Academic performance of students from entrance to graduation via quasi U-statistics: a study at a Brazilian research university

Rafael Pimentel Maia, Hildete Prisco Pinheiro* and Aluísio Pinheiro

Department of Statistics, University of Campinas, Campinas, Brazil

(Received 20 May 2014; accepted 27 July 2015)

We present novel methodology to assess undergraduate students’ performance. Emphasis is given to potential dissimilar behaviors due to high school background and gender. The proposed method is based on measures of diversity and on the decomposability of quasi U-statistics to define average distances between and within groups. One advantage of the new method over the classical analysis of variance is its robustness to distributional deviation from the normality. Moreover, compared with other nonparametric methods, it also includes tests for interaction effects which are not rank transform procedures. The variance of the test statistic is estimated by jackknife and p-values are computed using its asymptotic distribution. A college education performance data is analyzed. The data set is formed by students who entered in the University of Campinas, Brazil, between 1997 and 2000. Their academic performance has been recorded until graduation or drop-out. The classical ANOVA points to significant effects of gender, type of high school and working status. However, the residual analysis indicates a highly significant deviation from normality. The quasi U-statistics nonparametric tests proposed here present significant effect of interaction between type of high school and gender but did not present a significant effect of working status. The proposed nonparametric method also results in smaller error variances, illustrating its robustness against model misspecification.

Keywords: asymptotic theory; diversity measures; jackknife; nonparametric methods; interaction effect

1. Introduction

Assessment of undergraduate performance from entrance to graduation has been of great interest in the literature ([7,12,14,19,20] and references therein). For instance, since the implementation of a quota system in the Public Federal Universities in Brazil, the interest in the performance of undergraduate students has grown. The law states that (it started in 2013) all the courses/majors in Public Federal Universities must have at least 12.5% of their positions/spaces reserved to students who studied all High School years in public high schools (PuHS). In addition, this percentage should gradually get to 50% in 2017. It is important to point out that most of the

*Corresponding author. Email: hildete@ime.unicamp.br

© 2015 Taylor & Francis
middle-class students go to Private Schools (Elementary, Middle and High Schools) and there is a great competition to enter Public Universities. Therefore, the socioeconomic status is being measured indirectly by looking at the High School system of the students. There are highly competitive exams to get into the Public Universities in Brazil. The State Universities in São Paulo, Brazil, do not have a quota system, but they are implementing some affirmative action programs. The State University of Campinas (Unicamp), located in the state of São Paulo, is one of the top universities in Brazil and since 2005 implemented an affirmative action program, which allows students who studied all High School years in Public schools to receive a bonus in their final scores of Unicamp’s highly selective Entrance Exam. Unicamp has an average of over 15 candidates per undergraduate position offered each year (www.comvest.unicamp.br). For instance, the affirmative action program stated in 2005 at Unicamp gives a bonus of 30 points in their Entrance Exam Score (EES) for students coming from PuHS and 10 more points for those PuHS students who declared themselves as Black, Pardo1 or Indian. Note that indirectly the affirmative action program is adjusting for income level and race.

Most of the studies were based on classical analysis of variance or correlation analysis. In some cases when the normality assumption is not met, nonparametric tests based on ranks have been used [19], such as Kruskal-Wallis test for one-way analysis of variance, Friedman test for complete block designs or Durbin test for incomplete block design and others [4,6,8]. However, none of these tests accommodate interaction effect hypotheses. In this paper we propose a nonparametric methodology which allows the analysis of non-normal or unspecified distribution data as well as normal or other known distribution data. Furthermore, we add the possibility of interaction effects, which is not possible with the most popular nonparametric approaches. The proposed test statistics are asymptotically normal under both null and alternative hypotheses for a large class of models [16,17]. Moreover, here we extend the results presented in [1] for homogeneity tests based on quasi $U$-statistics by including tests for interaction effects.

Pedrosa et al. [14] proposes regression models to assess the performance of undergraduate students using as response variable the relative gain which is based on the relative rank of final (or last) recorded GPA (grade point average) and the entrance exam grade rank. The relative gain, proposed by Pedrosa et al. [14], is defined as follows. The students of each course/major and who entered in the same year (same ‘class’, say) are ranked twice. The student’s first rank is based on the EES; the second rank is based on the final (or last) GPA score. For instance, the student with the worst EES or worst GPA will be assigned rank 1, while the one with the highest score will have rank $n$, and so on. We thus have an initial and final rank for each student. The ranks are then divided by the total number of students in the same ‘class’ (same major, entering in the same year). The relative gain is then obtained by the difference between the final and the initial relative ranks. Therefore, the relative gain is a continuous variable symmetric around zero, limited between $-1$ and 1 and the tails of its distribution are lighter than the normal distribution, that is, its distribution is leptokurtic. Even for large sample sizes, the normality approximation may not be appropriate.

