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

[1] *Department of Molecular Biology and Genetics, Aarhus University, Denmark*

[2] *Department of Statistics,* University of Campinas, Brazil

Rafael Pimentel Maia[1],

Hildete P. Pinheiro[2]

Aluísio Pinheiro[2]

Abstract

We present novel methodology to assess undergraduate students' performance. The proposed methods are based on measures of diversity and on the decomposability of quasi U-statistics to define average distances between and within groups. They have been employed as an alternative to the classic analysis of variance especially when the assumption of normality is not met. The quasi U-statistics nonparametric method can handle tests for interaction and uses jackknife to get p-values for the tests. The nonparametric method also results in smaller error variances, illustrating its robustness against model misspecication.

Keywords: *asymptotic theory, diversity measures, jackknife, nonparametric methods, normal distribution, triangular distribution, U-statistics.*

1 Introduction

Assessment of undergraduate performance from entrance to graduation has been of great interest in the literature (Zachary and Schaeffer, 1984; Johnson, 1997; Pedrosa et al., 2007; Zwick, 2007 and references

*Corresponding author: *E-mail address*: hildete@ime.unicamp.br (H. Pinheiro)

therein). Most of the studies were based on classical analysis of variance or correlation analysis. In some cases where the normality assumption is not met, nonparametric tests based on ranks have been used (Zachary and Schaeffer, 1984), 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 (Kruskal and Wallis, 1952; Friedman, 1937; Hollander and Wolfe 1973), but none of these tests accommodate interaction effect tests. Some alternative methods, such as Bayesian analysis for the assessment of performance of students has been proposed in Johnson (1997), but it is not an alternative nonparametric method. These nonparametric methods are designed for linear effects, with no interactions. Johnson (1997) provides an alternative to classical ANOVA under the Bayesian paradigm.

We propose a nonparametric methodology which allows the analysis of nonnormal data as well as normal data. Furthermore, we add the possibility of interaction effects, which is not possible with more popular nonparametric approaches. The proposed test statistics are asymptotically normal under both null and alternative hypotheses for a large class of models (Pinheiro et al., 2009; 2011).

Pedrosa et al. (2007) 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* 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 entrance exam scores; the second rank is based on the final (or last) GPA scores. For instance, the student with the worst entrance exam grade 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 between -1 and 1 and symmetric around zero. As this variable, by construction, is limited between -1 and 1, the tails of its distribution are lighter than the normal distribution, i.e, its distribution is leptokurtic. Even for large sample sizes, the normality approximation may not be appropriate. For instance, in our application the sample size is greater than 8,000 and we still have problems with the normality assumption of the *relative gain* (p-values of tests for normality assumption are less than 0.0001). 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.

The proposed method is based on measures of diversity and decomposability of quasi U-statistics (Pinheiro et al., 2009; 2011) to properly decompose between and within group distances.

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 parametric and nonparametric approaches. In the parametric procedures, we assume that the *relative gain* follows either

normal or a triangular distribution. The data comes from the State University of Campinas (UNICAMP), a public institution, located in the State of São Paulo and one of the top research universities in Brazil. Unicamp is highly selective, with an average of over 15 candidates per undergraduate position offered each year (www.comvest.unicamp.br). The socioeconomic data of 8,225 students admitted to UNICAMP from 1997 through 2000 forms the study database.

In Section 2 we present the data that motivated the study. In Section 3 we present a short introduction about diversity measures, the development of a hypothesis test to evaluate the homogeneity between groups and the parametric and nonparametric approaches based on quasi U-statistics. Section 4 presents the application to the Unicamp data. A discussion follows in Section 5.

2 Material and Standard Methods

The dataset is composed by 8,225 students which have enrolled at Unicamp at years 1997, 1998, 1999 and 2000 in all Bachelor's degree majors. The academic situation of these students were classified as following: Graduates (students who have already graduated their courses - 76.8%), and Others (these are the ones 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 entrance exam scores 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 41.3% female), from all Brazilian regions and enrolled in 45 different majors from the areas of Health Sciences, Engineering and Exact Science, Social Science and Arts.

Students who declared having studied all or most of their High School years in Private Schools (PrS), were considered coming from Private School. Analogously, the ones who declared having studied all or most of their High School years in Public Schools (PuS) were considered coming from Public Schools. Table 1 shows the distribution according to type of High School and entry year. 30.7% of students who enrolled between 1997 and 2000 come from PuS.

Table 1: Sampling distribution per year according to type of High School

High School	Entrance year				Total	
	1997	1998	1999	2000		
	%	%	%	%	n	%
Private	68.4	69.2	70.6	68.9	5610	69.3
Public	31.6	30.8	29.4	31.1	2485	30.7
Total	100.0	100.0	100.0	100.0	8095*	100.0

*There was missing information for 130 students.

Another characteristic evaluated was if the student worked when enrolled at the university. In the sample 28.0% of the students declared they worked when enrolled at the university. Among the students that came from Public High Schools (PuHS) 48.5% declared to be working, while among the ones who came from Private High Schools (PrHS), we observed 18.6%.

