October 2013

Off-task Behavior and Gaming the System: An Analytic Approach

Alan D. Alezker
Worcester Polytechnic Institute

Follow this and additional works at: https://digitalcommons.wpi.edu/iqp-all

Repository Citation
Off-task Behavior and Gaming the System: An Analytic Approach

An Interactive Qualifying Project Report

Submitted to the Faculty of

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirement for the

Degree of Bachelor of Science

By

Alan Alezker

Date: May 9, 2013

Approved:

Professor Ryan S. Baker, Advisor
Abstract

Educational Data Mining is a growing field, mainly dealing with developing and investigating distinctive types of data that come from educational settings. By implementing these methods, analysts are able to better understand how a student learns and the surroundings in which they learn in.

Research has shown that when using intelligent tutoring software, some students engage in various types of off-task behavior [1]. A type of off-task behavior known as “gaming the system” has been shown to have the strongest effect on a student’s learning [1]. “Gaming the system” refers to finding loopholes in the tutoring system to help achieve a correct answer [2]. There are two very important types of gaming the system that have been observed, “harmful” and “non-harmful” gaming [2]. “Harmful” gaming refers to a student who exploits the system when they arrive at a problem that they have difficulty solving, whereas “non-harmful” gaming refers to a student who is exploiting the system for reassurance on a problem that they know how to solve [2]. The probability that a student was off-task [3], gaming in a fashion associated with worse learning, or gaming in a fashion not associated with worse learning are represented by the acronyms OT(off-task), GH(gaming harmful), GNH(gaming not harmful) respectively [3]. This study was conducted to draw correlations between these three variables and time. By analyzing how each of these dependent variables vary based on time, we can see if students are spending a greater portion of their time off-task and thereby learning less, trying to find loopholes in the tutoring software without learning, or exploiting the system to double-check their work.
Computing Correlations

I found the population mean and standard deviation of the GNHpredthresh, GHpredth, OTpredthresh, time, timeSD, time3SD, time5SD, and timeperact columns. I also took samples of size 50 from the time and GNHpredthresh columns and found the means of both. These means became data points of (time, GNHpredthresh) and I plotted the points from 10 of these samples in a scatterplot to see if there was a correlation between time and GNHpredthresh. The sample size and number of samples were both chosen arbitrarily as I was unsure at the time if this was a valid means of analysis. I later realized that this was a statistically inefficient process of analysis for this particular data set. The analysis of the random samples wasn’t an accurate representation of the population as a whole. In order to accurately give a graphical representation of the given data, I devised an alternative approach in which I would group the points into different intervals and average them.

Results

Before I switched to the alternative method, I used excel to construct a line of best fit which revealed that the time and GNHpredthresh were positively correlated. After constructing a line of best fine, I used a program within excel called StatTools to construct a box-whisker plot of one of my simple random samples. This plot showed that the mean is higher than the median for this specific random sample, which as a result also holds true for the population as a whole. This plot also shows the effect of outliers on the sample, since the values (18.02, 18.206, 20.049, and 40.258) are all significantly higher than the mean. This effect is also prevalent in the population as there 2 data points which are more than double any other data points. Outliers of
this magnitude have a significant effect on measures of variability so I decided to omit these two values when calculating the summary statistics.

**Alternative Method**

I ended up constructing three graphs based on this new approach, and added vertical standard error bars. (Figure 1, 2, 3). The purpose of these graphs was to discover a relationship between time and the dependant variable that was being analyzed (OTpredthresh, GHpredthresh, GNHpredthresh). The resulting correlation analysis showed that OTpredthresh had a clear positive correlation with respect to time, while analysis on the variables GHpredthresh and GNHpredthresh showed no significant statistical trend.
References


