THE POTENTIAL OF INTRODUCTORY STATISTICS TO PROMOTE DATA LITERACY AND ATTRACT UNDERREPRESENTED MINORITY STUDENTS TO DATA SCIENCE

Sayed A Mostafa and Tamer M Elbayoumi
Department of Mathematics & Statistics, North Carolina A&T State University, USA
sabdelmegeed@ncat.edu

We are in the age of "Big Data" where only few can do their work effectively without performing some sort of quantitative analysis or referring to empirical information. With the introductory statistics course being the main source of quantitative training for undergraduates, efforts should focus on designing introductory statistics to promote data literacy and help students develop statistical reasoning and acquire the data-analytical skills essential for the Data Science era. Such efforts should be particularly supported at minority serving institutions to help with closing the diversity gap in Statistics and Data Science. This study uses extensive students' data from a large minority serving institution in the United States to show the consequences of the consensus (traditional) introductory statistics course design on students' learning gains and experience, and to explore the potential of introductory statistics to promote data literacy and attract minority students to pursue Data Science education and/or careers.

INTRODUCTION

Data analysis is a central component of the Science, Technology, Engineering and Mathematics (STEM) fields as it is essential for justifying and validating scientific findings. The "Big Data" revolution has made data-analytical skills a top attribute that most employers seek in their applicants (NACE, 2019). The big data era has also led to the emergence of the fast-growing interdisciplinary STEM field known as "Data Science", which combines skills and concepts from statistics, mathematics, and computer science. According to a 2017 Business-Higher Education Forum (BHEF) and PwC joint report, the demand for data science skills is growing in all industries (BHEF-PwC, 2017). The report projected that by 2021, 69% of employers will give preference to candidates with data science and analytics (DSA) skills, whereas only 23% of college and university leaders expect their graduates to have those skills. This discrepancy suggests that the shortage of qualified job candidates with DSA skills will likely expand in coming years. To rectify this situation, BHFE and PwC recommend that institutions of higher education "champion data literacy for all: enable all students to become data literate and open more routes to data science" (BHEF-PwC, 2017).

Being required for most STEM majors and many non-STEM majors, the introductory statistics (Intro Stats) course represents the main, and sometimes only, source of quantitative training for undergraduates. Thus, efforts should focus on developing and enhancing Intro Stats courses to prepare students for using solid statistical reasoning in their majors and career fields and to help promote data literacy (BHEF-PwC, 2017). To help students develop statistical reasoning and acquire essential data-analytical skills, we should teach statistics, from the introductory level, as a way of thinking rather than an operational discipline, as traditionally done. It is crucial that data-infused reasoning and data-analytical skills are developed at early stages of college education and reach a broad group of students (e.g., Horton et al., 2015). This study uses rich data collected from Intro Stats students at a large Historically Black University (HBCU) in the United States to assess the students' learning gains and experience in the course and to explore the potential of Intro Stats to promote data literacy and attract minority students to data science.

THE STUDY SETTING

The main Intro Stats course at the study institution is an algebra-based "Introduction to Probability & Statistics" course which serves students from many STEM (~46%) and non-STEM (~54%) fields (18% Psychology; 14% Biology; 13% Kinesiology; 10% Animal Science; 7% Computer Science; 38% Other majors). Most students in the course are from groups that are known to be underrepresented in Statistics and Data Science (~82% are African Americans and ~69% are females). About 7 sections of the course are offered every Fall and Spring semester with near 45 students in each section. In total, the course serves more than 600 students annually. The course consists of 3.0 hours of lectures weekly. It is a semi-coordinated course where section instructors use

In: R Helenius, E Falck (Eds.), *Statistics Education in the Era of Data Science*
Proceedings of the Satellite conference of the International Association for Statistical Education (IASE),
Aug-Sept 2021, Online conference. iase-web.org/Conference_Proceedings.php

the same textbook and online homework system and cover about the same topics, but every section has its own syllabus, class activities, and grading policy. The course is taught in the "traditional manner" where the instructor uses whiteboard or PowerPoint slides or a combination to present definitions, formulas, and procedural steps, and guide students to manually solve practice problems.