Table 2 presents the distribution of the students according to the type of High School by Gender and by Work status. We do not observe large differences between men and women. In general, 70% declared to have studied most of the period in a PrHS. For the students which did not work at the time of entrance in the university, 78.1% came from PrHS and for students that worked, around half of them have studied in PuHS.

Table 2: Sampling distribution according to type of high school, by gender and e by worked.

High School	Gender		Worked	
	Male	Female	No	Yes
Private	70.1	68.8	78.1	46.4
Public	29.9	31.2	21.9	53.6
Total	100.0	100.0	100.0	100.0

Figure 1 presents the histogram of the relative gain as well as the curves of the theoretical normal and triangular distributions. Its mean is equal to zero by construction and varies from -1 to 1. We can see that the tails are lighter compared with the normal distribution since we have a variable limited to the interval (-1,1). Comparing with the triangular distribution, we can see that it has a greater concentration of zeros than the theoretical triangular distribution. Applying the Kolmogorov-Smirnov and Anderson-Darling tests for normality, we obtained, respectively, p-values < 0.01 and < 0.005, which means that the hypothesis of normality is not supported by the data.

Table 3: Descriptive statistics for the *relative gain* according to Gender, Type of High School and Work

Group	N^*	Mean	Standard Deviation	P-value for T-test	P-value for Wilcoxon Rank Sum Test
Female	3398	0.0613	0.3461	< 0.0001	< 0.0001
Male	4827	-0.0433	0.3594		
Private	5610	-0.0195	0.3511	< 0.0001	< 0.0001
Public	2485	0.0453	0.3674		
Worked	2279	-0.0028	0.3729	0.5442	0.4745
Did Not Work	5864	0.0027	0.3508		

*There were 130 missing information for type of High School and 82 missing information for Work.

Table 3 presents some descriptive statistics of the *relative gain* according to Gender, Type of High School and Work as well as results of two-sample t-tests and Wilcoxon rank sum tests for group comparisons. We

would like to point out that, even though the standard deviations seem to be equal or very close, they are not. The relative gain is in a relatively small scale (between -1 and 1) and the sample sizes are large (all greater than 2,000). So, when performing the two-sample t-tests we used the p-values for unequal variances, since the equality of variances was statistically rejected in all three cases. The PuHS students presented a mean relative gain significantly greater than those who studied in PrHS. When comparing genders, women have a significantly greater relative gain than men, but we do not observe any difference in the relative gain when comparing the students that worked to those who did not.

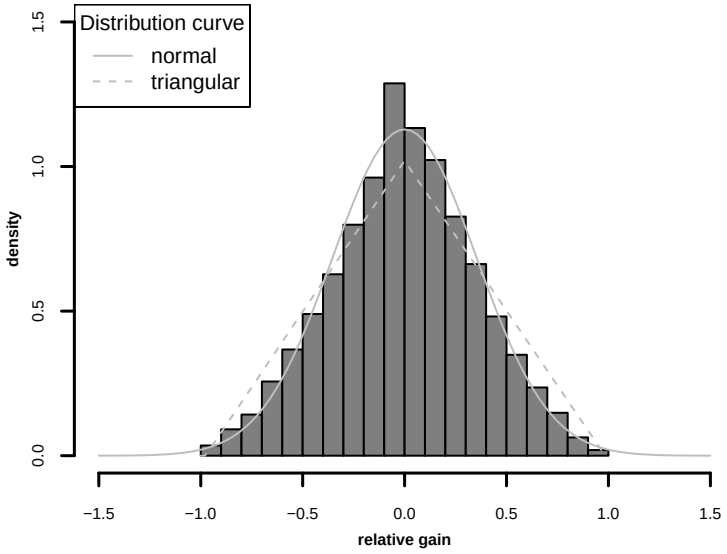

Figure 1: Histogram of the relative gain.

From Figure 2, 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.

Even though the assumption of normality does not seem to be true for the *relative gain*, we adjusted a linear model with Gender, type of High School and Work with all three main effects and all pairwise interactions. None of the interaction terms were significant and the analysis of variance table for the final model is presented on Table 4. The linear model used dummy variables (Female =1, Male=0; PuHS=1, PrHS-0 and Did not work=1, Worked=0) as codification for the design matrix. According to the parameter estimates, all the coefficients are positive, which means that Female has greater relative gain than Male, 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 looked at the results of the t-tests and Wilcoxon rank sum tests on Table 3, there was no difference in relative gain between those who worked and

