Employment status and education/employment relationship of PhD graduates from the University of Ferrara

Rosa Arboretti Giancristofaro\textsuperscript{a*}, Stefano Bonnini\textsuperscript{b} and Luigi Salmaso\textsuperscript{c}

\textsuperscript{a}Department of Mathematics, University of Ferrara, Italy; \textsuperscript{b}Centre for Modelling, Computing and Statistics, University of Ferrara, Italy; \textsuperscript{c}Department of Management and Engineering, University of Padova, Italy

(Received 22 April 2008; final version received 18 November 2008)

Two sample surveys of Post-Docs were planned and carried out at the University of Ferrara in 2004 and 2007 aimed at determining the professional status of Post-Docs, the relationship between their PhD education and employment, and their satisfaction with certain aspects of the education and research program. As part of these surveys, two methodological contributions were developed. The first concerns an extension of the non-parametric combination of dependent rankings to construct a synthesis of composite indicators measuring satisfaction with particular aspects of PhD programs [R. Arboretti Giancristofaro and L. Salmaso, \textit{Global ranking indicators with application to the evaluation of PhD programs}, Atti del Convegno “Valutazione e Customer Satisfaction per la Qualità dei Servizi”, Roma, 8–9 Settembre 2005, pp. 19–22; R. Arboretti Giancristofaro, S. Bonnini, and L. Salmaso, \textit{A performance indicator for multivariate data}, Quad. Stat. 9 (2007), pp. 1–29; R. Arboretti Giancristofaro, F. Pesarin, and L. Salmaso, \textit{Nonparametric approaches for multivariate testing with mixed variables and for ranking on ordered categorical variables with an application to the evaluation of PhD programs}, in Real Data Analysis, S. Sawilowsky, ed., a volume in Quantitative Methods in Education and the Behavioral Sciences: Issues, Research and Teaching, Ronald C. Serlin, series ed., Information Age Publishing, Charlotte, North Carolina, 2007, pp. 355–385]. The procedure was applied to highlight differences in the interviewed Post-Docs’ multivariate satisfaction profiles in relation to two aspects: education/employment relationship; employment expectations; and opportunities. The second consists of an inferential procedure providing a solution to the problem of hypothesis testing, where the objective is to compare the heterogeneity of two populations on the basis of sampling data [G.R. Arboretti, S. Bonnini, and F. Pesarin, \textit{A permutation approach for testing heterogeneity in two-sample categorical variables}, Stat. Comput. (2009) doi: 10.1007/S11222-008-9085-8]. The procedure was applied to compare the degrees of heterogeneity of Post-Doc judgments in the two surveys with regard to the adequacy of the PhD education for the work carried out.

Keywords: employment survey; performance indicators; heterogeneity tests

*Corresponding author. Email: rosa.arboretti@unife.it
1. Introduction

Recent years have seen universities multiplying their efforts to improve the quality of doctorates in line with recommendations expressed by the National Committee for the Evaluation of the University System, including those regarding the analysis of employment opportunities for Post-Docs.

Evaluating the quality of a PhD is somewhat complex and includes several dimensions of analysis (adequacy of educational content, success of entry into employment, etc.) and several figures involved in the education process at various levels. Considering a PhD to be a process characterized by initial inputs and final outcomes, its evaluation is based on process quality indicators and outcome indicators, taking elements of initial inputs (input characteristics of PhD researchers, resources available to them, characteristics of territorial context, etc.) into account.

Some universities have also carried out ad hoc surveys to obtain the opinions of PhD researchers on the education and research activities that characterize doctorate courses and on organizational aspects of the education programs. Still lacking, or at best sporadic, however, are surveys for the evaluation of PhD outcomes and employment opportunities (some experiments are still underway).

In this context, and as part of the work of the Statistical Monitoring Committee operating in the University of Ferrara, a sample survey of Post-Docs was planned and carried out in 2004 and again in 2007 aimed at determining their professional status, the relationship between PhD education and employment, and their satisfaction with certain aspects of the education and research program.

This paper highlights some interesting results from these innovative surveys. The main results of the descriptive statistical analysis are illustrated in Section 2. A peculiar methodological solution to the problem of synthesizing multivariate statistical information is described and applied in Section 3. The two-sample heterogeneity comparison test is introduced and performed to compare Post-Doc judgment heterogeneity between the two surveys in Section 4, and Section 5 is dedicated to the concluding remarks.

2. Post-Doc surveys at the University of Ferrara: some descriptive results

The University of Ferrara’s first Post-Doc survey involved a sample of 120 PhD graduates, selected from four cohorts of 288 Post-Docs who graduated between 2001 and 2004. The PhDs were grouped into three macro-areas: economic–legal (EL), medical–biological (MB) and scientific–technological (ST). A random sample of 30 Post-Docs was selected from each cohort, stratified by doctorate area with proportional allocation. The second survey (April 2007) involved a reduced sample of 60 Post-Docs who graduated between 2004 and 2006, again stratified by macro-area. The surveys were carried out by means of computer-assisted telephone interviews. The degree of satisfaction with various aspects of their work, education received and organization of the PhD course was marked on a scale of 1–4 (not at all, not very, quite, very satisfied).

This section reports the results of the analysis by doctorate title cohort relating to the employment status of Post-Docs and two aspects of Post-Doc satisfaction: education–employment relationship; employment expectations; and opportunities.

With reference to employment status of Post-Docs interviewed in the two surveys, some significant results can be highlighted (Figure 1). Employment in the academic field concerns little more than half of Post-Docs in the various cohorts. A percentage varying from 7% to 14% of Post-Docs interviewed 3 to 31/2 years after graduation reconcile work in the academic field with a job in the non academic field. The percentage of Post-Docs positioned exclusively in the non academic field is very significant: 3–3.5 years after graduation the percentage stands at around 43%.

