Inferential versus descriptive statistical approach in the analysis of Delphi performance: A case study.

Analisi delle performance del Delphi. Approccio inferenziale verso descrittivo: studio di un caso

Bolzan M., Auciello M., Pesarin F

Abstract This study aims at evaluating, by permutation methods, the performances of Delphi approach in the research to predict the future of the family in NorthEast of Italy in ten years. The usual descriptives indicators are: stability, consensus and convergence speed. In the work we intend to test – by permutation methods -three equivalent distinct statistical hypotheses: equality, convergence and combination.
Abstract *Questo studio mira a valutare, con metodi di permutazione le performance dell'approccio Delphi nella ricerca per predire il futuro della famiglia nel Nord-Est dell'Italia tra dieci anni. Gli indicatori descrittivi usuali sono: stabilità, consenso e velocità di convergenza. Nel lavoro intendiamo verificare - con metodi di permutazione - tre ipotesi statistiche distinte equivalenti: uguaglianza, convergenza e combinazione.*

Key words: Delphi Approach, performance Indicators; Peermutation methods ...

1 Introduction

The Delphi method is considered by many scholars to be the father of methods that are useful in participatory social research and for the construction of future scenarios on themes that, by nature, do not lend themselves to be analysed by traditional

Bolzan Mario
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: mario.bolzan@unipd.it

Auciello Massimiliano
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: massimilianoauciello@hotmail.it

Pesarin Fortunato
Dipartimento di Scienze Statistiche, Università degli Studi di Padova e-mail: fortunato.pesarin@unipd.it

quantitative approaches. It is set up as a multi-interview survey, carried out through a number of rounds with experts or, more generally, with privileged witnesses, who provide a series of opinions on the subject of the research [3, 2, 4, 1]. The iterativity of the Delphi procedure allows a certain degree of consensus to be reached amongst experts and facilitates a comparison and mutual exchange of knowledge whilst allowing the individual respondents to re-evaluate their positions and beliefs up to their adequate and acceptable convergence. The performance indicators of Delphi applications are measures of stability, consensus and convergence speed. With stability, we denote the situation in which the results of two successive Delphi interviews are not consistently different; in fact, this property can be used as a criterion to stop the procedure. It is measured by regression coefficients calculated on the values of the first and third quartiles of the last two rounds. The closer these two coefficients S_1 and S_3 are to zero, the more the evaluations of the panel of experts can be considered stable. The consensus measures the convergence of opinions, so if the process leads to a final interquartile range which is small enough (e.g. 20% of the domain), consensus is considered to have been reached. The final level of consensus expresses a static aspect that evaluates only the last round regardless of the stability of the answers. It is measured by the width of the last interquartile range, denoted with IQ. The convergence speed or the velocity of convergence (V of C), instead, evaluates the dynamic aspect of the process and is measured by the coefficient of variation (CV), calculated over all the rounds. The more these coefficients decrease towards zero, the greater the V of C is. The research question to which we intend to offer some answers of various levels of completeness is as follows: Is the information produced by the descriptive performance indicators of the Delphi process sufficient, or can the integration of inferential statistical instrumentation offer additional useful and pertinent information? The descriptive empirical approach has the considerable advantage of offering indications based on the contingent dimension of the phenomenon under study and on the basis of shared parameters; the inferential one provides information on the risk of false negatives in which conclusions of general value are drawn. In the case of Delphi surveys, the so-called *minimal sufficient statistics is the whole data set*. Then there is no one-dimensional statistic able to summarize the entire set of information contained in the data. To gather this information as best as possible, a plurality of statistics would be necessary. Indeed, if the data are n, the entire set of information is necessarily gathered from no less than n indicators; whose number, therefore, increases as n. It follows that with a smaller number of statistics one can only aspire to an incomplete (insufficient) and approximate representation with consequent loss of information. It should also be kept in mind that the group of experts involved in the survey is not constructed according to the criteria of representativeness produced by a probabilistic selection of the sample. Therefore, the data (the opinions expressed on the individual items) cannot be analysed with classical statistical methods (parametric tests); the application of non-parametric methods is instead required. The work concerns a survey using the Delphi approach conducted in 2016-2017 on the topic *tomorrow in the family* in Northeast Italy. It involved a group of 32 experts selected amongst professionals and scholars in the areas of more specific interest in the study of the family. The objective is to derive predictions based on the convergence of their opinions on the evolution of phenomena expressed by items (see Table 1) which reflect - specifically in this article - the conditions of parent' lives and which imagine a society and family placed in a future that is sufficiently advanced (10 years) [1].

