2021 United States Conference On Teaching Statistics

Expanding Opportunities

Designing Introductory Statistics to Attract Minority Students to Data Science

Sayed Mostafa[1] & Seongtae Kim[2]

Department of Mathematics & Statistics
North Carolina A&T State University

[1] Assistant Professor & Coordinator of Introductory Statistics

[2] Associate Professor & Coordinator of the Undergraduate Program in Statistics & Data Science

Session Outline

- **Part I**: The Status of Intro Stats at NC A&T (25 minutes)
  - Course design & content
  - Students gains from the course
  - GAISE recommendations in Intro Stats
  - Discussion
- **Part II**: Data Science Awareness & Aspirations among Intro Stats Students (30 minutes)
  - DS awareness & aspirations survey
  - The potential of Intro Stats to promote DS
  - Discussion
- **Part III**: Redesigning Intro Stats to Promote DS (20 minutes)
  - Revised course content
  - Virtual statistical computing lab in Intro Stats
  - Integration of DS knowledge and tools in the course
  - NSF grant
  - Discussion

About NC A&T

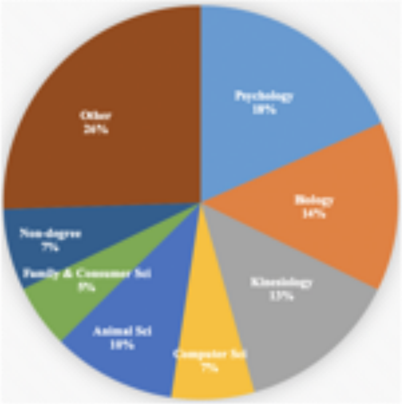
- North Carolina Agricultural & Technical State University (NC A&T) is the largest Historically Black College and University (HBCU) in the nation (>12,000 Fall 2020 enrollment)
- Top producer of Aferican American STEM graduates
- The A&T Four:

About NC A&T

- NC A&T is the largest HBCU in the nation
- Top producer of Aferican American STEM graduates
- Dr. Ronald McNair:

Introductory Statistics at NC A&T

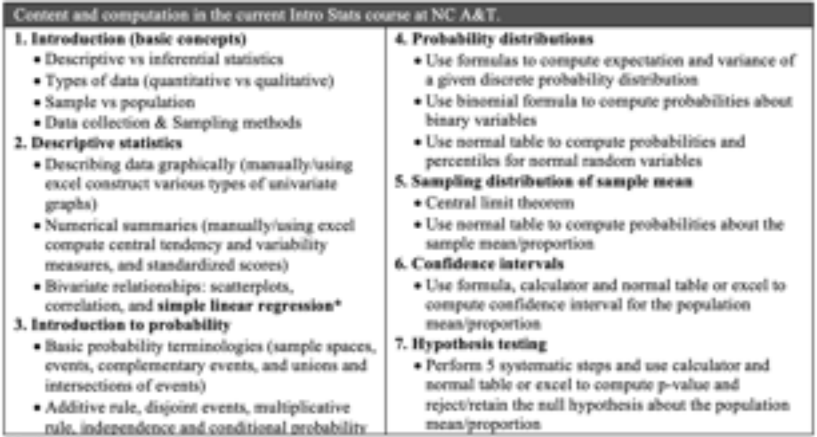
- "Introduction to Probability & Statistics" (MATH224)
- Algebra-based semi-coordinated 3.00 credits course
- Serves STEM (∼46%) and non-STEM (∼54%) majors



- About 7 sections (∼45 students in each section) each semester

Introductory Statistics at NC A&T

- Course Design & Content:



Content and computation in the current Intro Stats course at NC A&T.

