Teacher Education and Curriculum Studies
2016; 1(1): 20-27
http://www.sciencepublishinggroup.com/j/tecs
doi: 10.11648/j.tecs.20160101.13

SciencePG
Science Publishing Group

Statistical Validation of E-learning Assessment

Jesús Ortiz[1], Antonio Aznar[1], José I. Hernando[1], Adriana Ortiz[2], Jaime Cervera[1]

[1]Department of Building Structures, Architecture School, Technical University, Madrid, Spain

[2]Department of Nuclear Services, IDOM (Ingenieria y Direccion de Obrasy Montaje) International, Madrid, Spain

Email address:
antonio.aznar@upm.es (A. Aznar)

To cite this article:
Jesús Ortiz, Antonio Aznar, José I. Hernando, Adriana Ortiz, Jaime Cervera. Statistical Validation of E-learning assessment. *Teacher Education and Curriculum Studies.* Vol. 1, No. 1, 2016, pp. 20-27. doi: 10.11648/j.tecs.20160101.13

Received: September 5, 2016; **Accepted**: September 14, 2016; **Published**: October 10, 2016

Abstract: This paper focuses on how to fulfill the objectivity and reliability goals, as well as the efficiency of the e-learning evaluation tools, and their integration in a blended evaluation system. In order to contribute to these goals, a new branch of statistics, i.e. "Statistical Learning", has been chosen to support this study. The proposed techniques can be very simply implemented with little knowledge of arithmetic and with the help of a standard spreadsheet. These techniques can allow us to get the whole picture of the evaluation procedure output, in order to systematically sort the main categories of the different students, and to easily identify the outliers for further assessment.

Keywords: Computer-Based Assessment, On-Line Learning, Questionnaire, Computer-Assisted Learning

1. Introduction

Information and Communications Technologies (ICT) are making its way into the University programs and studies. Teachers appraise that on-line learning activities may provide important advantages related to a student's emotional engagement and motivation [1]. Most of these advantages are subjective and difficult to assess by objective methods. This is the main incentive to synergize the conventional and well-known face-to-face education with ICT.

The application of on-line learning methods in combination with traditional learning programs is becoming increasingly frequent, especially when monitoring and evaluating the student's work, which may help to explain the continuous development of numerous ICT-type evaluation tools within Universities.

Most ICT evaluation tools adopted by educators are generally embodied in on-line questionnaires. These b-learning tools offer students many advantages, particularly when planning their studies and in self-evaluation. Nevertheless, the difficulty lies upon assessing the contribution of this tools on the student's learning process [2].

In the present study, the results obtained by the randomized MOODLE (Modular Objet-Oriented Dynamic Learning Environment) questionnaires has been developed by professors from Technical University of Madrid [3-8]. In this paper, an evaluation and a comparison between the on-line and the face-to-face results has been carried out. It focuses on procuring a method capable of quantifying the efficiency of the aforementioned e-learning evaluation tools, which, needless to say, calls for objective and measurable procedures. Objectivity is the main reason for employing the new branch of statistics - "*Statistical Learning*" [9]- to support this study.

The aim of this paper is not to innovate on *Statistical Learning* techniques, but to show the enormous usefulness and simplicity of some of them, in order to meet the goal of systematically validate (or invalidate) this blended evaluation that includes continuous on-line assessment and one or several face-to-face examinations. In addition, the proposed techniques are presented in a very plain and self-contained way, so that no more than basic arithmetic is needed to fully comprehend them or put them into practice. No specific software is required other than EXCEL spreadsheet or its equivalent in OPENOFFICE; although the utilization of R-systems can be even easier.

These techniques allow us to get the whole picture of the evaluation procedure output, in order to systematically sort the main categories of the different students, and to easily identify the outliers for further assessment. Thus, the human factor needed to ensure the trustworthiness of the evaluation procedure, although always necessary, can be reasonably minimized. For these purposes the classic statistical approach [10] is definitely inadequate. The proposed new statistical methods are widely used and developed in other fields [11] from which the field of Education can learn and improve.

Education and learning activities using ITC have through the years attracted great social attention. Prof. Jascha Kessler [12] comments in a humorous way that half a century ago the first attempts to perform massive education by means of the then new technologies, failed because of the enormous workload added to the teacher as it required reading and grading thousands of papers. He also advocates for "educating" – etymologically "drawing out" – rather than "training". He is indeed right in the latter but fortunately wrong in the former as the ICT along with the Statistical Learning techniques are now able to help teachers to clearly understand "what the data says" and to avoid the risk of drowning in a sea of information.

2. Short Description of the On-Line Assessment Procedure

The authors of this work have the well-founded opinion that both teachers and students can greatly benefit from ICT and e-learning, if they are relied upon only for "training", keeping the "formative" aspect of the learning process entirely to the face-to-face human interaction.