2 Related statistical methodology

The methodological problem connected with Delphi data, due to its complexity, will be subdivided into some sub-problems in accordance with the hypotheses that are of interest to be analyzed with data observed at two or more time occasions.

The first response model for the analysis concerns a given variable Y observed at two different time occasions, $t = 1,2$, on subject i. In such a context we assume that responses behave on:

$$Y_{1i} = \mu + \eta_i + Z_{1i} \quad \text{and} \quad Y_{2i} = \mu + \eta_i + \Delta_i + \sigma_i Z_{2i},\, i = 1,\ldots,n, (1)$$

where: a) μ is a population constant; b) η_i represents the effect due to the set of covariates specific to subject i (e.g. competence, experience, skillness, and so forth), either observed or not; c) Z_{1i} and Z_{2i} are the so-called error components (*natural deviates*) specific of the adopted instrument, errors that are assumed to be independent between subjects but possibly not time independent within a subject; d) σ_i is the dispersion coefficient specific to subject i: it is > 1 if there is a divergence from occasion *t=2* with respect to that at *t=1*; it is < 1 if, vice versa, there is convergence; e) Δ_i are the Delphi effect descibing how subject i changes its response between two time observations and, although with different meanings, it stands both for *evolution* and *relevance*: it is positive if second response is stochastically larger than the first, otherwise it is negative.

Of course, all such coefficients (parameters) are unknown to the analyst and $[(\eta_i, \Delta_i, \sigma_i), i = 1,\ldots,n]$, three from each subject and each observed response, are to be analyzed. So, there are much more unknown parameters (about 3780) than there are observed subjects ($n = 32$). This implies the necessity for the statistical analysis of taking recourse to nonparametric approaches, because the parametric ones are absolutely impossible.

Considering response differences between two occasions, the model becomes: $Y_{2i} - Y_{1i} = \Delta_i + \sigma_i Z_{2i} - Z_{1i},\ i = 1,\ldots,n$. It is worth noting that this model results independent of constant μ and individual effects η_i; it depends on effects Δ_i and σ_i. To test for such coefficients, nonparametric permutation solutions for their *evolution* , i.e. for their *stability*, the hypotheses under analysis are $H_{0U} : (\Delta_i = 0,\, i = 1,\ldots,n)$ against $H_{1U} : (\exists \Delta_i \neq 0)$, and for their *convergence* (concentration) the hypotheses are $H_{0C} : (\sigma_i = 1,\, i = 1,\ldots,n)$ against $H_{1C} : (\exists \sigma_i \neq 1)$.

However, it is of particular importance to see, separately for each response variable or group of variables, wether the evolution effects, instead of merely different from zero, are positive (i.e. $\Delta > 0$) or negative (i.e. $\Delta < 0$), and those of convergence, instead different from one, are larger $(\sigma > 1)$ or smaller $(\sigma < 1)$ than stability, while identifying for both the direction. This requires that the related alternatives must be written as $H_{1U} : [(\exists \Delta_i < 0) \bigcup (\exists \Delta_i > 0)]$ and $H_{1C} : [(\exists \sigma_i < 1) \bigcup (\exists \sigma_i > 1)]$, respectively. Thus, the statistical problem must imply two separate tests for each variable, one for each of two aspects into which the alternatives are broken-down.

For the permutation tests with paired data we refer to the book by Pesarin and Salmaso (2010, pg. $13 \div 23$). That is: $T_L^* = \sum_{1 \leq i \leq n} [Y_{Li}(t_2) - Y_{Li}(t_1)] \cdot S_i^*$ where S_i^* is a random permutation of equally likely signs $(-1, +1)$ with $L = U$ or C.

