The Holy Grail and the Bad Sampling - A test for the homogeneity of missing proportions for evaluating the agreement between peer review and bibliometrics in the Italian research assessment exercises*

Alberto Baccini† Lucio Barabesi‡ Giuseppe De Nicolao§

Abstract

Two experiments for evaluating the agreement between bibliometrics and informed peer review - depending on two large samples of journal articles - were performed by the Italian governmental agency for research evaluation. They were presented as successful and as warranting the combined use of peer review and bibliometrics in research assessment exercises. However, the results of both experiments were supposed to be based on a stratified random sampling of articles with a proportional allocation, even if solely subsets of the original samples in the strata were selected owing to the presence of missing articles. Such a kind of selection has the potential to introduce biases in the results of the experiments, since different proportions of articles could be missed in different strata. In order to assess the "representativeness" of the sampling, we develop a novel statistical test for assessing the homogeneity of missing proportions between strata and we consider its application to data of both experiments. Outcome of the testing procedure show that the null hypotesis of missing proportion homogeneity should be rejected for both experiments. As a consequence, the obtained samples cannot be considered as "representative" of the population of articles submitted to the research assessments. It is therefore impossible to exclude that the combined use of peer review and bibliometrics might have introduced uncontrollable major biases in the final results of the Italian research assessment exercises. Moreover, the two experiments should not be considered as valid pieces of knowledge to be used in the ongoing search of the Holy Grail of a definite agreement between peer review and bibliometrics.

KEYWORDS: PEER REVIEW; BIBLIOMETRICS; RESEARCH ASSESSMENT EXERCISE; STRATIFIED RANDOM SAMPLING; HYPOTHESIS TESTING FOR THE HOMOGENEITY OF MISSING PROPORTIONS

*Funding: This work was supported by Institute For New Economic Thinking Grant ID INO17-00015.
†Department of Economics and Statistics, University of Siena, Italy; alberto.baccini@unisi.it
‡Department of Economics and Statistics, University of Siena, Italy
§Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Italy

1 Introduction

The definite proof that peer review substantially agrees with some kind of bibliometric indicators is the Holy Grail for research assessment designers (RADs), since simpler, cheaper and more "objective" bibliometric indicators could wholly replace the peer review (Pride and Peter, 2018). Many RADs and scholars would welcome that definite proof since finally objective algorithms could substitute unreliable human peer reviewers, prone to nepotism and opportunism. In this perspective, the Holy Grail is an evaluation without human evaluators, ideally tending to a "view from nowhere" (Goukrager, 2012). It is usual to read about the contraposition of "objective bibliometric data" and "subjective peer reviews". It is not surprising at all that many descriptions of the aims of national research assessments emphasize that point. For example, the second Polish research assessment exercise is "based on a parametric assessment to make the evaluation more objective and independent from its peer" (Kulczycki et al., 2017). Again, in a final report of the research assessment in Italy, some members of the panel claim "the need of strongly reducing the peer evaluation, since it introduces a subjectivity representing a bias that cannot be normalized" (AREA7 Rapporto finale, p.113). Others RADs would welcome the definite proof for a less radical and more practical reason: if peer review and bibliometrics agree, it is possible to combine them in a universal research assessment where some disciplines - notably the humanities - cannot simply be evaluated by indicators.

Unfortunately, the search of the Holy Grail of research assessment has not been fruitful so far. The most extensive research campaign to date has produced negative results, even suggesting that the Holy Grail of evaluation does not exist: "This work - according to the authors - has shown that individual metrics give significantly different outcomes from the REF peer review process, showing that metrics cannot provide a like-for-like replacement for REF peer review" (Wilsdon et al., 2015).

In Italy, the governmental agency for research evaluation (ANVUR) conducted two research campaigns in search of the Holy Grail. During the two last national research assessments, VQR1 for the years 2004-2010 and VQR 2 for 2011-2014, ANVUR realized two extensive experiments (hereafter EXP1 and EXP2) by comparing evaluations reached by bibliometrics and by informed peer review (IPR) based on large samples of journal articles. Apparently, results of EXP1 and EXP2 were promising for the search of the Holy Grail. Results of EXP1 were published not only as official reports, but also disseminated as working papers or scientific papers in refereed journals by scholars working for the agency. Ancaiani et al. (2015) claimed that the results of EXP1 support "the choice of using both techniques in order to assess the quality of Italian research institutions". Bertocchi et al. published five identical working papers where they interpreted the results of EXP1 as claiming that peer review and bibliometrics "are close substitutes" (for all Bertocchi et al. 2013). In the peer reviewed version of the papers they concluded that "the agencies that run these evaluations could feel confident about using bibliometric evaluations and interpret the results as highly correlated with what they would obtain if they performed informed peer review" (Bertocchi et al., 2015). The results of EXP2 were also presented as a success in the official report: since a "not-zero correlation" was found (ANVUR, 2017), "we can hence conclude that the combined used of bibliometric indicators for citations and journal impact may provide a useful proxy to assess articles quality" (Alfò et al., 2017).

