

Efficacy analytics audition project :: Oct 2016

Marco Pierbattista

November 4, 2016

The proposed project consists in developing an algorithm to analyse the data collected by a Pearson e-learning platform. The data consist in a table containing the information about the completion of a number of activities by a number of students from all over the world. The data sample contains information about 81432 activities carried out by a total number of 13158 students from 87 countries in the world. The data file is structured in a way that every student can have as many records as the number of the units completed by the student.

The data parameters available in the data matrix and referred to each unit are: Student ID; Student country; Whether the student is in a course given by a teacher or is an independent student ('t' or 'f' tag); The number or name of the unit at which the entry is referred; The average score obtained by the student on all activities in that unit (in percentage); The percentage of activity completed in the unit in object; A measure of how much the student followed the order suggested by the pedagogue in completing the unit activities (0 following the suggested order, 1 inverting the suggested order).

The first approach consists in extracting the actual number of students from the total number of unit entries and to find the number of students from each country. Figure 1 shows the number of students recorded in the data sample for each country with the fraction of students studying with and without a teacher. The country sample has been split in 4 groups to ease the data visualisation, and the number of students per country was sorted growing from the first to the fourth group. The plots show that there is a clear dichotomy in the data with 5 countries, namely Turkey, Spain, Colombia, and Poland (sorted from the largest to the smallest student population), having approximatively the 80% of the students of the sample. One has to note that for the countries with a non zero number of students studying alone, their percentage is of the order of few percent, with exception of Afghanistan, Alan Island, Thailand, and Venezuela with 12.5%, 20%, 16.7%, and 33.3%, respectively.

From now on we will neglect the country distinction between students and we will perform a data analysis in order to understand how the fact that the activities within a unit are accomplished under the tutorship of a teacher impacts the student results and so their learning ability. Figure 2 shows the distribution of the average score obtained by a student over all activities of a unit for students alone and followed by a teacher. The

comparison highlights that the students followed by a teacher obtain larger grades than students alone: the modes of the distribution differ by 0.1 indicating a larger probability to accomplish correctly the activities under a teacher tutorship. Moreover, the distribution of students followed by a teacher shows a continuous rising trend from the distribution mode to the 100% grade against the drop observed in the students studying alone between 80% and 100% grades. This important feature indicates that in presence of a teacher is actually active a learning process that leads the students to the 100% success in the activity completions while without a teacher there is a learning threshold that does not allow students to reach the complete success in performing activities. The distribution mode for both plots has been computed by excluding the last bin corresponding to 100% grade. Another element in favour of studying with a teacher can be obtained by noting that the 100% grade units represent a larger percentage of the total when a teacher is involved.

Figure 3 shows the distribution of the percentage of completed units with and without the support of a teacher. It enforces the conclusion of the previous paragraph showing that when a teacher is present, the number of totally completed activities reaches the 50% of the total against the 35% when no teacher is involved. Moreover, the number of not completed units is larger when a teacher is absent.

Figure 4 shows the distribution of the inversion parameter that indicates how much the student changed the pedagogical order with which the activities should be undertaken within a unit for units with and without teacher. The two distributions follow the very same trend suggesting that the presence of a teacher scarcely influence the students in departing from the suggested order. The inset figure of each panel show the very same distribution built by excluding all inversion rate equal to 0. They confirm that no appreciable difference shows up in the data.

We studied whether there is a correlation between the percentage of activities completed in one unit and the average grade of the unit activities. Figure 5 shows this correlation for the two cases where a teacher follows or not the learning process. No trend shows up in the data suggesting no linear correlations between the studied variables. The plots reflect the histograms of Figure 3 highlighting a central low region in the plane more poor of points. This indicates a weak dependency between the treated variables suggesting that there are no low average score when the students complete a very small fraction or the 100% of the unit activities. This could be motivated by saying that those who obtain low grades are prevalently those who start the unit activities without the necessary knowledge and whether stop soon after realising to do not have the knowledge (high points density at low % of completion and low grades) or dare to try up to the end completing a section with a equally low grade (high points density at high % of completion and low grades).

We studied whether there is a correlation between the activity completion inversion rate and the average grade of the activities. Figure 5 shows this correlation for the two cases where a teacher follows or not the learning process. Once again, no trend shows up in the data. The apparently weak non linear correlation showing that above a threshold value of average grade, quantified in ~ 0.5 , as the activity inversion rate increase the number of

students with large grade increase is probably fictitious and depending on the lack of data points for high inversion rate coefficients.