So, the response structure of such a test for evolution is: $T_U^* = \sum_i (\Delta_i + \sigma_i Z_{Bi} - Z_{Ai}) S_i^* = \sum_i \Delta_i S_i^* + \sum_i (\sigma_i Z_{Bi} - Z_{Ai}) S_i^*$. From this structure we see that: a) T_U^* essentially depends on coefficient Δ: if $\Delta > 0$ the observed value of the test is stochastically larger than that under H_0, vice versa if $\Delta < 0$; b) the error component $\sum_i (\sigma_i Z_{2i} - Z_{1i}) S_i^*$ is distributed around zero. So, when $\sigma_i = 1$, $i = 1, \dots, n$, and error components Z_{2i} and Z_{1i} are symmetrically distributed around zero, such a test is exact, i.e. it exactly controls first kind error rate. Otherwise the test is exact only asymptotically. A simulation study with error components strongly asymmetric has shown that, with sample sizes of $n = 30$ the test is practically exact.

Since the exact determination of the permutation distribution of test T^* it is necessary to consider all the $4.295 \cdot 10^9$ possible permutations of signs, to estimate its p-value it is usual to consider R random permutations. Such p-value estimates are $\hat{\lambda}_U^> = [\#(T_U^* > T_U^{oss}) + \frac{1}{2}\#(T_U^* = T_U^{oss})]/R$ and $\hat{\lambda}_U^< = [\#(T_U^* < T_U^{oss}) + \frac{1}{2}\#(T_U^* = T_U^{oss})]/R$, respectively, where T_U^{oss} is the observed value.

Of course, the global p-value is then $\hat{\lambda}_U = \min(\hat{\lambda}_U^<, \hat{\lambda}_U^>)$ to be compared with $\alpha/2$ if one wants that the global first kind error rate is α. And if it results that $\hat{\lambda}_U^< \leq \alpha/2$, then one concludes that data behavior stochastically conforms according to the alternative $H_{1U}^< : (\exists \Delta_i < 0)$ at α level, and so having identified both the presence of non-null effects and their direction.

Regarding the test on convergence, to put due emphasis on response variations around a suitable central point, i.e. $|Y_{1i} - \tilde{Y}_1|$ and $|Y_{2i} - \tilde{Y}_2|$, $i = 1, \dots, n$, respectively, it is worth noting that the related response models are: $|Y_{1i} - \tilde{Y}_1| = |\eta_i - \tilde{\eta} + Z_{1i} - \tilde{Z}_1|$ and $|Y_{2i} - \tilde{Y}_2| = |\eta_i - \tilde{\eta} + \Delta_i - \tilde{\Delta} + \sigma_i (Z_{2i} - \tilde{Z}_2)|$. Thus, the test statistic to take into consideration is: $T_C^* = \sum_i \left[|Y_{2i} - \tilde{Y}_2| - |Y_{1i} - \tilde{Y}_1| \right] \cdot S_i^*$; the observed value of which is: $T_C^{oss} = \sum_i \left[|Y_{2i} - \tilde{Y}_2| - |Y_{1i} - \tilde{Y}_1| \right]$. A specific simulation study, carried out to find the most suitable central point providing the best approximation for the null distribution, this results that is the sampling median: $\tilde{Y}_j = Med(Y_{ji}, i = 1, \dots, n)$, $j = 1, 2$.