Consequently with the previously described b-learning practice, we have developed a blended evaluation procedure that includes continuous on-line assessment and one or several examinations that guarantee the reliability of the evaluation process. Given a numerical grade for the exam(s) (E) and an additional final grade of the Continuous Evaluation (C) for a given student, we apply the following weighting criterion to obtain his final grade (x): $x = \frac{7}{10}E + \frac{3}{10}C$.

At first glance, this might be seen as unfair or inconsiderate of the hard work underlying the "C" mark but we prioritize to dismiss any possible criticism about the reliability of the whole evaluation system. Fortunately, we have found that the heavy weight applied to the first term of the equation does not at all discourage the majority of the alumni from following the on-line evaluation, parallel to the normal lectures.

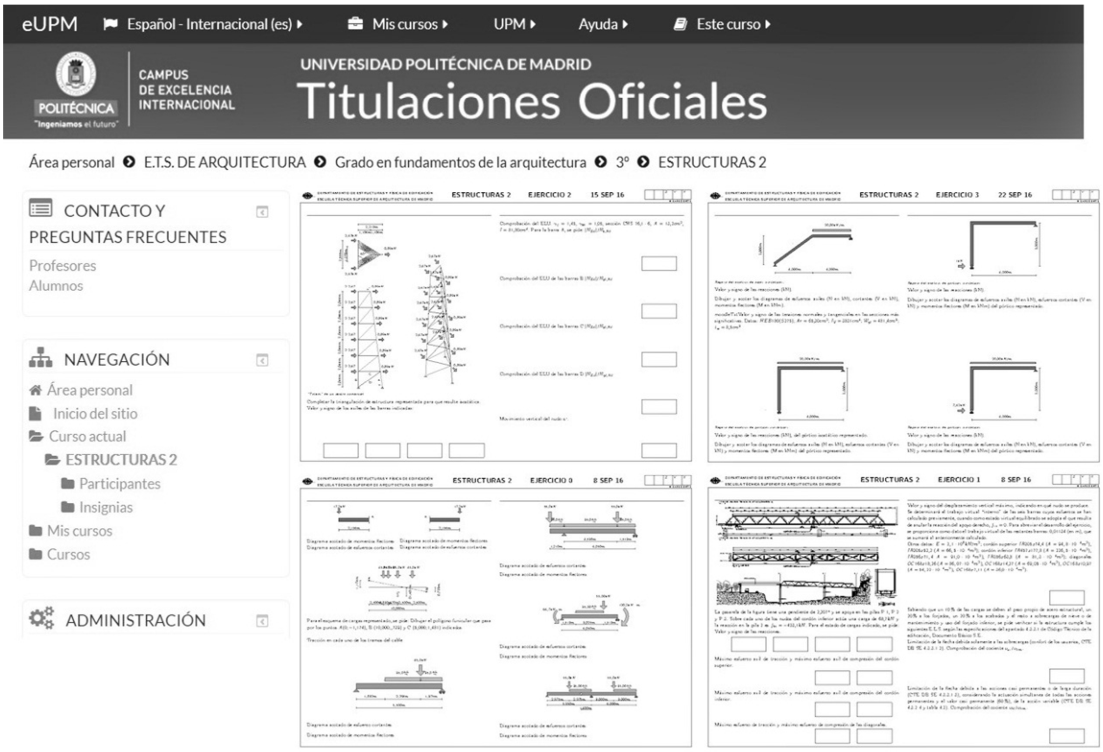
eUPM
Español - Internacional (es)
Mis cursos
UPM
Ayuda
Este curso
POLITÉCNICA
"Ingeniamos el futuro"
CAMPUS DE EXCELENCIA INTERNACIONAL
UNIVERSIDAD POLITÉCNICA DE MADRID
Titulaciones Oficiales
Área personal
E.T.S. DE ARQUITECTURA
Grado en fundamentos de la arquitectura
3º
ESTRUCTURAS 2
CONTACTO Y PREGUNTAS FRECUENTES
Profesores
Alumnos
NAVEGACIÓN
Área personal
Inicio del sitio
Curso actual
ESTRUCTURAS 2
Participantes
Insignias
Mis cursos
Cursos
ADMINISTRACIÓN
ESTRUCTURAS 2 EJERCICIO 2 15 SEP 16
ESTRUCTURAS 2 EJERCICIO 3 22 SEP 16
ESTRUCTURAS 2 EJERCICIO 0 8 SEP 16
ESTRUCTURAS 2 EJERCICIO 1 8 SEP 16

Figure 1. *Sample from one of the on-line exercises used. Data is randomized individually for each student so that the correct answers are different for each student. The contents of this particular pertain to building structural analysis, though not relevant from and education stand point.*

The standard platform MOODLE was chosen for several reasons: MOODLE has been implemented by Polytechnic University of Madrid who is responsible of its maintenance at an institutional level. It is a relatively modern educational tool, whereby it is expected not only to maintain some degree of stability but also its use could spread to other subjects and universities. Finally, it is a generic platform with open software which allows for the development of small applications that improve its performance and facilitate their adaptation to the specific requirements of each subject.

