SIS2016

Università degli Studi di Salerno
June 8th – June 10th, 2016

PROCEEDINGS

of the 48th scientific meeting of the
Italian Statistical Society

Editors: Monica Pratesi and Cira Pena

ISBN: 9788861970618

(SOL-20) Models for studying the mobility of students

S. Balia	Modelling inter-regional patient mobility: evidence from the Italian NHS. (Co-author(s): R. Brau, E. Marrocu)
A. D'Agostino	University mobility at enrollment: geographical disparities in Italy. (Co-author(s): G. Ghellini, S. Longobardi)
M. Enea	From South to North? Mobility of Southern Italian students at the transition from the first to the second level university degree.
F. Giambona	Measuring territory student-attractiveness in Italy. Longitudinal evidence.

CONTRIBUTED SESSION (CON)

(CON-01) Bayesian statistics (1)

F. Giummolè	Reference priors based on composite likelihoods. (Co-author(s): V. Mameli, L. Ventura)
B. Nipoti	On Bayesian nonparametric inference for discovery probabilites. (Co-author(s): J. Arbel, S. Favaro, Y. W. Teh)
R. Pappadà	Relabelling in Bayesian mixture models by pivotal units. (Co-author(s): L. Egidi, F. Pauli, N. Torelli)
C. Scricciolo	On Deconvolution of Dirichlet-Laplace Mixtures.

(CON-02) Statistical modeling

P. Faroughi	A New Bivariate Regression Model for Count Data with Excess Zeros. (Co-author(s): N. Ismail)
B. Francis	Dynamic latent class profiles in cross-sectional surveys: some preliminary results. (Co-author(s): V. Hoti)
P. M. Kroonenberg	The use of deviance plots for non-nested model selection in loglinear models, structural equations, three-mode analysis.
A. Lucadamo	Variable selection through Multinomial LASSO for PCMR. (Co-author(s): L. Greco)
O. Paccagnella	Integrating CUB Models and Vignette Approaches. (Co-author(s): S. Pavan, M. Iannario)

(CON-03) Demographics and social statistics (1)

D. Bellani	Gender egalitarianism, education and life-long singlehood: A multilevel analysis. (Co-author(s): G. Esping-Andersen, L. Nedoluzhko)
L. Colangelo	Fear of Crime and Victimization among Sexual Harassed Women: Evidence from Italy. (Co-author(s): P. Mancini)

S. De Cantis	A survival approach for the analysis of cruise passengers' behavior at the destination. (Co-author(s): M. Ferrante, A. Parroco, N. Shoval)
A. Di Pino	Retirement of the Male Partner and the Housework Division in the Italian Couples: Estimation of the Causal Effects. (Co-author(s): M. Campolo)
F. Lariccia	Many women start, but few continue: determinants of breastfeeding in Italy. (Co-author(s): A. Pinnelli)

(CON-04) Environmental statistics

F. Bono	Measuring sustainable economic development through a multidimensional Gini index. (Co-author(s): M. Giacomarra, R. Gialmo)
C. Calculli	Modeling multi-site individual corals growth. (Co-author(s): B. Cafarelli, D. Cocchi, E. Pignotti)
F. Di Salvo	GAMs and functional kriging for air quality data. (Co-author(s): A. Plaia, M. Ruggieri)
F. Durante	The Kendall distribution and multivariate risks.

(CON-05) Health statistics

E. di Bella	Dental care systems across Europe: the case of Switzerland. (Co-author(s): L. Leporatti, I. Krejci, S. Ardu)
F. Gasperoni	Multi-state models for hospitalizations of heart failure patients in Trieste. (Co-author(s): F. Ieva, G. Barbati)
F. Grossetti	Multi-state Approach to Administrative Data on Patients affected by Chronic Heart Failure. (Co-author(s): F. Ieva, S. Scalvini, A. M. Paganoni)
G. Montanari	Evaluation of health care services through a latent Markov model with covariates. (Co-author(s): S. Pandolfi)