The structure of the response model for the difference of absolute values of two deviates can be written as:

$$[|Y_{2i} - \tilde{Y}_2| - |Y_{1i} - \tilde{Y}_1|] = \varphi[(\Delta_i - \tilde{\Delta}),\ \sigma_i (Z_{2i} - \tilde{Z}_2) - (Z_{1i} - \tilde{Z}_1)],$$

where the function φ, whose specific structure is difficult to define precisely, indicates that when the null hypothesis $H_{0C} : (\sigma_i = 1,\ i = 1, \dots, n)$ is true, such a function depends on the difference of two pure errors $(Z_{2i} - \tilde{Z}_2) - (Z_{1i} - \tilde{Z}_1)$, on the quantities $(\Delta_i - \tilde{\Delta})$, and on the possible interactions of all such components. However, under the null hypothesis such differences are stationary even with non-constant dispersion. Under the alternative, i.e. in case of convergence, function φ

also depends on coefficients σ_i so it will be suitable to put into evidence the possible convergence as better as the σ_i are far from unity. It is worth noting, however, that the quantities $(\Delta_i - \tilde{\Delta})$ and the interactions may depend on which hypothesis between H_{0C} and H_{1C}, is true. From the one hand, this shows that two test statistics T_U^* and T_C^* are dependent in a way that is too difficult to study and so their joint analysis require the nonparametric combination of dependent permutation tests. From the other hand, it shows that the test T_C^* will be not exact but with a rate of approximation converging to zero as sample size n diverges. We also shown that the dependence of tests T_C^* and T_U^* is asymptotically irrelevant. Simulation trials with sample sizes around $n = 30$, with asymmetric variables while using the same sets of random signs S_i^* for both tests, have shown that their correlation coefficient is practically zero. It is also worth noting that, the joint analysis of two aspects U and C, two test are to be computed on the same random permutations of signs S_i^*.

Similarly to test T_U, when for test T_C it is of interest to also detect the direction of deviates, as with $H_{1C}^{<} : (\exists\sigma_i < 1)$ and $H_{1C}^{>} : (\exists\sigma_i > 1)$, the procedure is the same with obvious substitution of symbols.

The same simulation study has shown that the test T_C^*, being essentially approximate, is somewhat *liberal*, as its rejection probability under H_{0C}, instead of $\alpha = 0.10$ it was of $\alpha = 0.145$, since it suffer from the presence of $(\Delta_i - \tilde{\Delta})$. This requires to empirically adjust its p-value distribution as $[\hat{\lambda}_C]^\gamma$, in place of $\hat{\lambda}_C$ one would have if its null distribution under H_{0C} were exactly uniform. With the same conditions of the real problem under examination, the value of such coefficient is $\gamma \approx 1.2$.

The second model for response variable Y regards the case where it, for subject i, is observed at three time occasions: $t = 1,2,3$. In such a context, responses are assumed to behave according to: $Y_{ti} = \mu + \eta_i + \Delta_{ti} + \sigma_{ti} Z_{ti}$, $i = 1,\ldots,n$, $t = 1,2,3$, where, in particular, $\Delta_{1i} = 0$ and $\sigma_{1i} = 1$, $\forall i$, and where the various coefficients, with obvious modifications of symbols, have the same meaning of the former case.

In the context of observations repeated three times is of particular interest to test for the hypothesis of monotonic convergence if any, since is properly this aspect that plays the fundamental role of Delphi method. That is, with obvious meaning of the symbols, testing for $H_0 : |Y_1 - \tilde{Y}_1| \overset{d}{=} |Y_2 - \tilde{Y}_2| \overset{d}{=} |Y_3 - \tilde{Y}_3|$, against

$$H_1 : [|Y_1 - \tilde{Y}_1| \overset{d}{\leq} |Y_2 - \tilde{Y}_2| \overset{d}{\leq} |Y_3 - \tilde{Y}_3|] \bigcup [|Y_1 - \tilde{Y}_1| \overset{d}{\geq} |Y_2 - \tilde{Y}_2| \overset{d}{\geq} |Y_3 - \tilde{Y}_3|], \quad (2)$$

with at least one strict inequality in either branches and where $\tilde{Y}_t = Med(Y_{ti}, i = 1,\ldots,n)$, $t = 1,2,3$.