For Post-Docs with a more structured employment position (excluding those with a study grant or research allowance or unpaid work), a number of satisfaction indicators have been calculated
with regard to the relationship between education received during the PhD and work currently carried out (Figure 2). In the 2004 survey, 59% stated that 3.5 years after graduation, the work they carried out was very much in line with the training received during the PhD (23% quite satisfied with this coherence). In the same cohort, 68% of respondents felt that PhD training was

<table>
<thead>
<tr>
<th>Cohorts</th>
<th>2004 Survey</th>
<th>2007 Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non employment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Not employed (including unpaid collaborations)</td>
<td>3.3</td>
<td>6.7</td>
</tr>
<tr>
<td>Employment in the academic field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Employed</td>
<td>53.4</td>
<td>66.7</td>
</tr>
<tr>
<td>Employment only in the academic field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Employed</td>
<td>46.7</td>
<td>50.0</td>
</tr>
<tr>
<td>Employment only in a NON academic field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>% Employed</td>
<td>43.3</td>
<td>26.6</td>
</tr>
<tr>
<td>% Working prior to the qualification and employed in a non academic field</td>
<td>46.7</td>
<td>53.8</td>
</tr>
</tbody>
</table>

**Figure 2. Satisfaction with education/employment relationship by PhD graduation cohort. Those in training, with study grants, and/or unpaid collaborations are excluded.**
very much suited to the work they carried out (14% quite satisfied with suitability), with only one-fifth of respondents not at all or not very satisfied with education/employment coherence or adequacy. In the 2007 survey, these comforting opinions would appear to worsen considerably: the oldest cohort’s negative opinion of education/employment coherence or adequacy represents, respectively, 37% and 45% of Post-Docs.

3. Global satisfaction index

3.1 An extension of non-parametric combination ranking for ordered categorical variables

In this section, we define an appropriate synthesis indicator of a set of $k$ informative ordered categorical variables representing judgments of a specific quality aspect under evaluation (e.g. external effectiveness of educational processes within the university system) [1,2,3]. Let us denote the responses as a $k$-dimensional variable $Y = [Y_1, \ldots, Y_k]$, where the marginal variable $Y_i$ ($i = 1, \ldots, k$) can assume $m_i$ ordered modalities $v_1, \ldots, v_{m_i}$, or discrete scores, $h = 1, \ldots, m_i$, $m_i \in \mathbb{N}\{0\}$, $m_i > 1$. If $Y_i$ is a categorical variable, then a numerical transformation of modalities $v_1, \ldots, v_{m_i}$ into discrete scores is needed. Large values of $h$ correspond to higher satisfaction rates. From now on, to simplify the notation, we assume $m_i = m$ for every $i$, but it is not necessary that all variables have the same number of modalities/scores. For reasons of application, these variables are given different (non-negative) degrees of importance: $(0 < w_i \leq 1, i = 1, \ldots, k)$. Such weights are thought to reflect the different roles of the variables in representing indicators of the specific quality aspect under evaluation (e.g. indicators of a PhD researcher’s success in entering the labour market or academic field) and are provided by responsible experts or from the results of surveys previously carried out in the specific context.

The methodological problem we are facing concerns the construction of a global satisfaction index of $N$ statistical subjects starting from $k$ dependent variables on the same $N$ subjects, each representing a specific aspect under evaluation. If the observed scores $h$ on variable $Y_i$ are observed ranks, each individual satisfaction response is just the vector of ranks and the global satisfaction index is equivalent to a global ranking.

Two main aspects should be considered when attempting to find a global index or a global ranking of satisfaction.

(I) The search for a suitable combining function of two or more indicators or rankings.

(II) The consideration of extreme units of the global ranking. Bird et al. [5] pointed out that “... the principle that being ranked lowest ... does not immediately equate with genuinely inferior performance should be widely recognized; and reflected in the method of presentation (of ranking).”

The non-parametric combination (NPC) of dependent rankings [9] provides a solution to problem (I). The main purpose of the NPC ranking method is to obtain a single ranking criterion for the statistical units under study, which summarizes many partial (univariate) rankings.

Let us consider a multivariate phenomenon whose variables $Y$ are observed on $N$ statistical units. Using component variables $Y_i, i = 1, \ldots, k$, each one providing information about a partial aspect, we wish to construct a global index or combined ranking $T$:

$$T = \phi(Y_1, \ldots, Y_k; w_1, \ldots, w_k), \quad \phi : \mathbb{R}^{2k} \longrightarrow \mathbb{R}^1,$$

where $\phi$ is a real function that allows us to combine the partial dependent rankings and $(w_1, \ldots, w_k)$ is a set of weights that take the relative degrees of importance among the $k$ aspects of $Y$ into account.
We introduce a set of minimum reasonable conditions related to variables $Y_i, i = 1, \ldots, k$:

1. For each of the $k$ informative variables a partial ordering criterion is well established, i.e. “large is better.”
2. Regression relationships within the $k$ informative variables are monotonic (increasing or decreasing).
3. The marginal distribution of each informative variable is non-degenerate.