Table 1 summarizes the content of the course and the computing approach used for each topic. This course content is consistent with what is known in the statistics education literature as the "consensus" Intro Stats course (e.g., Cobb, 2015). As can be seen from Table 1, the course design puts heavy emphasis on the students learning the mechanics of statistical procedures, applying these procedures to artificial data, and performing the various computations with the assistance of a calculator and/or statistical distribution tables. Table 1 also contrasts the course design with the recommendations outlined in the Guidelines for Assessment and Instruction in Statistics Education (GAISE) College Report endorsed by the American Statistical Association (GAISE, 2016). Clearly, these GAISE recommendations are not reflected in the Intro Stats course design under study and there is a need to revise this course design to incorporate these recommendations. Indeed, due to the data science revolutionization of the practice of statistics, several calls arose asking for amending the undergraduate statistics curriculum starting with alternatives to the consensus introductory course (e.g., Cobb, 2015).

Table 1. GAISE recommendations in the Intro Stats course at the study institution.

Content and computation in the current Intro Stats course at study institution		GAISE Recommendations
1. Introduction (basic concepts) • Descriptive vs inferential statistics • Types of data (quantitative vs qualitative) • Sample vs population • Data collection & Sampling methods **2. Descriptive statistics** • Describing data graphically (manually/using excel construct various types of univariate graphs) • Numerical summaries (manually/using excel compute central tendency and variability measures, and standardized scores) • Bivariate relationships: scatterplots, correlation, and **simple linear regression*** **3. Introduction to probability** • Basic probability terminologies (sample spaces, events, complementary events, and unions and intersections of events) • Additive rule, disjoint events, multiplicative rule, independence, and conditional probability	**4. Probability distributions** • Use formulas to compute expectation and variance of a given discrete probability distribution • Use binomial formula to compute probabilities about binary variables • Use normal table to compute probabilities and percentiles for normal random variables **5. Sampling distribution of sample mean** • Central limit theorem • Use normal table to compute probabilities about the sample mean/proportion **6. Confidence intervals** • Use formula, calculator and normal table or excel to compute confidence interval for the population mean/proportion **7. Hypothesis testing** • Perform 5 systematic steps and use calculator and normal table or excel to compute p-value and reject/retain the null hypothesis about the population mean/proportion	1. Teach statistical thinking. - Teach statistics as an **investigative process** of problem-solving and decision-making. - Give students experience with **multivariable thinking**. 2. Focus on conceptual understanding. 3. Integrate **real data** with a context and purpose. 4. Foster active learning. 5. Use **technology** to explore concepts and analyze data. 6. Use assessments to improve and evaluate student learning.

*Optional/time-permitting topic.

DATA AND METHODS

The present study utilizes two types of students' data to evaluate the design and content of the Intro Stats course described above and to highlight the potential of Intro Stats to raise the level of data science awareness and aspiration among undergraduate students, especially, from minority groups. (i.e., females and African Americans). All data collection was approved by the university's IRB.

Students' Performance Data

The Intro Stats students' test scores on each of four tests and a comprehensive final exam were collected from 37 course sections taught by the course coordinator over a 10-year period (2006 to 2016). The four tests cover the course topics of descriptive statistics; probability; random variables & sampling distributions; and statistical inference, respectively. This data is used to evaluate students' performance across the different parts of the Intro Stats course. These tests were designed by the course instructor and are not validated assessments. To get a better assessment of students' learning

