Statistica Applicata - Italian Journal of Applied Statistics Vol. 35 (1)
doi.org/10.26398/IJAS.0035-002

MULTIVARIATE PERMUTATION MCNEMAR'S TEST WITH APPLICATION TO PERFORMANCE EVALUATION OF BASKET PLAYERS

Stefano Bonnini[1]
Department of Economics and Management, University of Ferrara, Ferrara, Italy

Livio Corain
Department of Management and Engineering, University of Padova, Padova, Italy

Fortunato Pesarin
Department of Statistical Sciences, University of Padova, Padova, Italy

Luigi Salmaso
Department of Management and Engineering, University of Padova, Padova, Italy

Abstract *The McNemar test can be considered the extension of the one-sample test on proportions to the case of two dependent samples or a special case of the sign test for paired data. In this paper we focus on the multivariate McNemar's test by considering an unusual but interesting application of Basket Analytics. The application is related to the evaluation of the effect of the field factor in the performance of basket players. The proposed method is based on the nonparametric combination of permutation tests.*

Keywords: *Basket Analytics, McNemar's Test, Multivariate Analysis, Permutation Test.*

1. INTRODUCTION

The McNemar test provides a nonparametric solution to a very popular problem. It can be considered the extension of the one-sample test on proportions to the case of two dependent samples or a special case of the sign test for paired data (McNemar, 1947). Medical applications are very widespread (Eliasziw and Donner, 1991; Gonen, 2004; Lachin, 1992). However the fields of application are numerous and very heterogeneous: computer science (Shao et al., 2021), marketing (Bonnini et al., 2014), genetics (Akazawa et al., 2021), engineering (Ibrahim et al., 2021), education (Stransky et al., 2021), behavioral ecology (Pembury Smith and Ruxton, 2020) and many others. McNemar's test is also suitable for comparing classification rates of multiple predictive models (Demsar, 2006; Durkalski et al., 2003; Leisenring et al., 2000; Lyles et al., 2005).

[1] Stefano Bonnini, email: stefano.bonnini@unife.it

Let us consider a binary response variable with paired observations. For example, let us take into account a sample of basket players and a dichotomous variable X representing the players' performance in a given season. X takes value 1 if the performance is good (or positive) and 0 if the performance is bad (or negative). We are interested in the distinction between *home* and *away* matches and the observed data can be represented by a 2×2 table whose rows correspond to good and bad performance in the *home* matches and the columns to good and bad performance in the *away* matches. The hypothesis that the performance of basket players is not affected by the so-called "field factor" is equivalent to the equality of the marginal probabilities of good performance in the *home* and *away* matches. We will see that, in order to test the significance of field factor's effect, we must compare the number of discordant paired observations. This is the typical goal of McNemar's test. Several versions and improvements of the test have been proposed over time to have powerful solutions suitable for the specific framework of the study, nature of the data and research objectives.

Methodological proposals have been published for the application of McNemar's test on clustered binary data. Some of these contributions are based on scalar adjustments of the test statistic as if the assumption of independence on two variables is satisfied and a further adjustment by a factor in order to keep the null distribution approximately correct (Donald and Donner, 1987, 1990; Donner, 1992). Others are focused on the ratio estimator (Obuchowski, 1998; Rao and Scott, 1992). Wu (2018) proposes a method for power calculation of the adjusted McNemar test with clustered data.

For multiple comparisons of dependent proportions Westfall et al. (2010) proposes a stepwise testing approach, by using discrete characteristics for exact McNemar's tests. This is a valid solution to several applications and is also suitable in case of missing values, tests with different sample sizes, and other non-standard or complex problems. In addition, to keep into account the dependence structure, an approximate bootstrap method is also proposed. These methods control the familywise error rate in the strong sense.

For the case of two independent samples of paired univariate dichotomous variables, we mention the contribution of Feuer and Kessler (1989). The case of binary crossover data was addressed by Becker and Balagtas (1993). Agresti and Klingenberg (2005) present solutions for the comparison of two independent multivariate binary vectors for an overall comparative evaluation of marginal incidence rates in two populations. A multivariate extension of the McNemar test is developed by Klingenberg and Agresti (2006), by discussing Wald and Score-

Type tests, Generalized Estimating Equations approach, Likelihood Ratio and Ordinary Score Test.

In this paper we focus on the multivariate McNemar test by considering an unusual but interesting application of Basket Analytics. This application concerns the evaluation of the effect of the field factor related to the performance of basket players. The proposed method is based on the nonparametric combination (NPC) of dependent permutation tests (Pesarin and Salmaso, 2010). The rest of the paper is organized as follows. In Section 2, we present the classic univariate version of the McNemar test. Section 3 is dedicated to introduce the application of Basket Analytics, concerning the performance evaluation of Basket players by comparing *home* and *away* performance. We will consider a review of the literature specialized on this topic in order to determine a suitable multivariate response that represents the performance of basket players. In Section 4 we describe the multivariate permutation McNemar test and we apply it to the problem of Basket Analytics. Conclusions are provided in Section 5.