Worthy of note is that we need not assume continuity of $Y_i, i = 1, \ldots, k$, so the probability of ex-equo can be positive. The real combining function $\phi$ is chosen from class $\Phi_1$ of real combining functions satisfying the following minimum properties:

1. $\phi$ must be continuous in all $2k$ arguments, in that small variations in any subset of arguments imply a small variation in the $\phi$-index;
2. $\phi$ must be monotone non-decreasing with respect to each argument:
   \[
   \phi(\ldots Y_i, \ldots; w_1, \ldots, w_k) \geq \phi(\ldots Y'_i, \ldots; w_1, \ldots, w_k)
   \]
   if $1 > Y_i > Y'_i > 0, \ i = 1, \ldots, k$;
3. $\phi$ must be symmetric with respect to permutations of the arguments, in that if, for instance, $u_1, \ldots, u_k$ is any permutation of $1, \ldots, k$, then:
   \[
   \phi(Y_{u_1}, \ldots, Y_{u_k}; w_{u_1}, \ldots, w_{u_k}) = \phi(Y_1, \ldots, Y_k; w_1, \ldots, w_k).
   \]

Property 1 is obvious. Property 2 means that if, for instance, two subjects have exactly the same values for all $Y$’s except the $i$th, then the one with $Y_i > Y'_i$ must have at least the same satisfaction $\phi$-index assigned to it. Property 3 states that any combining function $\phi$ must be invariant with respect to the order in which informative variables are processed.

The three described properties are satisfied by Fisher’s combining function: $\phi = -\sum_{i=1}^{k} w_i \log(1 - \lambda_i)$, where $\lambda_i = (Y_i + 0.5)/(m+1)$ are normalized scores defined in the open interval $[0,1]$ ($+0.5$ and $+1$ are added, respectively, to the numerator and denominator of $\lambda_i$). Of course, other combining functions may be of interest [6] but Fisher’s combining function seems to be more sensitive when assessing higher satisfaction than when assessing lower satisfaction, i.e. small differences in the lower satisfaction region seem to be identified with greater difficulty than those in the higher satisfaction region.

For problem (II), we propose an extension of the NPC ranking method to the case of ordered categorical variables based on extreme satisfaction profiles [2,3]. Extreme satisfaction profiles are defined a priori on a hypothetical frequency distribution of variables $Y_i, i = 1, \ldots, k$. Let us consider data $Y$, where the rule “large is better” holds for all variables. Observed values for the $k$ variables are denoted by $y_{ji}, i = 1, \ldots, k; j = 1, \ldots, N$. Examples of extreme satisfaction profiles are given below.

1. Maximum satisfaction profile is defined as follows.
   - Maximum satisfaction is obtained when all subjects have the highest value of satisfaction for all variables:
     \[
     f_{hi} = \begin{cases} 
     1 & \text{for } h = m, \\
     0 & \text{otherwise, } \forall i, i = 1, \ldots, k
     \end{cases}
     \]
     where $f_{hi}$ are the relative frequencies of categories $h, h = 1, \ldots, m$, for variable $Y_i, i = 1, \ldots, k$.  

• Minimum satisfaction is obtained when all subjects have the smallest value of satisfaction for all variables:

\[ f_{hi} = \begin{cases} 1 & \text{for } h = 1 \\ 0 & \text{otherwise} \end{cases}, \quad \forall i, i = 1, \ldots, k. \]

The weak satisfaction profile is defined as follows.

• Maximum satisfaction is obtained when subjects have the highest value of satisfaction with relative frequencies varying across the variables:

\[ f_{hi} = \begin{cases} u_i & \text{for } h = m \\ u_{hi} & \text{otherwise, where } \sum_{h=1}^{m-1} u_{hi} = (1 - u_i) \quad i = 1, \ldots, k. \end{cases} \]

• Minimum satisfaction is obtained when subjects have the smallest value of satisfaction with relative frequencies varying across the variables:

\[ f_{hi} = \begin{cases} l_i & \text{for } h = 1 \\ l_{hi} & \text{otherwise, where } \sum_{h=2}^{m} l_{hi} = (1 - l_i) \quad i = 1, \ldots, k, \end{cases} \]

where \( u_i \) and \( l_i \) represent realistic achievable targets that can be fixed observing past experience or motivational targets established by managers or organizers in the strategic and business planning.

In order to include the extreme satisfaction profiles in the analysis, we transform original values \( h, h = 1, \ldots, m \).

First, we separate the values of \( h \) corresponding to a judgment of satisfaction, say the last \( t, 1 \leq t \leq m \), from those corresponding to judgments of dissatisfaction, i.e. \( (m - t) \). For the last \( t \) values of \( h \) corresponding to a judgment of satisfaction, the transformed values of \( h \) are defined as:

\[ h + f_{hi} \times 0.5 \quad h = m - t + 1, \ldots, m; \quad i = 1, \ldots, k. \]

For the first \( (m - t) \) values of \( h \) corresponding to judgments of dissatisfaction, the transformed values of \( h \) are defined as:

\[ h + (1 - f_{hi}) \times 0.5 \quad h = 1, \ldots, m - t; \quad i = 1, \ldots, k. \]

This transformation is equivalent to the assignment to original values \( h, h = 1, \ldots, m \), of additive degrees of importance which depend on relative frequencies \( f_{hi} \) and which increase the original values \( h \) up to \( h + 0.5 \). The limit 0.5 is fixed in such a way that the increase in the original score \( h \), positively (negatively) related to the fraction of evaluators who choose the corresponding judgement, is less than one; hence, the transformation of \( h \) is less than \( h + 1 \).

Let us suppose, for example, that \( h = 1,2,3,4 \) and values 3 and 4 correspond to judgments of satisfaction. By applying the above transformation, the value of 3 tends towards the upper value 4, which represents higher satisfaction when \( f_{i3} \) increases. On the contrary, the value of 1 tends towards 2 (less dissatisfaction) when \( f_{i1} \) decreases. Figure 3 displays this example.