The platform meets the following basic requirements:

It offers students the possibility of an inmediate correction of their excercises with unlimited attempts.

It automates the evaluation process.

It has the ability to progressively include a large number of exercises.

In our approach the questionnaires are highly randomized and personalized.

It can be stated that the use of this tool has fostered a more active and collaborative learning, while involving students in the process. According to data from MOODLE, some students try to solve the same questionnaire many times (sometimes hundreds of times). Whereas there are students who obtain the highest possible score in their first or second attempt, others may take up dozens of attempts, which shows unequivocally that the use of the tool has been an incentive to correctly solve each exercise.

3. Statistical Learning Methods

3.1. Kernel Density Estimation

It's assumed that we have the final grades of the subject: $x_1, x_2, \ldots x_N$, one for each of the 'N' students, with x_k in the range $0 \leq x_k \leq 10$; being $x_k \geq 5$ the score needed to pass.

Recognizing that a certain uncertainty exists in the grades themselves, each of these grades can be seen as the mean of an stochastic variable with a maximum deviation, say, of ± ½, for a given student, that takes into account the set of possible random causes that may have influenced the student's score.

The graphic below shows a set of rectangles with base $x_k \pm 1/2$ and common height, 1, for a hypothetical group of 10 students ($x_1, x_2, \ldots x_{10}$ = 1.0, 2.5, 4.0, 5.0, 5.0, 5.5, 6.0, 7.0, 7.0, 9.0). This provides an overall picture of the results. Nonetheless, for a bigger set of students (this is, a bigger N, as it will be considered later in practical cases) this first graphic will become rather useless.

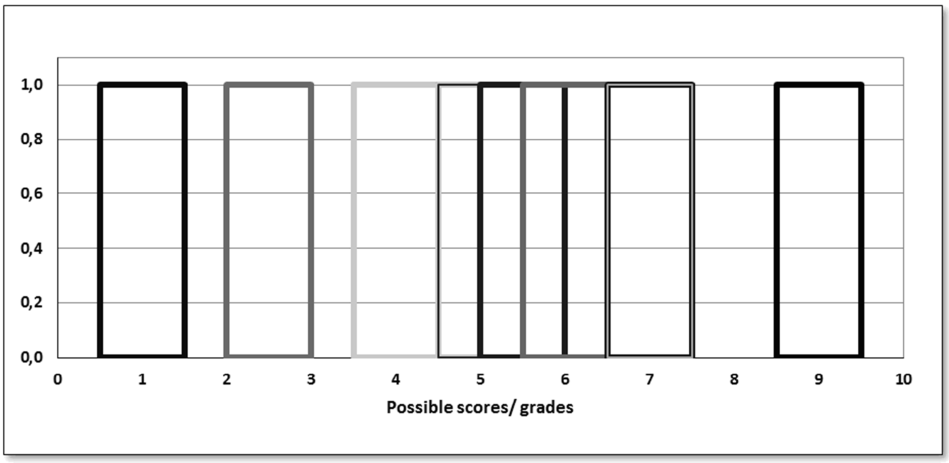

Figure 2. *Set of individual grades of a small theoretical group of students.*

Note that in this chart the rectangles pile up in areas where grades are more frequent, making it confusing. To counteract this drawback, at each point of the horizontal axis the ordinates of the overlapping rectangles are added up and divided by N (i.e., if they are averaged), so that the former graphic is transformed better displaying the clustering of the grades in a certain range and the dispersion in the rest:

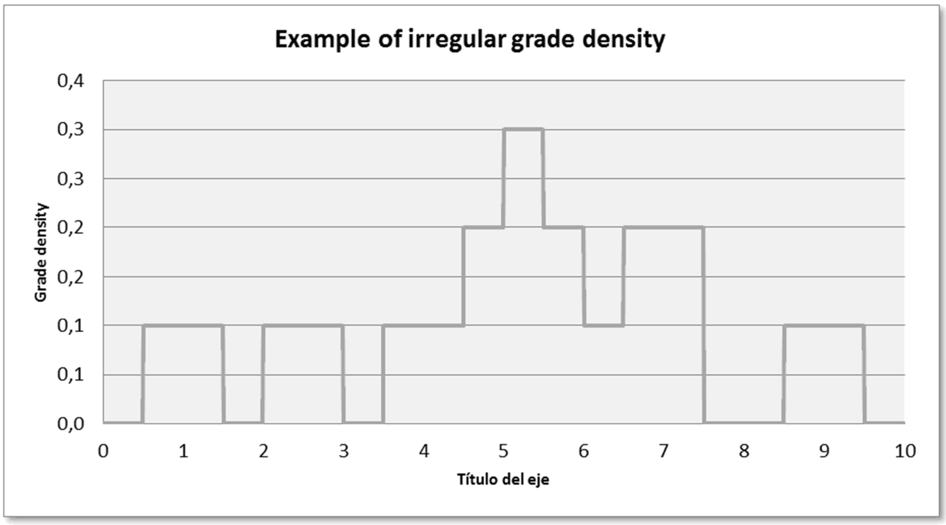

Figure 3. *Irregular density of grades. The horizontal axis shows the final grades (x).*