2. MCNEMAR TEST FOR PAIRED DATA WITH BINARY RESPONSES

The McNemar problem is also called *test for marginal homogeneity*. The reason of this name will soon be clear according to the following description. Let us assume that the dataset consists of n independent observations of the bivariate response variable (X_{i1}, X_{i2}), the determinations of which are $\{(x_{i1}, x_{i2}), i = 1, \ldots, n\}$, where the two marginal responses can take only two categories, conventionally denoted by 0 and 1. For example, the couple (X_{i1}, X_{i2}) could represent the presence/absence of two characteristics on the i-th statistical unit. Another example concerns classifications according to a dichotomous scale by two evaluators on n objects, subjects or items. Marginal homogeneity is equivalent to equality of the marginal distributions of the bivariate response or the agreement between the two evaluators. Data are assumed to be determinations of a bivariate *Bernoulli* random variable. The joint probability distribution can be represented as in Table 1, where θ_{rs} denotes the probability of occurrence of the couple (r, s), with $r, s \in \{0, 1\}$. The hypotheses under testing are $H_0 : \theta_{\bullet 1} = \theta_{1 \bullet}$ and $H_1 : \theta_{\bullet 1} \neq \theta_{1 \bullet}$.

The joint frequency distribution can be represented by Table 2, where f_{rs} denotes the absolute frequency of the couple (r, s) in the observed sample, with $r, s \in \{0, 1\}$. Note that this table, being not related to independent samples, is not properly a contingency table; hence the typical techniques for contingency tables cannot be applied.

The more similar $f_{00} + f_{01}$ and $f_{00} + f_{10}$ are (i.e. difference between f_{01} and

Table 1: Probability distribution of the bivariate Bernoulli random variable.

		X_1		
		0	1	
X_2	0	θ_{00}	θ_{01}	$\theta_{0\bullet}$
	1	θ_{10}	θ_{11}	$\theta_{1\bullet}$
		$\theta_{\bullet 0}$	$\theta_{\bullet 1}$	1

f_{10} close to zero) the greater the empirical evidence in favor of the hypothesis of marginal homogeneity (null hypothesis) and vice-versa. Hence, a suitable test statistic for such problem might be based on $(f_{01} - f_{10})$. For small sample sizes the test statistic (conditional on the marginal frequencies) might equivalently be

$$T = f_{01}.$$

In fact, the sum $f_{01} + f_{10} = n - f_{00} - f_{11} = s$ is fixed and the test assesses disparity of the discordants f_{01} and f_{10}. Therefore $f_{01} - f_{10} = 2f_{01} - s$ and, consequently, there is an exact linear relationship between the two test statistics. Thus, they lead to the same p-values. When $f_{01} + f_{10} \leq 20$, approximate distributions are not required and not valid, and the exact distribution of one of the two equivalent test statistics can be used for the inferential purpose. Under marginal homogeneity, T follows a Binomial distribution with parameters $f_{01} + f_{10}$ and 0.5, that is $T \sim Bin\left(f_{01} + f_{10}, 0.5\right)$. The null hypothesis is rejected for either small or large values of T. When $f_{01} + f_{10} > 20$ then

$$T = \left(f_{01} - f_{10}\right)^2 / \left(f_{01} + f_{10}\right)$$

is typically used as a test statistic (Kvam and Vidakovic, 2007).

Under H_0 it approximately follows a χ^2 distribution with 1 degree of freedom. Some authors take into account the discontinuity correction:

$$T = \left(|f_{01} - f_{10}| - 1\right)^2 / \left(f_{01} + f_{10}\right).$$

Table 2: Absolute frequency distribution of a bivariate binary response variable

		X_1		
		0	1	
X_2	0	f_{00}	f_{01}	$f_{00}+f_{01}=f_{0\bullet}$
	1	f_{10}	f_{11}	$f_{10}+f_{11}=f_{1\bullet}$
		$f_{00}+f_{10}=f_{\bullet 0}$	$f_{01}+f_{11}=f_{\bullet 1}$	n

But, from the practical point of view, some experts think that, thanks to the computational capabilities of modern computers, this correction becomes not relevant (Kvam and Vidakovic, 2007). Simple changes to the decision rule must be considered for the one-sided problem. This test was proposed by McNemar (1947). Some variations were presented by Bennett and Underwood (1970); Mantel and Fleiss (1975); McKinlay (1975); Ury (1975).

The McNemar test can also be seen as the extension of the one-sample test on proportion to the case of two dependent samples. It can be also considered a special case of the sign test for paired data.