Conclusions. We have analysed a data sample containing a 81432 entries from a Pearson e-learning platform about 13158 students from 87 countries. We found that around the 80% of all students come from 4 countries. Just 17 countries out of 87 shows student willing to learn without a teacher and in just 4 cases their percentage is larger than the characteristic value of a few % (large values go from 13% to 30%).

The sample shows that to undertake the learning process under the tutorship of a teacher increases the probability to accomplish the activities of the unit of study with larger grades and represent the main way to upgrade the knowledge up to the top level. This last finding suggest that to study under the tutorship of a teacher guarantees the continuity of the learning process from the absence of knowledge up to the perfect knowledge against the way of studying alone that allows just to reach a knowledge threshold lower than the top level. Moreover, the students followed by a teacher are more productive and tend to complete the proposed units at a largest extent.

The distribution of the rate of inversion of the proposed pedagogical activity order for students with and without a teacher shows that the absence of a teacher does not push students to invert the proposed order. We studied how the average grade changes as a function of the percentage of activities completed in a unit and of the rate of inversion of the proposed pedagogical activity order. We did not find any indication of a correlation between the analysed variables indicating that the ability of a student to learn, poorly depends on these parameters.

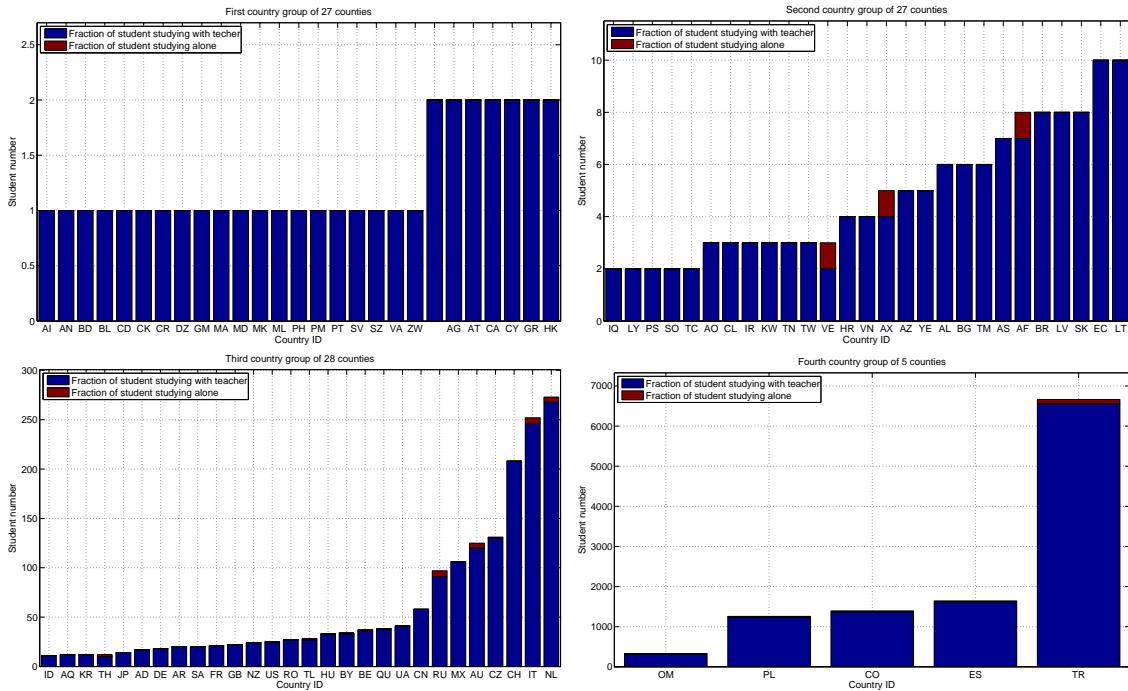


Figure 1: Histograms indicating the number of students for each country. The histograms are arranged from the lowest to the largest student number, from the top left to the bottom right panel, respectively. Students following courses with and without a teacher are indicated in blue and red for each country. The blank field between the counties ZW and AG in the top left panel refers to students with no specified country.