We pursue more robust methods for cases which can be applied for small or moderate sample sizes as well as any in which a normal assumption is not reasonable. Robust methods is key to evaluate the performance of students, since GPA scores can vary a lot from one area to another, for example, from Exact Sciences to Social Sciences, for instance. Even in the same course, we may have very different grades according to the class/lecturer. So, it is crucial to employ more robust measures of performance.

The proposed method is based on measures of diversity and decomposability of quasi $U$-statistics [16,17] to properly decompose between and within group distances. Furthermore, the method allows us to look at interaction effects by extending the decomposition of diversity measures using the idea in [18]. The main emphasis here is given to the sector of high school education from which college students come – private or public. A homogeneity test is proposed
for group comparisons using a nonparametric approach. The data comes from the State University of Campinas (Unicamp) as well as the data in [12]. We further the analysis in [12] in two ways. We present here a thorough analysis for the consolidated data set. Moreover, the analysis is more inferential in opposition to the more descriptive analysis on the preliminary data [12]. The data set consists of students who have enrolled at Unicamp from 1997 to 2000 with 76.8% who have already graduated their courses and 23.2% who have dropped out from the University.

In Section 2, we present a short introduction about diversity measures, the development of a hypothesis test to evaluate the homogeneity between groups and the extension of the procedure to a multifactor design, which can be a parametric or a nonparametric approach based on quasi $U$-statistics. The decomposition of diversity measures based on quasi $U$-statistics presented so far in the literature [1,16, 17] is analogous to a one way ANOVA. Here, we extend the quasi $U$-statistic method to multifactors as in [18]. In Section 3, we present the data that motivated the study. Section 4 presents an application to the Unicamp database with all the students (those who graduated and those who dropped out) as well as an application with only those who graduated. A discussion follows in Section 5.

2. Quasi U-statistics methods

2.1 Measures of diversity and decompositions

Several metrics have been constructed to measure distances in qualitative or quantitative variables ([2,9,11] and references therein). Measures of diversity have been widely used to measure variability in ecology, genetics, physics and many areas. Measures of diversity can be used to decompose the total diversity into within and between groups diversity due to a certain number of factors [18]. When we have a mixture of groups, one can be interested in knowing if the amount of diversity is due to the difference within or between groups. The classical ANOVA case may be thought as a special case of decomposition of measures of diversity. For instance, the total sample variance can be written as

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \left( \frac{n}{2} \right)^{-1} \sum_{i<j} (x_i - x_j)^2 / 2,$$

the mean of all the pairwise comparisons of half of the squared difference between $x_i$ and $x_j$, which in other words is a U-statistics of degree 2 with kernel $\phi(x_i, x_j) = (x_i - x_j)^2 / 2$ [5,10]. In the classical ANOVA case, the $X_i$’s are assumed to be independent with normal distribution, leading the distribution of the between and within mean sum of squares being chi-square and the ratio of them having an $F$-Snedecor distribution.

Here we will use a more general approach for which we do not need an assumption for the distribution of the $X_i$’s. Consider a symmetric and non-negative function $\phi(X_i, X_j)$, which is a measure of the difference between two individuals. Then define, for $g \neq g’ = 1, \ldots, G$,

$$E[\phi(X_i^g, X_j^g)] = \theta_g \quad \text{and} \quad E[\phi(X_i^g, X_j^{g’})] = \theta_{gg’},$$

where $\theta_g$ is the diversity within group $g$ and $\theta_{gg’}$ is the diversity between groups $g$ and $g’$. Estimators of $\theta_g$ and $\theta_{gg’}$ can be found using U-statistics [5]. If $\phi(x_i, x_j)$ is a convex function, then

$$\theta_{gg’} \geq \frac{\theta_g + \theta_{g’}}{2}. \quad (1)$$

The excess dissimilarity measure between groups $g$ and $g’$ is given by

$$D_{gg’} = \theta_{gg’} - \frac{1}{2}[\theta_g + \theta_{g’}], \quad (2)$$
that is, $D_{gg'}$ is the excess measure of diversity between groups $g$ and $g'$ compared to the average of the within diversities measures in groups $g$ and $g'$.

Additionally, the estimators of $\theta_g$ and $\theta_{gg'}$, for $g' \neq g = 1, \ldots, G$, are given by

$$\tilde{D}_g \equiv \hat{\theta}_g = \left( \frac{n_g}{2} \right)^{-1} \sum_{i<j} \phi(x_i^g, x_j^g),$$

and

$$\tilde{D}_{gg'} \equiv \hat{\theta}_{gg'} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} \phi(x_i^g, x_j^{g'}),$$

where $n_g$ is the number of individuals at the $g$th group. Note that $\hat{\theta}_g$ is a U-statistic of degree 2 and $\hat{\theta}_{gg'}$ is a two-sample U-statistic of degree (1,1), so that $E(\tilde{D}_g) = \theta_g$ and $E(\tilde{D}_{gg'}) = \theta_{gg'}$.

