametric techniques for handling missing data* (June 10, 2005 - Promotors: M. Aerts and G. Molenberghs)

Missing data arise in various settings, including surveys, clinical trials and epidemiological studies. With or without missing data, the goal of a statistical analysis is to make valid and efficient inferences about a population of interest. The issue of missing values complicates this process. Early on, modelling incomplete data relied on the use of parametric models. Recently, there is a general trend towards non- and semi-parametric approaches to relax assumptions on which parametric models typically rely. Non- and semi-parametric procedures in general will not be as efficient as model-based techniques when there is a posited model, and the model is appropriate. However, if the assumed model is not the correct one, inferences can be worse than useless, leading to misleading interpretations of the data.

In this work, a variety of non- and semi-parametric techniques are used to handle missing data problems. The material presented clearly shows the benefits of relaxing assumptions.

While starting off with a basic introduction into the field of missing data and non- and semi-parametric techniques, the successive parts of this work focus on different topics. A first part describes a kernel based imputation procedure which makes use of a non-parametric regression relationship between a partially observed response and fully observed covariate. The approach is related to the approximate Bayesian bootstrap method and can be seen as an extension of the local single imputation of Cheng (1994) to a proper local multiple imputation approach. An essential ingredient of the algorithm is the local generation of responses.

In a regression analysis, selecting an appropriate model from a candidate set of models is based on, e.g., the Akaike Information Criterion (AIC, Akaike 1973). If however observations are incomplete, the use of complete cases can lead to wrong model choices. In a second part, two modifications of the AIC-criterion are proposed. Firstly, inverse probability weighting is used to improve upon model selection. The method is applicable to both incomplete data and design-based samples. If the weights are unknown, they are estimated using generalized additive models with penalized regression splines. Whenever only a few complete cases are available by deleting every observation with at least one missing value, weighting is not adequate anymore and imputation can provide a solution. Therefore, secondly focus is on an imputation-based AIC-criterion where imputation is non-parametric in nature by using generalized additive models with penalized regression splines. The simulations also reveal potential benefits of model selection after smoothing for fully observed regression data. The use of these AIC versions is illustrated on a case study and contrasted with tree-based methods who deal with both missing values and design.

From the existing material to deal with dropout in longitudinal studies, it is clear that a sensitivity analysis should be part of any statistical analysis. Next to providing an overview of existing sensitivity tools, the third part of the thesis describes a non-parametric sensitivity tool called `kernel weighted influence'. It uses a `kernel based neighbourhood' concept to explore the global and local influence towards non-random missingness for types of observations instead of observations itself in a selection model framework Diggle and Kenward (1994). These sensitivity tools pick up a lot of different anomalies in the data, not only deviations from the MAR-assumption. A method to oppose missing at random versus missing not at random in a selection model framework is the likelihood ratio test. The bootstrap will be used in an attempt to generate the null distribution of the likelihood ratio test statistic opposing missing not at random versus missing at random in a selection model framework.

In a last part, generalized estimating equations are used to determine the force of infection for binary clustered data. The impact of missing data on the analysis is illustrated and inverse probability weighted estimating equations are proposed. The weights are estimated non-parametrically by a generalized additive model with penalized regression splines. Several other complications in the dataset are dealt with, including the constraint for the age-specific seroprevalence to be monotone increasing. Deriving confidence intervals under these constraints is done using the bootstrap. The application of these techniques in the context of veterinary epidemiology is new and therefore considered to be a motivation for interdisciplinary collaboration between statisticians and veterinary epidemiologists working in this field.

Next issue

We would like to publish in this *Newsletter* any statistical matter such as :

- information about universities, institutes (1 to 3 pages);
- lists of recent publications and technical reports;
- abstracts of recent PhD theses;
- news of members;
- forthcoming statistical events and announcements;
- short papers about teaching methods in statistics, statistics in the industry, official statistics, etc.

Suggestions are welcome: please, contact us.

Suitable information for the next issue, prepared as **(LA)TEX or WORD FILES**, should reach the editors of the Newsletter **BEFORE December 31, 2005**, preferable by e-mail to:

Mia.Hubert@wis.kuleuven.ac.be
lambert@stat.ucl.ac.be

Any change of job, address, phone number, ... ?

Please notify the Secretary of the Society:

Gentiane Haesbroeck:
Université de Liège
Institut de mathématique
Sart Tilman B37
B-4000 Liège
g.haesbroeck@ulg.ac.be