Integration of Data Science and Computing into Introductory Statistics

Sayed Mostafa & Tamer Elbayoumi

I. Introduction

Background

- Nolan and Temple Lang's (2010)[1] paper on "Computing in the Statistics Curriculum" led many statistics educators to advocate integrating computing in statistics courses starting with the Introductory Statistics (Intro Stats) course.
- The need for a computationally-infused statistics curriculum was further signified by the fast-growing demands on graduates with computational and data analytical skills who can work as data scientists.
- See the *Journal of Statistics & Data Science Education Special Issue on* "Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking"[2].

Objectives

We aim to

1. introduce an Intro Stats course design that integrates computing as a core component of the course and
2. evaluate the effectiveness of such design for
   - enhancing students' statistical gains,
   - boosting students' levels of data science (DS) awareness, aspiration, and readiness, and
   - improving students' overall course performance.

II. Computationally-Infused Intro Stats

Figure 1 – Proposed Intro Stats Course Design – Phase I (Fall 2022)

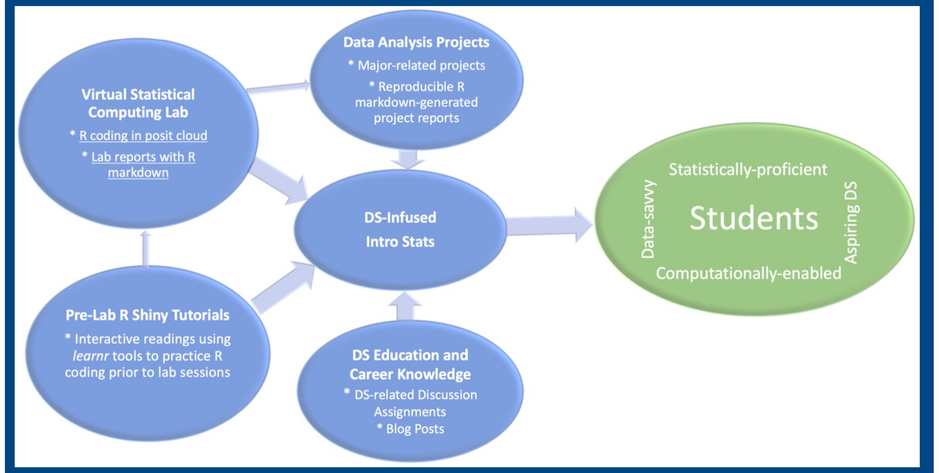
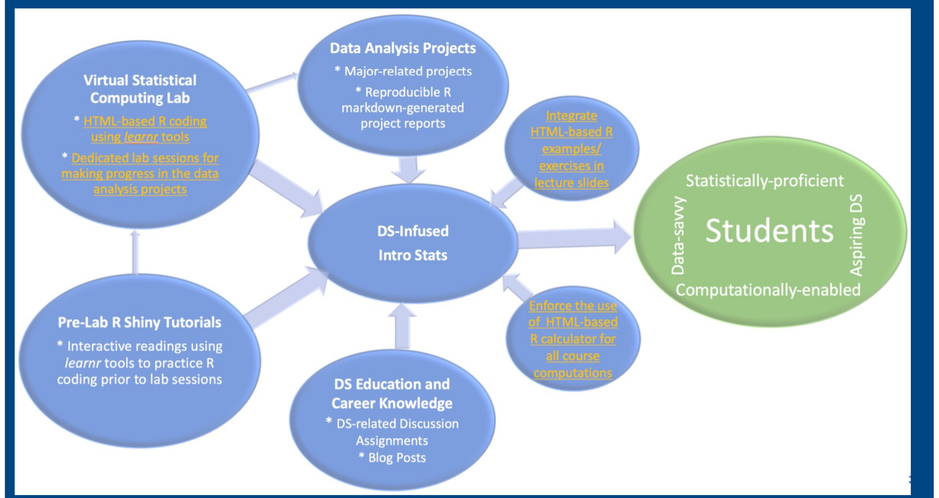


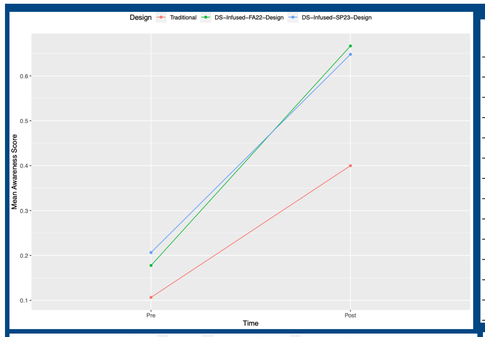
Figure 2 – Proposed Intro Stats Course Design – Phase II (Spring 2023)