The overall mean distance is defined as the total variability of the pooled sample and it can be estimated by

$$U_n = \sum_{i<j} \left( \frac{n}{2} \right)^{-1} \phi(x_i, x_j), \quad (3)$$

with $n = \sum_{g=1}^{G} n_g$. As shown in [15,16], we can decompose $U_n$ as

$$U_n = \left( \frac{n}{2} \right)^{-1} \left[ \sum_{g=1}^{G} \left( \frac{n_g}{2} \right) \tilde{D}_g + \sum_{g < g'} n_g n_{g'} \tilde{D}_{gg'} \right] = W_n + B_n,$$

where

$$W_n = \frac{1}{n} \sum_{g=1}^{G} n_g \tilde{D}_g,$$

and

$$B_n = \frac{1}{n(n-1)} \sum_{g < g'} n_g n_{g'} [2\tilde{D}_{gg'} - \tilde{D}_g - \tilde{D}_{g'}]. \quad (4)$$

Note that $W_n > 0$, so that

$$E[W_n] = \frac{1}{n} \sum_{g=1}^{G} n_g \theta_g > 0,$$

but $B_n$ may assume any real value and

$$E[B_n] = \frac{1}{n(n-1)} \sum_{g < g'} n_g n_{g'} (2\theta_{gg'} - \theta_g - \theta_{g'}) \geq 0,$$

by Equation (1).

Here we use $\phi(X_i, X_j) = (X_i - X_j)^2$, with $X_i$ and $X_j$ being the relative gain of students $i$ and $j$, respectively.
2.2 Hypothesis testing

We can test homogeneity between groups using the individual performance data. Intuitively, we can say that, under $H_0$: $D_{gg'} = 0$, $\forall \ 1 \leq g < g' \leq G$, and given Equations (1) and (2), we can write these hypotheses as

$$H_0 : 2\theta_{gg'} = \theta_g + \theta_{g'} \ \forall 1 \leq g < g' \leq G \ \text{vs.} \ H_1 : 2\theta_{gg'} > \theta_g + \theta_{g'} \ \text{for some} \ g \neq g'.$$

The average $(\theta_g + \theta_{g'})/2$ can be thought as a baseline. Under the null hypothesis, $\theta_{gg'} = (\theta_g + \theta_{g'})/2$, which means that the diversity between groups is not greater than the baseline, which means that the excess diversity between groups is zero, that is, $D_{gg'} = 0$. Large sample values of $B_n$ are indications of large values of $D_{gg'}$. Thence, one rejects $H_0$ for large values of $B_n$. Under mild regularity conditions $B_n$ is asymptotically normal [16,17].

Let $\phi_1(x) = E[\phi(X_i, X_j) \mid X_i = x]$ and $\phi_0 = E[\phi(X_i, X_j)]$. Then, write

$$\phi(X_i, X_j) = \phi_0 + \{\phi_1(X_i) - \phi_0\} + \{\phi_1(X_j) - \phi_0\} + \{\phi(X_i, X_j) - \phi_1(X_i) - \phi_1(X_j) + \phi_0\}$$

$$= \phi_0 + \psi_1(X_i) + \psi_1(X_j) + \psi_2(X_i, X_j),$$

which is the Hoeffding decomposition of $B_n$’s kernel $\phi$.

Note that, under $H_0$, we can rewrite $B_n$ as

$$B_n = \left(\frac{n}{2}\right)^{-1} \sum_{1 \leq i < j \leq n} \eta_{nij}\psi_2(X_i, X_j),$$

where $X_1, \ldots, X_n$ represent the ordered pooled sample, in which the first $n_1$ observations relate to group 1, the next $n_2$ to group 2 and so on; and

$$\eta_{nij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ come from different groups}, \\ - \frac{(n - n_g)}{(n_g - 1)} & \text{if } i \text{ and } j \text{ are both from the same group}. \end{cases}$$

Pinheiro et al. [16] showed that, under the null hypothesis of homogeneity between groups the asymptotic distribution of $B_n$ is normal, i.e.