Such a kind of testing requires a sort of "*multi-aspect*" method while considering all partial tests for paired observations: $T_{C,12}^*$, $T_{C,13}^*$, and $T_{C,23}^*$ where:

$$T_{C,hj}^* = \sum_i (|Y_i(t_j^*) - \tilde{Y}(t_j^*)| - |Y_i(t_h^*) - \tilde{Y}(t_h^*)|), \; 1 \leq h < j \leq 3.$$

Since all such partial tests are homogeneous (the same sample sizes, the same null distribution, and all significant for large values), for the global inference we use the so-called nonparametric "direct combination": $T_C^* = T_{C,12}^* + T_{C,13}^* + T_{C,23}^*$.

In analogy with the case of two occasions, it is worth observing that such a combination gives a solution to the so-called ”directional analysis of variance” problem [5], that is it permits to decide between two components of H_1:

$$H_1^< : [|Y_1 - \tilde{Y}_1| \overset{d}{\leq} |Y_2 - \tilde{Y}_2| \overset{d}{\leq} |Y_3 - \tilde{Y}_3|] \text{ and } H_1^> : [|Y_1 - \tilde{Y}_1| \overset{d}{\geq} |Y_2 - \tilde{Y}_2| \overset{d}{\geq} |Y_3 - \tilde{Y}_3|],$$

related to the monotonically increasing and, respectively, decreasing stochastic ordering of concentration. In particular, the combined test T_C^* becomes:

$$T_C^* = 2\sum_i[|Y_i(t_1^*) - \tilde{Y}(t_1^*)|] - 2\sum_i[|Y_i(t_3^*) - \tilde{Y}(t_3^*)|],$$

where is apparently only involved the data at times $t = 1$ and $t = 3$. It is, however, to underline that: 1) random permutations $\mathbf{t}^* = (t_1^*, t_2^*, t_3^*)$ of $\mathbf{t} = (1,2,3)$, to preserve the underlying within subjects dependence on observations, are common to three partial tests; 2) test T_C^* would be exact if in place of median estimates $\tilde{Y}_t$, $t = 1,2,3$, true medians Me_1, Me_2 and Me_3 were known; 3) the approximation rate, evaluated by a specific simulation study, is practically negligible as its convergence to zero is fast.

The nonparametric combination of $K \geq 2$ dependent tests is a useful method to make inference when a set of observed variables, for explanatory or interpretative reasons, can form a so-called section of information (for example, it is of interest to jointly see all the variables concerning the section regarding the family, and so forth). The use of such a method is unavoidable when the numbers of V variables and/or of P parameters in the response model are larger than sample size n. We invite readers to see Chapter IV of the book by Pesarin and Salmaso (2010) where the theory and related methodology is wholly discussed.

In this regards, let us suppose that the K partial tests are $(T_{C1}^*, \ldots, T_{CK}^*)$, the p-value of which are $(\lambda_{C1}, \ldots, \lambda_{CK})$. Their nonparametric combination can be done, for instance, by Fisher’s combination as:

$$T_F^* = -2\sum_{k=1}^K \log(\lambda_{Ck}^*),$$

where $\log(\cdot)$ are natural logarithms, to obtain the p-value of which it is required that the K test statistics are jointly calculated at each data permutation and common to all of them (for instance, in terms of the Delphi data, with $T_{Ck}^*[\mathbf{X}(\mathbf{t}^*)]$, $k = 1, \ldots, K$, where $\mathbf{t}^* = (t_1^*, t_2^*, t_3^*)$ are permutations of $\mathbf{t} = (1,2,3)$, for data observed at three times, and so forth).