For example, let us consider the data about the performance of basket players in the *2016/2017* Italian Championship (regular season). A reasonable measure of individual performance in a match is the ratio between the number of scored points (PTS) and the actual played time in minutes ($TIME$): $PER = PTS/TIME$. In the *2016/2017* regular season of the Italian Championship, the general mean value of PER with respect to all the players and all the matches was 0.35. Hence, to determine whether the individual performance of a given player over the regular season has been good/positive ($X = 1$) or bad/negative ($X = 0$) we can consider the average value of the individual index and compare it with the general average 0.35. Formally

$$X_i = \begin{cases} 1 & \text{if } \overline{PER}_i \geq 0.35 \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{PER}_i$ denotes the average of the values of PER over the regular season for the i-th player.

A typical goal of the performance analysis of athletes playing round robin tournaments is whether the field factor affects their performance. In other words, the question is whether the probability of good performance in *home* matches is equal to the probability of good performance in *away* matches. Let random variables X_H and X_A represent the individual performance in the *home* matches and in the *away* matches respectively. Let $\theta_H = P(X_H = 1)$ and $\theta_A = P(X_A = 1)$. We want to test $H_0 : \theta_H = \theta_A$ versus $H_1 : \theta_H \neq \theta_A$.

In Basketball, the distintion between functions and roles of the 5 different players of a team is not very evident and the tasks are often interchangeable. Anyway, there are some reference roles:

1. *point guard* (playmaker), with the task of calling the game patterns and dictating the rhythms of the ball

2. *shooting guard*, with the tasks of supporting the point guard, with whom he shares most of the characteristics and is usually the best shooter of the team

3. *small forward*, usually tall, fast and agile, he is interchangeable with the shooting guard and the power forward; he is important for the particular offensive peculiarities as well as for the defensive phase, especially in rebounding

4. *power forward*, occupies the same areas as the small forward, but he has a more marked physicality, less suited to running; he is one of the tallest players and is inclined to make space between the opposing defenders in the area, ready to receive and reject impacts with the opponents

5. *center* (pivot), typically the tallest and slowest player, has most of the points in his hands (especially in shots near the rim of the basket) and, in the defensive phase, he is the main protector of his team's area

In the individual performance analysis of Basketball players the role is clearly a possible confounding factor and the distinction between roles must be considered, for instance through a suitable stratification. Since the distinction of the 5 roles presented above could be not suitable because the roles are not always so distinct and well defined, a more general classification, very common in U.S.A., can be considered:

- *backcourt* players during ball possession, take care of playing the ball in the back court; this category includes point guard and shooting guard

Table 3: Absolute frequency distribution of *2016/2017* Italian Basket regular season sample of players according to their (binary) performance as a function of the $\overline{PER}$ index.

	Performance in home matches	
Performance in away matches	Bad	Good
Bad	8	7
Good	1	8

- *frontcourt* players are responsible of scoring in the offensive half of the court; this category includes small forward, power forward and center.

Data about the *2016/2017* Italian Basket regular season were collected. A stratified random sample of 24 players (12 backcourt and 12 frontcourt) from all the individuals who played at least 10 *home* and 10 *away* matches, was selected. For each of these 24 players, the seasonal average performance $\overline{PER}$ in the *home* matches and in the *away* matches was computed in order to obtain the couples of binary data (x_{iH}, x_{iA}), where x_{iH} indicates whether the average performance of the i-th player in *home* matches was good or not and x_{iA} indicates whether the average performance of the i-th player in *away* matches was good or not. A synthesis of sample data, in the form of 2×2 table, is shown in Table 3.

In *R*, for the application of McNemar test, the command *mcnemar.test(x)* is to be used, where x represents the 2×2 table like Table 3 or the equivalent for other problems. If the significance level of the test is set at $\alpha = 0.10$, since the p-value of the test is 0.0703, then the null hypothesis of equal probability of performance in the *home* and *away* matches is rejected in favor of the hypothesis that the probability of good performance changes according to the field factor (the one-sided p-value for $\theta_H > \theta_A$ is 0.0352).

3. PERFORMANCE EVALUATION OF BASKET PLAYERS

The analysis of the individual performance of basketball players has been the subject of a vast scientific literature. Among the most recent contributions, we mention Page et al. (2007), Cooper et al. (2009), Piette et al. (2010), Fearnhead and Taylor (2011), Ozmen (2012) and Deshpande and Jensen (2016). Some works focused on the prediction of the match outcomes (Brown and Sokol, 2010; Gupta,

2015; Loeffeholz et al., 2009; Lopez and Matthews, 2015; Ruiz and Perez-Cruz, 2015; West, 2006; Yuan et al., 2015). An interesting work about players positions and effectiveness of the shots from different areas of basketball court is that of Shortridge et al. (2014). Zuccolotto and Manisera (2020) present an overview of methods, models and *R* packages for the analysis of Basketball data.
In the considered case study, related to the Italian Basketball Championship regular season *2016/2017*, we select a stratified random sample according to the latter role classification.