(CON-06) Labor market statistics

A. Bianchi	Multifactor Partitioning: an analysis of employment and firm size. (Co-author(s): S. Biffignandi)
G. Busetta	Ugly Betty looks for a job. Will she ever find it in Italy?. (Co-author(s): F. Fiorillo)
G. Busetta	No country for foreigners: an analysis of hiring process in Italian labor market. (Co-author(s): M. Campolo, D. Panarello)
F. Crippa	Know your audience. Towards a partnership between employers and university. (Co-author(s): M. Zenga)
I. Vannini	Online Job Vacancies: a big data analysis. (Co-author(s): D . Rotolone, C. Di Stefano, A. P. Paliotta, D. F. Iezzi)

(CON-07) Robust statistics

F. Greselin Robust estimation of mixtures of skew-normal distributions. (Co-author(s): L. Gàrcía-Escudero, A. Mayo-Iscar, G. McLachlan)

M. Musio Renyi's Scoring Rules. (Co-author(s): A. F. Dawid)

A. Paganoni Robust classification of multivariate functional data. (Co-author(s): F. Ieva)

G. C. Porzio A robust estimator for the mean direction of the von Mises-Fisher distribution. (Co-author(s): T. Kirschstein, S. Liebscher, G. Pandolfo, G. Ragozini)

F. Palumbo Robust Partial Possibilistic Regression Path Modeling. (Co-author(s): R. Romano)

(CON-08) Sampling methods

A. Ghiglietti Adaptive Randomly Reinforced Urn design and its asymptotic properties.

D. Marella PC algorithm from complex sample data. (Co-author(s): P. Vicard)

S. Missiroli Optimal Adaptive Group Sequential Procedure for Finite Populations in the Presence of a Cost Function. (Co-author(s): E. Carfagna)

E. Pelle The Rao regression-type estimator in ranked set sampling. (Co-author(s): P. Perri)

M. Ruggiero Modelling stationary varying-size populations via Polya sampling. (Co-author(s): P. De Blasi, S. Walker)

(CON-09) Economic data analysis

M. Brunetti Getting older and riskier: the effect of Medicare on household portfolio choices. (Co-author(s): M. Angrisani, V. Atella)

E. Ciavolino Modelling the Public Opinion on the European Economy with the HO-MIMIC Model. (Co-author(s): M. Carpita)

G. D'Epifanio Indexing the Worthiness of Social Agents. To norm index on conventional specifications.

G. Guagnano An econometric model for undeclared work. (Co-author(s): M. Arezzo)

M. Mussini A spatial shift-share decomposition of energy consumption variation. (Co-author(s): L. Grossi)

(CON-10) Quantile methods

M. Bernardi Bayesian inference for L$_p$–quantile regression models. (Co-author(s): V. Bignozzi, L. Petrella)

V. Bignozzi On the L_p-quantiles and the Student t distribution. (Co-author(s): M. Bernardi, L. Petrella)

M. Marino M-quantile regression for multivariate longitudinal data. (Co-author(s): M. Alfò, M. Ranalli, N. Salvati)

D. Vistocco Comparing Prediction Intervals in Quantile and OLS Regression. (Co-author(s): C. Davino)

(CON-11) Statistical algorithms

N. Loperfido An Algorithm for Finding Projections with Extreme Kurtosis. (Co-author(s): C. Franceschini)

L. Scrucca Poisson change-point models estimated by Genetic Algorithms.

A. Stamm Maximum Likelihood Estimators of Brain White Matter Microstructure. (Co-author(s): O. Commowick, S. Vantini, S. K. Warfield)

(CON-12) Statistics for medicine

G. Barbati Competing risks between mortality and heart failure hospital re-admissions: a community-based investigation from the Trieste area. (Co-author(s): F. Ieva, A. Scagnetto, G. Sinagra, A. Di Lenarda)

C. Brombin Evaluating association between emotion recognition and Heart Rate Variability indices. (Co-author(s): F. Cugnata, R. M. Martoni, M. Ferrario, C. Di Serio)