Two of the authors of the present paper analyzed EXP1. Despite they were unable

to wholly replicate its results, since ANVUR did not disclose anonymized raw data, they documented many flaws (Baccini and De Nicolao, 2016a, 2016b, 2017a, 2017b, Benedetto et al., 2017). In particular, from the perspective of the sampling design, a stratified sampling of articles with proportional allocation was assumed by ANVUR experts. However, the results of EXP1 were not computed on the whole sample of articles, but solely on a subset of the sample owing to the presence of missing articles. This selection has the potential to introduce biases in the results of the experiment, since different proportions of articles could be missed in different strata. Since the design of EXP1 was repeated also for EXP2, the second experiment might also suffer from the same problem.

The present paper is targeted to evaluate the "representativeness" of the samples adopted in EXP1 and EXP2. Hence, on the basis of the previous discussion, the focus is assessing the homogeneity of missing proportions in the strata. To this aim, after developing a novel statistical test, we apply our methodological findings to data of EXP1 and EXP2. Thus, the article is organized as follows: in Section 2 we present the structure of EXP1 and EXP2 by briefly reminding the essential issues of the Italian research assessment exercises. In Section 3 the theoretical development of the test for assessing missing proportion homogeneity is presented. Section 4 reports the results of the proposed testing procedure. Section 5 briefly concludes the paper by discussing if EXP1 and EXP2 can be considered as a valid piece of evidence in favor of the agreement between peer review and bibliometrics.

2 A brief description of the Italian experiments

A brief contextualization of EXP1 and EXP2 is preliminarily needed for understanding their relevance and role in the two Italian research assessments (this description is largely based on Baccini and De Nicolao, 2016a). The aim of both VQR1 and VQR2 were to evaluate research institutions such as universities or departments, and research areas and fields both at national or institutional levels. Each university, department and research field was classified by calculating the average score obtained by the research outputs submitted by researchers. To this end, all the researchers with a permanent position in an university had to submit a fixed number (with few exceptions) of research outputs (3 in VQR1 and 2 in VQR2). Each research work was then evaluated as Excellent, Good, Acceptable, Limited in VQR1 and Excellent, Elevated, Fair, Acceptable, Limited in VQR2, and received a score (scores slightly changed between the two exercises).

Both VQR1 and VQR2 were organized in 16 widely defined research areas. The 16 areas were: Mathematics and Informatics (Area 1), Physics (Area 2), Chemistry (Area 3), Earth Sciences (Area 4), Biology (Area 5), Medicine (Area 6), Agricultural and Veterinary Sciences (Area 7), Civil Engineering and Architecture (Areas 8a and 8b), Industrial and Information Engineering (Area 9), Antiquities, Philology, Literary studies, Art History (Area 10), History, Philosophy, Pedagogy and Psychology (Areas 11a and 11b), Law (Area 12), Economics and Statistics (Area 13), Political and Social Sciences (Area 14). These areas originates from the traditional classification of research areas adopted in Italy. For each area, an evaluation panel was established with a number of panelists proportional to the number of research outputs to be evaluated. Each panel was organized in sub-panels, specialized for specific research fields, so a total of 44 sub-panels were defined in both VQR1 and VQR2.

Panels directly managed and evaluated subsets of research products submitted for evaluation in their area of expertise. In both research assessments, research evaluation was

analogously realized. Panels for the so-called "Bibliometric Areas", i.e. hard sciences, engineering and life sciences, evaluated papers mainly but not exclusively, through bibliometrics. The bibliometric algorithm changed between VQR1 and VQR2, but in both assessments it was based on the number of citations received by an article and on a journal indicator, e.g. the impact factor, of the journal in which it was published. In the case that the two indicators gave coherent indications, the algorithm generated a score for the article. Otherwise, if they disagreed (high number of citations and low impact factor or viceversa), the algorithm output was unable to attribute a defined score to the article and it was therefore classified as "IR" and evaluated by informed peer review.

Panels of the so called "non-bibliometric areas", i.e. Social Science and Humanities, excluding Economics and Statistics, evaluated submitted research products exclusively by IPR. Area 13 (Economics and Statistics) was an exception since the Area 13 panel developed a journal ranking where journals were classified as Excellent, Good, Acceptable, Limited (VQR1) or Excellent, Elevated, Fair, Acceptable, Limited (VQR2). All the articles published in one of the listed journals then received the score of the journal in which they were published. All other research outputs (books, chapters, articles published in journals not ranked by ANVUR) were evaluated by informed peer review.

A couple of anonymous reviewers chosen by one or two members of the sub-area panels, performed the informed peer review of the article, by using a predefined format (slightly different between the two research assessment and also between panels in the same assessment). One of the member of the panel who had chosen at least one of the referees, summarized then the two referee's reports and attributed one of the four scores to the journal article.

ANVUR coined the expression "evaluative mix" to denote this complex evaluative machinery that created many problems, documented for example by (Abramo and D'Angelo, 2016, 2017, Franceschini and Maisano, 2017). The main one is the possible biases induced by the adoption of different evaluation techniques. Indeed, if IPR produced scores systematically different from the ones produced by bibliometrics, this might have introduced systematic bias in the scoring system used for ranking institutions. Indeed, ANVUR realized EXP1 and EXP2 precisely for addressing that problem: a good agreement between bibliometric evaluation and evaluation performed by IPR might justify the adoption of the two different evaluation methods and preserve the comparability of results among areas, institutions, departments and research fields. Positive results of EXP1 and EXP2 were crucial for the soundness of Italian research assessment results: if peer review and bibliometrics do not agree and give significantly different results, the average scores of an institution might be distorted by the different percentage of scores attributed by IPR and by bibliometrics.