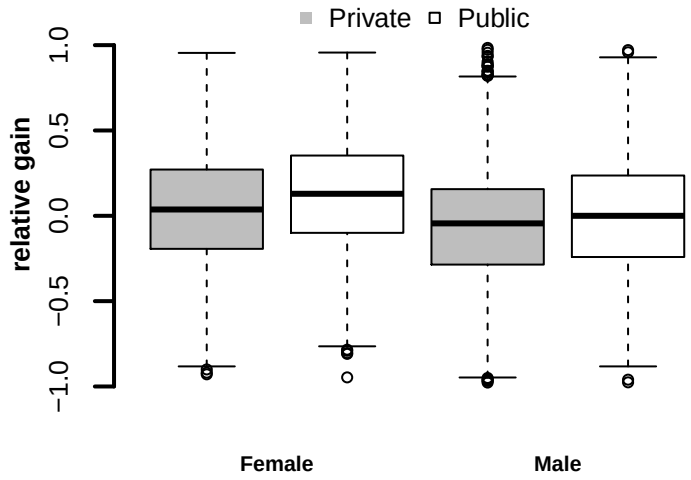

Figure 2: Box plots of the *relative gain* according to gender and type of High School.

those who did not work, whereas in the model it is statistically significant. This may be due to the fact that the model was adjusted for the other effects. Figure 3 shows the Q-Q plot of the residuals for the model in Table 4 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, presented p-value < 0.0001).

Table 4: Analysis of Variance

Souce	d.f.	Type III SS*	F-value	p-value
Gender	1	20.34	164.09	< 0.0001
Type of High School	1	8.03	64.77	< 0.0001
Work	1	0.72	5.80	0.0160

*Correspond to the variation attributable to an effect after correcting for any other effects in the model.

The normality 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.

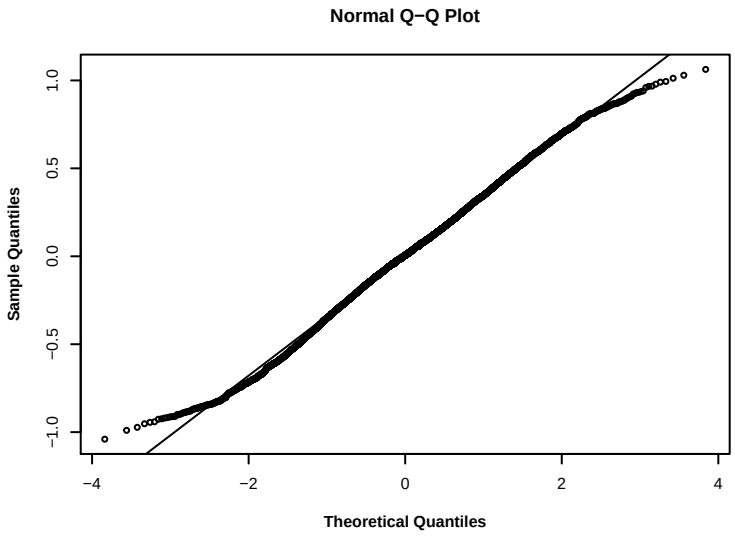

Figure 3: Q-Q plot of the residuals for the linear model presented on Table 4

3 Quasi U-statistics Methods

3.1 Measures of Diversity and Decompositions

Several metrics have been constructed to measure distances in qualitative or quantitative variables (Mahalanobis, 1936; Lalouel, 1980; Chakraborty and Rao, 1991 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 (Rao, 1982). 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

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

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 $\phi(x_i, x_j) = (x_i - x_j)^2/2$ (Lee, 1990; Hoeffding, 1948).

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 and we do not necessarily need an assumption for the distribution of the X_i's. Consider a symmetric and nonnegative function $\phi(X_i, X_j)$, which is a measure of the

difference between two individuals.

Define

$$E[\phi(X_i^g, X_j^g)] \;=\; \theta_g$$

and

$$E[\phi(X_i^g, X_j^{g'})] \;=\; \theta_{gg'},$$

where θ_g is the diversity within group g and $\theta_{gg'}$ is the diversity between groups g and g'. Estimators of θ_g and $\theta_{gg'}$ can be found using U-statistics (Hoeffding, 1948). If $\phi(x_i, x_j)$ is a convex function, then

$$\theta_{gg'} \geq \frac{\theta_g + \theta_{g'}}{2}. \tag{3.1}$$

The excess dissimilarity measure between groups g and g' is given by

$$\mathcal{D}_{gg'} \;=\; \theta_{gg'} - \frac{1}{2}[\theta_g + \theta_{g'}], \tag{3.2}$$

i.e., $\mathcal{D}_{gg'}$ is the excess measure of diversity between groups g and g' compared to the average of the within measures of diversities in groups g and g'.