M. Ferrante Socio-economic deprivation, territorial inequalities and mortality for cardiovascular diseases in Sicily. (Co-author(s): A. Milito, A. Parroco)

M. Giacalone The use of Permutation Tests on Large-Sized Datasets. (Co-author(s): A. Alibrandi, A. Zirilli)

(CON-13) Statistics for the education system

G. Boscaino Further considerations on a new indicator for higher education student performance. (Co-author(s): G. Adelfio, V. Capursi)

C. Masci Analysis of pupils' INVALSI achievements by means of bivariate multilevel models. (Co-author(s): A. Paganoni, F. Ieva, T. Agasisti)

A. Valentini Promoting statistical literacy to university students: a new approach adopted by Istat. (Co-author(s): G. De Candia, M . Carbonara)

(CON-14) Testing procedures

E. Cascini A Reliability Problem: Censored Tests.

G. De Santis Testing the Gamma-Gompertz-Makeham model. (Co-author(s): G. Salinari)

M. M. Pelagatti A nonparametric test of independence.

A. Pini Functional Data Analysis of Tongue Profiles. (Co-author(s): L. Spreafico, S. Vantini, A. Vietti)

A. Vagheggini On the asymptotic power of the statistical test under Response-Adaptive randomization. (Co-author(s): A. Baldi Antognini, M. Zagoraiou)

(CON-15) Time series analysis

C. Cappelli	Robust Atheoretical Regression Tree to detect structural breaks in financial time series. (Co-author(s): P. D'Urso, F. Di Iorio)
P. Chirico	Prediction intervals for heteroscedastic series by Holt-Winters methods.
M. Costa	Inequality decomposition for financial variables evaluation.
G. De Luca	Three-stage estimation for a copula-based VAR model. (Co-author(s): G. Rivieccio)

(CON-16) Forecasting methods

M. Andreano	Forecasting with Mixed Data Sampling Models (MIDAS) and Google trends data: the case of car sales in Italy. (Co-author(s): R. Benedetti, P. Postiglione)
V. Candila	Probability forecasts in the market of tennis betting: the CaSco normalization. (Co-author(s): A. Scognamillo)
S. Vantini	Daily Prediction of Demand and Supply Curves. (Co-author(s): A. Canale)

(CON-17) Bayesian statistics (2)

G. Marchese	Bayesian hierarchical models for analyzing and forecasting football results. (Co-author(s): P. Brutti, S. Gubbiotti)
L. Paci	Bayesian modeling of spatio-temporal point patterns in residential property sales. (Co-author(s): A. E. Gelfand, M. Beamonte, P. Gargallo, M. Salvador)
V. Vitale	Non-parametric Bayesian Networks for Managing an Energy Market. (Co-author(s): V. Guizzi, F. Musella, P. Vicard)

(CON-18) Business statistics

E. Bartoloni	How do firms perceive their competitiveness? Measurement and determinants.
C. Bocci	An evaluation of export promotion programmes with repeated multiple treatments. (Co-author(s): M. Mariani)
A. Righi	The inter-enterprise relations in Italy. (Co-author(s): A. Nuccitelli, G. Barbieri)

(CON-19) Clustering and classification

C. Drago	Dendrograms Stability Analysis of Sub-periods Time Series Clustering. (Co-author(s): R. Ricciuti)
G. Menardi	Stability-based model selection in nonparametric clustering.
T. Padellini	Topological signatures for classification. (Co-author(s): P. Brutti)

(CON-20) Demographics and social statistics (2)

M. Antonicelli	Ecolabels: informin or confusing customers? Evidences form the agrifood sector. (Co-author(s): D. Calace, D. Morrone, A. Russo, V. Vastola)
B. Arpino	What makes you feeling old? An analysis of the factors influencing perceptions of ageing. (Co-author(s): V. Bordone, A. Rosina)
G. De Santis	A (partial) solution to the intractability of APC models. (Co-author(s): M. Mucciardi)
G. Gabrielli	Partner reunification of first generation immigrants in Lombardy. (Co-author(s): A. Paterno, L. Terzera)