Typically, there are two approaches of performance analysis in Basketball Analytics: the *bottom-up* approach starts from the individual contributions of each athlete to predict the team's performance or the final result of a match; the *top-down* approach uses the overall contribution of the team to determine the individual contributions of players. Our contribution, although not specifically aimed at calculating the team's performance, is compatible with the *bottom-up* approach of which it could be a preliminary step. Since, the starting point and the raw data refer to the individual performance, let us consider some scientific contributions about performance measures of individual players.

The ratio between the number of scored points PTS and the played time in minutes $TIME$ mentioned in the previous section is a simple, reasonable but in many cases not adequate performance measure of a player in a match. Typical more sophisticated measures are:

- *Player Efficiency Rating (PER)*: it takes into account and weighs the number of 3-points shots, of 2-points shots and of free shots, the number of assists, the stolen balls, the blocks and other quantities. It is a reliable measure of performance only for the offensive phase and the reference values change season by season

- *Win Shares*: they measure the contribution of each single player to the team's overall victories, by distinguishing and summing the offensive and the defensive contribution. It is not suitable for small tournament such as the Italian Championship with a total of only 30 matches in the regular season.

- *Tendex*: proposed by the sports journalist Dave Heeren in 1959, the *Tendex Rating* is a measure of efficiency based on a weighted algebraic sum of partial indices such as PTS, number of rebounds, number of assists, number of stolen balls, number of blocks, turnovers, free throws made, field goals made and personal fouls. This index is used to determine the *Efficiency*

Rating used still today, especially in the United States, as an efficiency assessment index and based on the ratio between Tendex and number of played matches. It is very popular because it uses simple variables, usually included in the box-scores, and takes into account both offensive and defensive performance.

- *Performance Index Rating (PIR)*: it can be considered the European version of Tendex. In 1991 it appears, for the first time, in the Spanish ACB League. It is still used today to determine the Most Valuable Player (MVP) of the week in the Spanish National League and in the EuroLeague. It includes in the algebraic sum the same variables of Tendex with, in addition, Fouls Drawn and (with negative sign) Shots Rejected.

- *Offensive Efficiency Rating (OER)*: this is another very popular index defined by Dean Oliver as the number of points done by a player per 100 total possessions or simply the ratio between PTS and number of total possessions (PO).

The goal of this work is not to determine an optimal performance index but it is evident that each index has pros and cons and represents a partial aspect of a complex phenomenon. Consequently, the concept of *performance* of a basket player is multidimensional. In order to consider the multivariate nature of the response variable, we take into account the two most commonly used indices, PIR and OER, and we transform them with a logic similar to what we did with the PER index in order to compute a bivariate binary response variable representing the performance of a basket player in the regular season.

4. MULTIVARIATE EXTENSION OF MCNEMAR TEST: PERMUTATION SOLUTION

Let us consider the multivariate extension of the problem illustrated above. The dataset consists of multivariate paired data with q binary variables. The data are assumed to be determinations of the random variables (X_{1ih}, X_{2ih}) with $i = 1, \ldots, n$ and $h = 1, \ldots, q$. Let $\theta_{rs,h}$ denote the probability (or population proportion) of the couple (r, s) for the h-th response, with $r, s \in \{0, 1\}$ and $h = 1, \ldots, q$. The multivariate McNemar test can be defined as

$$H_0 : \bigcap_{h=1}^{q} \left[\theta_{01,h} = \theta_{10,h}\right],$$

against

$$H_1 : \bigcup_{h=1}^{q} \left[\theta_{01,h} <\neq> \theta_{10,h}\right],$$

where, in the overall alternative hypothesis, some of the partial hypotheses can be two-sided and some others one-sided. Each partial testing problem can be solved with the binomial test based on the test statistic $T_h = f_{01,h}$ which, when H_0 is true, follows a binomial distribution with parameters $f_{01,h} + f_{10,h}$ and 0.5, where $f_{rs,h}$ denotes the sample absolute frequency of the couple (r,s) for the h-th variable, with $r,s \in \{0,1\}$ and $h = 1,\ldots,q$.

Equivalently, we can consider the following data transformation

$$Y_{ih} = g\left(X_{1i,h}, X_{2i,h}\right) = \begin{cases} +1 & if \quad X_{1i,h} < X_{2i,h} \\ -1 & if \quad X_{1i,h} > X_{2i,h} \\ 0 & \quad otherwise, \end{cases}$$

and apply the permutation test for paired data based on the test statistic

$$T_h^* = \textstyle\sum_{i=1}^{n} Y_{ih} S_i^*$$

with $S_i^* = +1$ with probability 0.5 and -1 with probability 0.5 under H_0. The application of the NPC methodology for multivariate permutation tests provides a solution to this testing problem (Pesarin and Salmaso, 2010).

