Application of Nonparametric Techniques for Reducing False Positive and Negatives in Student Learning Effectiveness Assessment in Engineering Curricula

Jaewan Yoon
Old Dominion University, Norfolk, VA

Abstract

Frequently employed statistical assessment methodologies are likely to be based upon a routine assumption of normal probability distribution. Pronounced skewness expected in the student response contributed by highly unidirectional and passive characteristic of classroom instruction in STEM/Engineering curricula, differentials in student learning effectiveness depends on specific instruction delivery mechanism and /or level of learning readiness by the students, and most of all, the size of samples representative of student subgroups to be assessed, the very normal probability distribution assumption is routinely abbreviated and over-approximated without much needed verification. Consequently, assessment analysis and comparison would incorrectly amplify the likelihood of significance, and may lead to both false positives and negatives in student learning effectiveness assessment. The motivation of this research is to introduce a new pedagogical assessment framework based on Nonparametric techniques in junction with the Randomized Factorial Design (RFD) experimental design.

Keywords

STEM, Assessment, Statistical, Nonparametric, False Positives and Negatives

Introduction

Lemma of student learning effectiveness assessment in engineering curricula lies in identifying the reproducible patterns in student response, both desirable and depreciable, then successfully reflecting such trigger-feedback relationships in curricular and pedagogical improvements.

Proper and objective methodology employed in most of quantitative assessment hinges on statistics - temporal and longitudinal analysis for the same curriculum in one location or spatiotemporal comparisons study for multi-locations or evaluation of effectiveness in improvements over the prior-to-posterior. Subsequently, frequently employed statistical assessment methodologies are likely to be based upon a routine assumption of normal probability distribution.

Pronounced skewness expected in the student response contributed by highly unidirectional and passive[1,2,4] characteristic of classroom instruction in STEM (Science, Technology, Engineering and Math)/Engineering curricula, differentials in student learning effectiveness depends on specific instruction delivery mechanism and /or level of learning readiness by the students, and most of all, the size of samples representative of student subgroups to be assessed, the very normal probability distribution assumption is frequently abbreviated and over-approximated without much needed verification. Consequently, assessment analysis and comparison would incorrectly amplify the likelihood of significance, and may lead to both false positives and negatives in student learning effectiveness assessment.

The motivation of this research is to introduce a new pedagogical assessment framework based on Nonparametric techniques such as Wilcoxon's Rank Sum method in junction with the Randomized Factorial Design (RFD)[3] experimental design to obviate common problems of concluding false positive and negative conditions so that capturing true student feedbacks at system-level and correctly used to reproduce gains in student learning effectiveness in the future.

Typical Design of Student Learning Effectiveness Assessment

Most typical approach in determining whether the implemented curricula enhancement/modification was successful in enhancing students' learning effectiveness is to compare standard quantitative outcomes (written test, lab reports, exams, final grade) from post-implementation to the pre-implementation condition, i.e., was there any gain in mean scores under Without (=control) and With (=implementation). For smaller scale implementation cases. qualitative data such as observation, survey and interview are also augmented in addition to quantitative scores to supplement the final assessment conclusion.

Once measures of student learning effectiveness are determined prior to implementation, then assessment sequence and the number of replications for collecting assessment data are determined. For most rudimentary assessment, single trial of implementation (=With) is popular with using historical data (=Without) as the control to be compared to. In case when the size of recipient, students or classes, participating in the implementation study is large enough, then the recipient is further sub-grouped to define a control (=Without) and comparison(s) (=With). In either case, each assessment data, written test, lab reports, exams, final grade, are considered as independent event outcomes and discretized assessment for each will be conducted subsequently either by simply comparing their arithmetic means or by using incorrect statistics such as Student t-test (under the popular misconception and justification of "small sample distribution").

Weakness in aforementioned typical assessment procedures includes (i) relatively smaller numbers of students assessed, which often incurs a non-applicability condition for most of statistical procedures, (ii) insufficient number of implementation trials that is highly likely to introduce a strong temporal skewness in assessment data, (iii) frequently abbreviated and over-approximated Normal probability distribution assumption for the assessed data without a much needed verification, (vi) omission of important student learning attributes such as Gender/Age, Associate degree, Cumulative GPA, Current class load, Math SAT scores and other demographic attributes that influence one's learning effectiveness.

As results, identification and verification of possible gains in student learning effectiveness contributed by the implementation are often not assessed correctly (by false positive or false negative in gains) or assessed with embedded ambiguity and uncertainty contrast to initial intention and efforts.