The transformation of values \( h, h = 1, \ldots, m \), weighted by relative frequencies \( f_{ih} \), is applied to observed values \( y_{ji}, i = 1, \ldots, k; j = 1, \ldots, N \). For the last \( t \) values of \( h \) corresponding to a judgment
of satisfaction, the transformed values of $y_{ji}$ are defined as:

$$z_{ji} = y_{ji} + \sum_{h=m-t+1}^{m} I_h(y_{ji}) \times f_{ih} \times 0.5, \quad i = 1, \ldots, k; \quad j = 1, \ldots, N,$$

with

$$I_h(y_{ji}) = \begin{cases} 1 & \text{if } y_{ji} = h \\ 0 & \text{if } y_{ji} \neq h. \end{cases}$$

For the first $(m-t)$ values of $h$ corresponding to judgments of dissatisfaction, the transformed values of $y_{ji}$ are defined as:

$$z_{ji} = y_{ji} + \sum_{h=1}^{m-t} I_h(y_{ji}) \times (1 - f_{ih}) \times 0.5, \quad i = 1, \ldots, k; \quad j = 1, \ldots, N.$$
It is worth noting that $z_{i,\text{max}}$ represents the preferred value for each variable, and it is obtained when satisfaction is at its highest level according to the extreme satisfaction profile; $z_{i,\text{min}}$ represents the worst value, and it is obtained when satisfaction is at its lowest level according to the extreme satisfaction profile. Scores $\lambda_{ji}, i = 1, \ldots, m, j = 1, \ldots, N$ are one-to-one increasingly related to values $y_{ji}, z_{ji}$.

In order to synthesize the $k$ partial rankings based on scores $\lambda_{ji}, i = 1, \ldots, m, j = 1, \ldots, N$, using the NPC ranking method, we use a combining function $\phi$:

$$[T_j = \phi(\lambda_{j1}, \ldots, \lambda_{jk}; w_1, \ldots, w_k), j = 1, \ldots, N].$$

In order for the global index to vary in the interval $[0,1]$, we put:

$$S_j = \frac{T_j - T_{\text{min}}}{T_{\text{max}} - T_{\text{min}}}, j = 1, \ldots, N,$$

where

$$T_{\text{min}} = \phi(\lambda_{1,\text{min}}, \ldots, \lambda_{k,\text{min}}; w_1, \ldots, w_k),$$

$$T_{\text{max}} = \phi(\lambda_{1,\text{max}}, \ldots, \lambda_{k,\text{max}}; w_1, \ldots, w_k),$$

and $\lambda_{i,\text{min}}$ and $\lambda_{i,\text{max}}$ are obtained according to the extreme satisfaction profiles:

$$\lambda_{i,\text{min}} = \frac{(z_{i,\text{min}} - z_{i,\text{min}}) + 0.5}{(z_{i,\text{max}} - z_{i,\text{min}}) + 1}, i = 1, \ldots, k,$$

$$\lambda_{i,\text{max}} = \frac{(z_{i,\text{max}} - z_{i,\text{min}}) + 0.5}{(z_{i,\text{max}} - z_{i,\text{min}}) + 1}, i = 1, \ldots, k.$$

Note that value $T_{\text{min}}$ represents the unpreferred value of the satisfaction index since it is calculated from $(\lambda_{1,\text{min}}, \ldots, \lambda_{k,\text{min}})$, while $T_{\text{max}}$ represents the preferred value since it is calculated from $(\lambda_{1,\text{max}}, \ldots, \lambda_{k,\text{max}})$. $T_{\text{min}}$ and $T_{\text{max}}$ are reference values to evaluate the “distance” of the observed satisfaction values from the situation of highest satisfaction defined according to the extreme satisfaction profile.

### 3.2 Global satisfaction indicators in the University of Ferrara’s Post-Doc Survey

This section reports the results of the analysis applied to data collected by doctorate area from the University of Ferrara’s 2004 Post-Doc Survey relating to two aspects of Post-Doc satisfaction: education/employment relationship, and employment expectations and opportunities. With regard to the former, Post-Docs (excluding persons in training, with study grants, and/or unpaid collaborations) gave an evaluation of the coherence between education and employment, of use in employment of the skills acquired during the PhD, and of the adequacy of the PhD training for the work carried out. With regard to PhD researchers’ employment expectations after obtaining their qualification, three categorical variables (opportunities in the academic field, opportunities in the labour market, and openings in the scientific community) were considered.

The aim of the analysis is to verify the presence of differences in satisfaction profiles of Post-Docs belonging to different areas. Tables 1 and 2 show the percentages of very satisfied and quite satisfied Post-Docs in relation to variables regarding the two aspects under examination.

To construct a global satisfaction indicator, one for each aspect under examination, the NPC ranking methodology described in the previous section was applied considering the strong satisfaction profile. For the three doctorate areas (ML, EL, ST), Table 3 shows the median values
Table 1. Education/employment relationship.

<table>
<thead>
<tr>
<th>Education/employment relationship*, % very satisfied (% quite satisfied)</th>
<th>MB area (N = 20)</th>
<th>ST area (N = 23)</th>
<th>EL area (N = 23)</th>
<th>Areas total (N = 66)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education/employment coherence</td>
<td>30.0 (35.0)</td>
<td>60.9 (21.7)</td>
<td>91.3 (8.7)</td>
<td>62.1 (21.2)</td>
</tr>
<tr>
<td>Use of education at work</td>
<td>10.0 (45.0)</td>
<td>39.1 (17.4)</td>
<td>65.2 (30.4)</td>
<td>39.4 (30.3)</td>
</tr>
<tr>
<td>Adequacy of education in relation to employment</td>
<td>25.0 (30.0)</td>
<td>26.1 (30.4)</td>
<td>56.5 (34.8)</td>
<td>36.4 (31.8)</td>
</tr>
</tbody>
</table>

*Those in training, with study grants or unpaid collaborations, are excluded.