EXP1 and EXP2 have an identical structure and rationale. The bulk of ANVUR experiments consisted in the analysis of the agreement between the evaluation obtained through IPR and bibliometric algorithms. The statistical technique adopted by ANVUR was Cohen's kappa (Cohen, 1960), the most popular index of interrater agreement for nominal categories (Sheskin, 2003). High level of agreement should be interpreted as justifying the use of bibliometric and IPR in a same research assessment.

Both in EXP1 and EXP2, according to "Appendice B" of the Final Reports of both VQR1 and VQR2, ANVUR adopted a stratified random sampling with proportional allocation of the population constituted by the journal articles submitted to the research assessments (sample size was about 10% of the population size). Indeed, the Final Reports remark that: "The sample was stratified according to the distribution of the products among the sub-

Table 1: Population, sample and sub-sample sizes for scientific areas in EXP1.

Scientific Areas	Population	Sample	Sub-sample
Area 1 - Mathematics and Informatics	6758	631	438
Area 2 - Physics	15029	1412	1212
Area 3 - Chemistry	10127	927	778
Area 4 - Earth Sciences	5083	458	377
Area 5 - Biology	14043	1310	1058
Area 6 - Medicine	21191	1984	1602
Area 7 - Agricultural and Veterinary Sciences	6284	532	425
Area 8a - Civil Engineering	2460	225	198
Area 9 - Industrial and Information Engineering	12349	1130	919
Area 13 - Economics and Statistics	5681	590	590
	99005	9199	7597

Table 2: Population, sample and sub-sample sizes for scientific areas in EXP2.

Scientific Areas	Population	Sample	Sub-sample
Area 1 - Mathematics and Informatics	4631	444	344
Area 2 - Physics	10182	1008	926
Area 3 - Chemistry	6625	653	549
Area 4 - Earth Sciences	3953	388	320
Area 5 - Biology	10423	951	792
Area 6 - Medicine	15400	1293	1071
Area 7 - Agricultural and Veterinary Sciences	6354	630	489
Area 8b - Civil Engineering	2370	234	180
Area 9 - Industrial and Information Engineering	9930	890	739
Area 11b - Psychology	1801	175	133
Area 13 - Economics and Statistics	5490	498	498
	77159	7164	6041

areas of the various areas" (ANVUR 2017, Appendice B, p.1 our translation). For EXP1 we know the data of the population and of the sample at a sub-area level, while for EXP2 only data for areas are instead available. Each article of the samples received a score by the bibliometric algorithms and was also evaluated by IPR. Up to this point, the design of both experiments is apparently correct, even if a major problem arose during the procedure of bibliometric evaluation. As we have previously remarked, the bibliometric algorithms might result in an inconclusive classification IR for some articles for which the disagreement between citations and impact factor did not permit to automatically assign a score. Both in EXP1 and EXP2, all the articles classified as IR were dropped from the experiments. The consequent distortion in the sample was not accounted for by ANVUR, that just computed the agreement indexes for the articles in the sub-sample.

Tables 1 and 2 show the sizes of the article population, of the sample and of the reduced final sub-sample according to the stratification based on the areas for EXP1 and EXP2. Table 3 reports the same sizes for the stratification based on the sub-areas, which was solely available for EXP1.

For EXP1, the reduction of the sample was not disclosed neither in ANVUR's official reports nor in Ancaiani et al. (2015). Two of the authors of this paper documented (Baccini and De Nicolao, 2017b), with reference to Ancaiani et al. (2015), that the results of EXP1 were not computed on the whole random sample of articles, but on a subset of the sample. Serious concerns about the whole experiment were raised, by highlighting that unknown biases might have been introduced due to the missing items. The reply (Benedetto et al., 2017) concentrated not on biases but on "representativeness" of the selected subset (Baccini and De Nicolao, 2017a). Benedetto et al. (2017) wrote that: "the distribution

Table 3: Population, sample and sub-sample sizes for scientific sub-areas in EXP1.