$$(M_n V_n^*)^{-1/2}B_n \xrightarrow{D} N(0, 1) \ \text{as} \ n \to \infty, \quad (5)$$

where

$$M_n = \frac{(n - 1)}{2} \left\{ n(G - 1) + \sum_{g=1}^{G} \frac{(n - n_g)}{(n_g - 1)} \right\} = O(n^2),$$

$$V_n^* = U_n^{(2,2)} - U_n^{(3)}, \quad U_n^{(2,2)} = \left(\frac{n}{2}\right)^{-1} \sum_{1 \leq i < j \leq n} \phi_2(X_i, X_j),$$

and

$$U_n^{(3)} = \frac{1}{n(n - 1)(n - 2)} \sum_{i \neq j \neq l} \phi(X_i, X_j)\phi(X_i, X_l).$$
From the result given in Equation (5), one can also find the power of this test. Note that

\[ E_{H_1}[B_n] = \frac{1}{n(n-1)} \sum_{g < g'} \left( n_g n_{g'} E_{H_1}(2 \bar{D}_{gg'} - \bar{D}_g - \bar{D}_{g'}) \right) \]

\[ = \sum_{g < g'} \frac{n_g}{n} \frac{n_{g'}}{n-1} (2 \theta_{gg'} - \theta_g - \theta_{g'}) \]

\[ \rightarrow \sum_{g < g'} p_g p_{g'} (2 \theta_{gg'} - \theta_g - \theta_{g'}) \equiv \delta_n, \]

where \( p_g = n_g/n \) and \( p_{g'} = n_{g'}/(n-1) \), for all \( g, g' = 1, 2, \ldots, G \). Then,

\[ E_{H_1}[B_n] = \delta_n + O(n^{-2}). \]

Now let \( \delta_n = \Delta/n \). Then, \( \delta_n \rightarrow 0 \) as \( n_0 \rightarrow \infty \) and \( E(nB_n) = O(1) \), with \( n_0 = \min_g \{n_g\} \).

The power of the test will be given by

\[ P_{H_1} \left( \frac{nB_n}{\sqrt{V_n^*}} > q_\alpha \right) \rightarrow 1 - \Phi \left( q_\alpha - \frac{\Delta}{\gamma} \right), \]

where \( E_{H_1}[B_n] - \delta_n \rightarrow 0 \) and \( V_n^* \xrightarrow{p} \gamma^2 \) as \( n_0 \rightarrow \infty \).

The classical tests (both parametric and nonparametric) assess differences in location. Differences in scale are nuisance to the analysis leading only to loss in statistical power. The quasi-\( U \)-statistics are built upon differences on the distributions. Therefore, both scale and location sample differences contribute to rejecting the null hypothesis.

### 2.3 The multifactors problem

The results shown in the previous sections correspond to a one way ANOVA. Therefore, first in this section we will extend the quasi-\( U \)-statistics method to two factors. For more than two factors, the theory is easily generalized. Consider two factors, \( A_1 \) with \( s \) levels and \( A_2 \) with \( t \) levels, as in [13]. To obtain the combined effect of \( A_1 \) and \( A_2 \), consider the cross-classification of \( A_1 \) and \( A_2 \) as a single factor with \( s \times t \) levels. We can then decompose \( U_n \) by

\[ U_n = W_n(A_1, A_2) + B_n(A_1, A_2), \tag{6} \]

where \( B_n(A_1, A_2) \) is computed as Equation (4) with \( s \times t \) groups. Thus, the expected value for \( B_n(A_1, A_2) \) will be

\[ E[B_n(A_1, A_2)] = \frac{1}{n(n-1)} \sum_{1 \leq k < k' \leq s \times t} n_k n_{k'} (2 \gamma_{kk'} - \gamma_k - \gamma_{k'}), \tag{7} \]

for \( k, k' = 1, \ldots, s \times t \), and such that \( \gamma_{kk'} \) is the diversity between the groups \( k \) and \( k' \) and \( \gamma_k \) is the diversity within the group \( k \). Therefore, the null hypothesis to test the interaction effect between
the factors $A_1$ and $A_2$ would be

$$H_0 : 2\gamma_{kk'} = \gamma_k + \gamma_{k'} \quad \forall 1 \leq k \leq k' \leq s \times t.$$  

Moreover, as in the sum of squares for the classical analysis of variance, $B_n(A_1, A_2)$ can be decomposed as

$$B_n(A_1, A_2) = B_n(A_1) + B_n(A_2 | A_1), \quad (8)$$

with

$$B_n(A_2 | A_1) = B_n(A_1, A_2) - B_n(A_1).$$

Therefore, $B_n(A_2 | A_1)$ can be interpreted as a weighted average of the diversities between the levels of $A_2$ for each level of $A_1$. This represents the proportion of variability not explained by $A_1$ which is explained by $A_2$. If we define the expected value of $B_n(A_1)$ by

$$E[B_n(A_1)] = \frac{1}{n(n-1)} \sum_{1 \leq l < l' \leq s} n_{l} n_{l'} (2\theta_{l'l'} - \theta_l - \theta_{l'}),$$

for $l, l' = 1, \ldots, s$. The null hypothesis to test the effect of $A_2$ given $A_1$ can be defined by

$$H_0 : \sum_{1 \leq k < k' \leq s \times t} n_k n_{k'} (2\gamma_{kk'} - \gamma_k - \gamma_{k'}) = \sum_{1 \leq l < l' \leq s} n_{l} n_{l'} (2\theta_{l'l'} - \theta_l - \theta_{l'}),$$

for $k, k' = 1, \ldots, s \times t$ and $l, l' = 1, \ldots, s$.