Let X_i^g be the random variable representing the *relative gain* for individual i of group g, for $g = 1, ..., G$ and $i = 1, ..., n_g$. Define

$$\phi(x_i, x_j) = (x_i - x_j)^2,$$

the quadratic distance between the *relative gain* of individuals i and j. Estimators of θ_g, $\theta_{gg'}$ are given by

$$\hat{\theta}_g = \bar{D}_g = \frac{1}{\binom{n_g}{2}} \sum_{i<j} \phi(x_i^g, x_j^g)$$

and

$$\hat{\theta}_{gg'} = \bar{D}_{gg'} = \frac{1}{n_g n_{g'}} \sum_i \sum_j \phi(x_i^g, x_j^{g'}).$$

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), $E(\bar{D}_g) = \theta_g$ and $E(\bar{D}_{gg'}) = \theta_{gg'}$. The overall mean distance is defined as the total variability of the pooled sample and it can be estimated by

$$\bar{D}_n(0) \;=\; \sum_{i<j} \binom{n}{2}^{-1} \phi(x_i, x_j) \;=\; \binom{n}{2}^{-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \phi(x_i, x_j), \tag{3.3}$$

with $n = \sum_{g=1}^{G} n_g$. As in Pinheiro et al. (2005) and Pinheiro et al. (2009), we can decompose $\bar{D}_n(0)$ as

$$\begin{aligned}
\bar{D}_n(0) &= \binom{n}{2}^{-1} \left[\sum_{g=1}^{G} \binom{n_g}{2} \bar{D}_g + \sum_{g<g'} n_g n_{g'} \bar{D}_{gg'} \right] \\
&= \sum_{g=1}^{G} \frac{n_g}{n} \bar{D}_g + 2 \sum_{g \neq g'} \frac{n_g n_{g'}}{n(n-1)} \bar{D}_{gg'} - \sum_{g=1}^{G} \frac{n_g (n - n_g)}{n(n-1)} \bar{D}_g \\
&= D_n(W) + D_n(B),
\end{aligned}$$

where

$$D_n(W) = \frac{1}{n}\sum_{g=1}^{G} n_g \bar{D}_g$$

and

$$D_n(B) = \frac{1}{n(n-1)}\sum_{g\neq g'} n_g n_{g'} \left[2\bar{D}_{gg'} - \bar{D}_g - \bar{D}_{g'}\right]. \tag{3.4}$$

Note that

$$E(D_n(W)) = \frac{1}{n}\sum_{g=1}^{G} n_g \theta_g$$

and

$$E(D_n(B)) = \frac{1}{n(n-1)}\sum_{g\neq g'} n_g n_{g'}(2\theta_{gg'} - \theta_g - \theta_{g'}).$$

3.2 Hypothesis Testing

We can test homogeneity between groups using the individual performance data. Intuitively, we can say that under H_0: $\mathcal{D}_{gg'} = 0$, i.e., given (3.1) and (3.2), we can write these hypotheses as

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

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. Analogously, the excess between diversity $\mathcal{D}_{gg'} = 0$. Large absolute sample values of $D_n(B)$ are indications of large values of $\mathcal{D}_{gg'}$. Thence, one rejects H_0 for *large* absolute values of $D_n(B)$.

The nonparametric natural test statistic is $D_n(B)$ given in (3.4). Under mild regularity conditions this test statistic is asymptotically normal (Pinheiro, Sen & Pinheiro, 2009; 2010).

Note that, under H_0, we can rewrite $D_n(B)$ as

$$D_n(B) = \binom{n}{2}^{-1} \sum_{1\leq i<j\leq n} \eta_{nij}\psi_2(X_i, X_j),$$

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;

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

Using the Hoeffding decomposition of U-statistics (Hoeffding, 1948), Pinheiro et al. (2009) showed that, under the null hypothesis of homogeneity between groups the asymptotic distribution of $D_n(B)$ is normal, i.e,

$$n(V_n^*)^{-1/2} D_n(B) \xrightarrow{\mathcal{D}} N(0,1) \ \ as \ \ n \to \infty, \tag{3.5}$$

where

$$V_n^* = U_n^{(2,2)} - U_n^{(3)},$$

$$U_n^{(2,2)} = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \phi^2(X_i, X_j)$$

and

$$U_n^{(3)} = \sum_{i \neq i \neq l} \phi(X_i, X_j)\phi(X_i, X_l)/[n(n-1)(n-2)].$$

From the result given in (3.5), one can also find the power of this test. Note that

$$\begin{aligned} E_{H_1}(D_n(B)) &= \frac{1}{n(n-1)} \sum_{g<g'} \left[n_g n_{g'} E_{H_1}(2\bar{D}_{gg'} - \bar{D}_g - \bar{D}_{g'}) \right] \\ &= \sum_{g<g'} \frac{n_g}{n} \frac{n_{g'}}{(n-1)} (2\theta_{gg'} - \theta_g - \theta_{g'}) \\ &\longrightarrow \sum_{g<g'} p_g p_{g'} (2\theta_{gg'} - \theta_g - \theta_{g'}) \equiv \delta_n, \end{aligned}$$

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

$$E_{H_1}(D_n(B)) = \delta_n + O(n^{-2}).$$

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

The power of the test will be given by

$$P_{H_1}\left(\frac{nD_n(B)}{\sqrt{V_n^*}} > q_\alpha \right) \longrightarrow 1 - \Phi\left(q_\alpha - \Delta/\gamma\right),$$

where $E_{H_1}(D_n(B)) - \delta_n \to 0$ and $V_n^* \xrightarrow{p} \gamma^2$ as $n_0 \to \infty$.