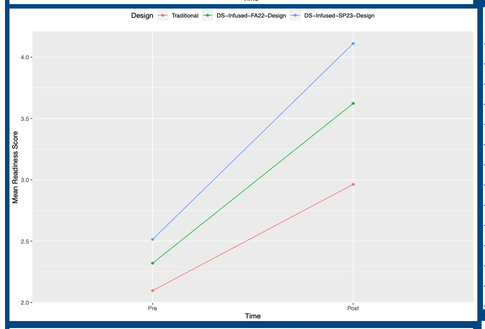


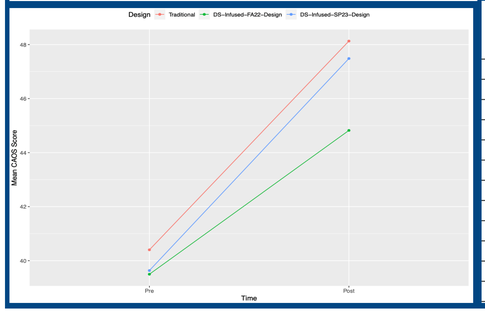
III. Methods

- **DS awareness, readiness & aspirations survey**
  - Students completed a DS awareness, readiness, and aspirations survey in Qualtrics (pre-survey and post-survey)
- **Statistical learning gains**
  - Students completed a revised version of the CAOS[3] (Comprehensive Assessment of Outcomes in Statistics) scale [pre-test and post-test]
- **Implementation:**
  - Phase I: 2 treatment sections and 2 control sections
  - Phase II: 4 treatment sections and 2 control sections

IV. Key Results







▶ Gains in DS awareness: regression on course design

Regression Term	Estimate	LCL	UCL	p.value	Sig.
Intercept	0.43	-0.13	0.99	0.1301	Not Sig.
Type: DS-Infused-FA22-Design	0.26	0.08	0.44	0.0050	**
Type: DS-Infused-SP23-Design	0.18	0.02	0.33	0.0230	_*
Sex: Male	-0.09	-0.24	0.06	0.2439	Not Sig.
Race: Not Black	-0.14	-0.32	0.04	0.1225	Not Sig.
PELL Recepient: Yes	-0.21	-0.40	-0.02	0.0290	_*
Rural: Yes	0.11	-0.10	0.32	0.3135	Not Sig.
Residency: Out-of-State	0.05	-0.11	0.21	0.5109	Not Sig.
STEM: Yes	-0.15	-0.29	0.00	0.0431	_*
AP Stat: Yes	0.09	-0.07	0.25	0.2813	Not Sig.
Pre-Course Cum GPA	0.00	-0.14	0.14	0.9858	Not Sig.
Attendance	0.00	0.00	0.01	0.6380	Not Sig.

▶ Gains in DS readiness: regression on course design

Regression Term	Estimate	LCL	UCL	p.value	Sig.
Intercept	0.94	-0.53	2.42	0.2087	Not Sig.
Type: DS-Infused-FA22-Design	0.43	-0.04	0.90	0.0740	Not Sig.
Type: DS-Infused-SP23-Design	0.84	0.46	1.22	0.0000	****
Sex: Male	-0.16	-0.54	0.22	0.4079	Not Sig.
Race: Not Black	-0.21	-0.67	0.25	0.3687	Not Sig.
PELL Recepient: Yes	-0.62	-1.11	-0.13	0.0130	_*
Rural: Yes	-0.58	-1.10	-0.06	0.0296	_*
Residency: Out-of-State	0.30	-0.09	0.70	0.1300	Not Sig.
STEM: Yes	-0.06	-0.44	0.32	0.7428	Not Sig.
AP Stat: Yes	-0.27	-0.68	0.14	0.1934	Not Sig.
Pre-Course Cum GPA	0.15	-0.19	0.49	0.3932	Not Sig.
Attendance	0.00	-0.01	0.01	0.9119	Not Sig.

▶ Statistical learning gains (Change): regression on course design

Regression Term	Estimate	LCL	UCL	p.value	Sig.
Intercept	0.25	-19.37	19.87	0.9800	Not Sig.
Type: DS-Infused-FA22-Design	-0.82	-7.34	5.70	0.8049	Not Sig.
Type: DS-Infused-SP23-Design	0.28	-5.08	5.65	0.9172	Not Sig.
Sex: Male	-1.54	-6.46	3.37	0.5373	Not Sig.
Race: Not Black	1.15	-5.53	7.84	0.7340	Not Sig.
PELL Recepient: Yes	0.69	-5.52	6.89	0.8277	Not Sig.
Rural: Yes	-5.74	-12.38	0.90	0.0901	Not Sig.
Residency: Out-of-State	-6.29	-11.71	-0.88	0.0229	-*
STEM: Yes	1.14	-3.68	5.95	0.6422	Not Sig.
Pre-Course Cum GPA	-1.96	-6.69	2.76	0.4139	Not Sig.
Attendance	0.18	0.01	0.36	0.0408	-*

V. Conclusions

Integration of DS tools/knowledge into Intro Stats was associated with

- significant gains in students' levels of DS awareness under both designs
- significant gains in students' levels of readiness for DS under the revised design (phase II) only
- significant drop in students' aspirations of DS under phase I design only
- modest (insignificant) statistical learning gains under both designs

VI. References

1. Nolan, D., and Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, 64, 97–107.
2. Horton, N.J. and Hardin, J.S. (2021). Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking. *Journal of Statistics and Data Science Education*, 29:sup1 S1-S3.
3. Tintle, N., Clar, J., Fischer, K., Chance, B., Cobb, G., Roy, S., Swanson, T. and 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.

North Carolina Agricultural and Technical State University

NSF Targeted Infusion Project Grant #HRD2106945

North Carolina A&T State University
Department of Mathematics & Statistics
https://introtostatncat.github.io
sabdelmegeed@ncat.edu & tmelbayoumi@ncat.edu