The procedure requires the examination of all 2^n possible permutations. In practice, when this number is large (2^{24} = 16 777 216), their complete examination may become unpractical. Thus, according to the literature (Pesarin, 2001; Pesarin and Salmaso, 2010), especially in the q-dimensional case, we suggest considering a random sample from the set of permutations consisting in carrying out B independent permutations. In other words, this is realized by a random generation of B sets of n-dimensional vectors of signs (note: the same permutation of signs jointly for all q variables). To emphasize that the B permutations are taken conditionally on the given dataset, this procedure is named "Conditional Monte Carlo" (CMC). Once the q partial tests are carried out, the related q partial significance level functions are to be combined by means of a suitable combining function through the NPC methodology (Pesarin, 2001). According to the null permutation distribution of the combined test statistic, the p-value can be computed and compared with the significance level α in order to take the final decision about either rejection or acceptance of the null hypothesis H_0. This method can be considered a particular case of the more general permutation test for multivariate paired observations. Suitable combining functions are:

- *Fisher* combining function: $T_F = -2\sum_h log(\lambda_h)$,
- *Liptak* combining function: $T_L = \sum_h \phi^{-1}(1-\lambda_h)$, ϕ^{-1} being the standard normal quantile function,
- *Tippett* combining function: $T_T = \max_h(1-\lambda_h)$,

where λ_h is the partial p-value.

The CMC procedure works as follows:

1. Compute the vector of observed values of the q partial test statistics as a function of the observed dataset $\mathbf{X}$: $\mathbf{T}_{obs} = [T_1(\mathbf{X}),\dots,T_q(\mathbf{X})]' = [T_{1(0)},\dots,T_{q(0)}]'$
2. Consider B random permutations and compute the values of the test statistics corresponding to each permuted dataset. For the b-th permuted dataset $\mathbf{X}^*_{(\mathbf{b})}$ (with $b=1,\dots,B$), the test statistics are: $\mathbf{T}^*_{\mathbf{b}} = [T_1(\mathbf{X}^*_{(\mathbf{b})}),\dots,T_q(\mathbf{X}^*_{(\mathbf{b})})]' = [T^*_{1(b)},\dots,T^*_{q(b)}]'$
3. Estimate the p-values according to the null permutation distribution: $\hat{\lambda}_h = \hat{L}_h(T_{h(0)})$, $\hat{\lambda}^*_{h(b)} = \hat{L}_h(T^*_{h(b)})$, with $\hat{L}_h(t) = [\sum_{r=1}^{B} I(T^*_{h(r)} \geq t) + 0.5]/(B+1)$ and $I(A)$ being the indicator function of the event A
4. Compute the observed value and the permutation values of the combined test statistic based on the combining function ψ, $T_\psi = \psi(\lambda_1,\dots,\lambda_q)$: $T_{\psi,obs} = \psi(\hat{\lambda}_1,\dots,\hat{\lambda}_q)$ and $T^*_{\psi(b)} = \psi(\hat{\lambda}^*_{1(b)},\dots,\hat{\lambda}^*_{q(b)})$
5. Estimate the p-value of the combined test according to the null permutation distribution: $\hat{\lambda}_\psi = \hat{L}_\psi(T_{\psi,obs})$

Since all partial tests are marginally unbiased, the combined test is unbiased. In other words, the probability of rejecting the null hypothesis in favor of the alternative, when the latter is true in at least one of q components, is greater than the significance level α (Pesarin and Salmaso, 2010). Even if each partial test is distributed according to the binomial law, the multivariate (global) test is not multinomial. Moreover, when $q > 2$, the asymptotic approximation of the multivariate distribution cannot be considered, because the dependence relations among component binomials cannot be restricted to the $q(q-1)/2$ pair-wise correlations coefficients (Joe, 1997; Pesarin, 2001). Indeed, also dependence three-wise, four-wise, etc. should be considered. Thus the described NPC by the CMC procedure based on B iterations is a suitable solution.

Let us consider again the example of the Italian Championship regular season *2016/2017*. The bivariate response variable is based on the indices, *PIR* and *OER*, transformed by a rationale similar to what we did with the *PER* index. In the *2016/2017* regular season of the Italian Championship, the mean value of *PIR* with respect to all the players and all the matches was 8.5. Hence, in order to determine whether the individual performance of a given player over the regular season has been good/positive ($X_1 = 1$) or bad/negative ($X_1 = 0$) with respect to *PIR*, we can consider the average value of the individual index and compare it with the general average 8.5. Formally

$$X_{1i} = \begin{cases} 1 & \text{if } \overline{PIR}_i \geq 8.5 \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{PIR}_i$ denotes the average of the values of *PIR* over the regular season for the i-th player. Similarly

$$X_{2i} = \begin{cases} 1 & \text{if } \overline{OER}_i \geq 0.84 \\ 0 & \text{otherwise,} \end{cases}$$

where $\overline{OER}_i$ denotes the average of the values of *OER* over the regular season for the i-th player and 0.84 is the average mean over all players.