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

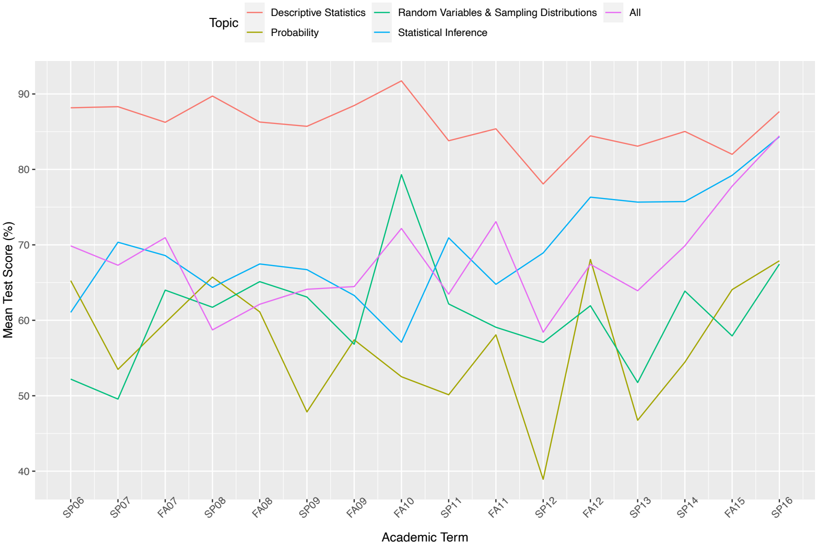
- Use formula, calculator and normal table or excel to compute confidence interval for the population mean/proportion

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

*Optional/time-permitting topic.

Students Performance in Intro Stats at NC A&T

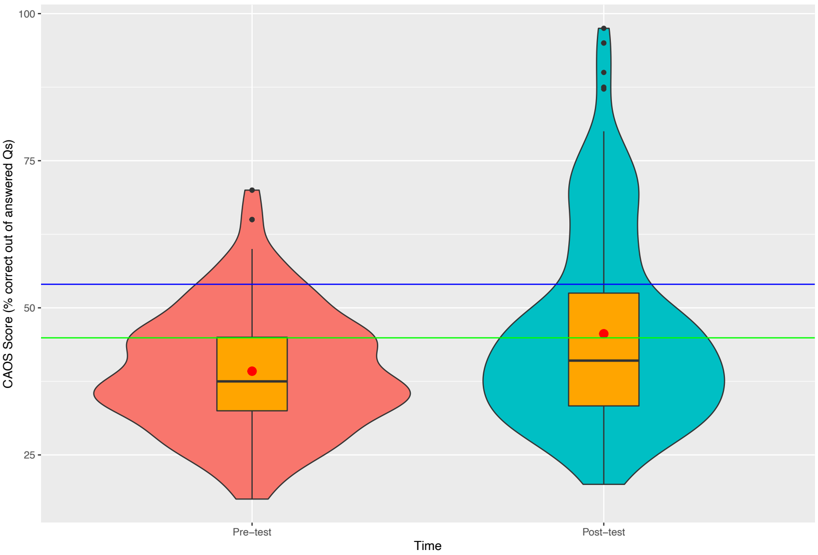


n ranges from 37 to 113 in different semesters

Students Learning Gains from Intro Stats

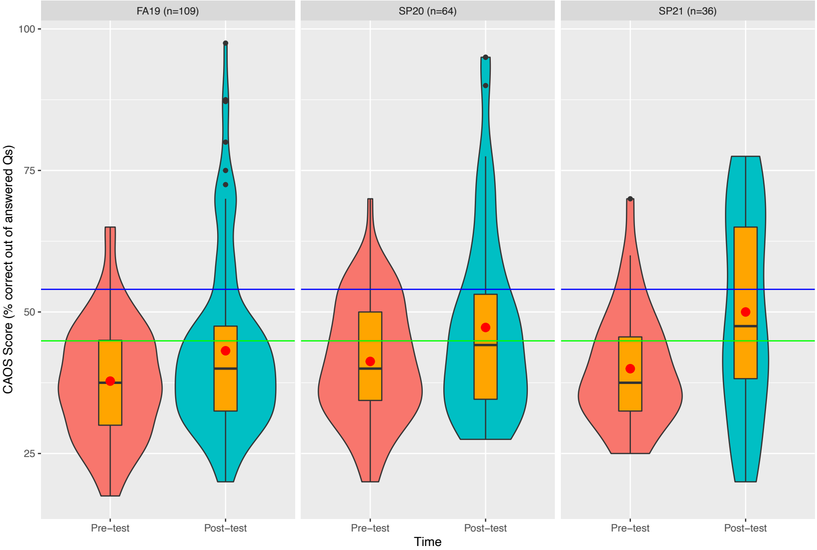
- The Comprehensive Assessment of Outcomes in Statistics (CAOS) test was used to measure students learning gains
- CAOS consists of 40 questions assessing concepts covered in the Intro Stats course (e.g., delMas et al., 2007)
- CAOS is commonly used for assessing students gains from Intro Stats (e.g., delMas et al. (2007); Tintle et al. (2018))
- Students in multiple sections of Intro Stats completed the test at the beginning and at the end of semester during Fall 2019, Spring 2020 and Spring 2021
- Students were encouraged to complete the pre- and post-test by offering some extra credit
- Student's response was considered valid if s/he completed both pre- and post-test and spent $>=$ 5 minutes on each test