Table 2. Employment expectations and opportunities.

<table>
<thead>
<tr>
<th>Employment expectations and opportunities, % very satisfied (% quite satisfied)</th>
<th>MB area (N = 41)</th>
<th>ST area (N = 41)</th>
<th>EL area (N = 32)</th>
<th>Areas’ total (N = 114)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Opportunities in the academic field</td>
<td>4.9 (22.0)</td>
<td>17.1 (22.0)</td>
<td>31.3 (43.8)</td>
<td>16.7 (28.1)</td>
</tr>
<tr>
<td>Opportunities in the labour market</td>
<td>4.9 (22.0)</td>
<td>4.9 (41.5)</td>
<td>3.1 (50.0)</td>
<td>4.4 (36.8)</td>
</tr>
<tr>
<td>Openings in the scientific community</td>
<td>31.7 (41.5)</td>
<td>31.7 (51.2)</td>
<td>37.5 (53.1)</td>
<td>33.3 (48.3)</td>
</tr>
</tbody>
</table>

Table 3. NPC ranking.

<table>
<thead>
<tr>
<th>Global index of satisfaction, median (first quartile–third quartile)</th>
<th>MB area</th>
<th>ST area</th>
<th>EL area</th>
<th>Areas’ total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education/employment relationship: MB, N = 20; ST, N = 23; EL, N = 23</td>
<td>0.34 (0.12–0.53)</td>
<td>0.47 (0.37–0.63)</td>
<td>0.66 (0.53–0.79)</td>
<td>0.53 (0.36–0.66)</td>
</tr>
<tr>
<td>Employment expectations and opportunities: MB, N = 41; ST, N = 41; EL, N = 32</td>
<td>0.26 (0.20–0.41)</td>
<td>0.32 (0.26–0.47)</td>
<td>0.41 (0.30–0.50)</td>
<td>0.32 (0.26–0.46)</td>
</tr>
</tbody>
</table>

and the interquartile range of the global satisfaction score expressed on a scale of 0–1 in relation to the three aspects regarding the education/employment relationship and the three aspects regarding prospects offered by the doctorate. The box plots shown in Figures 4 and 5 illustrate the distribution of the two global satisfaction indices in the three groups considered.

Comparison of the survey’s three PhD areas (EL, ST, MB) carried out using the non-parametric procedure discussed in Section 3.1 represented a significant aspect of the gathered data analysis. It aimed to highlight any differences in the interviewed Post-Docs’ multivariate and univariate satisfaction profiles. It should be pointed out, however, that subsequent interpretation of the different does require adequate caution, particularly due to the fact that the analyzed variables represent PhD course quality indicators. The variables’ subjective nature and the survey’s observational
context do in fact mean that particular care is required when interpreting the results of the statistical analysis. Post-Docs from the humanities and EL areas are generally expected to be less critical than colleagues from the other areas, whatever the objective quality of the PhD courses. Comparison results therefore require careful consideration by those running the PhD courses in order to identify appropriate means of interpretation.

Considering the information about the external effectiveness of PhDs taken from a combination of questions regarding coherence between study and employment, use of competencies acquired during studies, and adequacy of education in relation to employment, the high satisfaction field (very/quite satisfied) is characterized by differences between the areas, with a rising trend as we move from MB to ST and EL (Table 1). More positive opinions are found in relation to coherence (62.1% very satisfied and 21.2% quite satisfied), less positive for the level of use of acquired techniques (69.7% very/quite satisfied), and adequacy of education (31.8% not satisfied). Differences among the areas are also highlighted by the distribution of the global satisfaction index expressed on a scale of 0–1 (Table 3, Figure 4). In the EL area, in addition to a shifting of the distribution towards high values of satisfaction, there is less variability of scores, particularly when compared with the MB area.

Contributing to further qualification of the data regarding certain aspects of the PhD’s external effectiveness are results from the evaluation of satisfaction levels with regard to prospects offered by the PhD, i.e. opportunities in the academic field or labour market or openings in the scientific community (Table 2).
The low percentages of persons who consider themselves to be very satisfied with these aspects display a certain lack of confidence in the absorption ability of both the university (55.2% not very/not at all satisfied) and even more so the labour market in general (58.8% not very/not at all satisfied). A major feature in the differentiation of satisfaction profiles are opinions on the possibility of following an academic career, considered more realistic by Post-Docs in the EL area. Distribution of the global satisfaction index expressed on a scale of 0–1 (Table 3, Figure 5) stands level with relatively low values for all three areas, which also highlights substantial homogeneity in Post-Doc perceptions.

4. Two-sided non-parametric tests for heterogeneity comparisons

4.1 Methodology

It is often of interest to compare the concentration of two distributions. For example, the so-called “external effectiveness” of two different doctorate courses can be assessed by comparing the heterogeneity of the range of job opportunities offered by each. Obviously, the most effective course is the one that offers the widest and therefore most heterogeneous range of job opportunities. In this matter, from a statistical point of view, job opportunities represent the nominal categorical response variable, and the two populations consist of all students who have gained a doctorate in one of the two courses. A sample survey aimed at determining the type of job PhD graduates get as their first post can provide the initial data useful in resolving the above issue.

As far as the methodology is concerned, heterogeneity has been dealt with mostly from a descriptive point of view [7,8,10,11,13,15,17], i.e. with the aim of determining those indicators capable of providing appropriate measurements of the degree of heterogeneity of a distribution. The aim here is to establish an inferential procedure that allows for the solution of the above-mentioned issues and the goal of which is to compare the heterogeneity of two populations on the basis of sampling data [4].