(CON-21) Statistical inference

E. Kenne Pagui	Median bias reduction of maximum likelihood estimates in binary regression models. (Co-author(s): A. Salvan, N. Sartori)
N. Lunardon	On penalized likelihood and bias reduction. (Co-author(s): G. Adimari)
A. Maruotti	Population size estimation and heterogeneity in capture-recapture count data. (Co-author(s): O. Anan, D. Böhning)

(CON-22) Survey methods

A. Pinto	Italian consumers' food risks perception: an approach based on the correspondence analysis. (Co-author(s): E. Ruli, S. Crovato, L. Ventura, L. Ravarotto)
R. Salvatore	Spatial-temporal multivariate small area estimation. (Co-author(s): F. Cappuccio)
D. Toninelli	Is the Smartphone Participation Affecting the Web Survey Experience?. (Co-author(s): M. Revilla)

POSTER SESSION (POS)

M. Bernardi	Non-conjugate Variational Bayes Approximation. (Co-author(s): E. Ruli)
M. Bernardi	The Multivariate Fuzzy Skew Student–t distribution.
M. Bini	Quality of Educational Services, Institutional Image, Students' Satisfaction and Loyalty in Higher Education. (Co-author(s): L. Masserini, M. Pratesi)
L. Bisaglia	Estimation of INAR(p) models using bootstrap. (Co-author(s): M. Gerolimetto)
D. Bossoli	Effect of internet-based cognitive therapy on children anxiety disorders: results from a marginal logistic quantile regression.

The use of Permutation Tests on Large-Sized Datasets

L'uso dei Test di Permutazione su Grandi Datasets

Massimiliano Giacalone, Agata Zirilli and Angela Alibrandi [1]

Abstract The increasing availability of large-sized datasets produces a growing interest in permutation testing methods. They represent an effective solution for problems concerning the testing of multidimensional hypotheses, difficult to face in a parametric context. In this paper we propose an application of permutation test on a large amount of data in order to show its utility to analyze a big dataset array. The analysis was performed in order to assess the existence of significant differences, with reference to several variables, between two gastrointestinal illnesses.

Abstract *La crescente disponibilità di grandi set di dati comporta un notevole interesse per i metodi basati su test di permutazione. Essi rappresentano un'efficace soluzione per problemi inerenti la verifica di ipotesi multidimensionali, difficili da affrontare in un contesto parametrico. Nel presente lavoro proponiamo un'applicazione di tale metodologia ad una elevata numerosità di dati, al fine di dimostrarne l'utilità nell'analisi di un grande dataset. In particolare è stata indagata l'esistenza di differenze significative, in relazione a diverse variabili esaminate, tra due malattie gastrointestinali.*

Key words: Large-Sized Datasets, Permutation Tests, Application in Medical Field.

The statistical background of Large-Sized Datasets

In recent years, there is a growing interest in permutation testing methods due to the increasing availability of large-sized datasets and the consequent need to solve

[1] Massimiliano Giacalone, Naples University "Federico II", massimiliano.giacalone@unina.it
Agata Zirilli, Messina University, azirilli@unime.it
Angela Alibrandi, Messina University, aalibrandi@unime.it

more and more complex multivariate problems. Actually, permutation tests are essentially exact in a nonparametric conditional framework, where conditioning is on the pooled observed data set; it is generally a set of sufficient statistics in the null hypothesis. Many complex multivariate problems are difficult to handle outside the conditional framework and, in particular, outside the nonparametric combination (NPC) of dependent permutation tests (Arboretti and Brombin, 2014). While permutation tests and bootstraps have very wide-ranging application, both share a common potential drawback: as data-intensive resampling methods, both can be runtime prohibitive when applied to large or even medium-sized datasets. The data explosion over the past few decades has made this a common occurrence and it highlights the increasing need for faster and more efficient permutation tests and bootstrap algorithms (Opdyke, 2013). The permutation test essentially works by combining two important ideas: exchangeability and conditioning. More generally the exchangeability and other sorts of stochastic orders are keys to robust inference on large-sized datasets.