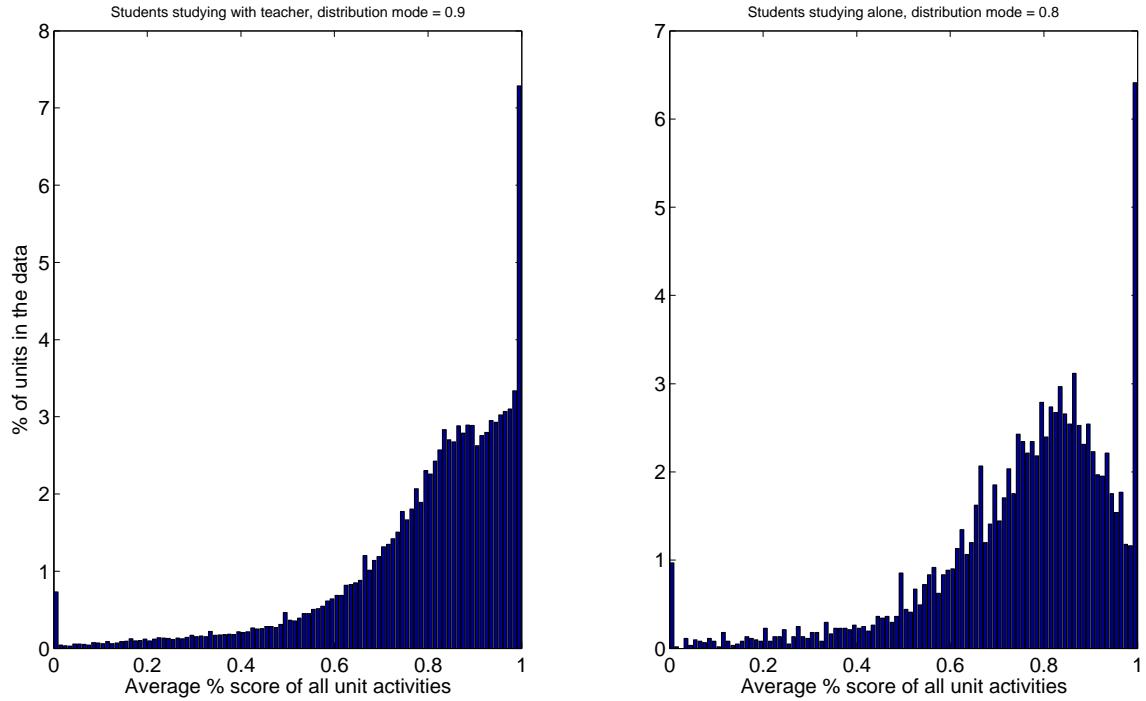


Figure 2: Distributions of the average percentage score obtained by the students while completing the activities of one unit. Left and right panels refer to student following courses with and without a teacher, respectively. The distribution modes have been obtained by excluding the scores equal to 1.