Let us consider a random sample of 24 players, stratified with respect to role (12 *backcourt* and 12 *frontcourt*). Data are shown in Table 4. We want to test if the proportion of good performances in the *home* matches is different from the proportion of good performances in the *away* matches for at least one of the two response variables.

The significance level of the test is set at $\alpha = 0.10$. The application of the combined permutation test, using $B = 10\,000$ CMC runs with *Fisher*, *Liptak* and *Tippett* combining function provides the p-values 0.003, 0.004 and 0.026 respectively (see Figure 1). Hence, according to all three combined permutation tests, the null hypothesis of equal performance in the *away* and *home* matches is rejected in favor of the alternative hypothesis that the performance depends on the field factor. Note that the partial p-values of the univariate tests of the two components of the bivariate response (*OER*-based and *PIR*-based performance) are 0.021 and 0.044 respectively, as shown in Figure 1. To attribute the significance of the overall test to one of the two partial tests or to both of them, the p-values of the two partial tests must be adjusted. This is necessary to avoid the probability of type I error in the overall test exceeding the nominal significance level α. The p-values of the two partial tests, adjusted with the well-known Bonferroni-Holm

Table 4: Sample data about PIR-based and OER-based performance of players of the Italian Championship in the *2016/2017* regular season, in the *away* and *home* matches.

Player		$X_1(PIR)$		$X_2(OER)$	
Name	*Role*	*Away*	*Home*	*Away*	*Home*
Alibegovic	backcourt	0	0	0	1
Bushati	backcourt	0	0	0	0
Cournooh	backcourt	0	1	0	1
Dowdell	backcourt	1	1	0	1
Forray	backcourt	0	0	0	0
Harvey	backcourt	1	0	0	1
Mian	backcourt	0	0	0	1
Obasohan	backcourt	0	0	0	0
Randolph	backcourt	0	1	0	1
Spanghero	backcourt	0	0	0	1
Vitali	backcourt	0	1	0	1
Tonut	backcourt	1	1	1	1
Abass	frontcourt	0	1	1	1
Cusin	frontcourt	0	1	0	1
Fesenko	frontcourt	1	1	1	0
Iannuzzi	frontcourt	0	1	0	1
Kangur	frontcourt	0	0	0	1
Mazzola	frontcourt	0	0	1	1
Pascolo	frontcourt	0	1	1	1
Sacchetti	frontcourt	1	1	1	1
Thomas A.	frontcourt	0	1	0	0
Watt	frontcourt	1	1	1	1
Wojciechowski	frontcourt	0	0	1	1
Viggiano	frontcourt	0	0	1	0

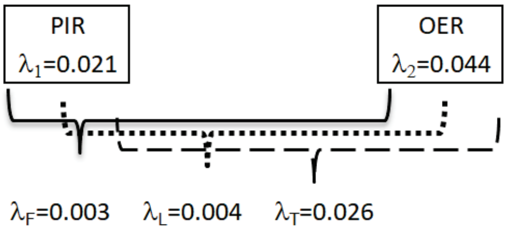

Figure 1: *P*-values of the combined permutation McNemar tests with Fisher, Liptak and Tippett combination for the two-tailed alternative hypothesis

method, are both significant (0.042 and 0.044 respectively). Hence, the performance of the players in the *home* matches is not equal to their performance in the *away* matches. This conclusion concerns both the *Performance Index Rating* and the *Offensive Efficiency Rating*.

It is worth noting that the method can also be applied to directional tests, i.e. with one-tailed alternatives. For example, the alternative hypothesis could be $H_1 : [P(X_{1H} = 1) > P(X_{1A} = 1)] \cup [P(X_{2H} = 1) > P(X_{2A} = 1)]$, where $(X_{1H} = 1)$ and $(X_{1A} = 1)$ mean that the seasonal performance according to *OER* in the *home* and *away* matches respectively is good and $(X_{2H} = 1)$ and $(X_{2A} = 1)$ have a similar meaning for *PIR*. In fact, it is reasonable to think that the performance at home is better than the performance away according to both partial indices. In other words, the probability of good performance at *home* is higher than the probability of good performance *away*. This multivariate test with restricted alternatives (one-tailed alternative hypotheses) admits a difficult asymptotic solution also for $q = 2$, where the normal approximation for the two marginal distributions would be assured but with an unknown approximation rate for finite n, such as $n = 24$. Therefore, the application of a parametric approach based on the assumption of (approximately) normal underlying distribution is not suitable because this assumption is not plausible with these sample sizes. Hence, in these conditions, the proposed solution is appropriate and valid because distribution-free and robust with respect to the departure from normality. For the one-tailed test with $B = 10\,000$, we obtained the partial p-values $\hat{\lambda}_{PIR} = 0.0201$ and $\hat{\lambda}_{OER} = 0.0112$, and the Liptak combined $\hat{\lambda}_{T_L} = 0.0013$. Hence, we have empirical evidence that the performance at home is better (home-field effect) and this is true for both the performance measures considered in the study.