For multiple factors, $U_n$ is the total variability, which is a mixture of several subgroups. When there are $k$ factors, $A_1, \ldots, A_k$, $W_n(A_1, \ldots, A_k)$ is the weighted average of the diversities within each group defined by the cross-classification of $A_1, \ldots, A_k$ and

$$B_n(A_1, \ldots, A_k) = U_n - W_n(A_1, \ldots, A_k).$$

Analogously,

$$B_n(A_1, \ldots, A_s | A_{s+1}, \ldots, A_k) = B_n(A_1, \ldots, A_k) - B_n(A_{s+1}, \ldots, A_k).$$

Thus, $B_n(A_i)$ can be interpreted as the main effect of factor $A_i$ for all $i = 1, 2, \ldots, k$, $B_n(A_1, \ldots, A_s)$ is the effect of the interaction or the combined effect of $A_1, \ldots, A_s$ ($s \neq k$), and $B_n(A_i | A_j)$ is the effect of $A_i$ conditioned on $A_j$, that is, it is the effect of $A_i$ taking out the effect of $A_j$.

The method proposed here can be parametric or nonparametric, that is, we may assume a known distribution for $X$ (the relative gain in this particular case) or we can work under a large class of distributions for $X$. In the parametric approach maximum likelihood estimators (MLE) of the moments of $X$ can be used for the estimation of the mean and variance of the test statistic $B_n$. Details and analytical results of the MLE for the normal and triangular distributions are shown in Appendix A of the supplementary material.

For the nonparametric approach, no assumption is made about the distribution of the relative gain, the variance of $B_n$ is estimated using the jackknife variance according to [3] and $p$-values are computed according to the asymptotic distribution given in Equation (5). Details of the algorithm is given in Appendix B of the supplementary material.
3. Database

The data set is composed by 8225 students which have enrolled at Unicamp at years 1997, 1998, 1999 and 2000 in all Bachelor degree’s majors. These students entered the system without any differential treatment due to Race or Public/Private High School system. The academic situation of these students were classified as following: Graduates (6316 students who have already graduated their courses – 76.8%), and Others (1909 students who dropped out from the University – 23.2%). Socio-economic and demographic characteristics of the students were provided by a questionnaire filled out by the students when they registered for their entrance exam. The EESs as well as their final GPA scores were also provided.

The students were, in their majority, between 16 and 24 years old (only 8.1% have more than 24 years of age), both genders (58.6% male and 41.4% female), from all Brazilian regions and enrolled in 45 different majors from the areas of Health Sciences, Engineering and Exact Sciences, Social Sciences and Arts.

Students who declared having studied all or most of their high school years in private schools, were considered coming from private school (PrHS). Analogously, the ones who declared having studied all or most of their high school years in public schools were considered coming from public schools (PuHS). Table 1 shows the distribution according to type of high school and entry year. In total 30.7% of all the students who enrolled between 1997 and 2000 come from PuHS.

There is no information about Race in the database as this question was only introduced in Unicamp’s questionnaire in 2003. Another characteristic of interest evaluated by the models is whether students worked at the time of their enrollment. In the sample 28% of the students declared they worked when enrolled at the university. Within students from PuHS and PrHS, 48.5% and 18.6% declared to be working at enrollment time, respectively. Table 2 presents the distribution of the students according to type of high school by gender and by working status.

The Spearman correlation coefficient between the initial and final relative rank is 0.249, indicating a weak correlation between both relative ranks. Tables 3 and 4 present some descriptive statistics of the relative gain according to gender, type of high school and working status as well as results of two-sample t-tests and Wilcoxon rank sum tests for group comparisons using the whole data set and only those who graduated, respectively. We would like to point out that, even though the standard deviations seem to be equal or very close, in many cases they are not (the equality of variances was statistically rejected in all three cases for Table 3 and for gender in Table 4). So, when performing the two-sample t-tests we use the p-values for unequal variances in Table 3 and for gender in Table 4. The PuHS students present a mean relative gain significantly greater than those who studied in PrHS. When comparing genders, women have a significantly greater relative gain than men. On the other hand, when comparing students who worked to those who did not, we do not observe any difference in the relative gain for the whole data set, but there is a significant result when we look only at the students who graduated, with a greater relative gain for those who did not work.