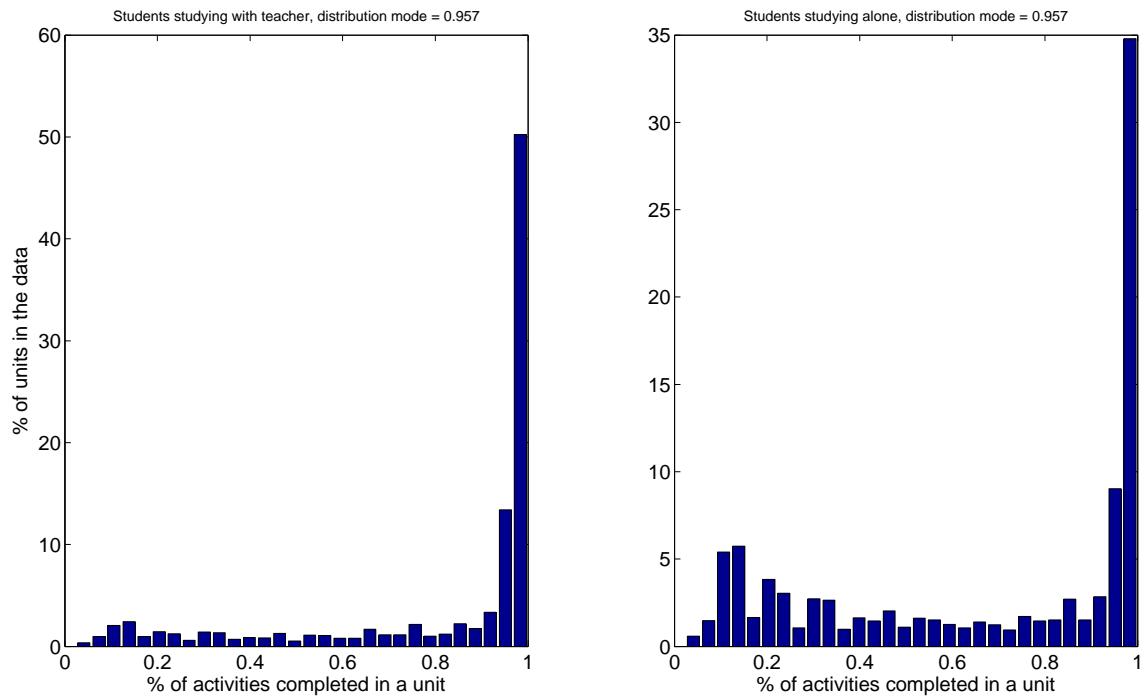


Figure 3: Distributions of the percentage of activities completed in one unit. Left and right panels refer to student following courses with and without a teacher, respectively. The distribution modes have been obtained by excluding the percentage equal to 100%.

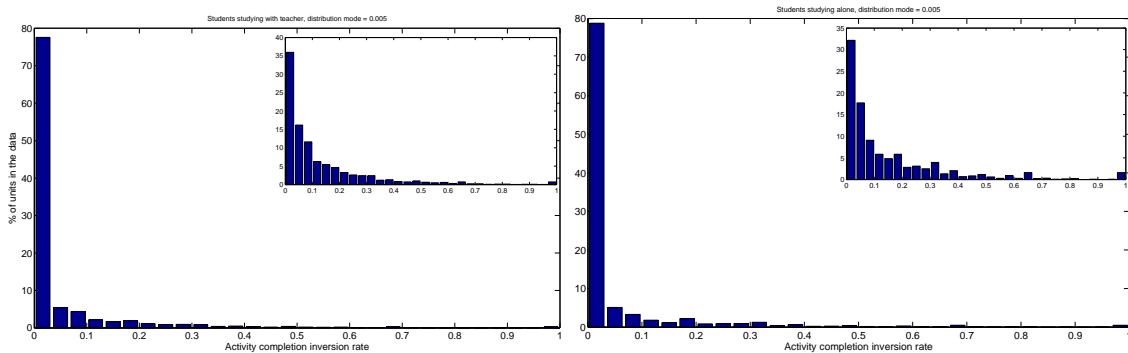


Figure 4: Distributions of the activity completion inversion rate. This index indicates how much the sequence of completion of the activity within one unit, departs from the sequence suggested by a pedagogue. 0 means no departure, 1 means total departure and inversion of the sequence. Left and right panels refer to student following courses with and without a teacher, respectively. Each insight panel shows the same distribution in the large panel obtained by not including the rates equal to 0.