3 Analysis of the results

The so-called Delphi effect is configured as the interaction of two distinct but not exclusive contributions: that of the median convergence of expert evaluations, here verified through the non-parametric equality test (U), and that in the distribution of the same assessments verified through the convergence test (C) of the pairwise comparisons of the surveys. The hypothesis H_0 of equality verifies if the distributions related to the surveys agree in median between them - two by two and amongst all - that is, the tendential idea (expressed by the median of the first survey) is also confirmed in the subsequent interviews. The inevitable and reasonable adjustments of the distribution median, typical of the Delphi method, are not sufficiently large to be significantly relevant, and the interviewees, although from different directions and positions, offer indications during the three rounds that are recognised in the same orientation and opinions expressed by the median value of the distribution. The hypothesis H_0 of convergence states that the dispersion around the median does not decrease significantly, so its rejection refusal is to be read as a confirmation that in the course of the rounds, there is sensitive and gradual convergence and consolidation towards the central value of the distribution. In the case of statistical significance, the experts involved during the surveys move away gradually but consistently from their initial positions and approach the final median. On the basis of model (2), the two contributions on which the hypothesis test is started are not directly separable or evaluable in a strictly separate manner (each one, in fact, conditions the other) and to an increasingly lesser extent as the number of observations increases. This last consideration clashes with a qualifying feature of the Delphi approach in which it is conducted through a very limited number of experts and surveys. Therefore, because of the impossibility of distinguishing the two components of the Delphi effect, the combined Fisher test is applied to measure the joint effect of the two contributions described above. The result of the significance of the latter supports and integrates the summary information offered by the reading of the performance indicators. In Table 1 reports the description by summarising the measures of the performance parameters and the level of significance of the tests.

Table 1 tab:1 Performance parameters stability, convergence speed and consensus for each items and the level of significance of the test on equality (U), convergence (C) and combined with the Fisher (F) test of the pairwise comparisons of the rounds, as well as the multi-aspect one per item

Area 1. Parents (six items)	S(*)	VC	C	1-2 U	1-2 C	**1-2 F**	2-3 U	2-3 C	**2-3 F**	1-3 U	1-3 C	**1-3 F**
1. Parents (father and mother) will devote themselves to training their children.	−	++	++	.519	.068	**.384**	.051	.006	**.004**	.194	.002	**.006**
2. The father will be present in the training and leisure activities of the children (school, sports, associations, etc.).	+	++	++	.287	.103	**.333**	.103	.014	**.027**	.070	.000	**.000**
3. The mother will be able to organise work and family life to be more present in the children's activities of education and free time.	+	−	+	.320	.040	**.166**	.315	.056	**.219**	.217	.004	**.014**
4. For the mother, the organisation of family life will be conditioned by professional rhythms and commitments.	+	+	++	.187	.032	**.090**	.064	.302	**.233**	.053	.005	**.006**
5. The father will try to organise professional commitments according to the organisation of family life.	++	+	++	.273	.033	**.125**	.174	.144	**.251**	.130	.011	**.037**
6. Parents will invest in the role as educators of their children.	++	++	++	.145	.003	**.015**	.090	.110	**.126**	.476	.000	**.002**

[*] S: Stability; V of C: Speed of convergence; C: Consensus.

Legend common to the three indicators ++: excellent perfect; +: good partially goode; −: absence.
Consider that the initial domain for the response scale is 100, so a final interquartile range less than 20 is considered good.

1. Stability: $S_1 = 10$, $S_3 = -10$. There is none; the assessments converge quickly. V of C: $v_1 = 5$, $v_3 = -5$, excellent. Consensus: $IQ = 10$, excellent.
2. Stability: $S_1 = 0$, $S_3 = -10$. Is in the first quartile but not in the third one, which tends to decrease. V of C: $v_1 = 0$, $v_3 = -2.5$, excellent. Consensus: $IQ = 10$, excellent.
3. Stability: $S_1 = 5$, $S_3 = -1.3$. There is in the first quartile, but the third one seems stabler. V of C: $v_1 = -2.5$, $v_3 = -5.6$. It is not good that the coefficients are both negative. Consensus: IQ = 13.7. Good. There is a change in course starting from the second round. The certain and important element, however, is the final consensus that being less than 15 is good.
4. Stability: $S_1 = 10$, $S_3 = 0$. Not good on the first quartile but perfect on the third one.V of C: $v_1 = 5$, $v_3 = 1.3$. The first quartile quickly converges towards the median, whereas the third one stabilises. Consensus: $IQ = 10$, excellent.
5. Stability: $S_1 = 0$, $S_3 = 0$. Perfect. V of C: $v_1 = 0$, $v_3 = -2.5$. Good; the third quartile decreases rapidly. Consensus: $IQ = 10$, excellent.
6. Stability: $S_1 = 0$, $S_3 = -1.3$. Good on the third quartile and perfect on the first one. V of C: $v_1 = 2.5$, $v_3 = -5$ Excellent. Consensus: $IQ = 10$, excellent.