Table 1. Sampling distribution per year (%) according to type of high school.

| Type of high school | Entrance year |
|---------------------|---------------|
|                     | 1997 | 1998 | 1999 | 2000 |
| Private             | 68.4 | 69.2 | 70.6 | 68.9 |
| Public              | 31.6 | 30.8 | 29.4 | 31.1 |
| Total               | 100.0| 100.0| 100.0| 100.0|

Note: There was missing information for type of high school for 130 students.
Table 2. Sampling distribution according to type of high school, gender and working status from classes 1997 to 2000.

| Gender | Working status | No | Yes | No | Yes | Total |
|--------|----------------|----|-----|----|-----|-------|
|        |                |    |     |    |     |       |
| Type of high school | | | | | | |
| Private |                | 1945 | 392 | 2600 | 648 | 5585 |
|          |                | (24.1%) | (4.9%) | (32.3%) | (8.0%) | (69.3%) |
| Public  |                | 579  | 422 | 696  | 780 | 2477 |
|          |                | (7.2%) | (5.2%) | (8.6%) | (9.7%) | (30.7%) |
| Total   |                | 2524 | 814 | 3296 | 1428 | 8062 |
|          |                | (31.3%) | (10.1%) | (40.9%) | (17.7%) | (100%) |

Note: There were 163 missing values for Type of High School or Working Status.

Table 3. Descriptive statistics and results of statistical tests for the relative gain according to gender, type of high school and working status for all students.

| Group          | Level | N   | Mean  | SD   | t-test | Wilcoxon test |
|----------------|-------|-----|-------|------|--------|---------------|
| Gender         | Female | 3398 | 0.0613 | 0.3461 | < 0.0001 | < 0.0001 |
|                | Male   | 4827 | −0.0433 | 0.3594 |        |               |
| Type of high school* | Private | 5610 | −0.0195 | 0.3511 | < 0.0001 | < 0.0001 |
|                | Public | 2485 | 0.0455 | 0.3674 |        |               |
| Working status* | Yes    | 2279 | −0.0028 | 0.3729 | 0.5442 | 0.4745 |
|                | No     | 5864 | 0.0027 | 0.3508 |        |               |

*There were 130 missing values for type of high school and 82 for working status.

Table 4. Descriptive statistics and results of statistical test for the relative gain according to gender, type of high school and working status (only for the students who graduated).

| Group          | Level | N   | Mean  | SD   | t-test | Wilcoxon test |
|----------------|-------|-----|-------|------|--------|---------------|
| Gender         | Female | 2825 | 0.1061 | 0.3200 | < 0.0001 | < 0.0001 |
|                | Male   | 3387 | 0.0034 | 0.3356 |        |               |
| Type of high school | Private | 4362 | 0.0438 | 0.3250 | < 0.0001 | < 0.0001 |
|                | Public | 1850 | 0.1221 | 0.3368 |        |               |
| Working status  | Yes    | 1533 | 0.0925 | 0.3391 | 0.0007 | 0.0010 |
|                | No     | 4679 | 0.0588 | 0.3271 |        |               |

Note: There were 6212 observations (excluding 104 missing values) for type of high school or working status.

Looking at Figure 1 (the whole data set), it seems that differences in relative gain between types of high schools is slightly greater among women (mean difference = 0.0771) than among men (mean difference = 0.0563), suggesting a possible interaction between gender and type of high school. In Figure 2 (only those who graduated), the situation is reversed. The differences in relative gain between types of high schools is slightly greater among men (mean difference = 0.0823) than among women (mean difference = 0.0722).

In addition, we adjusted a classical linear model for all the students with all three main effects (type of high school, gender and work) and all pairwise interactions. None of the interaction terms were significant and the analysis of variance table for the final model is presented in
Figure 1. Box plots of the relative gain according to gender and type of high school (all students).

Figure 2. Box plots of the relative gain according to gender and type of high school (only for graduated students).

Table 5. The linear model uses dummy variables (Female = 1, Male = 0; PuHS = 1, PrHS = 0 and didnotwork = 1, worked = 0) as codification for the design matrix. According to the parameter estimates, all the coefficients are positive, which means that female students has greater relative gain than male students, students coming from PuHS have greater relative gain than those from PrHS and those who did not work have a greater relative gain than those who worked. When we look at the results of the $t$-tests and Wilcoxon rank sum tests in Table 3, there is no
Table 5. Analysis of Variance (all students).

| Source                | d.f. | Type III SS\(^a\) | F-value | p-value |
|-----------------------|------|--------------------|---------|---------|
| Gender                | 1    | 20.34              | 164.09  | < .0001 |
| Type of high school   | 1    | 8.03               | 64.77   | < .0001 |
| Working status        | 1    | 0.72               | 5.80    | .0160   |

\(^a\)Correspond to the variation attributable to an effect after correcting for any other effects in the model.