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.

3.3 The Multifactors Problem

The results shown in the previous sections correspond to a one way ANOVA. Consider two factors, A_1 with s levels and A_2 with t levels, as in Nayak and Gastwirth (1989). For more than two factors, the theory is easily generalized. To obtain the combined effect of A_1 and A_2, consider the crossed classification of A_1 and A_2 as a single factor with $s \times t$ levels, and we get the decomposition of $D_n(0)$ as

$$D_n(0) = D_n(W)(A_1, A_2) + D_n(B)(A_1, A_2). \tag{3.6}$$

As in the sum of squares for the classical analysis of variance, $D_n(B)(A_1, A_2)$ can be decomposed as

$$D_n(B)(A_1, A_2) = D_n(B)(A_1) + D_n(B)(A_2 \mid A_1), \tag{3.7}$$

with

$$D_n(B)(A_2 \mid A_1) = D_n(B)(A_1, A_2) - D_n(B)(A_1).$$

Therefore, $D_n(B)(A_2|A_1)$ 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.

For multiple factors, $D_n(0)$ is the total variability, which is a mixture of several subgroups. When there are k factors, $A_1, ..., A_k$, $D_n(W)(A_1, ..., A_k)$ is the weighted average of the diversities within each group defined by the cross classification of $A_1, ..., A_k$ and

$$D_n(B)(A_1, ..., A_k) = D_n(0) - D_n(W)(A_1, ..., A_k).$$

Analogously,

$$D_n(B)(A_1, ..., A_s|A_{s+1}, ..., A_k) = D_n(B)(A_1, ..., A_k) - D_n(B)(A_{s+1}, ..., A_k).$$

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

3.4 Parametric and Nonparametric quasi U-statistics methods

The methods proposed can be parametric or nonparametric, i.e, we may assume a known distribution for X (the *relative gain* in this particular case) or work under a large class of distributions for X. In the parametric approach, two cases are explored: the *relative gain* is assumed to be normal and triangular. Maximum likelihood estimators (MLE) of the moments of the *relative gain* are used for the estimation of the mean and variance of the test statistic $D_n(B)$.

For the normal distribution we can find all the moments of X and apply the property of invariance of MLE to get the the estimation of the mean and variance of $D_n(B)$. To obtain the MLE for the triangular distribution, Kotz and Dorp (2004) suggest the use of the routines *BSearch* and ABSearch combined. The MLE in the triangular case can be obtained using the software *MLE-Estimator* (www.seas.gwu.edu/dorpjr/tab4/publications_book.html). Details and analytical results of the MLE for the normal and triangular distributions are shown in Appendix A.

In the nonparametric case, no assumption is made about the distribution of the *relative gain*, the variance of $D_n(B)$ is estimated by jackknife, and p-values are computed according to resampling methods (Davison and Hinkley, 1999). Details of the algorithm is in Appendix B.

4 Application

We consider the *relative gain* to follow a normal distribution with parameters μ_g and σ_g, say for each group. Alternatively, a triangular [-1,0,1] distribution is assumed. Note that the standard deviation of a triangular [-1,0,1] is 0.4082, which is greater than the observed standard deviation of the sample (0.3536). For instance, the frequency of zeros in the data is higher than what is expected in the triangular distribution.

The variance of $D_n(B)$ is computed by a jackknife resampling technique (Appendix B). We also considered here the interaction between Gender and Type of High School (described in Section 3.3).

Table 5 displays the estimates of $D_n(B)$, their standard deviations and the p-values for the three methods. For the Normal and the Triangular approaches, there is significant difference between genders at a 0.05 level (p-values equal 0.0259 and 0.0044, respectively). In the nonparametric approach, we observe a significant difference between Types of High School (p-value equals 0.0002) and Genders (p-value < 0.0001). We do not observe differences between the students which worked or not (p-values > 0.05). Using the expansion of quasi U-statistics for multifactors (discussed in Section 3.3), which is comparable to the ANOVA results given on Table 4, there are significant effects of Gender and Type of High School (p-values are < 0.0001) when adjusted for the other variables. On the other hand, we do not find a significant effect of Work adjusted by Gender and Type of High School (p-value=0.3153) and there is a significant effect of the interaction between Gender and type of School (p-value < 0.0001), which is different from the conclusions by the classical analysis of variance on Table 4.

From Table 5, we can also see that the estimated standard errors (SE's) by jackknife are smaller than those obtained by the parametric methods (using the normal or triangular distribution). 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 it should be. But here the tails are lighter and the empirical distribution are more concentrated than the normal or triangular distribution. For this reason the SE's for the nonparametric method is smaller than the parametric ones, showing here an advantage over the parametric procedure when the assumptions are not met.