The function $\hat{p}(x)$ shown above is actually a bumpy approach to the real probability density $p(x)$ of the final score for a generic student of this theoretical instance. The stepped shape of the graph above is due to the fact that it has been obtained by adding the rectangular areas shown in Figure 2. If we modify the initial rectangular areas for smoother shape, we will improve the appearance of the resulting graph. Indeed, this image can be significantly improved if the rectangles of the first graphic are replaced by a“ gaussian bell” (Hahn & Shapiro, 1967) centered on each students' mark. This function can be expressed as:

$$K(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mathrm{x}_k)^2}{2\sigma^2} \right) \tag{1}$$

In this expression, σ represents the standard deviation. In this case, it can be used as an instrumental parameter, which it is chosen freely in order to obtain a conveniently smoothed curve. In fact, testing different σ values we obtain the graphics that are shown in the figures below.

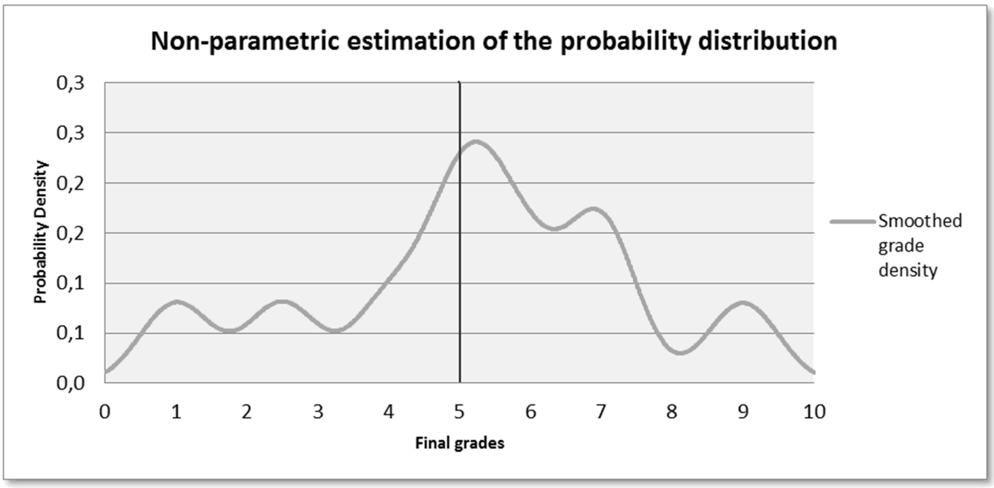

Figure 4. Marginally smoothed density of grades, using band-width σ = 1/2. The vertical axis shows the estimated probability density, $\hat{p}(x)$. The vertical red line signals the pass grade. The horizontal axis shows the final grades (x).

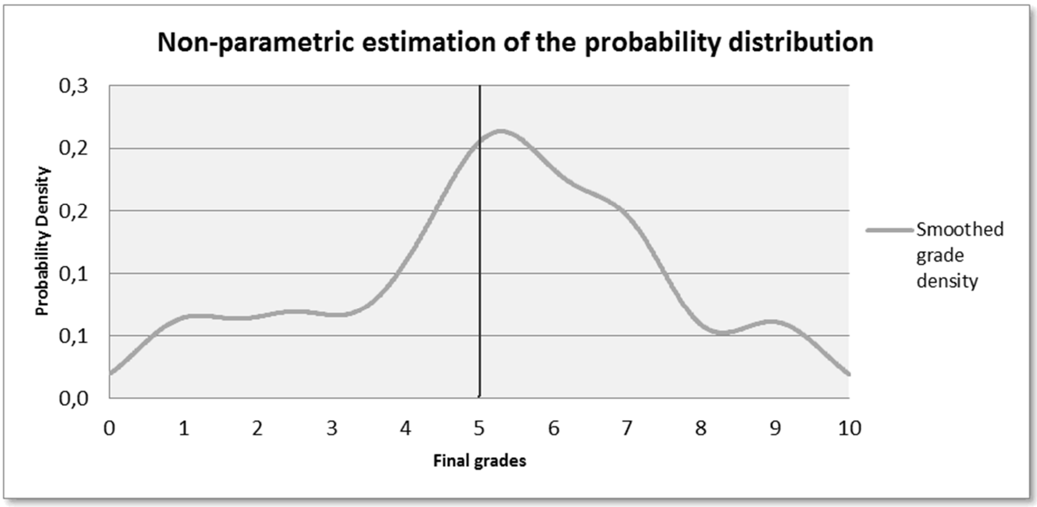

Figure 5. A more suitable smoothing by increasing the band-width (σ = 2/3). It can be observed: in the left a small plateau; in the middle the predominance of borderline grades (students who passed with the minimum grade required) and the gratifying exception of excellence in the right side of the curve.