The paper briefly shows an application of NPC test for the analysis of Big Data; since medical data were analyzed, theoretical, methodological and applicative aspects have been fruitfully integrated with specific competences from medicine field (Peek et al., 2014; Rezzani, 2013). We apply a permutation test on a large amount of patients (about 1700) in order to assess the existence of significant differences between patients affected by two gastrointestinal illnesses: Crohn's Disease (CD) and Ulcerative Colitis (UC). Our research showed the utility of the NPC test into analyze a large dataset array.

2. Permutation test

2.1 Methodology

In this context we introduce the theoretical aspects of Non Parametric Combination (NPC) test, based on permutation solution (Pesarin and Salmaso, 2010; Pesarin, 2001; Corain and Salmaso, 2004). Permutation tests represent an effective solution for problems concerning the testing of multidimensional hypotheses, that are difficult to face in a parametric context. This multivariate and multistrata procedure allows to reach effective solutions concerning problems of multidimensional hypotheses verifying within the non parametric permutation inference (Pesarin, 2001); it is used in different application fields that concern verifying of multidimensional hypotheses with a complexity that cannot be managed in parametric context. In comparison to the classical approach, NPC test is characterized by several advantages: it does not request normality and homoschedasticity assumption; it draws any type of variable (Pasarin and Samaso, 2006; Klingenberg et al., 2008); it assumes a good behaviour also in presence of missing data; it is powerful in presence of low sampling size (Brombin and Salmaso, 2009); it resolves multivariate problems without the necessity to specify the structure of dependence among variables; it allows stratified analyses; it allows to test multivariate restricted alternative hypothesis (allowing the verifying of the directionality for a specific alternative hypothesis); it resolves problems in which observations number is smaller than variables number (Finos and Salmaso, 2006; Basso et al., 2007). All these properties make NPC test very flexible and widely applicable in several fields; in particular we cite recent applications in medical context and in genetics (Zirilli and Alibrandi, 2009; Zirilli and Alibrandi, 2011; Zirilli and Alibrandi, 2012; Bonnini et al., 2006; Arboretti et al., 2005; Salmaso, 2005; Finos et al., 2004; Bonnini et al., 2003; Callegaro et al., 2003; Di Castelnuovo et al., 2000). By means of mentioned procedure it is preliminarily possible to define a set of k one-dimensional permutation test, denominated partial test, through which the marginal contribution of every answer-variable can be examined in the comparison among groups. The partial tests are non-parametrically combined through CMC (Conditional Monte Carlo) procedure in combined tests, using an opportune combination function (generally Fisher, Tippett or Liptak); these tests globally verify the existence of differences among the multivariate distributions of the groups. We supposed that K variables are noticed on N observations (dataset $N{\times}K$) belonging to C groups and that an appropriate K-dimensional distribution exists. The null hypothesis postulates the equality in distribution of k-dimensional distribution among all C groups (1) against the alternative hypothesis (2):

$$H_0 = [\bigcap_{i=1}^{k} X_{1i} \overset{d}{=} \ldots \overset{d}{=} X_{Ci}] = [\bigcap_{i=1}^{k} H_{0i}] \quad (1) \qquad \text{against} \qquad H_1 = \bigcup_{i=1}^{k} H_{1i} \quad (2)$$

Let's assume that, without loss of generality, the partial tests assume real values and they are marginally correct, consistent and significant for great values; the NPC test procedure (based on CMC resampling) develops into the following two-phases algorithm (Fig. 1). The hypotheses systems are verified by the determination of partial tests (1st order) that allow to evaluate the existence of statistically significant differences. The partial tests are combined, in a non parametric way (using a combined function as Fisher, Liptak or Tippett) in a second order test that globally verifies the existence of differences among the multivariate distributions. A procedure of conditioned resampling CMC (Pesarin, 2001) allows to estimate the p-values, associated both to partial tests and to second order tests.

