Nonparametric methods for stratified C-sample designs: a case study

Rosa Arboretti, Riccardo Ceccato, Luigi Salmaso

1. Introduction

The analysis of C-sample designs in the presence of stratification is a problem frequently faced by practitioners.

In the industrial field a variety of stratified analysis scenarios present themselves. Take, for example, a company that wishes to assess the performance of three different formulas for a new dishwasher detergent. Multiple dishwashers are used and multiple washes are carried out. At the end of each wash, an expert provides an evaluation of the cleaning performance of the formula. When analyzing the resulting data, the effect of using one dishwasher instead of another cannot be ignored, so each dishwasher is considered to be a separate stratum. Likewise, in the healthcare field it is quite common for multiple drugs to be tested on patients of different age groups. Each age group is again considered to be a stratum.

In this paper we focus on a scenario from the field of education. We are interested in assessing how the performance of students from different degree programs at the University of Padova changes, in terms of university credits and grades, when compared with their entrance exam results. In other words, we want to assess whether people who achieved the best results in this exam perform best during their academic career.

The entrance exam can have three possible outcomes (i.e. it is an ordinal variable). This is therefore a typical stochastic ordering problem (Basso et al., 2009; Basso and Salmaso, 2011; Bonnini et al., 2014), that is a problem in which the main interest lies in evaluating the null hypothesis $Y_1 \overset{d}{=} \ldots \overset{d}{=} Y_C$ against the alternative hypothesis $Y_1 \overset{d}{\leq} \ldots \overset{d}{\leq} Y_C$ and $\mathbb{E}[\psi(Y_1)] \leq \ldots \leq \mathbb{E}[\psi(Y_C)]$, where at least one inequality is strict, and $\psi(\cdot)$ is an increasing function (Pesarin and Salmaso, 2010). Our aim is in fact to assess whether by comparing increasing entrance exam outcomes, the $C = 3$ corresponding distributions of the student's performance measure Y are stochastically ordered.

A few nonparametric methods have been proposed in the literature to address these problems. Among them, Jonckheere's test (Jonckheere, 1954; Terpstra, 1952) is one of the first nonparametric solutions to test for ordered alternatives and is based on use of the Mann-Whitney test (Mann and Whitney, 1947) to perform all the possible $[C \times (C-1)]/2$ pairwise comparisons between C groups. Neuhäuser et al. (1998) also proposed a modification of this test that appears to be more powerful than the original test with small sample sizes (Shan et al., 2014). Additionally, permutation-based solutions involving the Non-Parametric Combination (NPC) technique (Pesarin and Salmaso, 2010; Klingenberg et al., 2009; Finos et al., 2007, 2008) were introduced.

We propose a further extension of the NPC technique to address stochastic ordering problems in the presence of stratification. Indeed, the impact of the student's choice of degree program cannot be ignored, therefore stratification must be considered in the testing procedure.

In section 2 we are going to describe the proposed permutation-based approach. In section 3 we apply it to the case study of interest related to university education. Finally, section 4

Rosa Arboretti, University of Padua, Italy, rosa.arboretti@unipd.it, 0000-0003-1263-0440
Riccardo Ceccato, University of Padua, Italy, riccardo.ceccato@unipd.it, 0000-0002-8629-8439
Luigi Salmaso, University of Padua, Italy, luigi.salmaso@unipd.it, 0000-0001-6501-1585

FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Rosa Arboretti, Riccardo Ceccato, Luigi Salmaso, *Nonparametric methods for stratified C-sample designs: a case study*, pp. 17-22, © 2021 Author(s), CC BY 4.0 International, DOI 10.36253/978-88-5518-304-8.05, in Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci, *ASA 2021 Statistics and Information Systems for Policy Evaluation. Book of short papers of the opening conference*, © 2021 Author(s), content CC BY 4.0 International, metadata CC0 1.0 Universal, published by Firenze University Press (www.fupress.com), ISSN 2704-5846 (online), ISBN 978-88-5518-304-8 (PDF), DOI 10.36253/978-88-5518-304-8

provides the results and conclusions.