Students Learning Gains from Intro Stats



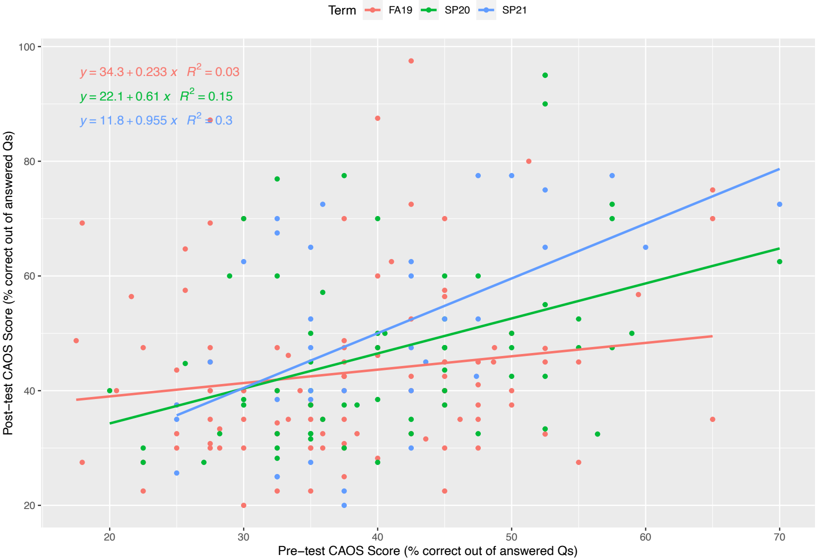
n=209 ; Horizontal Lines = Pre/Post Test National Averages (delMas et al., 2007)

Students Learning Gains from Intro Stats



Horizontal Lines = Pre/Post Test National Averages (delMas et al., 2007)

Students Learning Gains from Intro Stats

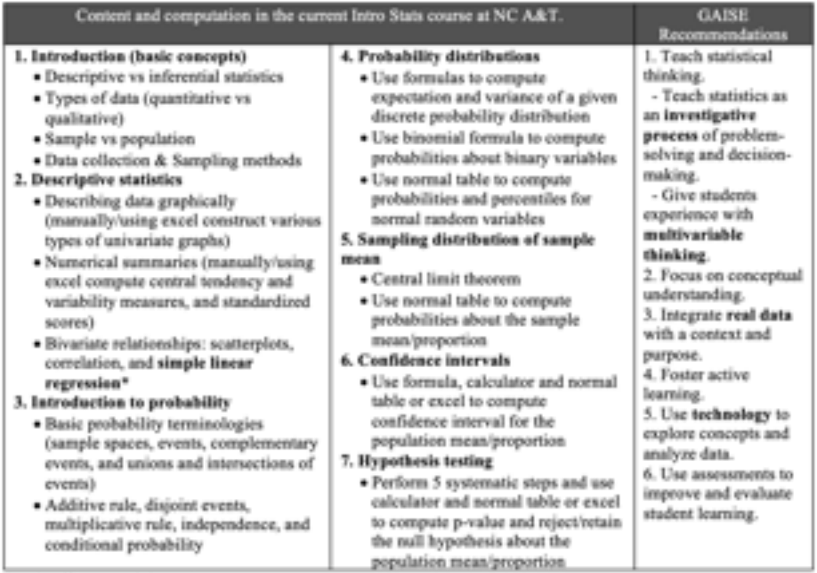


GAISE Recommendations

The Guidelines for Assessment and Instruction in Statistics Education (GAISE):

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.