The method we propose consists of determining suitable test statistics and a general methodology based on the ordering of probabilities of various categories and on a non-parametric test that, for the one-sided test, is similar to the one proposed by Pesarin [14] for issues of stochastic dominance.

The concept of heterogeneity is mostly used in descriptive statistics. Homogeneity notoriously means a statistical phenomenon’s aptitude for always being manifested in the same class or category. Heterogeneity is characterized by the absence of homogeneity and is at a maximum if the variable is equally distributed among all the categories.

From here on, we shall take the inferential problem into consideration. This involves comparing the sample heterogeneity of a categorical variable $X$ in two populations, i.e. testing, in light of two sample data, the hypothesis that heterogeneity in one population is greater than in the other (one-sided test) or the hypothesis that two heterogeneities are not equal (two-sided test). To this end, let $X_{jh}, h = 1, \ldots, n_j$, be independent and identically distributed sample data of size $n_j \geq 1$ from the $j$th population, $j = 1, 2$. It is assumed that the values of random variable $X$ can fall within one of the $k$ categories $A_1, A_2, \ldots, A_k$ with probability distribution $\{p_i \geq 0, \Sigma_i p_i = 1, i = 1, 2, \ldots, k\}$.

From a formal point of view, given two populations $X_1$ and $X_2$, if Het$(X_j)$ indicates the degree of heterogeneity of population $X_j$ $(j = 1, 2)$, the problem of hypothesis testing with one-sided alternatives can be expressed as:

$$H_0 : \text{Het}(X_1) = \text{Het}(X_2)$$

against

$$H_1 : \text{Het}(X_1) > \text{Het}(X_2).$$
Let us take the following normalized indices into consideration: Gini’s $G = \sum_i p_i (1 - p_i)/(1 - 1/k)$ [8], Shannon’s $S = -\sum_i p_i \log p_i / \log k$ [17], Rényi’s for $\alpha = 3$ and $\alpha \to \infty$, $R_3 = -\log(\sum_i p_i^3)/[2 \log(k)]$ and $R_\infty = -\log \left[\sup_i (p_i)/\log(k)\right]$, respectively [16], and Frosini’s $F = 1 - [(k(k-1)\sum_i (p_i - 1/k)^2)^{1/2}$ [7]. The choice of $R_3$ over $R_2$, which is perhaps Rényi’s most used index in the literature, is dictated by the fact that $R_2$ is one-to-one related with $G$, and therefore the two indices imply the same inferential conclusions when applying permutation test theory and methods, given they are permutationally equivalent. Similarly, Frosini’s index of heterogeneity is a monotonic transformation of Gini’s, because $F = 1 - (1 - G)^{1/2}$; hence, they are permutationally equivalent.

Let $p_{(i)}$, $i = 1, 2, \ldots, k$, indicate the underlying parameters of $X$ arranged in non-increasing order: $p_{(1)} \geq p_{(2)} \geq \cdots \geq p_{(k)}$. We note that indices $G$, $S$, $R_3$ and $R_\infty$ are order-invariant, i.e. their values do not change if they are calculated with ordered parameters $p_{(i)}$ in place of proper parameters $p_i$. If then $p_{j(i)}$, $i = 1, 2, \ldots, k$, indicates the ordered probabilities for population $j, j = 1, 2$, the fact that the indices of heterogeneity are order-invariant allows us to express heterogeneity through ordered parameters. To this end, let us observe that: two populations where $\{p_{1(i)} = p_{2(i)}, i = 1, 2, \ldots, k\}$, i.e. with the same ordered distribution, are equally heterogeneous. Moreover, if $\{p_{1(i)} = p_{2(i)}, i = 1, 2, \ldots, k\}$, then data of the two samples are exchangeable and so the permutation testing principle applies.

The null hypothesis can then be equivalently written as:

$$H_0 : p_{1(i)} = p_{2(i)}, \quad \text{for } i = 1, 2, \ldots, k.$$ 

As far as we are concerned, in differences of heterogeneity it is reasonable to take the difference of two sampling indices into consideration as a test statistic:

$$T_I = I_1 - I_2,$$

where index $I$ stands for $G$, $S$, $R_3$ and $R_\infty$, and, of course, $I_j$ indicates the sampling value of $I$ calculated for population $j, j = 1, 2$. Clearly, the tests will be significant for large values, i.e. large values observed in the test statistic can lead to the rejection of the null hypothesis in favour of the alternative. In order to apply the tests using the usual approach, it is necessary to make reference to their sample null distributions. These can be known in principle if the set of underlying null parameters $(p_{(1)}, p_{(2)}, \ldots, p_{(k)})'$ were known. However, since in practice, this knowledge is not available, we must act subject to a proper sampling estimate under $H_0$ of the marginal vector of ordered probabilities. That is, we must consider $\hat{p}_{j(i)} = f_{j(i)} = n_{j(i)}/n_j, j = 1, 2, i = 1, \ldots, n_j$, because the vectors of the probabilities $(p_1, p_{2}, \ldots, p_{k})'$ as well as those of the ordered probabilities $(p_{(1)}, p_{(2)}, \ldots, p_{(k)})', j = 1, 2$, are unknown. Indeed, this question is not easy to compute exactly. Therefore, instead of the true ordering of unknown parameters $(p_{(1)}, p_{(2)}, \ldots, p_{(k)}); j = 1, 2$, we use its estimate (data-driven solution) based on ordering the observed frequencies (empirical ordering):

$$f_{j(1)} \geq f_{j(2)} \geq \cdots \geq f_{j(k)} \iff \hat{p}_{j(1)} \geq \hat{p}_{j(2)} \geq \cdots \geq \hat{p}_{j(k)}, \quad j = 1, 2,$$

thus obtaining the following ordered table.