5 Discussion

The main contribution of this work is to propose new methods to analyze the performance of college students from entrance to graduation, especially in cases where the assumption of normality is not met and the use of classical analysis of variance may be compromised. The quasi U-statistics nonparametric procedure performed very well, compared to the classical analysis of variance and the quasi U-statistics parametric methods. Our proposed method is based on measures of diversity and makes use of all pairwise comparisons of individuals and U-statistics theory results to perform statistical tests. We apply the quasi U-statistic tests

Table 5: Analysis of diversity using parametric and nonparametric approach.

Effects	$D_n(B)$	Normal		Triangular		Nonparametric	
		$\widehat{SE}$	p-value	$\widehat{SE}$	p-value	$\widehat{SE}$	p-value
School	0.0017	0.0026	0.2602	0.0018	0.1833	0.0005	0.0002∗
Gender	0.0049	0.0026	0.0259∗	0.0019	0.0044∗	0.0008	< 0.0001∗
Worked	-0.00002	0.0026	0.4976	0.0018	0.4966	0.0001	0.3716
School | Gender + Work	0.0019	-	-	-	-	0.0005	< 0.0001*
Gender | School + Work	0.0049	-	-	-	-	0.0008	< 0.0001*
Work | School + Gender	0.0001	-	-	-	-	0.0002	0.3153
School | Gender	0.0018	-	-	-	-	0.0005	0.0001∗
Gender | School	0.0050	-	-	-	-	0.0008	< 0.0001∗
School x Gender	0.0067	-	-	-	-	0.0009	< 0.0001∗

∗ significant with 5%

using a parametric and a nonparametric approach. Nonparametric methods are more robust, since they do not make any assumptions about the distribution of the data. Resampling methods, such as jackknife, have been used to obtain p-values for the nonparametric statistical tests. It is quite simple to obtain estimates of variances of the test statistic by jackknife (see details in Appendix B). On the other hand, in the parametric quasi U-statistic method, we have to be careful with the assumption made for the distribution of the data, since it can provide over-estimation of the variances if the distribution of the data is leptokurtic. Moreover, the quasi U-statistics methods will be more powerful when the normality assumption is not met and within-group variances differ considerably.

Acknowledgements

R. P. Maia was supported by CAPES-Brazil, H.P. Pinheiro's research was partially supported by CNPq-Brazil (306240/2009-2) and FAPESP (2011/15047-7), A. Pinheiro's research was supported by FAPESP (2008/51097-6 and 2009/14176-8) and CNPq (306993/2008-2 and 480919/2009-7). We thank 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, intern at COMVEST and Silvio de Souza, database specialist at DAC.

References

Davison, A.C. and Hinkley, D.V. (1999). *Bootstrap methods and their application.* Cambridge University Press.

Chakraborty and Rao (1991). Measurement of genetic variation for evolutionary studies. In: Rao, C.R., Chakraborty, R. (Eds.), *Handbook of Statistics*, vol. 8. North Holland, Amsterdam.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32 (200)**, 675-701.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* , **19**, 293– 325.

Hollander, M. and Wolfe, D.A. (1973). *Nonparametric Statistics.* John Wiley, New York, NY.

Johnson, V.E. (1997). An alternative to traditional GPA for evaluating student performance. *Statistical Sciencd*, **12 (4)**, 251-269.

Kotz, S. and Dorp, J.R. va (2004). *Beyond Beta, other continuous families of distributions with bounded support and applications.* World Scientific Press, Singapure.

Kruskal and Wallis (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association.* **46 (260)**, 583-621.

Lalouel, J. (1980) Distance analysis and multidimensional scaling. In: Mielke, J., Crawford, M. (eds), *Current Development in Anthropological Genetics: Theory and Methods.* vol. I. Plenum, New York.

Lee, A.J. (1990). *U-statistics - Theory and Practice.* Marcel Dekker, New York, NY.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India* **2(1)**, 49-55.

Nayak, T.K. and Gastwirth, J.L. (1989). The use of diversity analysis to asses the relative influence factors affecting the income distribution. *Journal of Business and Economic Statistics*, **7(4)**, 453–460.

Pinheiro, A., Pinheiro, H.P. and Sen, P.K. (2005). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference*, **130(1-2)**, 325–339.

Pinheiro, A., Sen, P.K. and Pinheiro, H.P. (2009). Decomposability of High-Dimensional Diversity Measures: Quasi U-Statistics, Martingales and Nonstandard Asymptotics *Journal of Multivariate Analysis* **100**(8), 1645–1656.

Pinheiro, A., Sen, P.K. and Pinheiro, H.P. (2011). A class of asymptotically normal degenerate quasi U-statistics. *Annals of the Institute of Mathematical Statistics* **63**, 1165–1182.

Pedrosa, R.H.L., Dachs, J.N.W., Maia, R.P., Andrade, C.Y. and Carvalho, B.S. (2007). Academic Performance, Student's Background and Affirmative Action at a Brazilian Research University. *Higher Education Management and Policy.* **19**(3), 1–20.

Rao, C.R. (1982). Diversity: Its measurement, decomposition, apportionment and analysis. *Sankya A*, **44**, 1–21.

