SiDOR
Statistical Inference
Decision & Operations Research Group
Universidade Vigo

Seminars on Statistics

November 22–23, 2018

Facultade de Ciencias Económicas e Empresariais
Aula Seminario 8

Thursday, November 22th

16:00-17:00

Modelling count time series in a state-dependent under-reporting scheme

Alejandra Cabaña (Universitat Autònoma de Barcelona)

Abstract: Since the introduction of the Integer-Valued AutoRegressive (INAR) models in [1], the interest in the analysis of count time series has been growing. The main reason for this increasing popularity is the limited performance of the classical time series analysis approach when dealing with discrete valued time series. With the introduction of discrete time series analysis techniques, several challenges appeared such as unobserved heterogeneity, periodicity, under-reporting, ... The problem of under-reported data is still in a quite early stage of study in many different fields. This phenomenon is very common in many contexts such as epidemiological and biomedical research. It might lead to potentially biased inference and may also invalidate the main assumptions of the classical models.

The model we will present here considers two discrete time series: the observed series of counts Y_t which may be under-reported, and the underlying series X_t with an INAR(1) structure

$$X_t = \alpha \circ X_{t-1} + W_t$$

where $0 < \alpha < 1$ is a fixed parameter and W_t is Poisson(λ). The *binomial thinning* operator is defined as $\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} Z_i$, where Z_i are i.i.d Bernoulli random variables with probability of success equal to α.

The way we allow Y_t to be under-reported is by defining that $Y_t = X_t$ with probability $1 - \omega$ or it is $q \circ X_t$ with probability ω. This process $\{Y_n\}$ represents an under-reported phenomenon coming from the latent INAR(1) process, where parameters ω and $q$are the frequency and the intensity of under-reporting, respectively.

For a more general model, suppose now that the states of under-reporting follow a binary discrete-time Markov chain. This new situation only adds one further parameter to the previous situation, and is more flexible for modelling.

We derive the autocorrelation structure of both models, and compute maximum likelihood estimators for the parameters via a modification of Viterbi algorithm.

Several examples of application of the models in the field of public health will be discussed, using real data.

References

[1] Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive INAR(1) process. *Journal of Time Series Analysis*, 8, 261–275.

[2] Heiligers B. (1994). *E*-optimal designs in weighted polynomial regression. *Ann. Stat.*, , 22, 917–929.

[3] Moriña, D., Puig, P., Ríos, J., Vilella, A. and Trilla, A. (2011). A statistical model for hospital admissions caused by seasonal diseases. *Statistics in Medicine*, **30**, 3125–3136.

17:00-18:00

Powerful tests for stochastic and umbrella orderings

Mª Carmen Pardo (Universidad Complutense de Madrid)

Abstract: A wide variety of practical problems can be studied as problems of inference under order relations. For example, in a study of a chemotherapy agent for cancer patients, it is reasonable to expect that the time to the first dose-limiting toxicity (DLT) event is stochastically decrease with the increase of dose. This means that the time to DLT is ordered, stochastically, by dose. However, when we consider the relationship between dose and patient survival time, it is reasonable to expect that survival time will increase with dose to a certain level due to efficacy of the drug and then decrease with dose because of death due to the toxicity of the drug. In this case, the survival time shows a U-shape order in its distribution functions with dose, the flip side of an umbrella order.

In this talk, following [3], we propose two families of test statistics for testing stochastic and umbrella ordering. We focus on empirical likelihood and chi-squared-based test statistics and their asymptotic distributions are obtained. Next, we carry out a simulation study to compare the power of some new members of the two families against of a couple of known statistics ([1],[2]) in several scenarios. Furthermore, some real data are used to illustrate our approaches. Finally, we make concluding remarks and point out future research.

References

[1] Davidov, O. and Herman, A. (2010). Testing for order among K populations: theory and examples. *The Canadian Journal of Statistics*, 38(1), 97-115.

[2] El Barmi, H. and McKeague, I. W. (2013). Empirical likelihood based tests for stochastic ordering. *Bernoulli*, 19, 295-307.

[3] Zhang, J. and Wu, Y. (2007). k.-Sample tests based on the likelihood ratio. *Computational Statistics & Data Analysis*, 51(9), 468-691.