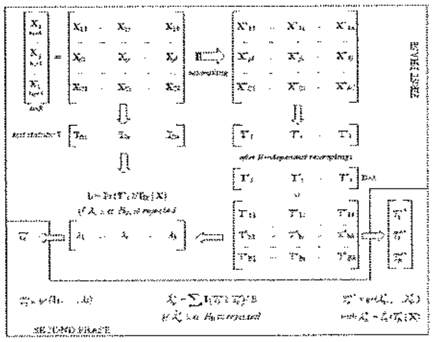

Figure 1: Two Phases Algorithm of NPC Procedure

2.2 Application

In this paper the NPC procedure was applied to a large amount of patients (1722) followed in the gastroenterology unit of the "G.Martino" University Hospital in Messina. The analysis was performed in order to assess the existence of significant differences between two gastrointestinal illnesses: Crohn's Disease (CD) and Ulcerative Colitis (UC). For each patient (in the respect of anonymity) we acquired information about: diagnosis age, gender, smoking habit, use of immunosuppressive therapy and its duration, treatment with biological drugs and its duration, hospitalization, adverse events, infections, cancers, diabetes, hypertension, heart failure, kidney failure, pulmonary failure, neuropathy, liver disease, Charlson Index (the most widely used index to predict the ten-year mortality for a patient who may have comorbid conditions; its score are 1, 2, 3 or 6, depending on the risk of dying), surgery, final exitus and follow-up time. The hypotheses system is the following:

$$H_0 : \left\{ diagnosis\_age_1 \overset{d}{=} .diagnosis\_age_2 \right\} \cap .... \cap \left\{ follow-up\_time_1 \overset{d}{=} follow-up\_time_2 \right\}$$

$$H_1 : \left\{ diagnosis\_age_1 \overset{d}{\neq} .diagnosis\_age_2 \right\} \cup .... \cup \left\{ follow-up\_time_1 \overset{d}{\neq} follow-up\_time_2 \right\}$$

where 1 and 2 are the two examined gastrointestinal illnesses (results in Table 1).

Table 1: NPC test for comparisons between CD and UC

VARIABLES	CD	UC	p-value
Diagnosis age	43.80±21.59	45.87±20.83	**0.028**
Gender (M% / F%)	49.3 / 50.7	57.7 / 42.3	**0.001**
Smoking habit (yes % / no %)	43.9 / 56.1	35.2 / 64.8	**0.000**
Immunosuppressive therapy (yes % / no %)	38.4 / 61.6	26.4 / 73.6	**0.000**
Duration of immunosuppressive therapy	12.27±8.67	14.92±9.84	**0.003**
Biological drugs (yes % / no %)	18.4 / 81.6	8.7 / 91.3	**0.000**
Duration of treat. with biological drugs	12.23±7.33	16.03±9.41	**0.005**
Hospitalization (yes % / no %)	45.6 / 54.4	27.8 / 72.2	**0.000**
Adverse events (yes % / no %)	17.9 / 82.1	9.2 / 90.8	**0.000**
Infections (yes % / no %)	7.0 / 93.0	7.6 / 92.4	0.627
Cancers (yes % / no %)	3.6 / 96.4	2.5 / 97.5	0.275
Diabetes (yes % / no%)	5.6 / 94.4	10.6 / 89.4	**0.000**
Hypertension (yes % / no %)	20.3 / 79.7	19.6 / 80.4	0.737
Heart failure (yes % / no %)	5.7 / 94.3	5.7 / 94.3	0.985
Kidney failure (yes % / no %)	2.1 / 97.9	2.6 / 97.4	0.507
Pulmonary failure (yes % / no %)	3.6 / 96.4	5.1 / 94.9	0.155
Neuropathy (yes % / no %)	1.7 / 98.3	1.6 / 98.4	0.885
Liver disease (yes % / no %)	1.6 / 98.4	1.3 / 98.7	0.607
Charlson (%) * score 0	87.6	93.6	**0.012**
scores 1 / 2 / 3	8.6 / 2.9/ 1	12.8 / 3.3 / 0.3	
Surgery (yes % / no %)	19.7 / 80.3	5.3 / 94.7	**0.000**
Final exitus (S /D)	97.9 / 2.1	97.6 / 2.4	0.664
Follow-up time	34.57±4.54	34.92±4.16	0.056
COMBINED p-value			***0.000***