Zachary, J.F. and Schaeffer, D.J. (1994) Correlations between preventerinary admissions varaibels and academic success in core courses during the first year of the veterinary curriculum. *Journal of Veterinary Medical Education*, vol. 21 (2), 1-6.

Zwick, R. (2007) *College Admission Testing.* Report for the National Association for College Admission Counseling - Guiding the way to higher education.

Appendix A

Let

$$\mu_g = E[X_i^g],\ \ \mu_{g2} = E[(X_i^g)^2],\ \ \mu_{g3} = E[(X_i^g)^3] \ \text{ and } \ \mu_{g4} = E[(X_i^g)^4]$$

be the first, second, third and fourth moments of X_i^g. Then,

$$\begin{aligned}
\phi_1(x_1^g) &= E[\phi(X_1^g, X_2^g) \mid X_1^g = x_1^g] = (x_1^g)^2 - 2x_1^g\mu_g + \mu_{g2}, \\
E[\phi_1(X_1^g)] &= E[(X_1^g)^2 - 2X_1^g\mu_g + \mu_{g2}] = 2\sigma_g^2, \\
E(\phi_1^2) &= \mu_{g4} - 4\mu_g\mu_{g3} + 3\mu_{g2}^2.
\end{aligned}$$

Therefore,

$$\sqrt{n_g}(\bar{D}_g - 2\sigma^2) \xrightarrow{D} N\left(0, 4\xi_1\right), \tag{A.1}$$

where

$$\xi_1 = \mu_{g4} - 4\mu_g\mu_{g3} + 3\mu_{g2}^2 - 4\sigma_g^4.$$

For $\bar{D}_{gg'}$, which is a U-statistic of degree $(1,1)$, let

$$\begin{aligned}
\theta_{gg'} &= E(\phi(X_1^g, X_2^{g'})) = \mu_{g2} - 2\mu_g\mu_{g'} + \mu_{g'2}, \\
\phi_{10} &= E[\phi(X_1^g, X_2^{g'}) \mid X_1^g = x_1^g] = x_1^g - 2x_1^g\mu_{g'} + \mu_{g'2}, \\
\phi_{01} &= E[\phi(X_1^g, X_2^{g'}) \mid X_2^{g'} = x_2^{g'}] = \mu_{g2} - x_2^{g'}\mu_g + x_2^{g'}.
\end{aligned}$$

Then,

$$\xi_{10} = E[\phi_{10}(X_1^g) - \theta_{gg'}^2] = \mu_{g4} - 4\mu_{g3}\mu_{g'} + 2\mu_{g2}\mu_{g'2} - 4\mu_g\mu_{g'}\mu_{g'2} + \mu_{g'2}^2 - \theta_{gg'}^2$$

and

$$\xi_{01} = E[\phi_{01}(X_2^{g'}) - \theta_{gg'}^2] = \mu_{g'4} - 4\mu_{g'3}\mu_g + 2\mu_{g'2}\mu_{g2} - 4\mu_g\mu_{g'}\mu_{g2} + \mu_{g2}^2 - \theta_{gg'}^2.$$

Therefore,

$$Var(\bar{D}_{gg'}) = \frac{1}{n_g}\xi_{10} + \frac{1}{n_{g'}}\xi_{01}. \tag{A.2}$$

As $D_n(B)$ is a linear function of U-statistics ($\bar{D}_g$ and $\bar{D}_{gg'}$), its variance will be dependent on the moments of X_i^g.

For instance, if $G = 2$,

$$Var(D_n(B)) = \frac{n_1^2 n_2^2}{n^2(n-1)^2} 4 \left[\frac{\xi_{10}}{n_1} + \frac{\xi_{01}}{n_2} + \frac{\xi_1}{n_1} + \frac{\xi_1}{n_2}\right] + O\left(\frac{1}{n}\right).$$

For the normal distribution, the MLE estimators of the mean and variance are

$$\begin{aligned} \hat{\mu} &= \bar{X}, \\ \hat{\sigma}^2 &= (n-1)S^2/n. \end{aligned}$$

As $\hat{\sigma}^2$ is biased for σ^2, we will use S^2 as an estimator of σ^2, since it is an unbiased estimator for σ^2. The estimators of the second, third and fourth moments are

$$\begin{aligned} \hat{\mu}_2 &= S^2 + \bar{X}^2, \\ \hat{\mu}_3 &= 3\bar{X}S^2 + \bar{X}^3, \\ \hat{\mu}_4 &= 3S^4 + 6S^2\bar{X}^2 + \bar{X}^4. \end{aligned}$$

If Z has a triangular distribution limited in $[a, b]$ with mode m, then

$$f(z) = \frac{2(z-a)}{(b-a)(m-a)} \boldsymbol{I}(a \le z \le m) + \frac{2(b-z)}{(b-a)(b-m)} \boldsymbol{I}(m \le z \le b) \tag{A.3}$$

The k-th moments of Z are

$$\begin{aligned} \mu_k &= c_1 \left(\frac{m^{k+2}}{k+2} - \frac{am^{k+1}}{k+1} - \frac{a^{k+2}}{k+2} + \frac{a^{k+2}}{k+1}\right) \\ &+ c_2 \left(\frac{b^{k+2}}{k+1} - \frac{b^{k+2}}{k+2} - \frac{bm^{k+1}}{k+1} + \frac{m^{k+2}}{k+2}\right), \end{aligned} \tag{A.4}$$

for all $k = 1, 2, \ldots$, with $c_1 = \frac{2}{(b-a)(m-a)}$ and $c_2 = \frac{2}{(b-a)(b-m)}$.