Proposed Randomized Factorial Design (RFD) Assessment Framework

To obviate a number of weakness identified, a Randomized Factorial Design (RFD) assessment framework is proposed. RFD consists with a standard Randomized Complete Block (RCB) experimental design[6], augmented by Factorial multicolinearity to incorporate effects from student attributes in assessing student learning effectiveness. Randomized Complete Block (RCB) experimental design is composed of Treatment and Block components.

Treatment[4] component is defined by dividing students enrolled in implemented course, either by temporal sequence, i.e., different semester and/or year, or same semester with multiple sections including a minimum of one base reference section as "control," or spatially distributed over multi-institutes, etc.

Block component is composed of different levels of classroom instruction, both topic variation and delivery mechanism settings such as traditional instruction vs. online/open access delivery or traditional instruction vs. traditional augmented with innovative pedagogical method, and etc. (Blocks). Elements of student learning effectiveness assessment would consist of common quantitative measures including test, quiz, lab report, exam scores and final grade.

Secondary Block component consists of student learning attributes/demographic attributes, which are often true trigger factors for making implementation successful or not yet they are often not incorporated into assessment framework at all. Suggested student demographic factors[4] to evaluate whether such factors contribute toward the student learning effectiveness gain under implemented classroom instruction settings are cumulative GPA, SAT Math score and High school GPAs (degree program-specific demographic factors) and student age, gender, ethnicity (person-specific demographic factors), and additional relevant factors can be incorporate as needed with flexibility. Use guideline of selected student demographic information should comply with the FERPA guideline[7]. Factorial multicolinearity[5] analysis outcomes are then evaluated by using similar procedural and componental sequence to isolate and identify individual or colinear effect toward student learning effectiveness gain.

Then the proposed Randomized Factorial Design (RFD) assessment framework is expressed by

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_k + \tau\beta\gamma_{ijk} + \varepsilon_{ijk}$$

$$(i = 1, \ldots, \text{Treatments};\ j = 1, \ldots, \text{Blocks};\ k = 1, \ldots, \text{Blocks})$$

where

y_{ijk} = Student response under (i,j,k)th Treatment and Block effects influenced by demographic student learning attributes
μ = Overall mean, Central Tendency (C.T.) (overall student learning effectiveness)
τ_i = i th Treatment effect (spatiotemporal sequence of implementation)
β_j = j th Block effect (topic and delivery mechanism variations for the implementation)
γ_k = k th Block effect (student learning attributes/demographic attributes)
$\tau\beta\gamma_{ijk}$ = Multicolinear Response on (i,j,k) th combination (significant compounding factors incurring gains in student learning effectiveness)
ε_{ijk} = Random error due to (i,j,k) th combination where $\varepsilon \sim NID(0,\sigma^2)$ by Gaussian

Markov theorem

The main approach of the RFD assessment framework is to statistically analyze and assess student learning effectiveness by using Test of Hypotheses (T.H.) so that students' learning can be objectively measured at reproducible system level instead of subjective "did quite well in general, or not look good and so forth." Since student enrollment to subject courses is generally handled centrally by institute's Registrar's Office, and the instructor(s) would have no control or knowledge as to who is registering in which section, one can assume that student composition to a given class/section to be assessed is a random process and ensure compliance to the Family Educational Rights and Privacy Act (FERPA)[7].

Advantage of the RFD approach includes a flexibility to construct spatiotemporally "unbalanced" treatment and block levels so that variations in frequency of class offering or total number of sections offered each time can be assessed and evaluated without depreciating the validity of analysis results.

Assessment outcomes are first verified for Normality using standard Shapiro-Wilk W-statistics[8] at 95% level of confidence at α=0.05. If Normality is confirmed, Multiple Means Comparison (MMC) method with Duncan's Multiple Range Test (MRT) [9] can be used to compare the student learning effectiveness under specific treatment and block level combinations to measure differential gains from the implementation by using a series of cascade Hypothesis.

H_0: All C.T.[contributed by Treatment $_i$ | Block $_j$ | Block $_k$] are statistically equal
H_a: At least one or more differs (at i, j, k and multicolinear factorialization)

Nonparametric Techniques for Reducing False Positive and Negatives in Student Learning Effectiveness Assessment

One of most frequent misapplication of statistical procedures used in assessment process lies in abbreviated and over-approximated Normal probability distribution assumption imposed on the assessed data without much needed verification. For example, paired Student t-test procedure employed in overwhelming percentage of student learning effectiveness assessment process, that is used to compare means of gain between Without (=control) vs. With (=implementation) under the popular misconception and justification of "small sample distribution," which is only valid if the very small sample intrinsically came from a normally distributed population or system with its variance σ^2 unknown.