* No patient has Charlson score equal to 6

3 Results and Final Remarks

In this paper we want to show as the permutation tests are very helpful for large-sized data analysis in many applicative contexts. In large data sets consisting of 1000 observations, performance of the permutation test appears equivalent to that of the asymptotic test (Potter, 2005); on the other hand, the NPC test, based on permutation solution, can be appropriately applied when the assumption for asymptotic tests are fulfilled (Ludbrook and Dudley, 1998). In addition, unlike the classical non parametric tests, the NPC method entails testing a global null hypothesis consisting of the intersection of $K > 1$ partial sub-hypotheses. In essence, the global null states that all of its constituent sub-hypotheses are true. The global alternative hypothesis is the union of K sub-alternatives (Pesarin and Salmaso, 2010). In this way NPC provides in multivariate context the combined p-value, by means of an adequate combining function. From the applicative point of view, we have great interest in evaluating this combined p-value because it provides a result that takes into account the contribution of all examined variables; on the other hand, no other non-parametric test provides the advantage of a combined p-value. This particular feature justifies our choice of the NPC test as methodically appropriate solution. In particular we applied permutation tests to perform comparison between a large number of patients affected by Crohn's Disease and Ulcerative Colitis. Both of these illness are inflammatory bowel diseases, involving more than 100,000 people in Italy; they often arise in young people, go on for a lifetime and manifest alterations of the intestinal canal, causing relationship and working problems. Examining the results achieved by applying NPC tests, we have to notice the high significance of the combined test, that provides guarantee affirming that patients with CD and UC significantly differ between them, in relation to the set of examined variables. Focusing our attention on partial tests, we can see that some variables significantly discriminate the two different subpopulations; in particular the UC patients, in comparison with the CD patients, have a higher diagnosis age, do not show a marked smoking status, the proportion of patients treated with immunosuppressants or with biological drugs is lower than the CD patients, even if the duration of such therapies is longer. CD patients have a higher rate of hospitalization; probably it is related to the significant greater occurrence of adverse events (rather than UC). Diabetes is more present in the sub-population of UC patients. Analyzing the Charlson score we can highlight that UC patients have a more severe clinical situation than CD patients. Finally, the CD patients are more frequently subjected to surgery compared to UC.

Until a few years ago the use of Big Data was not received particular attention from researchers. Today the conspicuous availability of large amounts of data and the need of their analysis required an adjustment of data processing methodologies, with careful attention to all the sources of variation in data. In this context, the non-parametric procedures, such as permutation tests, are widely applicable in virtue of the numerous optimal properties of which they are characterized.

Acknowledgment: We have to thank Professor Walter Fries (Director of Gastroenterology Unit of the "G.Martino" University Hospital in Messina) for providing of data and for his medical support into the realization of this paper.