2. Methodology

Firstly, let us further describe the stochastic ordering problem. The main interest lies in evaluating the system of hypotheses:

$$\begin{cases} H_0 : Y_1 \overset{d}{=} \ldots \overset{d}{=} Y_C \\ H_1 : Y_1 \overset{d}{\leq} \ldots \overset{d}{\leq} Y_C \text{ and at least one strict inequality } \overset{d}{<}, \end{cases}$$

where the symbol $\overset{d}{=}$ denotes equality in distribution and $\overset{d}{<}$ denotes stochastic dominance, i.e. $Y_1 \overset{d}{<} Y_2$ if and only if $F_1(z) \geq F_2(z), \forall z$ and $\exists I : F_1(z) > F_2(z), z \in I$ with $Pr(I) > 0$, where F_j is the cumulative distribution function. An alternative way to write this is:

$$\begin{cases} H_0 : F_1 = F_2 = \cdots = F_{(C-1)} = F_C \\ H_1 : F_1 \geq F_2 \geq \cdots \geq F_{(C-1)} \geq F_C \text{ and at least one strict inequality.} \end{cases} \tag{1}$$

NPC-based solutions generally consider a particular decomposition. The hypotheses are split in order to recreate the conditions of a set of two-sample problems as follows:

$$\begin{cases} H_0 : \bigcap_{i=1}^{C-1} H_{i0} = \bigcap_{i=1}^{C-1} [(F_1 = \cdots = F_{i)}) = (F_{(i+1)} = \cdots = F_C)] \\ H_1 : \bigcup_{i=1}^{C-1} H_{i1} = \bigcup_{i=1}^{C-1} [(F_1 = \cdots = F_i) > (F_{(i+1)} = \cdots = F_C)]. \end{cases}$$

where the null hypothesis H_0 is the intersection of a number of partial hypotheses and the alternative hypothesis H_1 is the union of $C-1$ sub-hypotheses.

For each pair of sub-hypotheses H_{i0} and H_{i1}, the first i and the last $(C-i)$ samples are pooled, so that two new samples X_1 and X_2 are achieved, with sizes N and M. The sub-problem can therefore be rewritten as:

$$\begin{cases} H_{i0} : X_1 \overset{d}{=} X_2 \\ H_{i1} : X_1 \overset{d}{<} X_2. \end{cases}$$

Each sub-hypothesis is then tested separately, using appropriate permutation tests. The adopted test statistic can differ according to the nature of the data, but a common and versatile choice is the modified version of the Anderson-Darling test statistic:

$$T = \sum_j^n [\hat{F}_1(X_j) - \hat{F}_2(X_j)] / \{\bar{F}(X_j)[1 - \bar{F}(X_j)]\}^{\frac{1}{2}} \tag{2}$$

where $X = \{X_1, X_2\}$ is the pooled sample, $\hat{F}_1(t) = \sum_j^N \mathbb{I}(X_{j1} \leq t)/N$, $\hat{F}_2(t) = \sum_j^M \mathbb{I}(X_{j2} \leq t)/M$, $\bar{F}(t) = \sum_j^n \mathbb{I}(X_j \leq t)/n$, $n = N + M$, $t \in \mathcal{R}^1$ and $\mathbb{I}(\cdot)$ is the indicator function which is 1 if $(\cdot)$ is satisfied and 0 otherwise.

According to the NPC algorithm (Pesarin and Salmaso, 2010), B permuted datasets are independently generated for each sub-problem and the related values of the test statistic $T_b^*, b = 1, \ldots, B$ are calculated to simulate the null distribution of T. Partial p-values (λ_i) and $\lambda_{ib}^*, b = 1, \ldots, B$ estimating their distributions can therefore be achieved. It is worth noting that the same permutation design is adopted for each sub-problem, to implicitly take into account the existing dependency among sub-problems.

A combination step now needs to be performed. The partial p-values $\lambda_i, i = 1, \dots, C-1$ related to the $C-1$ sub-problems $\{H_{i0} \text{ vs } H_{i1}\}$ are combined using an adequate combining function, such as Fisher's combining function $T''_F = -2 \cdot \sum_{i=1}^{C-1} \log(\lambda_i)$. The same is done for each of the B vectors $\lambda^*_{ib}, i = 1, \dots, C-1$. The elements of the new resulting vector represent the second-order test statistics, from which it is finally possible to achieve the global p-value λ'' to assess the system of hypotheses 1.