Scientific Areas	Sub-areas	Population	Sample	Sub-sample
Area 1 - Mathematics and Informatics	Informatics	1636	164	129
	Mathematics	1337	121	94
	Analysis and Probability	1994	179	125
	Applied Mathematics	1791	167	90
Area 2 - Physics	Experimental Physics	1531	139	119
	Theoretical Physics	5350	499	423
	Physics of Matter	3741	349	307
	Nuclear and Sub-Nuclear Physics	467	45	41
	Astronomy and Astrophysics	2719	270	236
	Geophysics	329	28	18
	Applied Physics, Teaching and History	892	82	68
Area 3 - Chemistry	Analytical Chemistry	3013	276	218
	Inorganic and Industrial Chemistry	3076	283	248
	Organic and Pharmaceutical Chemistry	4038	368	312
Area 4 - Earth sciences	Geochemistry etc.	1385	123	107
	Structural Geology	1052	96	81
	Applied Geology	628	56	43
	Geophysics	2018	183	146
Area 5 - Biology	Integrated Biology	3454	325	264
	Morfo-functional Sciences	2432	216	179
	Biochemistry and Molecular Biology	4419	410	339
	Genetics and Pharmacology	3738	359	276
Area 6 - Medicine	Experimental Medicine	3651	347	277
	Clinical Medicine	10578	968	802
	Surgical Sciences	5767	554	429
	Public Health	1195	115	94
Area 7 - Agricultural and Veterinary Sciences	Agricultural Sciences	4566	387	318
	Veterinary	1718	145	107
Area 8 - Civil Engineering and Architecture	Infrastructural Engineering	1131	99	86
	Structural Engineering	1329	126	112
Area 9 - Industrial and Information Engineering	Mechanical Engineering	1390	125	104
	Industrial Engineering	837	81	66
	Nuclear Engineering	1259	117	95
	Chemical Engineering	2186	201	166
	Electronic Engineering	2359	210	166
	Telecommunication Engineering	1469	135	110
	Bio-engineering	1158	110	88
	Informatics	1632	145	120
	Infrastructural Engineering	59	6	4
Area 13 - Economics and Statistics	Economics	2361	235	235
	History	147	37	37
	Management	1750	175	175
	Statistics	1423	143	143

of the subsample across the scientific areas is fairly uniform and proportional to the one resulting from the full sample, i.e. the subsample can be considered as representative of the population of reference in terms of its distribution among scientific areas. [...] Furthermore, [...] the ex-post distribution of bibliometric evaluations is pretty similar in the reference population and in the subsample, confirming that the subsample is a correct representation of the population of reference, also, and perhaps more importantly, in terms of bibliometric results. We hence conclude that [...] the evaluation of concordance has been performed on a sample that is fully representative of the original population of articles to be evaluated".

Despite the problem was known, ANVUR proceeded for EXP2 by adopting the same strategy: stratified sampling of papers with proportional allocation, dropping of papers with inconclusive bibliometric score (IR), calculation of the agreement without disclosing information about the biased selection of papers (ANVUR, 2017).

In order to gain a basic qualitative intuition of the problems induced by the such a selection of papers, it suffices to observe that ANVUR removed from both EXP1 and EXP2 the more problematic articles for which the bibliometric algorithm was unable to reach a score. We cannot exclude that these articles were also the more problematic to be evaluated by peer reviewers. If this is true, ANVUR conducted both experiments on sub-samples "more favorable" to agreement than the complete samples.

From a statistical point of view, drawbacks would arise if the removal of articles from the sample occurred in a non-proportional way between the strata. Thus, in the next section, we derive a procedure for testing the homogeneity of missing proportions between the strata.

3 Testing the homogeneity of missing proportions

Let us suppose a population of N units partitioned into L strata. Moreover, let us assume that N_l is the size of the l-th stratum, i.e. $N = \sum_{l=1}^{L} N_l$. A stratified sampling is carried out by drawing n_l units in each stratum according to simple random sampling without replacement and $n = \sum_{l=1}^{L} n_l$. In the following, with a slight abuse, we also adopt the notation $\boldsymbol{n} = (n_1, \ldots, n_L)$.

In this setting, each unit of the l-th stratum may be missed with probability $\pi_l \in [0, 1]$ - independently with respect to the other units. Thus, the size of missing units in the l-th stratum, say M_l, is a random variable (r.v.) distributed according to the Binomial law with parameters N_l and π_l, i.e. the probability function (p.f.) of M_l turns out to be

$$p_{M_l}(m) = \binom{N_l}{m} \pi_l^m (1 - \pi_l)^{N_l - m} \mathbf{1}_{\{0,1,\ldots,N_l\}}(m) \ ,$$

where $\mathbf{1}_A$ represents the indicator function of the set A.

Let us assume that the r.v. X_l represents the size of missing units of the l-th stratum in the sample. By supposing that the units are missing independently with respect to the sampling, the distribution of the r.v. X_l given the event $\{M_l = m\}$ is the Hypergeometric law with parameters n_l, m and N_l, i.e. the corresponding conditioned p.f. is given by

$$p_{X_l|\{M_l=m\}}(x) = \frac{\binom{m}{x}\binom{N_l - m}{n_l - x}}{\binom{N_l}{n_l}} \mathbf{1}_{\{\max(0, n_l - N_l + m), \ldots, \min(n_l, m)\}}(x) \ .$$

On the basis of these findings and by using the result by Johnson et al. (2005, p.377),

the r.v. X_l is distributed according to the Binomial law with parameters n_l and π_l, i.e. the p.f. of X_l turns out to be

$$p_{X_l}(x) = \binom{n_l}{x} \pi_l^x (1-\pi_l)^{n_l - x} \mathbf{1}_{\{0,1,\ldots,n_l\}}(x)$$

for each $l = 1, \ldots, L$. Obviously, the X_l's are independent r.v.'s. For subsequent use, we also consider the random vector $\boldsymbol{Y} = (X_1, \ldots, X_L)$.