gains from Intro Stats, we also used the Comprehensive Assessment of Outcomes in Statistics (CAOS) test in several Intro Stats sections in Fall 2019, Spring 2020, and Spring 2021, where students completed the same test both at the beginning and at the end of semester. The CAOS test consists of 40 questions assessing concepts covered in the Intro Stats course (e.g., delMas et al., 2007; Tintle et al., 2018). Both the pre- and post-test were taken online and outside of class through the ARTIST Website (https://app.gen.umn.edu/artist/), with some extra credit participation points given as an incentive for completion, but not for performance. We collected valid responses from a total of 135 students who completed both pre- and post-test and spent at least 10 minutes, but no more than 60 minutes, on each test. The conditions on test time were applied to eliminate students who did not engage sufficiently with the test questions or who spent an excessive amount of time on the test, possibly looking up answers (see delMas et al., 2007).

Data Science Awareness and Aspirations Survey

With data science being a relatively new field, most undergraduate students are unaware of the career opportunities it offers. We designed a survey for Intro Stats students to collect data about their levels of awareness and aspiration of data science. The survey questionnaire consisted of 15 questions, and it was completed online as a quiz in the learning management system Blackboard. The survey was administered in several sections of Intro Stats in Fall 2019 and Spring 2020 towards the end of semester to gauge students' awareness and aspirations of data science without any intervention from the course instructors. In Spring and Summer 2021, Intro Stats students took the same survey twice –once at the beginning of semester and another time towards the end of semester after they were given a 45-minute informational presentation about the data science discipline. The presentation was delivered by the section instructor or course coordinator during normal class session near the middle of semester. Students were encouraged to complete the survey outside of class for some extra credit participation points.

Statistical Analysis

All data were processed and analyzed using the R software (R Core Team, 2017). For comparing students' pre- and post-test scores on the CAOS test, we use the Wilcoxon Signed Rank test due to the relatively low term sample size and non-normality of the change in scores. A 0.05 significance level is used throughout to determine the statistical significance of results.

RESULTS

Intro Stats Students Struggle with Probability and Inference Topics

Fig. 1 displays the mean test scores for 4 tests each covering one of the 4 main course topics, and the mean test score on the cumulative final exam (i.e., All). Fundamental statistical concepts that are known to be challenging and likely to be misunderstood by both students and professionals include the concepts of randomness, probability, variability, sampling distributions, statistical tests, and confidence intervals (e.g., Tintle et al., 2018 & Reaburn, 2014). This is because these concepts are traditionally taught, at the study institution as well as many other institutions, in a manner that emphasizes the procedural steps and computations rather than conceptual understanding. Indeed, this "traditional manner" is what most instructors of Intro Stats at the study institution adopt when teaching these concepts. As a result, most of our students exit the course lacking a correct understanding of these important concepts. It is clear from Figure 1 that students tend to perform quite well on the topic of descriptive statistics (class average is consistently above 80%), but their performance drops dramatically on the second test covering basic probability concepts (class average is consistently below 60%), a pattern that continues but improves slightly for the remaining course topics (random variables, sampling distributions and inference). To note, these results are not unique to the students at the study institution; they apply to all students who are taught these concepts in the "traditional manner" described above (e.g., delMas et al., 2007).

Students Do Not Gain Much from the "Consensus" Intro Stats Course

The CAOS test results are summarized in Table 2. We present separate results from data obtained in each of the 3 semesters of data collection and the results from data combined across semesters. The per semester results are reported to account for possible effects of the varied course delivery methods used in these semesters (fully in-person instruction in Fall 2019, in-person

instruction followed by sudden transition to remote instruction due to COVID-19 in Spring 2020, and blended instruction in Spring 2021). In each of the 3 semesters, we observed an improvement in post-test mean % correct compared to the pre-test mean % correct. However, this improvement was statistically significant only in Spring 2021, but this result should be interpreted with caution as the sample size was only 19 students who chose to complete both pre- and post-tests, and therefore, the relatively high improvement of 9.52% may be due to selection bias (i.e., if only high performing students chose to participate). Combining the data from the 3 semesters, the average improvement is 4.26% (p-value = 0.023). Although the overall average improvement is statistically significant, a 4.26% improvement in students' correct answers is not of sufficient practical importance. Indeed, such improvement is 4.84 points below the national average of 9.10% obtained from the sample of 763 students representing 20 higher education institutions (4 two-year/ technical colleges, 10 four-year colleges, and 6 universities) from 14 states (delMas et al., 2007). In fact, even the national estimate suggests that the students' gain from the consensus Intro Stats course is quite modest.