Given that stratification needs to be included, we propose firstly applying this procedure to each of the S strata, testing S systems of hypotheses:

$$\begin{cases} H_{0s} : F_{1s} = F_{2s} = \dots = F_{(C-1)s} = F_{Cs} \\ H_{1s} : F_{1s} \geq F_{2s} \geq \dots \geq F_{(C-1)s} \geq F_{Cs} \text{ and at least one strict inequality.} \end{cases} \quad (3)$$

After applying the aforementioned NPC-based approach to each stratum, the global p-values $\lambda_s{''}, \forall s = 1, \dots, S$ (and the $\lambda^*_{sb}{''}$ estimating their distributions) are thus retained. Then we adopt a further combination step, using the Fisher combining function, and retrieve a final p-value λ'''. In this way, by comparing λ''' to the desired significance level α, we are able to solve the global stochastic ordering problem H_0 vs H_1.

Given that multiple systems of hypotheses H_{s0} vs $H_{s1}, \forall s = 1, \dots, S$ are assessed, we then apply an appropriate multiplicity correction to control the false discovery rate (FDR). Our choice is the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

3. A case study

Let us now focus on the real stratified C-sample problem at hand. As mentioned before, we are interested in evaluating the performances of students from different degree programs at the University of Padova. In particular, we want to understand if the university credits gained at the end of the first year (Y^a), the credits gained at the end of the third year (Y^b) and the final average grade (Y^c) somehow depend on the results achieved by the student in the entrance exam. In other words, we try to indirectly assess the efficacy of this exam in evaluating and selecting future students. The analysis is performed using R (R Core Team, 2020).

Let us briefly describe the data. The total sample size is 3083 students. Firstly, the degree programs are grouped into 4 classes (identified by their Italian subject titles):

- ING_INFORMAZIONE_NON_PROFES (S1)
- ING_CIVILE_AMBIENTALE_L7 (S2)
- ING_INFORMAZIONE_L8 (S3)
- ING_INDUSTRIALE_L9 (S4).

The different classes represent different strata (i.e. $S = 4$) and have different sample sizes (see Figure 1). The variable reporting the outcome of the entrance exam has three modalities (i.e. $C = 3$), namely INSUFFICIENTE, SUFFICIENTE and PIU'_CHE_SUFFICIENTE (Insufficient, Sufficient and More Than Sufficient). For the sake of simplicity, we are going to refer to them as INS, SUF and PIU in our notation. In Figure 1, the possible outcomes are ordered from worst to best.

For each response variable $Y^j, \forall j \in \{a, b, c\}$, we want to assess if $Y^j_{\text{INS}} \overset{d}{\leq} Y^j_{\text{SUF}} \overset{d}{\leq} Y^j_{\text{PIU}}$, with at least one strict inequality, taking into account the effect of the degree program class.

Looking at credits gained at the end of the first year, a first descriptive analysis (see Figure 2) appears to support the alternative hypothesis. Indeed, in all strata, students achieving INS at the entrance exam appear to perform worse than students achieving SUF, and students achieving PIU at the entrance exam tend to perform better than students achieving SUF.

Similar conclusions can be drawn about both credits gained at the end of the third year (see Figure 3) and the average grade at the end of the academic career (see Figure 4).

Applying our testing procedure, we managed to confirm these hypotheses. We set $B = 10000$ and used the test statistic in Equation 2 and Fisher's combining function. When looking at Y^a (see Table 1), all the partial p-values and the global p-value proved to be substantially smaller than 1%. The only exceptions were ING_CIVILE_AMBIENTALE_L7 (S2) and ING_INFORMAZIONE_L8 (S3), for which the descriptive analysis shows that the order among entrance exam outcomes is less evident.

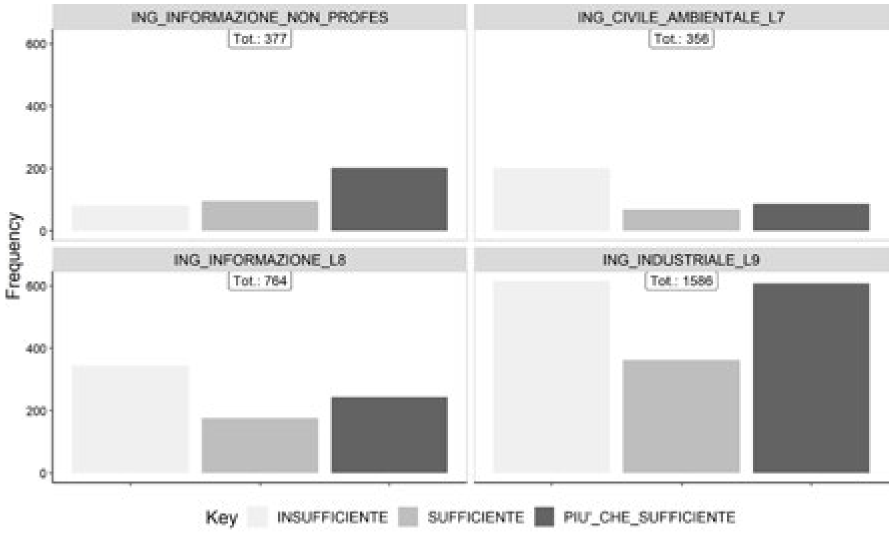

Figure 1: Description of the sample.