Figure 3. Q–Q plot of the residuals for the linear model presented in Table 5.

difference in relative gain between those who worked and those who did not work, whereas in the model it is statistically significant. This may be due to the fact that the model is adjusting for the other effects.

Figure 3 shows the Q–Q plot of the residuals for the model in Table 5 and we can see that the tails of the distribution are significantly lighter than the normal. The test for normality of the residuals rejects the null hypothesis of normality (both tests, Anderson-Darling and Kolmogorov-Smirnov, present \(p\)-values \(< .0001\) ). The test results clearly show that the normality assumption is not satisfied by the data set. This motivates us to pursue the analysis with alternative methods which are robust to distributional deviations from normality, but, at the same time we would like to test main effects and interactions. In the following sections, we present a nonparametric method which attains both robustness and model flexibility.

4. Application

As neither normality nor triangular (−1,0,1) distributions are reasonable for the dataset here (see Figure 4), we will use the nonparametric method, since it is a more robust procedure.
Figure 4. Histogram for the relative gain with the normal and triangular distribution curves.

Table 6. Analysis of diversity for all the students using the nonparametric approach.

| Effects          | $B_n$  | $\hat{SE}$ | $p$-value |
|------------------|--------|------------|-----------|
| SC†              | 0.0017 | 0.0005     | .0002*    |
| GE†              | 0.0049 | 0.0008     | <.0001*   |
| WS†              | −0.00002 | 0.0001  | .3716     |
| SC | GE + WS     | 0.0019 | 0.0005     | <.0001*   |
| GE | SC + WS     | 0.0049 | 0.0008     | <.0001*   |
| WS | SC + GE     | 0.0001 | 0.0002     | .3153     |
| SC | GE         | 0.0018 | 0.0005     | .0001*    |
| GE | SC         | 0.0050 | 0.0008     | <.0001*   |
| SC × GE         | 0.0067 | 0.0009     | <.0001*   |

Notes: *significant on a 5% level test.
†SC, GE and WS mean School, Gender and Working Status, respectively.

Tables 6 and 7 display the estimates of $B_n$, their standard deviations and the $p$-values obtained from the nonparametric method for all the students and only those who graduated, respectively. In Table 6, we observe a significant difference between types of high school ($p$-value = .0002) and genders ($p$-value < .0001). We do not observe significant differences between the students who worked or not ($p$-value = .3716). Using the expansion of quasi $U$-statistics for multifactors (discussed in Section 2.3), which is comparable to the ANOVA results given in Table 5, there are significant effects of gender and type of high school ($p$-values are < .0001) when adjusted for the other variables. On the other hand, we do not find a significant effect of working status adjusted by gender and type of high school ($p$-value = .3153) and there is a significant effect of the interaction between gender and type of high school ($p$-value < .0001), which is different from the conclusions of the classical analysis of variance in Table 5. As observed in Figure 1 (all students), the difference in the relative gain between students who came from PuHS and PrHS, among female students is slightly greater than the difference among male students. As
for Figure 2 (only those who graduated), the situation is reversed, with the difference in relative gain between type of High Schools slightly greater among men than among women (see details in Section 3).

Although there are small numerical differences between Tables 6 (all students) and 7 (only those who graduated), the conclusions in terms of significance of the effects are the same.

Typically, when the assumptions of normality are not met, the tails of the distribution are heavier than the normal and the SE’s, assuming normality, tend to be smaller than they should be. Here the tails are lighter and the empirical distribution are more concentrated than the normal distribution. For this reason the SE’s for the nonparametric method is smaller, showing here an advantage over parametric procedures when the assumptions are not met.

5. Discussion

We present a nonparametric method as alternative to analyze continuous response variables when the assumption of normality is not met and the use of analysis of variance may be compromised. Furthermore, the proposed method, compared with classical nonparametric methods, has the advantage of allowing the analysis of more than two factors and also their interactions, making it more flexible, robust and more competitive to the analysis of variance. It is a general tool that can be applied in many different fields.

We propose the new methodology to analyze the performance of college students from entrance to graduation with a real data set from the University of Campinas, Brazil. The leptokurtic nature of academic performance data makes the robustness of the nonparametric method even more relevant. The difficulties on the analysis of this kind of data are twofold. First of all, given the complexity and diversity of grading systems, the mere use of numerical measures may inadvertently lead to weighting systems biased according to unknown discrepancies among the subjects. These discrepancies may be in scale, in location or in both. The use of ranks helps ameliorate the effects of the grading system. The second problem regards the specification of the underlying distribution. Both normal and triangular can be postulated as feasible options. We amply illustrate that these are not adequate distributions for the data set in question. The nonparametric quasi $U$-statistics decomposition can be used as a solution for unspecified distributions in general.