Let us consider the null hypothesis of missing proportion homogeneity between strata $H_0 : \pi_l = \pi, \forall l = 1, \ldots, L$ versus the alternative hypothesis $H_1 : \pi_l \neq \pi, \exists l = 1, \ldots, L$. Thus, for a given realization of the random vector $\boldsymbol{Y}$, say $(x_1, \ldots, x_L) \in \mathbb{N}^L$ such that $\sum_{l=1}^L x_l = s$, the likelihood function under the alternative hypothesis is

$$L_1(\pi_1, \ldots, \pi_L) \propto \prod_{l=1}^L \pi_l^{x_l} (1-\pi_l)^{n_l - x_l} \mathbf{1}_{[0,1]^L}(\pi_1, \ldots, \pi_L) \ ,$$

while the likelihood under the null hypothesis is

$$L_0(\pi) \propto \pi^s (1-\pi)^{n-s} \mathbf{1}_{[0,1]}(\pi) \ .$$

Hence, the likelihood estimator of $(\pi_1, \ldots, \pi_L)$ under the alternative hypothesis is given by $(\widehat{\pi}_1, \ldots, \widehat{\pi}_L)$ where $\widehat{\pi}_l = X_l / n_l$, while the likelihood estimator of π under the null hypothesis is given by $\widehat{\pi} = S/n$ where $S = \sum_{l=1}^L X_l$.

The likelihood-ratio test statistic could be adopted in order to assess the null hypothesis. However, in the present setting the large-sample results are precluded since the sample size n is necessarily bounded by N and the data sparsity could reduce the effectiveness of the large-sample approximations. A more productive approach may be based on conditional testing (for more details on this issue, see Lehmann and Romano, 2005, chapter 10). First, we consider the χ^2-test statistic - asymptotically equivalent in distribution to the likelihood-ratio test statistic under the null hypothesis - which in the present setting reduces to

$$T := T(\boldsymbol{Y}) = \sum_{l=1}^L \left(\frac{n_l(\widehat{\pi}_l - \widehat{\pi})^2}{\widehat{\pi}} + \frac{n_l((1-\widehat{\pi}_l) - (1-\widehat{\pi}))^2}{1-\widehat{\pi}} \right) = \sum_{l=1}^L \frac{n_l(\widehat{\pi}_l - \widehat{\pi})^2}{\widehat{\pi}(1-\widehat{\pi})} \ .$$

It should be remarked that the r.v. S is sufficient for π under the null hypothesis. Hence, in such a case, the distribution of the random vector $\boldsymbol{Y}$ given the event $\{S = s\}$ does not depend on π. Moreover, under the null hypothesis, this conditional distribution is the multivariate Hypergeometric law with parameters s and $\boldsymbol{n}$, i.e. the p.f. of $\boldsymbol{Y}$ given $\{S = s\}$ turns out to be

$$p_{\boldsymbol{Y}|\{S=s\}}(\boldsymbol{x}) = \frac{\prod_{l=1}^L \binom{n_l}{x_l}}{\binom{n}{s}} \mathbf{1}_A(\boldsymbol{x}) \ ,$$

where $\boldsymbol{x} = (x_1, \ldots, x_L)$ and

$$A = \{\boldsymbol{x} : x_l \in \{\max(0, n_l - n + s), \ldots, \min(n_l, s)\}, \sum_{l=1}^L x_l = s\} \ .$$

Therefore, by assuming the conditional approach, an exact test may be carried out. Indeed, if t represents the observed realization of the test statistic T, the corresponding P-value

turns out to be

$$P(T \geq t \mid \{S = s\}) = \sum_{\boldsymbol{x} \in C_t} p_{\boldsymbol{Y}|\{S=s\}}(\boldsymbol{x}) \ ,$$

where $C_t = \{\boldsymbol{x} : \boldsymbol{x} \in A, T(\boldsymbol{x}) \geq t\}$. It should be remarked that the previous P-value may be approximated by means of a Monte Carlo method by generating realizations of a Hypergeometric random vector with parameters s and $\boldsymbol{n}$. The generation of each realization requires $(L-1)$ Hypergeometric random variates - for which suitable algorithms exist, see e.g. Hörmann et al. (2004) - and hence the method is practically feasible.

The previous testing procedure may be generalized to the case of strata partitioned in sub-groups and testing the homogeneity of missing proportions in the sub-groups is also required. Hence, let us now suppose that the l-th stratum is partitioned into G_l sub-groups and let us assume that N_{lk} is the size of the k-th sub-group in the l-th stratum, i.e. $N_l = \sum_{k=1}^{G_l} N_{lk}$ and $N = \sum_{l=1}^{L} \sum_{k=1}^{G_l} N_{lk}$. In addition, a stratified sampling is carried out by drawing n_{lk} units in the k-th sub-group of the l-th stratum according to simple random sampling without replacement, in such a way that $n_l = \sum_{k=1}^{G_l} n_{lk}$ and $n = \sum_{l=1}^{L} \sum_{k=1}^{G_l} n_{lk}$. Moreover, we also assume that $\boldsymbol{n}_l = (n_{l1}, \ldots, n_{lG_l})$ for $l = 1, \ldots, L$. Finally, as to a frequently-adopted notation in this section, if $\boldsymbol{a} = (a_1, \ldots, a_k)$ and $\boldsymbol{b} = (b_1, \ldots, b_m)$ represent two vectors, their concatenation is defined as $(\boldsymbol{a}, \boldsymbol{b}) = (a_1, \ldots, a_k, b_1, \ldots, b_m)$.