GAISE Recommendations in Intro Stats at NC A&T



Content and computation in the current Intro Stats course at NC A&T.		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.

Discussion

- How similar is the Intro Stats course design at your institution to the design used at NC A&T?
- To what extent are the GAISE recommendations reflected in your Intro Stats course design?
- Have you ever attempted to measure students learning gains from the Intro Stats course?
  - What scale did you use (CAOS or other)?
  - How do your results differ from the ones presented in this session?

Data Science at NC A&T

- NCA&T offers several data science tracks to prepare students **from any major** become data scientists:
    - **Undergraduate Certificate in Data Science & Analytics**
    - **Post-Baccalaureate Certificate in Data Analytics**
    - **MS in Data Science and Engineering**
    - **PhD in Data Science & Analytics**

Data Science at NC A&T

Undergraduate Certificate in Data Science & Analytics:

- Curriculum Requirements:

A student seeking the Undergraduate Certificate in Data Science and Analytics (DSA) must complete 15 credit hours of DSA-related undergraduate coursework:

- Two DSA core courses (6 credit hours): STAT 324 (Stat Methods for Data Analysis) and MATH 365 (Intro to Data Science) or COMP 365.
- Two DSA electives (6 credit hours): from STAT, BIOL, COMP, CST, ISEN, MGMT, or PHYS
- A DSA-related capstone project (3 credit hours).

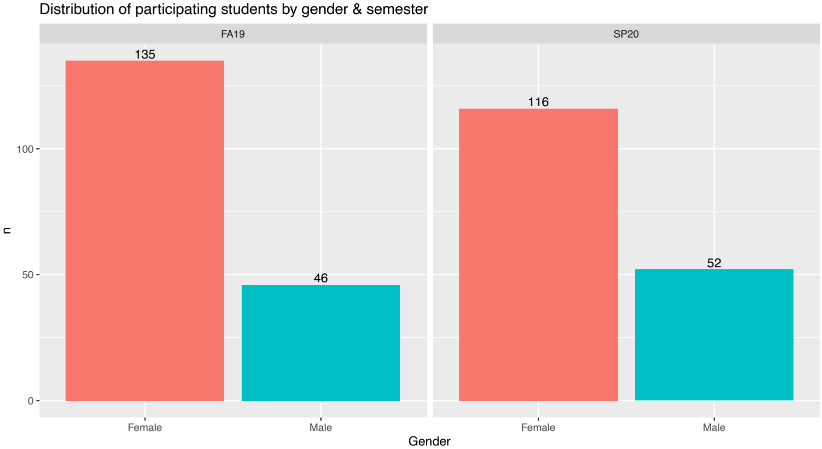
Data Science at NC A&T

- NC A&T's Students in the ASA's DataFest (2017):

University of Waterloo
Canada
University of Toronto
Berkeley
Chapman University
UCLA
UC Riverside
Cal Poly
Pomona College
USC
Nebraska
The University of New Mexico
Miami University
Emory University
Purdue University
Duke
Davidson
Meredith
WF
Elon
RIT
Middlebury
VT
Amherst College
UMass