In Statistical Learning the $K(x)$ function is known as “kernel”; σ is now called “bandwidth”. The aforementioned Gaussian function is the most popular kernel. In order to correctly tune the bandwidth some theoretical and empirical rules can be used. For our purpose, instead, the best option is that made by the professor him or herself, who is able to recognize in the graphic what he or she senses more plausible.

These graphics can be very easily developed with a spreadsheet and might be a very valuable tool for the assessment of the evaluation method, as it is confirmed by the real cases considered in the next section.

3.2. Real Cases: Multimodal Distributions

Classical statistical methods applied to the interpretation of a given set of grades provide statistical sample “moments”: mean (m), variance or its equivalent, the standard deviation (s), the third moment (related with the “skewness”) and the forth one (related with the 'kurtosis') [10]. These parameters are automatically calculated for a given data sample by a standard spreadsheet. But using these parameters to fit a classical density function, such as the normal or Gaussian function, formulated in the previous section, dramatically fails to match the expected structure of the data we are dealing with; this is confirmed by a simple “visual” comparison or, more precisely, by appropriate statistical tests.

In contrast, the next figure displays how the simple non-parametric technique, explained in section 3, captures a great deal of compelling features so that they might be rightly compared with an “x-ray” of the groups that comprise the whole course.

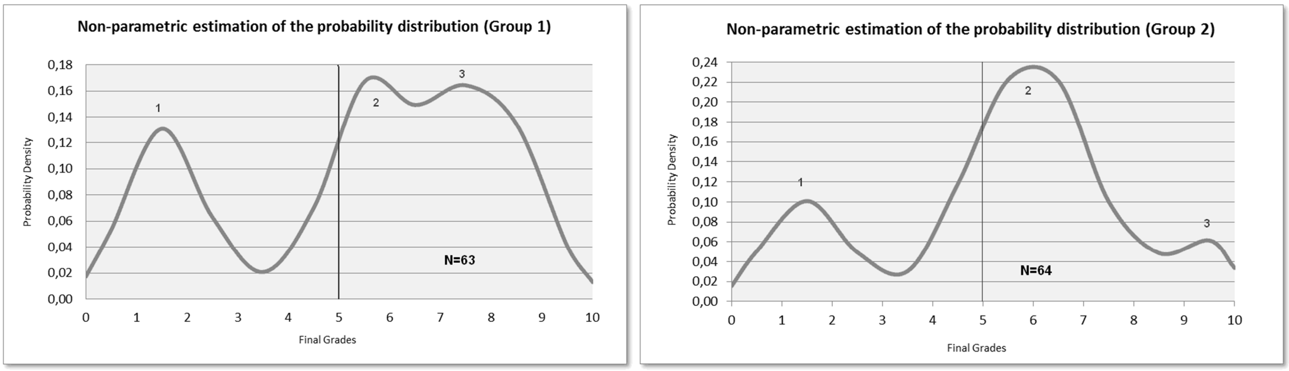

Figure 6. *Kernel probability density estimations of final grades (x) for two different student groups of the same course unit, each with approximately 60 students (the vertical red line signals the minimum grade necessary to pass). In both cases a trimodal distribution is displayed with band-width (σ =s/4= 2/3). These graphics provide a lot of information that classic parametric distributions are unable to capture. The vertical axis shows the estimated probability density* $\hat{p}(x)$ *and the horizontal axis the final grades x.*

These graphics distinctly delineate three categories of students, with their respective proportions in each group:

Those who fail the subject.

A dominant subset of students that barely manage to pass.

And a certain number of students that excel.

The average score and the standard deviation provide little information to explain the appearance of the aforementioned categories of students. The area right of the vertical red line outsizes the area to the left in both groups, in other words, both total mean grades are clearly over the minimum passing grade.

In both graphics the first peak indicates a considerable amount of students that fail with a considerably low mark, which is overall very similar in both groups. As it is shown in section 6, these are in essence students that do not follow the intense ongoing e-training throughout the course. The subset of “average students” stands out in the second group unlike the first group, where the areas under the peaks (2) and (3) are comparable.

Yet, the second group shows an unusual aspect. In Figure 6, the reader can appreciate how there are very few brilliant students but very much so, whereas in the first group the number of brilliant students is higher overall, although less brilliant. This can be seen in that the third peak in the second graph has a much smaller ordinate than the second peak (in contrast with the first graph); nonetheless its abscissa is much bigger when compared with the first graph.