Similarly to the setting considered in the simpler case, each unit in the k-th sub-group of the l-th stratum may be missed with probability π_{lk} - independently with respect to the other units. Hence, if X_{lk} represents the size of missing sampled units in the k-th sub-group of the l-th stratum, the r.v. X_{lk} is distributed according to the Binomial law with parameters n_{lk} and π_{lk}. More precisely, the p.f. of X_{lk} is given by

$$p_{X_{lk}}(x) = \binom{n_{lk}}{x} \pi_{lk}^{x} (1 - \pi_{lk})^{n_{lk}-x} \mathbf{1}_{\{0,1,\ldots,n_{lk}\}}(x)$$

for each $l = 1, \ldots, L$. In turn, the X_{lk}'s are independent r.v.'s.

In this framework, let us consider the global null hypothesis of missing proportion homogeneity between all the sub-groups $H_0^g : \pi_{lk} = \pi, \forall k = 1, \ldots, G_l,\ l = 1, \ldots, L$ versus the alternative hypothesis $H_1^g : \pi_{lk} \neq \pi, \exists k = 1, \ldots, G_l, l = 1, \ldots, L$. In complete analogy with the simpler case, by assuming that $\boldsymbol{Y}_l = (X_{l1}, \ldots, X_{lG_l})$ for $l = 1, \ldots, L$, a global test statistic for assessing H_0^g is given by

$$T_g := T_g(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L) = \sum_{l=1}^{L} \sum_{k=1}^{G_l} \frac{n_{lk}(\widehat{\pi}_{g,lk} - \widehat{\pi}_g)^2}{\widehat{\pi}_g(1 - \widehat{\pi}_g)} \ ,$$

where $\widehat{\pi}_{g,lk} = X_{lk}/n_{lk}$ and $\widehat{\pi}_g = S/n$, where - consistently with respect to the adopted notation - we assume that $S = \sum_{l=1}^{L} X_l$ and $X_l = \sum_{k=1}^{G_l} X_{lk}$. Under the null hypothesis H_0^g, the distribution of the random vector $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L)$ given the event $\{S = s\}$ is the multivariate Hypergeometric law with parameters s and $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L)$, i.e. the corresponding conditioned p.f. is given by

$$p_{(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_L)|\{S=s\}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = \frac{\prod_{l=1}^{L} \prod_{k=1}^{G_l} \binom{n_{lk}}{x_{lk}}}{\binom{n}{s}} \mathbf{1}_B(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \ ,$$

where $\boldsymbol{x}_l = (x_{l1}, \ldots, x_{lG_l})$ for $l = 1, \ldots, L$, while

$$B = \{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) : x_{lk} \in \{\max(0, n_{lk} - n + s), \ldots, \min(n_{lk}, s)\}, \sum_{l=1}^{L} \sum_{k=1}^{G_l} x_{lk} = s\} \ .$$

Hence, if t_g represents the observed realization of the test statistic T_g, the corresponding P-value is given by

$$P(T_g \geq t_g \mid \{S = s\}) = \sum_{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \in C_{t_g}} p_{(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L) \mid \{S=s\}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \ ,$$

where $C_{t_g} = \{(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) : (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \in B, T_g(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) \geq t_g\}$. In turn, this P-value may be approximated by means of a Monte Carlo method by generating realizations of a Hypergeometric random vector with parameters s and $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_L)$. The generation of each realization requires $(\sum_{l=1}^{L} G_l - 1)$ Hypergeometric random variates.

If H_0^g is rejected, the null hypothesis of missing proportion homogeneity between the sub-groups within each stratum $H_0^s : \pi_{lk} = \pi_l, \forall k = 1, \ldots, G_l, \forall l = 1, \ldots, L$ could be considered, jointly with the collection of the single L null sub-hypotheses $H_0^{s,1}, \ldots, H_0^{s,L}$, where $H_0^{s,l} : \pi_{lk} = \pi_l, \forall k = 1, \ldots, G_l$. The null hypotheses H_0^s and $H_0^{s,1}, \ldots, H_0^{s,L}$ may be simultaneously assessed by considering the test statistics T_s and $T_{s,1}, \ldots, T_{s,L}$, where

$$T_s := T_s(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L) = \sum_{l=1}^{L} T_{s,l} \ ,$$

while

$$T_{s,l} := T_{s,l}(\boldsymbol{Y}_l) = \sum_{k=1}^{G_l} \frac{n_{lk}(\widehat{\pi}_{s,lk} - \widehat{\pi}_{s,l})^2}{\widehat{\pi}_{s,l}(1 - \widehat{\pi}_{s,l})}$$

and $\widehat{\pi}_{s,lk} = X_{lk}/n_{lk}$ and $\widehat{\pi}_{s,l} = X_l/n_l$. It should be remarked that the random vector $\boldsymbol{Y} = (X_1, \ldots, X_L)$ is sufficient for $(\pi_1, \ldots, \pi_L)$ under the null hypothesis H_0^s. In such a case, the distribution of the random vector $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L)$ given the event $\{\boldsymbol{Y} = \boldsymbol{x}\}$ does not depend on $(\pi_1, \ldots, \pi_L)$. Moreover, under the null hypothesis H_0^s and conditioning to the event $\{\boldsymbol{Y} = \boldsymbol{x}\}$, the L random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L$ are independently distributed according to multivariate Hypergeometric laws. In addition, the distribution of the random vector $\boldsymbol{Y}_l$ given the event $\{\boldsymbol{Y} = \boldsymbol{x}\}$ is the multivariate Hypergeometric law with parameters x_l and $\boldsymbol{n}_l$. Hence, it turns out that the conditioned p.f. is