5. CONCLUSIONS

A solution to a multivariate version of the well-known McNemar test, has been proposed. The method is based on the NPC of dependent permutation tests. The case study relates to Basket Analytics. Specifically, the goal is to evaluate the performance of Basket players of the Italian Championship (*2016/2017* regular season) in order to test the so-called field effect. In other words, the goal is to test whether, according to a given list of response variables, the proportion of good performant players in the *away* matches is equal to the proportion of good performant players in the *home* matches or not.

The proposed non parametric test is flexible, robust, unbiased and consistent with respect to departure from assumptions in at least one component of the multivariate distribution of the response. It is particularly interesting to underline that the NPC procedure does not require any specific assumption about the dependence structure of the dichotomous components of the multivariate response. Indeed, the dependence structure is implicitly considered without the need of modelling or estimating any unknown population nuisance parameters (Pesarin and Salmaso, 2010).

References

Agresti, A. and Klingenberg, B. (2005). Multivariate tests comparing binomial probabilities, with application to safety studies for drugs. In *Applied Statistics*. 54:691-816.

Akazawa, K., Kagara, N., Sota, Y., Motooka, D., Nakamura, S., Miyake, T., and Shimazu, K. (2021). Comparison of the multigene panel test and oncoscanâ¢ for the determination of her2 amplification in breast cancer. In *Oncology Reports*. 46(4):1-8.

Becker, M. and Balagtas, C. (1993). Marginal modeling of binary cross-over data. In *Biometrics*. 49:997-1009.

Bennett, B. and Underwood, R. (1970). On mcnemar's test for the 2×2 table and its power function. In *Biometrics*. 26:339-343.

Bonnini, S., Corain, L., Marozzi, M., and Salmaso, L. (2014). *Nonparametric Hypothesis Testing: Rank and Permutation methods with Applications in R*. Wiley, Chichester.

Brown, M. and Sokol, J. (2010). An improved lrmc method for ncaa basketball prediction. In *Journal of Quantitative Analysis in Sports*. 6(3):1-23.

Cooper, W., Ruiz, J., and Sirvent, I. (2009). Selecting non-zero weights to evaluate effectiveness of basketball players with dea. In *European Journal of Operational Research*. 195:563-574.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple datasets. In *Journal of Machine Learning Research*. 7:1-30.

Deshpande, S. and Jensen, S. (2016). Estimating an nba player's impact on his team's chances of winning. In *Journal of Quantitative Analysis in Sports*. 12:51-72.

Donald, A. and Donner, A. (1987). Adjustments to the mantel-haenszel chi-square statistic and odds ratio variance estimator when the data are clustered. In *Statistics in Medicine*. 6:491-499.

Donald, A. and Donner, A. (1990). A simulation study of the analysis of sets of 2x2 contingency tables under cluster sampling: estimation of a common odds odds ratio. In *Journal of the American Statistical Association*. 85:537-543.

Donner, A. (1992). Sample size requirements for stratied cluster randomization designs. In *Statistics in Medicine*. 11:743-750.

Durkalski, V., Palesch, Y., Lipsitz, S., Philip, F., and Rust, P. (2003). The analysis of clustered matched-pair data. In *Statistics in Medicine*. 22:2417-2428.

Eliasziw, M. and Donner, A. (1991). Application of the mcnemar test to non-independent matched pair data. In *Statistics in Medicine*. 10(12): 1981-1991.

Fearnhead, P. and Taylor, B. (2011). On estimating the ability of nba players. In *Journal of Quantitative Analysis in Sports*. 7(3):11.

Feuer, E. and Kessler, L. (1989). Test statistic and sample size for a two-sample mcnemar test. In *Biometrics*. 45:629-636.

Gonen, M. (2004). Sample size and power for mcnemar's test with clustered data. In *Statistics in Medicine*. 23(14):2283-2294.

Gupta, A. (2015). A new approach to bracket prediction in the ncaa men's basketball tournament based on a dual-proportion likelihood. In *Journal of Quantitative Analysis in Sports*. 11:53-67.

Ibrahim, A., Kashef, R., and Corrigan, L. (2021). Predicting market movement direction for bitcoin: A comparison of time series modeling methods. In *Computers Electrical Engineering*. 89:106905.