References

1. Arboretti Giancristofaro R., Brombin C.: Overview of NonParametric Combination-based permutation tests for Multivariate multi-sample problems. Statistica (2014)
2. Arboretti Giancristofaro R., Marozzi M., Salmaso L.: Repeated measures designs: a permutation approach for testing for active effects, Far East Journal of Theoretical Statistics, Special Volume on Biostatistics, vol. 16, 2, pp.303-325 (2005)
3. Basso D., Chiarandini M., Salmaso L.: Synchronized permutation tests in *I*×*J* designs, Journal of Statistical Planning and Inference, 137, pp. 2564-2578 (2007)
4. Bonnini S., Corain L., Munaò F., Salmaso L.: Neurocognitive Effects in Welders Exposed to Aluminium: An Application of the NPC Test and NPC Ranking Methods, Statistical Methods and Applications, Journal of the Statistical Society, 15, 2, pp.191-208 (2006)
5. Bonnini S., Pesarin F., Salmaso L.: Statistical Analysis in biomedical studies: an application of NPC Test to a clinical trial on a respiratory drug, in Atti del Congresso Nazionale della Società Italiana di Biometria, pp.107-110. (2003)
6. Brombin C., Salmaso L.: Multi-aspect permutation tests in shape analysis with small sample size, Computational Statistics & Data Analysis (doi: 10.1016/j.csda.2009.05.010) (2009)
7. Callegaro A., Pesarin F., Salmaso L.: Test di permutazione per il confronto di curve di sopravvivenza, Statistica Applicata, 15, 2, pp.241-261 (2003)
8. Corain L., Salmaso L. (2004), Multivariate and Multistrata Nonparametric Tests: the NPC method, Journal of Modern Applied Statistical Methods, 3, 2, pp.443-461
9. Di Castelnuovo A., Mazzaro D., Pesarin F., Salmaso L.: Test di permutazione multidimensionali in problemi d'inferenza isotonica: un'applicazione alla genetica, Statistica, 60, 4, pp.691-700 (2000)
10. Finos L., Pesarin F., Salmaso L., Solari A.: Nonparametric iterated procedure for testing genetic differentiation, in Atti XLIII Riunione Scientica SIS, CLEUP, Padova (2004)
11. Finos L., Salmaso L.: Weighted methods controlling the multiplicity when the number of variables is much higher than the number of observations. Journal of Nonparametric Statistics, 18, 2, pp.245-261 (2006)
12. Klingenberg B., Solari A., Salmaso L., Pesarin F.: Testing marginal homogeneity against stochastic order in multivariate ordinal data. Biometrics (DOI: 10.1111/j.1541-0420.2008.01067.x), 65, pp.452 - 462 (2008)
13. Ludbrook J., Dudley H.: Why Permutation test are superior to t and F test in Biomedical Research. The American Statistician, 52 (2),127-132 (1998)
14. Opdyke J. D.: Bootstraps, Permutation Tests, and Sampling Orders of Magnitude Faster Using SAS, Computational Statistics-WIREs, Vol. 5, Issue 5, 391-405 (2013)
15. Pesarin F.: Multivariate Permutation Test, Wiley & Sons, Chichester, England (2001)
16. Pesarin F., Salmaso L.: Permutation Tests For Univariate And Multivariate Ordered Categorical Data, Austrian Journal of Statistics, 35, pp.315-324 (2006)
17. Peek N., Holmes J. H., Sun. J.: Technical Challenges for Big Data in Biomedicine and Health: Data Sources, Infrastructure, and Analytics. Yearb Med Inform.; 9(1): 42–47, doi: 10.15265/IY-2014-0018 (2014)
18. Pesarin F., Salmaso L.: Permutation Tests for Complex Data. Theory, Applications and Software. John Wiley & Sons (2010)
19. Potter DM., A permutation test for inference in logistic regression with small- and moderate-sized data sets. Stat Med. 15;24(5):693-708 (2005)
20. Rezzani A.: Big Data. Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati, Apogeo Education (2013)
21. Salmaso L.: Permutation tests in screening two-level factorial experiments, Advances and Applications in Statistics, 5, 1, pp.91-110 (2005)
22. Zirilli A, Alibrandi A.: A permutation approach to evaluate hyperhomocysteinemia in epileptic patients. In: Supplemento ai rendiconti del circolo matematico di Palermo. VII International Conference in "Stochastic Geometry, Convex Bodies, Empirical Measures and application to mechanics and Engineering train-transport", pp.369-378 (2009).
23. Zirilli A, Alibrandi A.: A permutation solution to compare two hepatocellular carcinoma markers. JP Journal of Biostatistics, 5, 2, pp.109-119 (2011).
24. Zirilli A, Alibrandi A.: The alteration of t, t-muconic acid and s-phenilmercapturic acid levels due to benzene exposure: an application of NPC test. JP Journal of Biostatistics, 7, 2, pp.91-104 (2012)