We observe that the order is achieved separately for each sample, and as it is based on frequencies rather than classes, it may be that the $i$th column of Table 4 refers to two diverse classes for the two samples. In other words class $(i)$ corresponds to the class whose observed frequency occupies the $i$th position in the ordered sequence and can be different for the two samples. Obviously, the order imposed by the frequencies presents a random component and may vary depending on sampling variations. Therefore, under $H_0$, data are not exactly exchangeable as they would be if the true order of population parameters were known and used. The exchangeability property can only be
Table 4. Probabilities ordered by frequencies.

<table>
<thead>
<tr>
<th>Population</th>
<th>Classes (1)</th>
<th>Classes (2)</th>
<th>⋯</th>
<th>Classes (k)</th>
<th>Sampling dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>n₁(1)</td>
<td>n₁(2)</td>
<td>⋯</td>
<td>n₁(k)</td>
<td>n₁</td>
</tr>
<tr>
<td>X₂</td>
<td>n₂(1)</td>
<td>n₂(2)</td>
<td>⋯</td>
<td>n₂(k)</td>
<td>n₂</td>
</tr>
<tr>
<td></td>
<td>nₙ(1)</td>
<td>nₙ(2)</td>
<td>⋯</td>
<td>nₙ(k)</td>
<td>N</td>
</tr>
</tbody>
</table>

obtained asymptotically. Permutation solutions are therefore approximate for finite sample sizes and asymptotically exact.

Using the data in Table 4, the observed value of test statistic $T^o_I$ is calculated. For each permutation of the data set, a new permuted table is obtained (as in Table 5), with different values from those of the observed table but with fixed marginal frequencies.

Using data from the permuted Table in the calculations of the test statistic, the permutation values $T^*_I$ are obtained. Calculating the values that can be obtained making all possible permutations, the permutation distribution of each test statistic is obtained. Alternatively, it is possible to extract a random sample from the set of all permutations, thus obtaining conditional Monte Carlo estimates.

In this way, it is possible to estimate the $p$-value according to the following formula, where $B$ is the number of considered permutations:

$$\hat{\lambda}_I = \frac{\#(T^*_I \geq T^o_I|X)}{B},$$

for the test based on index $I$, where $\#(T^*_I \geq T^o_I|X)$ indicates the number of times permutation values are not lower than the observed value, conditionally on the given data set $X$. The dependency on $X$ is equivalent to the dependency on the space generated by all the possible permutations of $X$, namely the orbit associated with $X$. Therefore, according to the tests’ general deciding rule, if the $p$-value is less than or equal to the fixed significance level, the null hypothesis is rejected in favour of the alternative, otherwise it cannot be rejected.

With a two-sided alternative hypothesis, the problem can be expressed as follows:

$$H_0 : \text{Het}(X_1) = \text{Het}(X_2)$$

against

$$H_1 : \text{Het}(X_1) \neq \text{Het}(X_2).$$

The test statistic can be based on the absolute value or on the square of difference of the sampling indices:

$$T^2_I = (I_1 - I_2)^2,$$

and the testing procedure is the same as for the one-sided test.

Table 5. Absolute frequencies after permutation of data.

<table>
<thead>
<tr>
<th>Population</th>
<th>Classes (1)</th>
<th>Classes (2)</th>
<th>⋯</th>
<th>Classes (k)</th>
<th>Sampling dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>X₁</td>
<td>n₁*(1)</td>
<td>n₁*(2)</td>
<td>⋯</td>
<td>n₁*(k)</td>
<td>n₁</td>
</tr>
<tr>
<td>X₂</td>
<td>n₂*(1)</td>
<td>n₂*(2)</td>
<td>⋯</td>
<td>n₂*(k)</td>
<td>n₂</td>
</tr>
<tr>
<td></td>
<td>nₙ*(1)</td>
<td>nₙ*(2)</td>
<td>⋯</td>
<td>nₙ*(k)</td>
<td>N</td>
</tr>
</tbody>
</table>
Table 6. Absolute frequencies of the “satisfaction with adequacy of education in relation to employment” variable.

<table>
<thead>
<tr>
<th>Satisfaction with adequacy of education in relation to employment</th>
<th>Around 1 year after graduation</th>
<th>Around 3 years after graduation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2004</td>
<td>2004</td>
</tr>
<tr>
<td>Not at all</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Not very</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Quite</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Very</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
<td>17</td>
</tr>
</tbody>
</table>

4.2 Comparative evaluation of Post-Doc satisfaction with education/employment relationship

Here we consider the data found in Table 6, regarding the distribution of absolute frequencies of the “satisfaction with adequacy of education in relation to employment” variable for the cohorts who graduated in 2004 and 2006 (survey 2007) and in 2001 and 2003 (survey 2004). It is of interest to compare the heterogeneity of opinions of PhD graduates interviewed in 2004 and 2007. To have comparability, the analysis is performed within each cohort: “around 1 year after graduation” and “around 3 years after graduation.”

The hypothesis testing problem is the following:

\[ H_0 : \text{Heterogeneity}(2004) = \text{Heterogeneity}(2007) \]
\[ H_1 : \text{Heterogeneity}(2004) < \text{Heterogeneity}(2007). \]

In order to answer the question, for each cohort, we apply a two-sample heterogeneity test for one-sided alternatives.