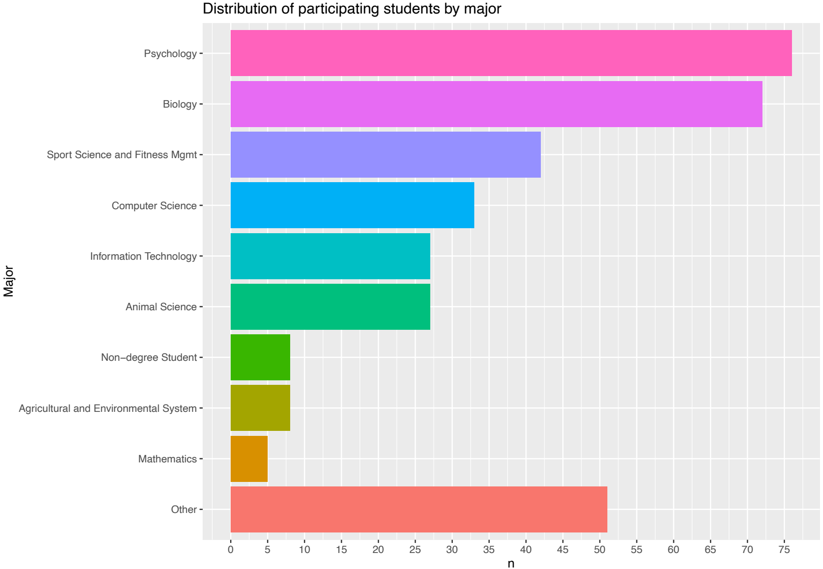
NCA&T Students' Awareness of Data Science

- With DS being a relatively new field, most undergraduate students are unaware of the career opportunities it offers!!
- We surveyed the NC A&T Intro Stats students to collect data about their awareness and aspirations of DS.

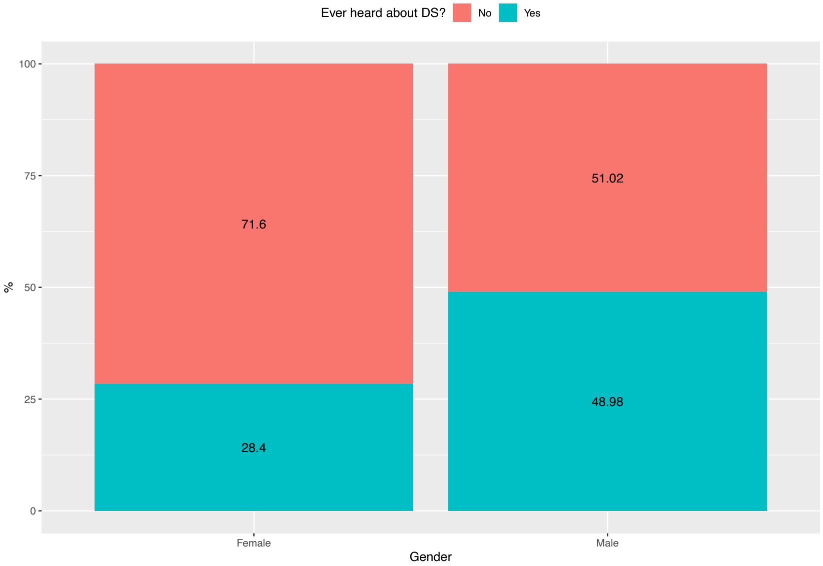


NCA&T Students' Awareness of Data Science

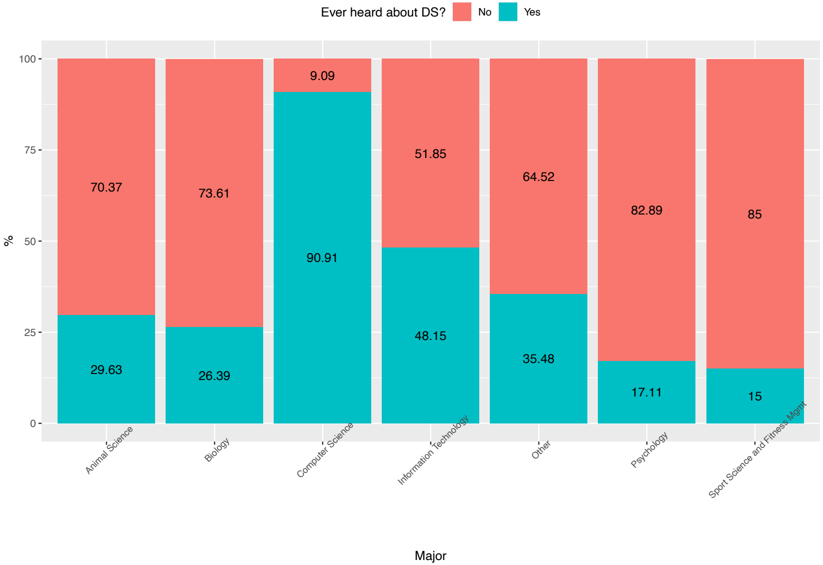
- We surveyed the NC A&T Intro Stats students to collect data about their awareness and aspirations of DS.



