negligible consequences when evaluating the model accuracy. A first problem concerns the choice of the evaluation metric to be used for performance measurement. The use of common measures, such as the error rate, yields to misleading results because they depend on the class distribution. More appropriate metrics are based on different propensity towards false negatives and false positives (e.g. precision, recall or ROC curves). Although these evaluation metrics share some drawbacks, the research activity focusing on this issue is very fruitful and several advances have been made. In fact, the evaluation of the accuracy of a classifier in unbalanced learning is subject to a more serious problem than the choice of an adequate error metric, concerning the estimate of such accuracy: whatever evaluation metric is chosen, its expression depends on the unknown probability distribution underlying the data, and hence estimation of this quantity has to be considered. In most of the literature about imbalanced classification, the empirical analysis consists in estimating the classifier over a training set and assessing its accuracy on a test set. However, in real data problems, there are not enough examples from the rare class for both training and testing the classifier and the scarcity of data leads to high variance estimates of the error. It stands to reason that poor estimates of the classifier's performance may lead to erroneous conclusions about the quality of the classifier and proposing more and more sophisticated learning methods becomes a wild-goose chase if we are not able to evaluate their accuracy. In this work we show how smoothed bootstrap techniques may be effectively used in imbalance learning to improve the quality of the estimates of the accuracy.

A Semi-Theoretician's Mid-Day Confession: The True Meaning of i.i.d. in (Applied) Statistics

Xiao-Li Meng
Department of Statistics, Harvard University

Doing good (applied) statistics is inherently – and increasingly – difficult. The size of the data and the complexity of their structure are increasing, as are the depth and specificity of the investigation goals. Yet the available time for conducting the study is decreasing due to intensive competition, especially for funding. Statistical consultants are therefore increasingly asked to perform magic, such as providing scientifically valid causal conclusions for a deeply stratified subgroup based on a handful of weighted samples with wildly varying weights. And by the way, the analysis must be done in one week and the methods must be implementable by (and explainable to) analysts whose statistical experience might come largely from reading output from SAS/SPSS/Stata. This is not a cynical observation, but rather a real challenge that we as statisticians must face in order for our profession to remain at the core of quantitative scientific investigation. In this talk, I will report my own experiences both crying and smiling as a member of a team of statistical consultants for the National Latino and Asian American Study (NLAAS), a recent survey of psychiatric epidemiology, which measured over 5000 variables and embedded experiments on different survey instruments. In particular, I will report on the success and failure of using Bayesian modeling and multiple imputation to deal with the respondents' untruthful self-reporting regarding health service use, as detected by the embedded experiments.
The true meaning of i.i.d. will become clear only by the end of my talk; unless, of course, you have already deciphered it from this abstract...

Network-WIDEE Statistical Modeling and Prediction of Computer Traffic

George Michailidis
University of Michigan

Computer network use is becoming increasingly widespread, both in terms of number of users and variety of applications. In order to provide consistently high quality service, network engineers and other professionals must monitor several aspects of the network, including the traffic intensity on the links that comprise the network. As networks grow, this type of monitoring has potential to become burdensome in terms of resources required. Motivated by the prospect of monitoring only a small subset of links, we explore the problem of using observed traffic measurements on selected links to predict the traffic on other, unobserved links. The characteristics of such unobserved links are learned through auxiliary data. Although more expensive to obtain, this extra data set provides the necessary information to represent important structure in the network, and can significantly improve the results of prediction as compared with more naive approaches. In addition, we introduce an adjusted control chart methodology that shows possible applications of our prediction results in situations where all links may be observed.

The Mathematical Model of an Assembly Link Checking Stand

Jiří Michálek
Institute of Information Theory and Automationof the ASCR, Prague, Czech Republic