There are empirical and theoretical rules for the optimum bandwidth selection, such as those given in [11], which depend mainly on the sample standard deviation (s) and the type of the kernel ($K(x)$); secondarily, they may also depend on the number of data (N), the skewness and the kurtosis of the sample. But, according to our practice, these criteria generate excessive smoothing for multimodal distributions. In the previous figure a recommended bandwidth is indicated (¼ of s) that is generally adequate for the purposes here pursued, when the Gaussian kernel is employed.

3.3. Parametric and Non-parametric Regression

Under the assumption that the total final marks (x_{k}) are well correlated and unbiased with respect to continuous final marks (C_k), it can be written approximately as:

$$\mathrm{x}_{\mathrm{k}} \approx a{\cdot}C_k \qquad (2)$$

Where “ a ” (corresponding to a regression coefficient) which is the same for all the students ($1 \le \mathrm{k} \le N$).

If the value of “ a ” is found to be close to unity, this may be a sign that the automated assessment process is working reasonably well, although there will be unavoidable non-zero deviations ($\varepsilon_{\mathrm{k}} = x_k - aC_k$) that should be properly checked.

One simple way of estimating an appropriate value for the hypothetical constant ' a ' is to multiply both sides of the expression ‘ $X_{\mathrm{k}} \approx a{\cdot}C_k$ ’ by the factor ‘ C_{k} ’, which results:

$$x_k C_k \approx a\mathrm{C}_{\mathrm{k}}^2 \qquad (3)$$

And then do the sum from 1 to N in both sides of the equation:

$$\sum_{k=1}^{N} x_k C_k \approx a \sum_{k=1}^{N} \mathrm{C}_\mathrm{k}^2 \Rightarrow a \approx \frac{\sum_{k=1}^{N} x_k C_k}{\sum_{k=1}^{N} \mathrm{C}_\mathrm{k}^2} \tag{4}$$

This operation can be effortlessly calculated on a spreadsheet and thereafter the differences $\varepsilon_\mathrm{k} = x_k - aC_k$ can be checked numerically or graphically.

It has been ascertained (Hastie et al, 2001) that the above value for the constant "a" minimizes the sum of squares $\varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_N^2$, which usually is considered as an appropriate measure of the set of residuals $\{\varepsilon_\mathrm{k}\}$ as a whole, under the hypothesis of a linear and unbiased relationship between the two sets of data that are linked.

The spreadsheet adjusts automatically to fit more sophisticated "regression functions" such as polynomials (up to a certain degree) which may allow for a better approach to the sampled data, and enable some a posteriori useful interpretations, which will be described in greater detail in subsequent sections. Nevertheless, with the polynomial approximations offered by the spreadsheet we cannot obtain a complete flexibility in adjusting the curve to the data cloud, due to the inclusion of parameters (such as the "a" constant).

Instead, as an alternative, the non-parametric joint probability density, $\hat{p}(x,C)$, can be estimated in a similar way to $\hat{p}(x)$ in section 3 of this article, and it can be used to calculate the marginal expected value of x given a specific value of C. This results in the well-kown Nadarya-Watson weighted average [9, 11]:

$$\hat{x}(C) = \sum_{k=1}^{N} w_k x_k \tag{5}$$

Where the "weights" (w_k) continuously depend on the distance between the "C" variable and the sampled data C_k. These last calculations can be also performed on a spreadsheet. In the examples shown in the following section, the parametric approach is employed because of its aforementioned automated results.

4. Results and Discussion: Reliability Analysis of the Continuous "E-assessment"

Each dot in the figures below depicts the status of a single student from the same two student groups considered in section 3.2 of this paper. These graphics are easily obtained by copying the list of final grades from the Continuous Evaluation (C_k) and examination grades (E_k), calculating the final grades ($\mathrm{x_k} = \frac{7}{10}E_k + \frac{3}{10}C_k$) on the spreadsheet, and plotting these data as a cloud of points.

In the graph below, the horizontal axis (abscissae) represents the Continuous Evaluation grades and the vertical axis (ordinates) represents the final grades. The red horizontal line indicates the minimum passing grade, so that the passed and failed students are clearly differentiated.

In order to get clearly classified the passing and failing cases (which are separated by the red horizontal straight line in the next graphics), the total marks have been chosen as ordinates (vertical axis), whereas the continuous final marks define the abscissae (horizontal axis). Moreover, this representation of the results manages somewhat to reduce the data dispersion and considerably facilitate the professor's final decisions.