If $Z_1, Z_2, \ldots, Z_n$ are independent r.v. with p.d.f given by (A.3) and $\mathbf{Z} = (Z_{(1)}, Z_{(2)}, \ldots, Z_{(n)})$ is the vector of order statistics, the distribution of $\mathbf{Z}$ is given by

$$f(z_{(1)}, z_{(2)}, \ldots, z_{(n)}) = \left(\frac{2}{b-a}\right)^n \left\{\prod_{i=1}^{r} \frac{z_{(i)} - a}{m-a} \prod_{i=r+1}^{n} \frac{b - z_{(i)}}{b-m}\right\}, \tag{A.5}$$

where r is such that $Z_{(r)} \leq m < z_{(r+1)}$, $z_{(0)} \equiv a$ and $z_{(n+1)} \equiv b$.

Therefore, for fixed values of a and b, satisfying $a < z_{(1)}$ and $b > z_{(n)}$, we have that

$$\max_{a \leq m \leq b} L(a, m, b \mid \mathbf{z}) = \left(\frac{2}{b-a}\right)^n \{M(a, b, \hat{r}(a, b))\}, \tag{A.6}$$

where

$\hat{r}(a,b) = \arg \max_{r \in \{1,\ldots,n\}} M(a,b,r)$ and $M(a,b,r) = \prod_{i=1}^{r-1} \frac{z_{(i)} - a}{z_{(r)} - a} \prod_{i=r+1}^{n} \frac{b - z_{(i)}}{b - z_{(r)}}$.

The MLE of m (as a function of a and b) is given by $\hat{m}(a,b) = Z_{(\hat{r}(a,b))}$. Note that the function $\hat{r}(a,b)$ indicates in which order statistic the MLE of m is achieved as a function of the lower and upper limit, a and b, respectively.

From (A.6) we have

$$\max_{S(a,m,b)} [\log\{L(a, m.b; \mathbf{z})\}] = \max_{a < z_{(1)}, b > z_{(n)}} [\log\{n \log 2 + G(a,b)\}], \tag{A.7}$$

where $S(a,b) = \{(a,m,b) \mid a < z_{(1)}, b > z_{(n)}, a \leq m \leq b\}$ and $G(a,b) = \log\{M(a,b,\hat{r}(a,b))\} - n\log(b-a)$.

Note that $G(a,b)$ is defined only for values of $a < z_{(1)}$ and $b > z_{(n)}$. So, the three dimensional problem of maximization of $L(a,m,b;\mathbf{z})$is reduced to a two dimensional problem of maximization of $G(a,b)$ under $a < z_{(1)}$ and $b > z_{(n)}$. From the likelihood structure, one can see that for all values of m such that $z_{(1)} < m < z_{(n)}$, the likelihood $L(a,m,b;\mathbf{z}) \to 0$ (and therefore $\log L(a,m,b;\mathbf{z}) \to \infty$) when $a \uparrow z_{(1)}$ or $b \downarrow z_{(n)}$.

Appendix B

Algorithm for the quasi U-statistic nonparametric method

Step 1 - Compute the relative gain according to the introduction.

Step 2 - Separate the data set according to the groups of interest.

Step 3 - Compute observed $\bar{D}_g$'s, $\bar{D}_{gg'}$'s and then $D_n(B)$ from the data, call it $D_n(B)_{obs}$.

Step 4 - Apply the jackknife re-sampling method to obtain the SE of $D_n(B)$.

Step 4.1 - Compute the jackknife's replication $D_n(B)(-i)$ by removing from the sample the $i-th$ individual and then apply Step 3 again, for all individuals in the sample, one by one.

Step 4.2 - Compute the $D_n(B)(\cdot) = \sum_{i=1}^{n} \frac{D_n(B)(-i)}{n}$.

Step 4.3 - The Jackknife Standard Error (SE_J) is then estimated by

$$SE_J = \left[\frac{n-1}{n} \sum_{i=1}^{n} [D_n(B)(-i) - D_n(B)(\cdot)]^2\right]^{1/2}.$$

Step 5 - Compute the p-value (one-sided test), (i.e, $P[D_n(B) > D_n(B)_{obs}]$), using the normal approximation (3.5) for $D_n(B)$ and Standard Deviation given by the SE_J calculated in step 4.