Friday, November 23th

10:00-10:30

Goodness-of-fit tests for disorder detection in NGS experiments

Norman Jiménez Otero (Universidade de Vigo)

Abstract: Next-generation sequencing experiments (NGS) are often performed in biomedical research nowadays, leading to methodological challenges related to the high-dimensional and complex nature of the recorded data. In this work we review some of the issues which arise in disorder detection from NGS experiments, that is, when the focus is the detection of deletion and duplication disorders for homozygosity and heterozygosity in DNA sequencing. A statistical model to cope with guanine/cytosine bias and phasing and prephasing phenomena at base level is proposed, and a goodness-of-fit procedure for disorder detection is derived. The method combines the proper evaluation of local p-values (one for each DNA base) with suitable corrections for multiple comparisons and the discrete nature of the p-values. A global test for the detection of disorders in the whole DNA region is proposed too. The performance of the introduced procedures is investigated through simulations. A real data illustration is provided.

10:30-11:00

Non-parametric statistical inference for comparing conditional ROC curves

Arís Fanjul Hevia (Universidade de Santiago de Compostela)

Abstract: The comparison of two or more Receiver Operating Characteristic (ROC) curves is a commonly accepted way of comparing the accuracy and the behaviour of different diagnostic procedures. Several methods may be found in the literature concerning the comparison of two or more ROC curves. Along with the diagnostic variable it is usual to observe other covariates, but that extra information has been hardly ever considered for the comparison of this kind of curves despite the fact that the discriminatory capability of these curves can be influenced by this extra information. With this idea, a new non-parametric test is proposed for the comparison of conditional ROC curves. Simulations are run to analyse the practical performance of the test and an application to real data is presented to illustrate the procedure.

11:00-11:15

Speeding up R code

José Carlos Soage (Universidade de Vigo)

Abstract: The speed of execution of R codes, mainly loops, is usually slow, since speed is not the main objective of this language. Such a lengthy computational time may be an

important issue when performing simulation studies as well as for the development of R packages which implement intensive algorithms (like numerical optimization or resampling methods, for example). In this presentation we will evaluate alternative ways to reduce the execution time of R codes, through vectorized and parallelized examples.

11:15-11:30

The conditional Efron-Petrosian estimator

Natalia Pérez Veiga (Universidade de Vigo)

Abstract: In this work we consider the problem of estimating the conditional distribution function of a doubly truncated variable given a one-dimensional continuous covariate. For that aim we introduce a conditional version of the Efron-Petrosian estimator. The case in which the truncating variables may depend on the covariate is investigated. An iterative algorithm to compute the estimator is proposed. A simulation study is performed, leading to some conclusions regarding optimal smoothing. An illustrative application to the analysis of AIDS incubation times is reported.

11:30-12:15

Coffee break

12:15-12:45

Penalised-based estimation of the conditional time-dependent ROC curve

María Xosé Rodríguez-Álvarez (BCAM - Basque Center for Applied Mathematics)

Abstract: Prior to using a diagnostic biomarker in a routine clinical setting, the rigorous evaluation of its diagnostic accuracy is essential. The receiver operating characteristic (ROC) curve is the measure of accuracy most widely used for continuous biomarkers. However, the possible impact of extra information about the patient (or even the environment) on diagnostic accuracy also needs to be assessed. In addition, in many circumstances the aim of a study may involve prognosis rather than diagnosis. The main difference between diagnostic and prognostic biomarkers is that, with prognostic biomarkers, a time dimension is involved. This is the case of survival studies, where the status of an individual varies with time (e.g, death and alive). To assess the accuracy of continuous prognostic biomarkers for time-dependent disease outcomes, time-dependent extensions of the ROC curve have been proposed. This work presents a novel penalised likelihood-based estimator of the cumulative-dynamic time-dependent ROC curve. The proposal allows to account for the possible modifying effect of covariates on the accuracy of the biomarker. The validity of the approach is supported by simulation, and applied to the evaluation of biomarkers for early prognosis of death after discharge in patients who suffered an acute coronary syndrome.

12:45-13:15

Recent advances in nonparametric estimation from doubly truncated data

Jacobo de Uña-Álvarez (Universidade de Vigo)

Abstract: Doubly truncated data often appear in Survival Analysis and Epidemiology, among other fields. Under double truncation, the observed values belong to a random interval which may vary from individual to individual. The naive analysis of such data can be systematically biased due to the double truncation issue, so suitable corrections are needed. The seminal paper of Efron and Petrosian (*J. Amer. Statist. Assoc.*, 1999) gave rise to a number of developments in this area of reasearch. However, several technical and practical issues remain still unsolved, and the sampling scheme itself and the corresponding potential biases are often overlooked. In this talk I will review some recent advances in nonparametric estimation from doubly truncated data, including: estimation of the lifetime cumulative distribution, correlation analysis, regression and multi-state models. Simulations and real data illustrations will be provided, as well as discussion of the asymptotic theory needed for inference purposes.