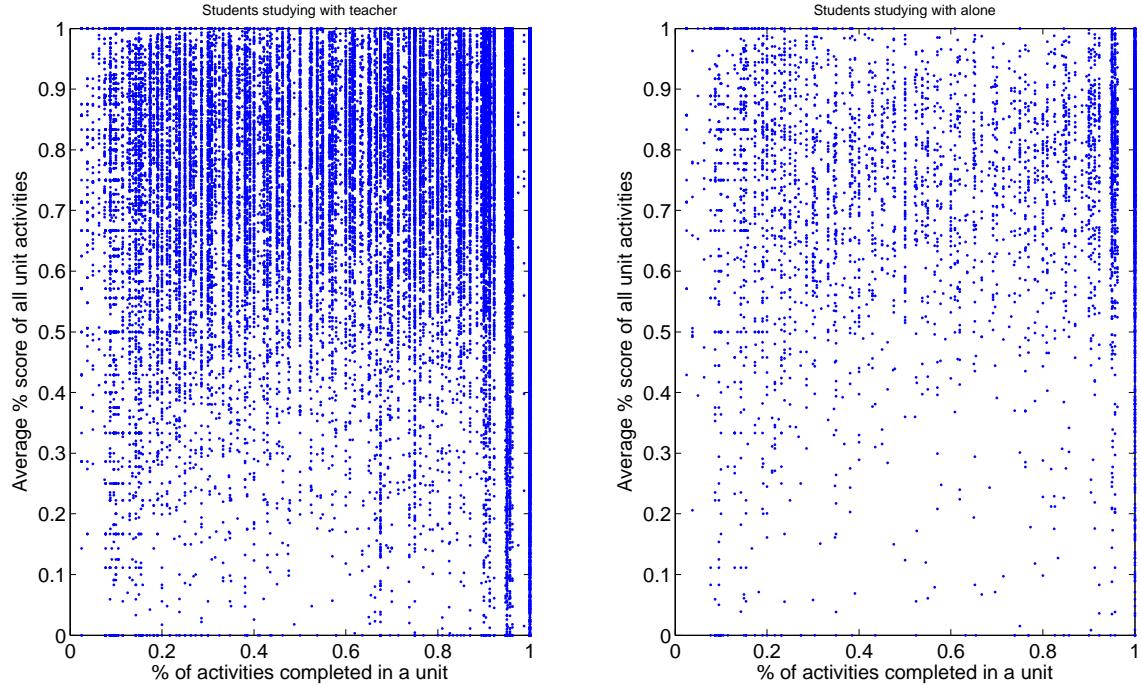


Figure 5: Variation of the average grade of all activities in one unit with the percentage of the activities completed in that unit. Left and right panels refer to student following courses with and without a teacher, respectively.

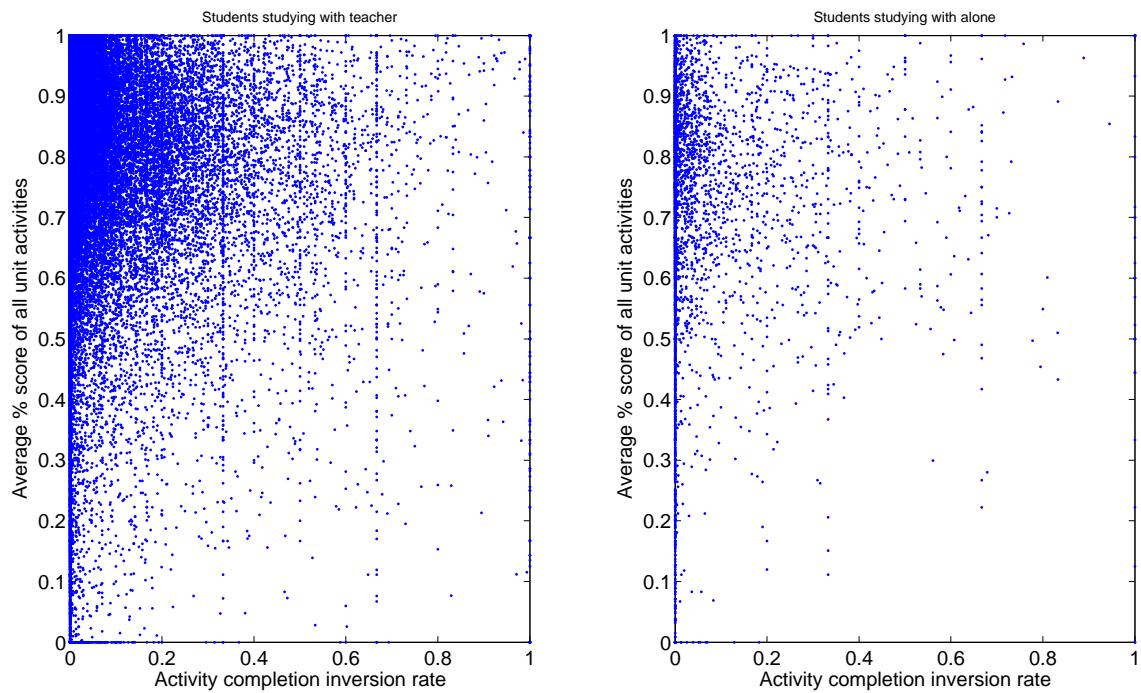


Figure 6: Variation of the average grade of all activities in one unit with the activity completion inversion rate in that unit. Left and right panels refer to student following courses with and without a teacher, respectively.