Performing four tests, \( T_S, T_G, T_{R3}, \) and \( T_{R\infty} \), using \( B = 50,000 \) conditional Monte Carlo simulations applied to the data in Table 6, in the “around 1 year after graduation” case, we obtain the following \( p \)-values:

\[ \hat{\lambda}_S = 0.2244 \text{ for Shannon’s } T_S, \hat{\lambda}_G = 0.1626 \text{ for Gini’s } T_G, \hat{\lambda}_{R3} = 0.1506 \text{ for Rényi’s } T_{R3}, \text{ and } \hat{\lambda}_{R\infty} = 0.1774 \text{ for Rényi’s } T_{R\infty}; \]

in the cohort “around 3 years after graduation”, we obtain the following \( p \)-values:

\[ \hat{\lambda}_S = 0.0608 \text{ for Shannon’s } T_S, \hat{\lambda}_G = 0.0563 \text{ for Gini’s } T_G, \hat{\lambda}_{R3} = 0.0460 \text{ for Rényi’s } T_{R3}, \text{ and } \hat{\lambda}_{R\infty} = 0.0871 \text{ for Rényi’s } T_{R\infty}. \]

In this case, if \( \alpha = 0.10 \), the four tests lead to rejection of the null hypothesis only for the second cohort, and so, in light of our results, we can say that the heterogeneity of opinions of PhD graduates interviewed 3 years after graduation in 2007 is significantly greater than that of PhD graduates of the same cohort interviewed in 2004. In any case, we cannot say the same thing around 1 year after graduation.

5. Concluding remarks

Indications as to the external effectiveness of PhDs, understood to mean the usefulness and usability of a PhD qualification for insertion into the academic or employment field, can be derived from a combination of questions regarding coherence between employment and study and the adequacy of education in relation to employment. On the whole, the satisfaction field (very/quite satisfied) was high in the 2004 survey, but less favourable in the 2007 survey. More positive opinions are found in relation to coherence. Room for improvement can also be identified with regard to adequacy of education (45% not at all or not very satisfied in 2007).

Indications can be taken from suggestions made by the interviewed Post-Docs for improvements to certain aspects of the PhD course: teaching improvements in particular were requested,
indicating the need for more specific, high-level teaching and more structured teaching methods with tests and exams. The second most recurring theme concerned greater involvement of teachers and course organizers and better links with the academic staff. In third place were suggestions for improvements to research activities through, for example, participation in workgroups, research projects set up at the outset of the PhD, better exchanges with other PhD students, writing courses for projects, and greater support with publications; comments were frequently made about increasing foreign exchanges and making them obligatory. Other suggestions concerned improvements to structures, funds, future prospects, administrative support, and the PhD student selection criteria.

Armed with this information, the organization of better structured education programs in the PhD field would seem to be desirable, at the same time encouraging awareness of the qualification outside academic circles. The organization of PhDs into PhD Colleges as well as improvements to the teaching program is a move in the right direction, with one objective being a more organic and wide-ranging relationship between university PhDs and outside professional and research environments. Publicizing the PhD qualification should therefore be accompanied by specific partnership initiatives between academic and non-academic fields. In a recent paper, the NCEUS actually highlighted the fact that finding employment is still an uncertainty for PhD graduates. The percentage of Italy’s workforce represented by scientists and engineers, a datum that implicitly represents the production system’s technology level, is one of the lowest among major European countries [12] highlighting the fact that companies lack experience when it comes to employing highly qualified personnel.

Considering the information on the external effectiveness of PhDs derived from a combination of questions regarding coherence between study and employment, use of competencies acquired during studies, and the adequacy of education in relation to employment, satisfaction is characterized by differences between the areas, with a rising trend as we move from MB to ST and EL. More positive opinions are found in relation to coherence and less positive for the level of use of acquired techniques and adequacy of education.

Considering satisfaction judgments about employment expectations and opportunities, there is generally a fairly good opinion about openings in the scientific community. Only Post-Docs of the EL area are moderately satisfied with opportunities in the academic field, while satisfaction with opportunities in the labour market is quite low for all the interviewed PhDs.

The developed global satisfaction index allowed us to obtain a comprehensible reduction in the observed data and to conclude that the global satisfaction of EL Post-Docs with the education/employment relationship is globally greater than that of the other areas and that MB Post-Docs are the most unsatisfied. Taking employment expectations and opportunities into account, lower satisfaction and less variability in scores among areas are found.

This paper also presented a proposal that consists of finding appropriate test statistics and a general methodology of hypothesis testing based on the ordering of probabilities, in which the objective is to compare the heterogeneity of two populations on the basis of sampling data. The test statistic consists of the comparison of the sampling indices of heterogeneity calculated for two samples and can vary according to the index of heterogeneity considered. The fact that the probabilities of the two compared distributions are unknown parameters, and therefore the ordering of probabilities can only be estimated on the basis of sampling data, implies that the proposed solutions are approximate. The observed relative frequencies were used as estimates of the probabilities. The choice of using the non-parametric test proves to be both practical and efficient. It is easy to apply and requires few, weak assumptions without knowing the distribution of either the data or the test statistics. Application of the test to the data of the Post-Doc surveys carried out in Ferrara leads to the conclusion that the heterogeneity of opinions of PhD graduates 3 years after graduation in 2007 is significantly greater than that of PhD graduates of the same cohort in 2004. This is not true 1 year after graduation.
A simulation study made it possible to assess both the degree of approximation in the null hypothesis and the power behaviour of the proposed non-parametric tests of heterogeneity [4]. The rejection rates increase as the homogeneity of distributions increases. Among the test statistics considered, the one based on Rényi’s index of order 3 seems to register higher rejection rates under $H_0$ but slightly higher power under $H_1$. The rejection rates under the alternative hypothesis are in any case satisfactory for all the considered tests.

References


[12] MIUR (Ministero dell’Istruzione, dell’Università e della Ricerca) (Ministry of Education, University and Research), Lo stato della ricerca nelle Università (The state of research in Italian Universities), Università obiettivo valutazione, 1 Analisi e commenti, Atenei, 2003, pp. 89–113.