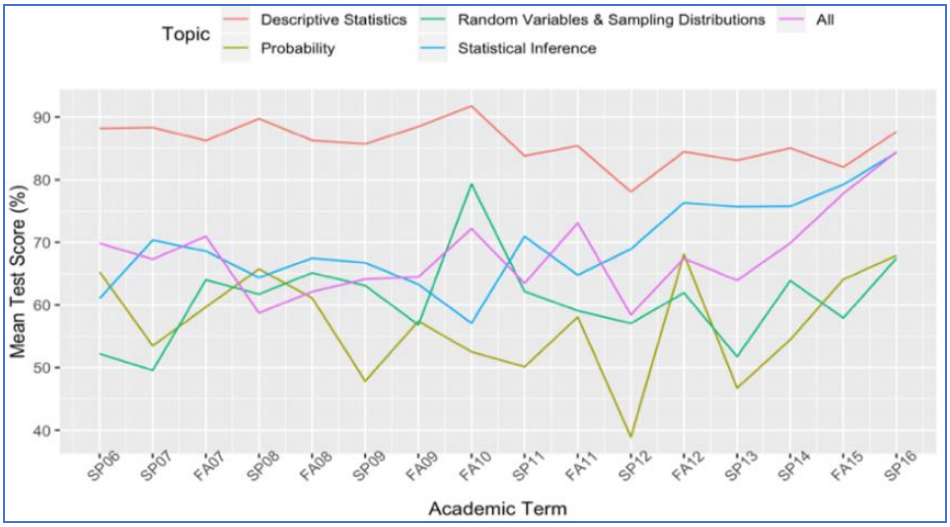

Figure 1. Mean test scores in introductory statistics by course topics (2006-2016)

Table 2. CAOS test results in the Intro Stats course compared to results from a national sample.

Term	Valid Responses	Pretest mean % correct (SD[1])	Posttest mean % correct (SD[1])	Increase in mean % correct (SD[1]) [P-value[2]]
Fall 2019	62	40.17 (10.22)	42.36 (13.90)	2.19 (15.28) [0.241]
Spring 2020	52	41.77 (14.95)	46.56 (17.15)	4.79 (15.96) [0.054]
Spring 2021	19	42.76 (11.57)	52.29 (19.93)	9.52 (15.70) [0.017]
All	133	**41.16** (10.48)	**45.42** (16.39)	**4.26** (15.69) [0.010]
National[3]	763	**44.90**	**54.00**	**9.10** (11.96) [<0.001]

[1]SD is the standard deviation of test scores or change in test scores. [2]P-values are based on results from right-sided Wilcoxon signed rank tests comparing the pretest and posttest scores. [3]Results are taken from [delMas et al., 2007, sec 5.5] and the corresponding p-value was obtained from paired t-test.

Low Data Science Awareness and Aspiration Levels Among Minority Students

The results from the data science awareness and aspirations survey taken by Intro Stats students at the minority serving institution (MSI) under study are summarized in Tables 3 and 4 below. The results indicate that only 34.20% of student participants (n=348) had heard about data science. Additionally, of those who had heard about data science, **only 9.88%** knew that the school offers data science courses and only **4.91%** knew that the school offers a data science certificate/degree. The survey also revealed that male students are more likely to be aware of data science than female students (48.98% vs 28.40%). Despite the low level of data science awareness among our Intro Stats students, the survey showed that a significant portion of those students (both males and females) are eager to consider taking data science courses or even completing a data