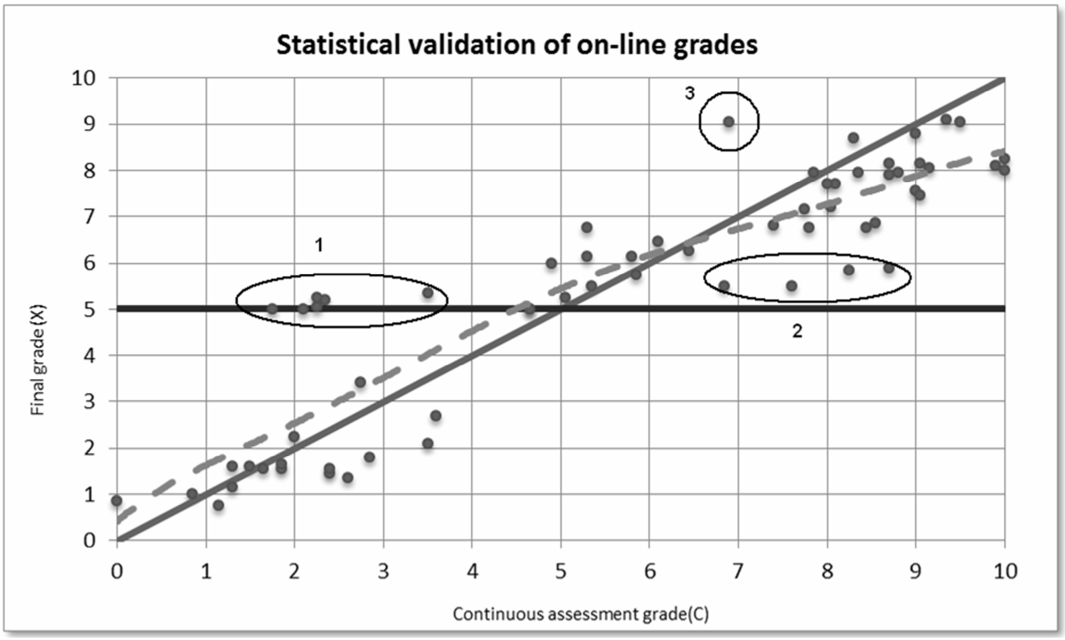

Figure 7. Correlation check of the on-line marks and the final ones. The vertical axis shows the final grade and the horizontal axis shows the on-line assessment grade.

The bisector red line corresponds to a theoretical exact match between C_k and E_k, and consequently between C_k and x_k. In other words, this would mean the student obtained the same grade in the Continuous Evaluation and in the exam. In this theoretical circumstance the examination would not have been necessary at all.

Of course, in reality the blue dots very rarely exactly coincide with the red bisector, although in Figure 7 the difference obtained is reasonably low, as typical deviations are no more than ± 1. More significantly, in this group all students that pass the continuous assessment also obtain a final passing grade. Might we infer from this that the final exam could be eliminated? Our opinion about this issue can be summarized in the Latin adage “Si vis pacem, para bellum” or, “If you wish for peace, prepare for war”’. A final examination is always essential for the complete evaluation of a student’s knowledge on any subject.

The first thing that stands out from a first look at Figure 7 is that this particular set of students have worked well overall and have benefited from the e-learning assessment. Otherwise, the figure would display a chaotic set of dots with no relation, which could very well lead to question or even dismiss this evaluation system.

In a more careful look at Figure 7, the reader may appreciate a number of other things (these are equally numbered in the figure and the paragraphs):

There is a group of six outliers that pass the final score (ordinate ≥ 5) but their continuous assessment grades (abscissae) reflect a poor academic performance throughout the course. The teacher of this group of students might be interested in looking into these specific anomalous cases as cheating could be involved during the final exam.

A dual case is observed in a set of four students, which appear on the opposite side of the bisector line, both above the red horizontal limit. Likewise, the teacher might look for evidence of cheating in the on-line exercises. This is commonly seen of students who have help in completing the exercises throughout the course or students that have not done them themselves.

Exceptions always exist in any large group of students like the one dot shown in Figure 7 under (3). These type of exceptions should always be welcomed as they represent students who disregards the Continuous Evaluation of the subject and decides to prepare exclusively for the exam in his or her own way.

The green dashed line can be obtained with the spreadsheet as a parametric regression curve (polynomial up to a certain degree) which can be very similar to the non-parametric curve defined in the previous section.

On the right side of the axis, the green dashed line is slightly lower than the red bisector. Today’s students, when they find any difference in the way they are asked in the exam to solve a given exercise in contrast with the on-line version, whether it’s the presentation of the data (e.g. graphically instead of numerically) or just the fact of having to hand write the solution, may very well be negatively influenced when solving the exam.

Finally, Figure 8 shows the same analysis for a different group of the same course, specifically the one described in section 4 of this article. There is a greater deviation of the green dashed line from the red bisector, which could imply that the teacher of this particular group has been stricter when his or her exams and in assigning final grades. This can be seen in the three students just under the red horizontal line. There are two cases that should deserve greater attention from the teacher as, if not for the final exam, the students would have passed the subject, possibly undeservingly.