Joe, H. (1997). *Multivariate Methods and Dependence Concepts*. Chapman Hall, London.

Klingenberg, B. and Agresti, A. (2006). Multivariate extensions of mcnemar's test. In *Biometrics*. 62:921-928.

Kvam, P. and Vidakovic, B. (2007). *Nonparametric Statistics with Applications to Science and Engineering*. Wiley, Hoboken, New Jersey.

Lachin, J. (1992). Power and sample size evaluation for the mcnemar test with application to matched case-control studies. In *Statistics in Medicine*. 11(9):1239-1251.

Leisenring, W., Alonzo, T., and M.S., P. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs. In *Biometrics*. 56:345-351.

Loeffeholz, B., Bednar, E., and Bauer, K. (2009). Predicting mba games using neural networks. In *Journal of Quantitative Analysis in Sports*. 5(1):1-17.

Lopez, M. and Matthews, G. (2015). Building and ncaa men's basketball predictive model and quantifying its success. In *Journal of Quantitative Analysis in Sports*. 11(1):5-12.

Lyles, R., Williamson, J., Lin, H., and Heilig, C. (2005). Extending mcnemar's test: Estimation and inference when paired binary outcome data are misclassified. In *Biometrics*. 61:287-294.

Mantel, N. and Fleiss, J. (1975). The equivalence of the generalized mcnemar tests for marginal homogeneity in 2^3 and 3^2 tables. In *Biometrics*. 31:731-735.

McKinlay, S. (1975). A note on the chi-square test for pair-matched samples. In *Biometrics*. 31:731-735.

McNemar, Q. (1947). A note on the sampling error of the difference between correlated proiportions and percentages. In *Psychometrika*. 12: 153-157.

Obuchowski, N. (1998). On the comparison of correlated proportions for clustered data. In *Statistics in Medicine*. 17:1495-1507.

Ozmen, U. (2012). Foreign player quota. experience and efficiency of basketball players. In *Journal of Quantitative Analysis in Sports*. 8:1-18.

Page, G., Fellingham, G., and Reese, C. (2007). Using box-scores to determine a position's contribution to winning basketball games. In *Journal of Quantitative Analysis in Sports*. 3(4):1-16.

Pembury Smith, M. and Ruxton, G. (2020). Effective use of the mcnemar test. In *Behavioral Ecology and Sociobiology*. 74, 133.

Pesarin, F. (2001). *Multivariate Permutation tests with Applications in Biostatistics*. Wiley, Chichester.

Pesarin, F. and Salmaso, L. (2010). *Permutation Tests for Complex Data: Applications and Software*. Wiley, Chichester.

Piette, J., Anand, S., and Zhang, K. (2010). Scoring and shooting abilities of nba players. In *Journal of Quantitative Analysis in Sports*. 6(1):1-24.

Rao, J. and Scott, A. (1992). A simple method for the analysis of clustered binary data. In *Biometrics*. 48(2):577-585.

Ruiz, F. and Perez-Cruz, F. (2015). A generative model for predicting outomes in college basketball. In *Journal of Quantitative Analysis in Sports*. 11(1):39-52.

Shao, E., Liu, C., Wang, L., Song, D., Guo, L., Yao, X., and Hu, Y. (2021). Artificial intelligence-based detection of epimacular membrane from color fundus photographs. In *Scientific reports*. 11(1):1-10.

Shortridge, A., Goldsberry, K., and Adams, M. (2014). Creating space to shoot: quantifying spatial relative field goal efficiency in basketball. In *Journal of Quantitative Analysis in Sports*. 10:1-11.

Stransky, J., Bassett, L., Bodnar, C., Anastasio, D., Burkey, D., and Cooper, M. (2021). A retrospective analysis on the impacts of an immersive digital environment on chemical engineering studentsâ moral reasoning. In *Education for Chemical Engineers*. 35:22-28.

Ury, H. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. In *Biometrics*. 3:643-650.

West, B. (2006). A simple and flexible rating method for predicting success in the ncaa tournament outcomes. In *Journal of Quantitative Analysis in Sports*. 2(3):1-16.

Westfall, P., Troendle, J., and Pennello, G. (2010). Multiple mcnemar tests. In *Biometrics*. 66(4):1185-1191.

Wu, Y. (2018). Power calculation of adjusted mcnemar's test based on clustered data of varying cluster size. In *Biometrical Journal*. 60(6):1190-1200.

Yuan, L., Liu, A., Yeh, A., and Kaufman, A. (2015). A mixture-of-modelers approach to forecasting ncaa tournament outcomes. In *Journal of Quantitative Analysis in Sports*. 11(1):13-27.

Zuccolotto, P. and Manisera, M. (2020). *Basketball Data Science: with applications in R*. Chapman Hall/CRC, -.