science certificate/degree program if they were provided with more information about the field and related career opportunities (44.54% and 39.26%, respectively, see Table 3). The survey also revealed that data science awareness and aspiration vary by student's major (see Table 4); Computer Science students are likely to hear about data science early in college (90.91%) and 30.30% of them plan to take data science courses before graduation, whereas only few of Animal Science, Biology, Kinesiology and Psychology majors enrolled in Intro Stats have heard about data science by their sophomore year (29.63%, 26.39%, 9.09% and 17.11%, respectively) and almost none of them plan to take data science courses before graduation. These data suggest that the Intro Stats course is a fertile ground that can be used for attracting a large pool of students to data science and STEM.

Table 3. Data Science (DS) awareness and aspirations among Intro Stats students by gender.

Gender	Ever heard about DS?		Total	Plan to take DS course(s) before graduation?			Total
	No # (row %)	Yes # (row %)		No # (row %)	Maybe if given more info # (row %)	Yes # (row %)	
Females	179 (71.60)	71 (28.40)	250	132 (52.59)	107 (42.63)	12 (4.78)	251
Males	50 (51.02)	48 (48.98)	98	36 (37.11)	48 (49.48)	13 (13.40)	97
Total	229 (65.80)	119 (34.20)	348	168 (48.28)	155 (44.54)	25 (7.18)	348
	If ever heard about DS, do you know if the university offers DS courses?			Plan to complete DS certificate/degree?			
Females	228 (92.31)	19 (7.69)	247	148 (58.96)	95 (37.85)	8 (3.19)	251
Males	82 (84.54)	15 (15.46)	97	46 (46.94)	42 (42.86)	10 (10.20)	98
Total	310 (90.12)	34 (9.88)	344	194 (55.59)	137 (39.26)	18 (5.16)	349

Table 4. Data Science (DS) awareness and aspirations among Intro Stats students by major.

Major	Ever heard about DS?		Total	Plan to take DS course(s) before graduation?			Total
	No # (row %)	Yes # (row %)		No # (row %)	Maybe if given more info # (row %)	Yes # (row %)	
Animal Sci	19 (70.37)	8 (29.63)	27	22 (81.48)	4 (14.81)	1 (3.70)	27
Biology	53 (73.61)	19 (26.39)	72	33 (45.83)	36 (50.00)	3 (4.17)	72
Computer Sci	3 (9.09)	30 (90.91)	32	8 (24.24)	15 (45.45)	10 (30.30)	33
IT	14 (51.85)	13 (48.15)	26	9 (33.33)	15 (55.56)	3 (11.11)	27
Kinesiology	20 (90.91)	2 (9.09)	19	13 (68.42)	6 (31.58)	0 (0.00)	19
Psychology	63 (82.89)	13 (17.11)	76	43 (56.58)	31 (40.79)	2 (2.63)	76
Other	57 (62.64)	34 (37.36)	94	40 (42.55)	48 (51.06)	6 (6.38)	94

Intro Stats Can Promote Data Science Among Minority Students

Noting that of those who have heard about data science, 37.61% have heard about it from one of their classes, the authors attempted to evaluate the potential impact of a small class intervention –in the form of a short informational presentation given by course instructor/coordinator about the data science field during one of the class sessions for in-person sections or as a pre-recorded video for online sections– in raising the level of students' awareness and aspirations of data science. The levels of data science awareness and aspiration before and after the intervention are summarized in Table 5.

The results in Table 5 clearly show that the intervention led to substantial improvement in students' awareness of data science, especially for female students. However, the improvement in students' aspiration of data science was insubstantial. Indeed, the percentage of females eager to consider taking data science courses or completing a certificate/degree has declined considerably after the intervention without corresponding proportional increase in the percentage of females planning to take data science courses or complete a certificate/degree. The somewhat negative impact of the intervention on students' aspiration of data science, especially for females, might be explained by the fact that 25.33% of female students in the sample have Kinesiology major, which is a non-technical major, and most of them were driven away from data science when they learned about its need for technical programing and quantitative skills (% of "No" answer for data science course aspiration increased from 31.58% to 68.42% among female kinesiology students after the intervention). It should be noted that the results in Table 5 should be interpreted with caution due to the small sample size and possible bias in sample selection due to the voluntary nature of participation in the survey.