The item, the three parameters provide mostly satisfactory measurements; no item was found to achieve negative performance, except for 3. The consensus is at least good for all items. In the first summary, it is observed that the performance indicators on the evolution of the items offered a generally satisfactory picture, even if apparently contradictory situations were recorded on individual items, as in items 1 and 3, where either stability or the V of C takes on unsatisfactory dimensions or, in any case, is in line with the other parameters.

The analysis of the results of the test application on significance shows that the hypothesis H_0 of equality is almost always accepted even with high levels for all three possible comparisons of the three measurements 1-2, 2-3 and 1-3. This information allows us to confirm that in the median, the three distributions are statistically equivalent, indicating that although the procedure registered different levels of stability amongst the six items - from absent in item 1 to very good in items 5 and 6 - this did not significantly alter the central effect expressed by the medians of the three rounds; this proves that the basic opinion of the group of experts tended to remain the same, albeit with variations. The first contribution of the Delphi effect can be said to be confirmed for these six items. It should be noted, however, that the descriptive analysis of stability is only based on the results of the second and third rounds, whereas with the application of the tests, all three distributions are compared with information that is therefore more exhaustive and differentiated. The analysis of the significance levels of the Ho of the convergence hypothesis (indicated in the columns with C) records less-systematic trends than the previous one. If we want to summarise and conclude, only in comparison 1-3 is the hypothesis Ho is systematically rejected for each of the six items, whereas in the previous comparisons, alternating positions are recorded.

It should be noted that the consensus expresses a measure of the reduction in the range of variation (through the interquartile range) only of the last round, whereas in the inferential analysis, three are jointly examined and compared. From a targeted reading of the results of significance, it is therefore clear that the three iterations were necessary and also sufficient to detect the monotonic convergence as expected -and desired- precisely by the procedure of successive interviews envisaged in Delphi, otherwise indicated just as Delphi effect. This convergence would not have been achieved, or at least partially only, had we stopped at the first two surveys. This information and conclusion would not have been achieved on the basis of only reading the consensus results. A more careful analysis, however, leads to the consideration of how the latter indicator (third column of Table 1) is always satisfactory in partially confirming the presence of reciprocal and partial integration of the two pieces of information. The significance levels of the combination Ho hypothesis (in Table 1 with F) of the two previous components are rejected and always have a high level of significance only in the comparison between the first and third rounds; in the two other comparisons, the situation is variegated, as the hypothesis Ho is mostly accepted. The high levels of significance of F make it clear that the joint effects of the two distinct contributions of the Delphi effect are consistent and translate to the effective confirmation of information coming from the Delphi procedure applied and described here. Joint effects are still not evident from the first two rounds (results everywhere are predominantly not significant with respect to comparisons between 1-2 and 2-3) to further confirm that we can reiterate the need for the three interviews to converge to expendable results. Finally, the information produced by the inferential analysis offers reflections and indications that the analysis of the indicators may, to some extent, be direct but not defined.

References

1. Bolzan M.(2018), *Domani in Famiglia*. Franco Angeli, Collana Strutture e Culture Sociali, Milano: pp. 1-226. (2017)
2. Dajani J.S., Sincoff M.Z., Talley W.K. (1979), *Stability and agreement criteria for the termination of Delphi studies*, Technol. Forecast. Soc. Chang. 13 pp. 83–90.
3. Glenn J.C. (1972), *Futurizing Teaching vs Futures Course, Social Science Record.* Syracuse university, Vol IX, n.3, Spring.
4. Pacinelli A. (2007), *Metodi per la ricerca sociale finalizzata.* Franco Angeli, Milano.
5. Pesarin F. Salmaso L. (2010),*Permutation test for complex data: Theory, Application and Software*. Wiley, Chichester, UK.