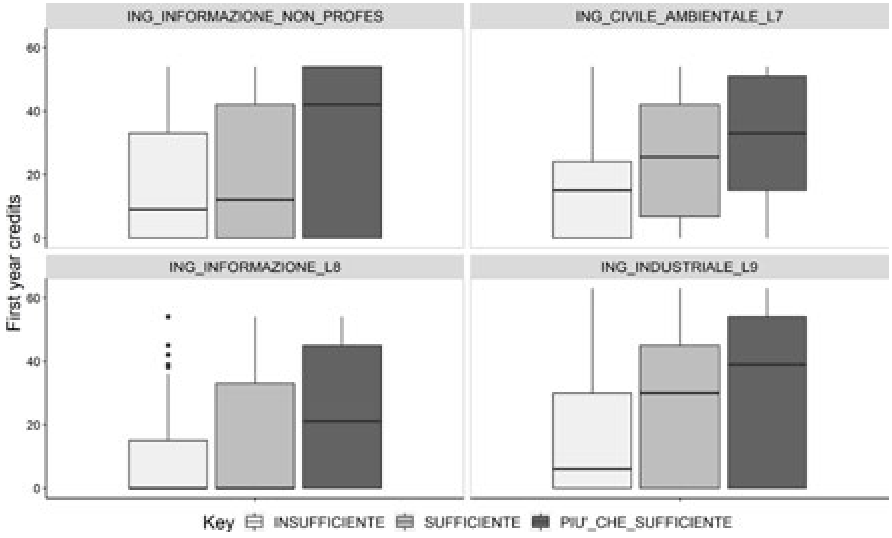

Figure 2: Credits at the end of the first year.

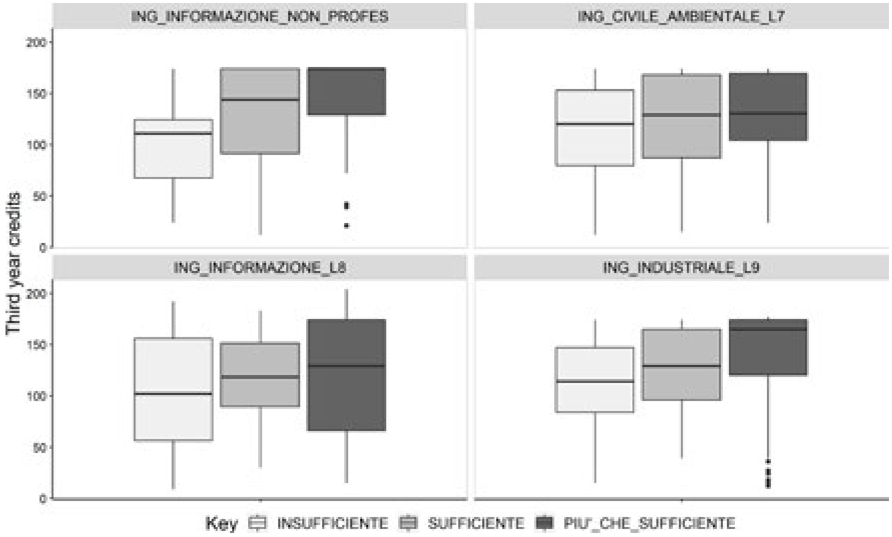

Figure 3: Credits at the end of the third year.