Finally, the results show that the misspecification of the underlying distribution may lead to wrong conclusions, since the results with the proposed method and the analysis of variance diverged. Although the solutions proposed here demand more computer power than their linear
counterparts, the method can be easily implemented in any computer language, specially in R, and they are within grasp for any researcher on the field.

Acknowledgments

We thank the referees for their valuable comments and the following people for their direct support regarding databases used for this study: Renato L. Pedrosa, Associate Coordinator of the Center for Advanced Studies at Unicamp, Mauricio Kleink, Coordinator of the Admissions Committee (Comvest) at Unicamp, Antonio Faggiani, Coordinator of the Academic Records Division (DAC) at Unicamp, Renato Hirata, database specialist at Comvest, Gleyson R. do Nascimento and Flavio Ferraresi, trainees at Comvest and Silvio de Souza, database specialist at DAC.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

R. P. Maia was supported by CAPES-Brazil, H.P. Pinheiro’s research was partially supported by CNPq-Brazil [305988/2012-3] and FAPESP [2011/15047-7], A. Pinheiro’s research was supported by FAPESP [13/00506-1] and CNPq [308439/2014-7 and 401478/2012-2].

Supplemental data

Supplemental data for this article can be accessed at 10.1080/02664763.2015.1077939.

Note

1. *Pardo* is commonly used in Brazil to refer to individuals with varied racial ancestries.

References

[1] T.B. Bordin, H.P. Pinheiro, and P. Pinheiro, Homogeneity tests among groups for microsatellite data, J. Appl. Stat. 9 (2011), pp. 1951–1962.
[2] R. Chakraborty and C.R. Rao, Measurement of genetic variation for evolutionary studies, Handb. Stat. 8 (1991), pp. 271–316. doi:10.1016/S0169-7161(05)80165-8.
[3] A.C. Davison and D.V. Hinkley, Bootstrap Methods and Their Application, Cambridge University Press, Cambridge, UK, 1999.
[4] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (1937), pp. 675–701.
[5] W. Hoeffding, A class of statistics with asymptotically normal distribution, Ann. Math. Statist. 19 (1948), pp. 293–325.
[6] M. Hollander and D.A. Wolfe, Nonparametric Statistics, John Wiley, New York, NY, 1973.
[7] V.E. Johnson, An alternative to traditional GPA for evaluating student performance, Stat. Sci. 12(4) (1997), pp. 251–278.
[8] W.H. Kruskal and W.A. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. 46 (1952), pp. 583–621.
[9] J. Lalouel, Distance analysis and multidimensional scaling, in Current Development in Anthropological Genetics: Theory and Methods, vol. I, J. Mielke, and M. Crawford, eds., Plenum, New York, NY, 1980.
[10] A.J. Lee, U-statistics – Theory and Practice, Marcel Dekker, New York, NY, 1990.
[11] P.C. Mahalanobis, On the generalized distance in statistics, P. Natl. Insti. Sci. India 2(1) (1936), pp. 49–55.
[12] R.P. Maia, H.P. Pinheiro, and A. Pinheiro, Heterogeneidade do desempenho de alunos da Unicamp, do ingresso à conclusão, Cadernos de Pesquisa 39(137) (2009), pp. 645–660.
[13] T.K. Nayak and J.L. Gastwirth, The use of diversity analysis to assess the relative influence factors affecting the income distribution, J. Bus. Econ. Stat. 7(4) (1989), pp. 453–460.
[14] R.H.L. Pedrosa, J.N.W. Dachs, R.P. Maia, C.Y. Andrade, and B.S. Carvalho, *Academic performance, student’s background and affirmative action at a Brazilian research university*, High. Educ. Manag. Policy 19(3) (2007), pp. 1–20.

[15] A. Pinheiro, H.P. Pinheiro, and P.K. Sen, *Comparison of genomic sequences using the Hamming distance*, J. Stat. Plan. Infer. 130(1–2) (2005), pp. 325–339.

[16] A. Pinheiro, P.K. Sen, and H.P. Pinheiro, *Measures: Quasi U-statistics, martingales and nonstandard asymptotics*, J. Multivariate Anal. 100(8) (2009), pp. 1645–1656.

[17] A. Pinheiro, P.K. Sen, and H.P. Pinheiro, *A class of asymptotically normal degenerate quasi U-statistics*, Ann. Inst. Math. Stat. 63 (2011), pp. 1165–1182.

[18] C.R. Rao, *Diversity: its measurement, decomposition, apportionment and analysis*, Sankya A. 44 (1982), pp. 1–21.

[19] J.F. Zachary and D.J. Schaeffer, *Correlations between pre-veterinary admissions variables and academic success in core courses during the first year of the veterinary curriculum*, J. Vet. Med. Educ. 21(2) (1994), pp. 1–6.

[20] R. Zwick, *College Admission Testing*. Report for the National Association for College Admission Counseling – Guiding the way to higher education, 2007.