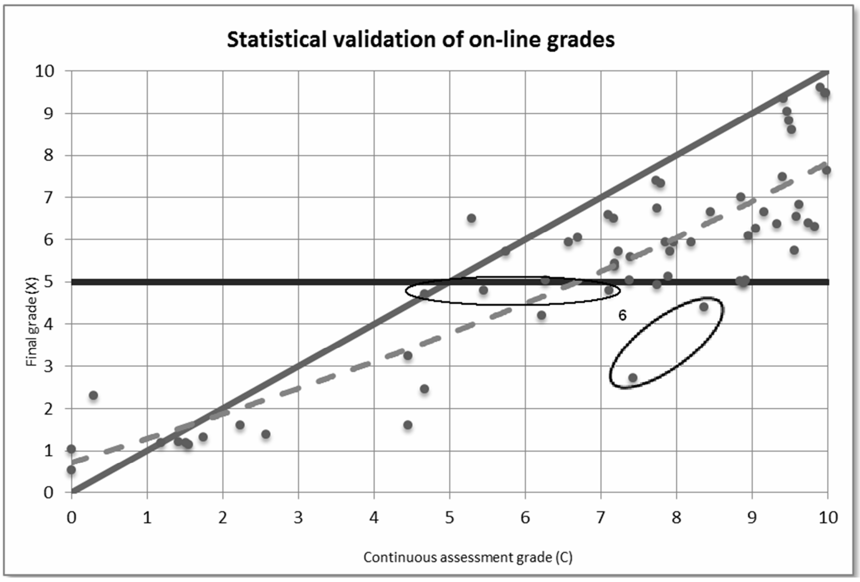

Figure 8. *Equal analysis for the second group of students from the same course and subject. The vertical axis shows the final grade and the horizontal axis shows the on-line assessment grade.*

5. Conclusions

The first and most important conclusion of this article is that simple Statistical Learning techniques can become extremely useful in handling and interpreting the great amounts of numerical data offered by a continuous e-assessment system.

The proposed techniques and analysis can be easily implemented with little knowledge of arithmetic and a standard spreadsheet. These simple techniques provide immediate graphic results which can highlight unforeseen aspects that a more conventional statistical study is unable to capture and that can enormously help in professorial activities. For these reasons, the graphic results obtained can be suitably compared with an 'x-ray' of the learning process and can provide very useful information when comparing different groups of a same course unit.

Finally, the results obtained by the statistical method used (reflected in the second type of graphics presented) are able to validate or invalidate the suitability of this type of blended evaluation procedures that combine continuous on-line assessment and one or several examinations. This includes the analysis of the reliability of the on-line evaluation data, with no need to check every student, except for a reduced number of outliers that can be identified and looked into.

References

[1] Jung, I., Choi, S., Lim, C. & Leem, J. (2002). *Effects of Different Types of Interaction on Learning Achievement, Satisfaction and Participation in Web-Based Instruction.* Innovations in Education and Teaching International. Vol. 39. No 2. pp. 153-162.

[2] Jones, N. & Sze Lau A. M. (2010). Blending learning: widening participation in higher education. Innovations in Education and Teaching International. Vol. 47, No 4. pp. 405-416.

[3] Aznar, A. & Hernando, J. I. (2011). *Herramienta informática de auto-corrección mediante MOODLE.* EvalTrends Proceeding book. Pp. 24-34.

[4] Aznar, A., Hernando, J. I., Cervera, J. & Ortiz, J. (2012). *Educational self-correcting application towards continuous assessment for e-learning in analysis of building structures.* Educación y Futuro. Vol. 2: pp. 2-15.

[5] Aznar, A. & Hernando, J. I. (2014). *A New Automatic On-Line Evaluation for Graphics Applied to Building Structures. EDULEARN14-Proceedings* 1. IATED. pp. 3061–3068.

[6] Aznar, A., Hernando, J. I., Ortiz, H. & Cervera, J. (2014). *Toward the Possibility of Automatic Evaluation of On-Line Graphics*. ICEILT Proceeding book. pp. 388-395.

[7] Aznar, A. & Hernando, J.I. (2014). *Novel Educational Assessment for Bulding Structures: Automatic Evaluation of On-Line Graphics*. *IETC Proceedings book*. Pp. 814-821.

[8] Aznar, A. & Hernando, J. I. (2015). *Novel educational assessment for building structures: Automatic evaluation of on-line graphics.* Social and Behavioral Sciences. Elsevier. Vol 176. pp. 602-609.

[9] Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning Springer*. Vol. 1. New York: Springer.

[10] Hahn, G. H. & Shapiro, S. S. (1967). *Statistical models in engineering*. 130-134.

[11] Fan, J. & Yao Q. (2005). *Non-linear Time Series: Nonparametric and Parametric Methods*. Springer.

[12] Kessler, J. (2012). *Brave new world without teachers, or learning, or thinkers*. Letters, Financial Times, August, 18.