Therefore, such unverified and forced Normality assumption often produce false Positive (=gain, significant *p*-value) and Negatives (=no gain, insignificant *p*-value) in student learning effectiveness assessment in Engineering curricula. In addition to a relatively small sample size that contributes to common non-Normal condition in assessed data is intrinsic characteristic of Engineering curricula. Compared to other disciplines, Engineering curricula exhibit strong bi- or tri-modal distribution characteristics in student performance, which often directly negate the normal approximation applicability. Based on author's observation, Positive gains in student learning effectiveness induced by innovative pedagogical implementation usually occur at the second quartile, between 25 to 50 percentiles in 300-level classes, if it occurs. On the other hand, 1st, 3rd and 4th quartiles often remain static in their responses regardless the implementation.

In case of non-normality, a robust, median-based, pairwise nonparametric statistics, Wilcoxon Rank Sum statistics[10,11,12], can be employed alternatively. Nonparametric method, in comparison to Normal distribution-based counterparts, compares C.T.s in two continuous distributions in their distribution shape similitude and is not constrained to a pre-defined “mound-shape” conformity. Thus nonparametric method facilitates resilience on non-normal conditions (such as bi- and tri-modal student performance distribution cases common in Engineering classes) with a far smaller sample size requirement (a minimum of eight compared to a minimum of thirty for the Central Limit Theorem used to apply a Normal approximation), and compensates the problem of unintended false Positive and Negatives conclusion in student learning effectiveness assessment in Engineering curricula.

Conclusion

Nonparametric method can be applied to both normally and non-normally distributed populations or systems without any depreciation so that whole assessment strategy can be developed from the beginning with nonparametric method augmented with the proposed Randomized Factorial Design (RFD) assessment framework as the main analysis apparatus. Its smaller sample size requirement will facilitate flexibility in assessment design and logistics, and provide much needed clarity in assessment conclusions by effectively eliminating any false Positive and Negatives – the very lemma of assessment process to identify the reproducible patterns in student learning effectiveness response, both desirable and depreciable, then successfully reflecting such trigger-feedback relationships in curricular and pedagogical improvements.

References

1 Catalano, G.D. and Catalano, K., “Transformation: From Teacher-Centered to Student-Centered Engineering Education,” Journal of Engineering Education, Vol. 88 (1), pp. 59-64, 1999

2 Catalano, G.D., “Some Ideas on the Teaching of Engineering Science: A Student-Centered Approach,” Journal of Engineering Education, Vol. 84 (1), pp. 21-24, 1995

3 Chaturvedi, S., R. Prabhakaran, F. McKenzie, J. Yoon, P. Katsioloudis, 2014, "Engineering Laboratory Instruction in Immersive Virtual Environment (ENLIIVEN),"Final Project Outcome Report, NSF Project 104389

4 Chaturvedi, S., J. Yoon, R. McKenzie, P. Katsioloudis, H. Garcia and S. Ren, 2012, “Implementation and Assessment of a Virtual Reality Experiment in the Undergraduate Themo-fluids Laboratory,” Outstanding Contributions to Student Learning Through Laboratory Experiences, Session T623: Technical Division Experimentation & Lab-Oriented Studies, ASEE Annual Conference, San Antonio, TX, June 10-13, 2012

5 Yoon, J., 2015, “Optimal Site Characterization and Selection Criteria for Oyster Restoration using Multicolinear Factorial Water Quality Approach,” General Oceanography II, OS31A-1981, Annual Fall 2015 Meeting, American Geophysical Union.

6 Montgomery, D.C. and G.C. Runger, 2014, Applied Statistics and Probability for Engineers, 6th ed., John Wiley & Sons.

7 Family Educational Rights and Privacy Act (FERPA), 1974, 20 U.S.C. § 1232g; 34 CFR Part 99, August 21, 1974.

8 Shapiro, S.S. and M.B. Wilk, 1965, “An Analysis of Variance Test for Normality,” Biometrika, 5:591-611.

9 Duncan, D B., 1955, Multiple range and multiple F tests, Biometrics 11:1–42.

10 Wilcoxon, F. 1945, "Individual comparisons by ranking methods". Biometrics Bulletin, 1, 80-83.

11 Kuels, M., 1952, "The Use of the Studentized Range in Connection with an Analysis of Variance," Euphytica, 1:112-122.

12 Sprent, P. and N.C. Smeeton, 2007, Applied Nonparametric Statistical Methods, Fourth Edition, Texts in Statistical Science Series, Chapman & Hall/CRC.

Name of the paper's First Author

Jaewan Yoon is a tenured Associate Professor and University Professor in the department of Civil and Environmental Engineering at Old Dominion University, Norfolk, Virginia. Engineering research interests: Meso-scale numerical water quality transport and fate phenomena, Stochastic system modeling theory and experimental design. Educational research interests: Statistical assessment research, Pedagogical method improvement in engineering curricula, STEM projects and reproducibility.