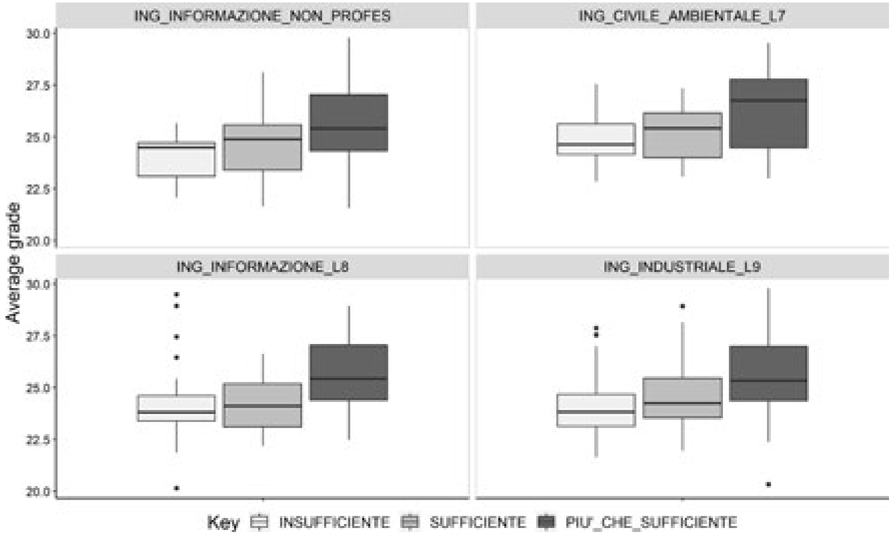

Figure 4: Average grade at the end of the academic career.

Table 1: Table of p-values for Y^a, Y^b and Y^c.

Response	Global	S1	S2	S3	S4
Y^a	1e-4	1e-4	1e-4	1e-4	1e-4
Y^b	1e-4	2e-4	0.1471	0.1185	2e-4
Y^c	1e-4	4e-4	2.9e-3	8e-4	6e-4

4. Conclusions

In this paper we presented a new solution to C-sample stochastic ordering problems in the presence of stratification, focusing on its application to a case study from the field of education.

Our proposal takes advantage of the Non-Parametric Combination (NPC) procedure (Pesarin and Salmaso, 2010), a versatile permutation-based methodology allowing us to solve several different complex problems, such as stochastic ordering. We apply this technique to evaluate the presence of stochastic ordering in each of the S existing strata and then use an appropriate combining function to assess the stochastic ordering in all the samples.

The application of this procedure allowed us to assess the efficacy of the University of Padova's entrance exams in evaluating and selecting future students. Indeed, it emerged that students with the worst results in the entrance exam tended to perform the worst during their academic career, in terms of both university credits achieved at the end of the first and third years and in terms of the final average grade, independently of the chosen degree program. The only exception was people from ING_CIVILE_AMBIENTALE_L7 and ING_INFORMAZIONE_L8. For these two strata, when the credits at the end of the third year were considered, it was not possible to find enough evidence in favor of the stochastic ordering hypothesis.

Overall, this approach appears to be significantly promising and a simulation study has been planned to further explore its performances.

References

Basso, D., Pesarin, F., Salmaso, L., Solari, A. (2009). *Permutation tests for stochastic ordering and ANOVA: theory and applications with R*. Springer Science & Business Media, New York, (NY).

Basso, D., Salmaso, L. (2011). A permutation test for umbrella alternatives. *Statistics and Computing*, **21**(1), pp. 45–54.

Benjamini, Y., Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), pp. 289–300.

Bonnini, S., Prodi, N., Salmaso, L., Visentin, C. (2014). Permutation approaches for stochastic ordering. *Communications in Statistics-Theory and Methods*, **43**(10-12), pp. 2227–2235.

Finos, L., Salmaso, L., Solari, A. (2007). Conditional inference under simultaneous stochastic ordering constraints. *Journal of statistical planning and inference*, **137**(8), pp. 2633–2641.

Finos, L., Pesarin, F., Salmaso, L., Solari, A. (2008). Exact inference for multivariate ordered alternatives. *Statistical Methods and Applications*, **17**(2), pp. 195–208.

Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives. *Biometrika*, **41**(1/2), pp. 133–145.

Klingenberg, B., Solari, A., Salmaso, L., Pesarin, F. (2009). Testing marginal homogeneity against stochastic order in multivariate ordinal data. *Biometrics*, **65**(2), pp. 452–462.

Mann, H. B., Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, **18**(1), pp. 50–60.

Neuhäuser, M., Liu, P.-Y., Hothorn, L. A. (1998). Nonparametric tests for trend: Jonckheere's test, a modification and a maximum test. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **40**(8), pp. 899–909.

Pesarin, F., Salmaso, L. (2010). *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons, Hoboken, (NJ).

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, (AT).

Shan, G., Young, D., Kang, L. (2014). A New Powerful Nonparametric Rank Test for Ordered Alternative Problem. *PloS one*, **9**(11), pp. 1–10.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, **14**(3), pp. 327–333.