Table 5. DS awareness and aspirations among Intro Stats students by gender (before vs after intervention).

Gender	Ever heard about DS? Yes # (row %)			Plan to take DS course(s) before graduation? Maybe if given more info # (row %)		Yes # (row %)		Total
	Before	After						
	Yes # (row %)	Yes, from Intro Stats	Yes, from outside Intro Stats	Before	After	Before	After	
Females	25 (33.33)	50 (66.67)	15 (20.00)	39 (52.00)	24 (32.00)	10 (13.33)	10 (13.33)	75
Males	15 (60.00)	14 (56.00)	7 (28.00)	16 (64.00)	9 (36.00)	4 (16.00)	10 (40.00)	25
Total	40 (40.00)	64 (64.00)	22 (22.00)	55 (55.00)	33 (33.00)	14 (14.00)	20 (20.00)	100
	Do you know if the university offers DS courses?							
	Yes # (row %)	Yes, I knew this from Intro Stats	Yes, I knew this from outside Intro Stats	Plan to complete DS certificate/degree?				
Females	10 (13.51)	42 (56.76)	4 (5.41)	37 (49.33)	27 (36.00)	5 (6.67)	7 (9.33)	75
Males	6 (24.00)	13 (52.00)	5 (20.00)	17 (68.00)	12 (48.00)	3 (12.00)	4 (16.00)	25
Total	16 (16.00)	55 (55.00)	9 (9.00)	54 (54.00)	39 (39.00)	8 (8.00)	11 (11.00)	100

DISCUSSION

The traditional Intro Stats curriculum and design, centered around the normal distribution for teaching statistical inference, have been used for more than a decade in most colleges and universities within and outside the United States (e.g., Tintle et al., 2018). This paper highlighted some of the undesirable consequences of such curriculum on students' learning of statistics at a large MSI in the United States. The paper also presented results that show that the Intro Stats course has great potential to promote data science among minority students, especially females. More data collection is needed to strengthen the findings of the paper in this regard. Future studies should also focus on developing designs for the Intro Stats course to enhance the statistical and quantitative skills of and promote data science literacy among students, especially those underrepresented in Statistics and Data Science.

REFERENCES

BHEF-PwC: Business Higher Education Forum and PwC. (April, 2017). *Investing in America's data science and analytics talent: The case for action*. pwc.com/us/dsa-skills

Cobb, G. (2015). Mere Renovation is Too Little Too Late: We Need to Rethink our Undergraduate Curriculum from the Ground Up. *The American Statistician*, *69*, 266-282.

delMas, R.C., Garfield, J., Ooms, A. & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28-58.

GAISE College Report ASA Revision Committee (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report*. http://www.amstat.org/education/gaise

Horton, N.J., Baumer, B.S. & Wickham, H. (2015). Setting the stage for data science: integration of data management skills in introductory and second courses in statistics. *CHANCE*, *28*(2):40-50.

NACE: National Association of Colleges and Employers. (2019). *Job Outlook.* https://www.naceweb.org/talent-acquisition/candidate-selection/employers-want-to-see-these-attributes-on-students-resumes/

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Reaburn, R. (2014). Introductory statistics course tertiary students' understanding of p-values. *Statistics Education Research Journal*, *13*(1), 53-65.

Tintle, N., Clar, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T. & Vanderstoep, J. (2018). Assessing the Association Between Precourse Metrics of Student Preparation and Student Performance in Introductory Statistics: Results from Early Data on Simulation-Based Inference vs. Nonsimulation-Based Inference. *Journal of Statistics Education*, *26*(2), 103-109.