$$p_{(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_L) \mid \{\boldsymbol{Y}=\boldsymbol{x}\}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L) = \prod_{l=1}^{L} p_{\boldsymbol{Y}_l \mid \{\boldsymbol{Y}=\boldsymbol{x}\}}(\boldsymbol{x}_l) \ ,$$

where

$$p_{\boldsymbol{Y}_l \mid \{\boldsymbol{Y}=\boldsymbol{x}\}}(\boldsymbol{x}_l) = \frac{\prod_{k=1}^{G_l} \binom{n_{lk}}{x_{lk}}}{\binom{n_l}{x_l}} \mathbf{1}_{B_l}(\boldsymbol{x}_l) \ ,$$

while

$$B_l = \{\boldsymbol{x}_l : x_{lk} \in \{\max(0, n_{lk} - n + x_l), \ldots, \min(n_{lk}, x_l)\}, \sum_{k=1}^{G_l} x_{lk} = x_l\} \ .$$

Hence, $T_{s,1}, \ldots, T_{s,L}$ are conditionally independent r.v.'s - even if they do depend on T_s.

Assuming the conditional approach, $(L+1)$ exact tests may be jointly carried out. Indeed,

if t_s and $t_{s,l}$ represent the observed realizations of the test statistics T_s and $T_{s,l}$ respectively, the corresponding P-values are

$$P(T_s \geq t_s \mid \{\boldsymbol{Y} = \boldsymbol{x}\}) = \sum_{(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L)\in C_{t_s}} p_{(\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_L)|\{\boldsymbol{Y}=\boldsymbol{x}\}}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L)$$

and

$$P(T_{s,l} \geq t_{s,l} \mid \{\boldsymbol{Y} = \boldsymbol{x}\}) = \sum_{\boldsymbol{x}_l\in C_{t_{s,l}}} p_{\boldsymbol{Y}_l|\{\boldsymbol{Y}=\boldsymbol{x}\}}(\boldsymbol{x}_l) \ ,$$

where we assume that $C_{t_s} = \{(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L) : \boldsymbol{x}_l \in B_l, T_s(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_L) \geq t_s\}$ and $C_{t_{s,l}} = \{\boldsymbol{x}_l : \boldsymbol{x}_l \in B_l, T_{s,l}(\boldsymbol{x}_l) \geq t_{s,l}\}$. Obviously, these P-values simultaneously hold. Finally, it should be remarked that the previous P-values may be approximated by means of a Monte Carlo method by generating realizations of L independent Hypergeometric random vectors with parameters x_l and $\boldsymbol{n}_l$ for $l = 1,\ldots,L$. Thus, in such a case the generation of each realization requires $(\sum_{l=1}^{L} G_l - L)$ Hypergeometric random variates.

4 Data and results

We have applied the testing procedures developed in the previous section to the data of EXP1 and EXP2 by considering the areas as the strata (see Tables 1 and 2) and to the data of EXP1 by considering the sub-areas as the sub-groups (see Table 3). In order to avoid the obvious criticism that the tests will reject the null hypotheses since Area 13 displays no missing articles, we have performed the testing procedures after having eliminated data for this area. At first, we have considered the null hypothesis H_0 of missing proportion homogeneity between strata both for EXP1 and EXP2. As to EXP1, the null hypothesis H_0 can be rejected since the P-value corresponding to the test statistic T was less than 10^{-6}. As to EXP2, the rejection of the null hypothesis H_0 is in turn justified since the P-value corresponding to the same test statistic was less than 10^{-6}. As previously remarked, in the case of EXP1 the data were also available for the sub-groups - in addition to strata. Hence, the simultaneous procedure for testing the homogeneity of missing proportions between the sub-areas of the scientific areas may be applied. In such a case, the P-value corresponding to the test statistic T_s was given by 0.00001, while the P-values corresponding to the test statistics $T_{s,l}$ were 0.00002 for Area 1, 0.02401 for Area 2, 0.01740 for Area 3, 0.24973 for Area 4, 0.16275 for Area 5, 0.07389 for Area 6, 0.03962 for Area 7, 0.68388 for Area 8 and 0.96818 for Area 9. Hence, the simultaneous null hypothesis H_0^s of missing proportion homogeneity between the sub-areas within each scientific area should be also rejected. The rejection of H_0^s is mainly induced by the Areas 1, 2, 3 and 7, i.e. the areas for which the null hypotheses $H_0^{s,l}$ could be reasonably rejected.