The mathematical model of an assembly link checking stand Jiří Michálek Institute of Information Theory and Automation Czech Academy of Science The construction of the proposed model is motivated by a real case of a checking stand where valves coming from the assembly link are tested and repaired respectively. The assembly link presents one-flow production of valves that are coming to the test station. When a valve passes the test the valve is marked OK and released to expedition. In case a valve does not pass the test the first attempt of repair follows. After repairing the valve goes through the test again and if the test is passed the valve is released. But when the test is not fulfilled the second attempt of repaire is starting and so on until the valve is OK. In the real situation the number of possible repaires is not limited but in the model this number is limited by three attempts for simplicity. These valves that could not be repaired during three attempts are declared as scraps and excluded from the next operations. The mathematical model is based on Markov and semi-Markov chains. The embedded Markov chain has 6 states in a simplier form considered in this contribution. These states are: state T for testing, state OK after passing the test, three states R1, R2 and R3 for successive repaires. State S is for scraps. The values of transition probabilities for the transition matrix were determined by estimates obtained from the automatic collection of data performed at the checking stand. The embedded Markov chain is nonperiodic and irreducible having the unique class of ergodic states. In accordance with the real situation a random amount of time spent in every state must be considered. Possible probability distributions are like log-normal or Weibull, also in special situations normal distribution could be acceptable. The matrix of transition probabilities together with probability distributions of random times define in a unique way the corresponding semi-Markov chain. This semi-Markov chain is in fact a regenerative process because after every state T the probabilistic structure of the process is the same. The most interesting characteristics are the mean length of time interval between two states OK and simmilarly the mean time between two scraps. Further interesting characteristics are the probability of accurence of a scrap or number of scraps within a time period, e.g. within a shift. The model can serve as a suitable tool for observing changes in its characterictics under changes in input transition probabilities and parameters of time distributions.

Quasi-Quantile Regression

Ivan Mizera
University of Alberta

Quantile regression, arguably "one of the most important development in statistics in recent years" (according to the new edition of the Ramsay's and Silverman's Functional Data Analysis book), is becoming an increasingly popular data-analytic tool. After reviewing several areas of applications of this methodology, illustrating the type of achievable insights on data examples and emphasizing the involvement of modern methods of convex optimization, we concentrate on one of its "twilight zones" (in the terminology of the Koenker's monograph on the subject): fitting of parametric models of quantile regression in the circumstances typically leading, in the mean-regression context, to what is generally referred to as generalized linear models. In particular, we focus on quantile regression of counts, where the existing methodology still exhibits somewhat peculiar aspects; we try to adapt the ideology of quasi-likelihood to quantile regression there, in the hope of obtaining more conventional, but still valid fitting strategies.

Multi-Treatment Regression Approach for the Time Evolution of Heavy Metals Concentration in a Electrodialtic Removal Method

Elsa Moreira, *João Tiago Mexia*
Research Center in Mathematics and Applications, Faculty of Sciences and Technology, Nova University of Lisbon

The electrodialytic removal of heavy metals from waste materials impregnated with chemicals is a remediation process that promotes the re-use. The electrodialytic removal of Cu, Cr and As was tested in wood chips treated with chromated copper arsenate (CCA). The method uses a low-level direct current as the cleaning agent, in the presence of an extracting solution. Several experiments were conducted using different extracting solutions and initial current intensity, which will be considered as different treatments. A polynomial model was fitted to the time evolution of each metal concentration in the electrolytes. Based on this modeling and aiming to choose the best treatment in jointly removing the three metals, five experiments were selected in order to be analyzed under multi-treatment regression approach. In this approach, instead of a sample for each treatment, there is a linear regression in the same variables, both controlled and dependent. Then, instead of the action of the treatments on the sample mean values, the action on the regression coefficients is studied. ANOVA algorithms and multiple comparison methods are adapted aiming to perform the comparison between the coefficients of different regressions. Data was unbalanced, that is, the model matrix $\boldsymbol{X}$ for the values of the controlled variables was not the same in all regressions. Thus, a special covariance matrix (diagonal by blocks) had to be considered in order to perform the hypotheses testing. The results point to the choice of the oxalic acid 2.5% and a current ranging from 20 to 40 mA.

Probabilities of Extreme Half-Spaces

Stephan Morgenthaler
Ecole polytechnique fédérale de Lausanne

Debbie Dupuis
HEC, Université de Montréal

In this talk we will present a method based on the generalized Pareto distribution for estimating probabilities of extremal half-spaces involving multivariate observations. If the multivariate distribution is elliptical we can use affine transformations between half-spaces. This allows us to combine estimators based on a variety of one-dimensional projections of the multivariate data set, rather than using a single projection. More reliable estimators result from this practice. Furthermore, the technique suggests some natural estimates of the sampling variance. An application to risk assessment in financial data will be given.

International Symposium on Business and Industrial Statistics

Portorož (Portorose) Slovenia July 5 – 9, 2010

ISBIS-2010 BOOK OF ABSTRACTS

ISBIS-2010

International Symposium on Business and Industrial Statistics

BOOK OF ABSTRACTS

[illegible] Business and Industrial Statistics

Portorož (Portorose) Slovenia July 5 – 9, 2010