5 Conclusion

This paper investigated the "representativeness" of the sampling procedure adopted in EXP1 and EXP2 by ANVUR. We have rigorously formulated the problem as one of testing missing proportion homogeneity between the strata of a population. After having developed the appropriate statistical tests, we have applied the theoretical findings to the data of the two experiments conducted by ANVUR. Results of the testing procedure show that the null

hypotesis of missing proportion homogeneity between the scientific areas should be rejected for both experiments. The null hyphotesis of homogeneity should be rejected also when the sub-areas for each scientific area are considered. As a consequence, the sampling selection adopted by ANVUR for EXP1 and EXP2 cannot be considered as "representative" of the population of articles submitted to the research assessment. Indeed, "representativeness" could be solely guaranteed if the missing proportions in the strata induced by article elimination were homogenous.

Results of this paper are relevant from two points of view. From the point of view of the Italian research assessments exercises, they demonstrate that the results of the experiments cannot be considered at all as validating the use of the dual metod of evaluation adopted. At the state of current knowledge, it cannot be excluded that the use of a dual method of evaluation introduced uncontrollable major biases in the final results of the assessment. Since all evidence drawn from data in the official research reports shows that peer reviewers' scores were, on average, lower than bibliometric ones, aggregate results for research fields, departments and universities might be affected by the proportion of research outputs evaluated by the two different tecniques: the higher the proportion of research outputs evaluated by peer review, the lower the aggregate score. From the point of view of the search of the Holy Grail of research assessment designers, this paper documents that the experiments conducted by ANVUR do not bring a valid contribution to the discussion about agreement between peer review and bibliometrics. Therefore, the papers describing Italian experiments conducted by ANVUR and authored by ANVUR collaborators should not be cited as valid pieces of knowledge.

References

Abramo, G. and D'Angelo, C.A. (2016). Refrain from adopting the combination of citation and journal metrics to grade publications, as used in the Italian national research assessment exercise (VQR 2011–2014). *Scientometrics*, 109(3), 2053-2065.

Abramo, G. and D'Angelo, C.A. (2017). On tit for tat: Franceschini and Maisano versus ANVUR regarding the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, 11(3), 783-787.

Alfò, M., Benedetto, S., Malgarini, M. and Scipione, S. (2017). On the use of Bibliometric information for assessing articles quality: an analysis based on the third Italian research evaluation exercise, STI 2017. Paris.

Ancaiani, A., Anfossi, A.F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., et al. (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. *Research Evaluation*, 24(3), 242-255.

ANVUR (2013). *Valutazione della qualità della ricerca 2004-2010. Rapporto finale.*

ANVUR (2017). *Valutazione della qualità della ricerca 2011-2014. Rapporto finale.*

Baccini, A. and De Nicolao, G. (2016a). Do they agree? Bibliometric evaluation versus informed peer review in the Italian research assessment exercise. *Scientometrics*, 108(3), 1651-1671.

Baccini, A. and De Nicolao, G. (2016b). Reply to the comment of Bertocchi et al. *Scientometrics*, 108(3), 1675-1684.

Baccini, A. and De Nicolao, G. (2017a). Errors and secret data in the Italian research assessment exercise. A comment to a reply. *RT. A Journal on Research Policy and Evaluation*, 5(1).

Baccini, A. and De Nicolao, G. (2017b). A letter on Ancaiani et al. 'Evaluating scientific research in Italy: the 2004-10 research evaluation exercise'. *Research Evaluation*, 26(4), 353-357.

Benedetto, S., Cicero, T., Malgarini, M. and Nappi, C.s. (2017). Reply to the letter on Ancaiani et al. 'Evaluating Scientific research in Italy: The 2004–10 research evaluation exercise'. *Research Evaluation*, 26(4), 358-360.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C.A. and Peracchi, F. (2013). Bibliometric evaluation vs. informed peer review: Evidence from Italy. Unpublished manuscript, CSEF working papers, Naples.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C.A. and Peracchi, F. (2015). Bibliometric evaluation vs. informed peer review: Evidence from Italy. *Research Policy*, 44(2), 451-466.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46.

Franceschini, F. and Maisano, D. (2017). Critical remarks on the Italian research assessment exercise VQR 2011–2014. *Journal of Informetrics*, 11(2), 337-357.

Goukrager, S. (2012). *Objectivity*. Oxford: Oxford University Press.

Hörmann, W., Leydold, J. and Derflinger, G. (2004) *Automatic Nonuniform Random Variate Generation*. Berlin: Springer.

Johnson, N., Kemp, A. and Kotz, S. (2005) *Univariate Discrete Distributions*, 3rd ed. New York: Wiley.

Kulczycki, E., Korzeń, M. and Korytkowski, P. (2017). Toward an excellence-based research funding system: Evidence from Poland. *Journal of Informetrics*, 11(1), 282-298.

Lehmann, E.L. and Romano J.P. (2005) *Testing Statistical Hypotheses*, 3rd ed. New York: Springer.

Pride, D. and Peter, K. (2018, September 10–13, 2018). In E. Méndez, F. Crestani, C. Ribeiro, G. David and J. Correia Lopes (Eds.), *Peer Review and Citation Data in Predicting University Rankings, a Large-Scale Analysis*. Paper presented at the Digital Libraries for Open Knowledge. 22nd International Conference on Theory and Practice of Digital Libraries, Porto, Portugal. New York: Springer.

Sheskin, D.J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures*. London: Chapman & Hall.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., et al. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. HEFCE.