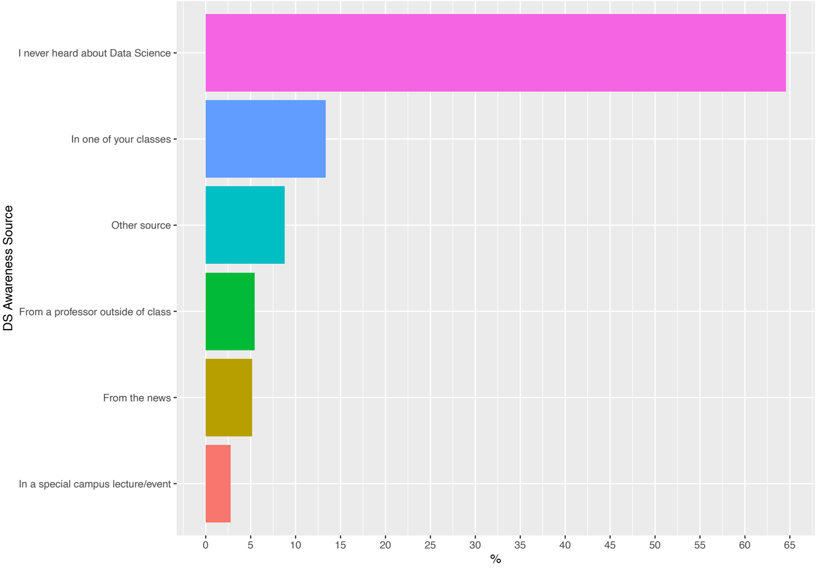
NCA&T Students' Awareness of Data Science by Gender



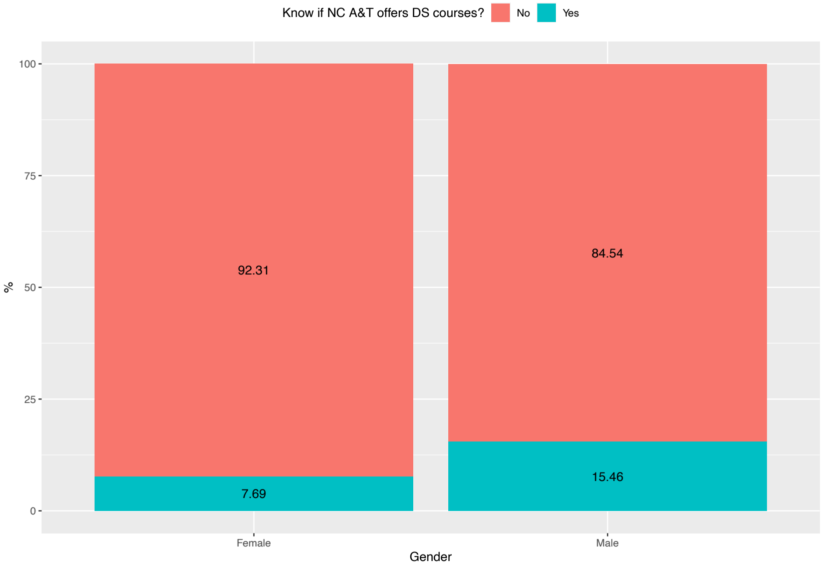
NCA&T Students' Awareness of Data Science by Major



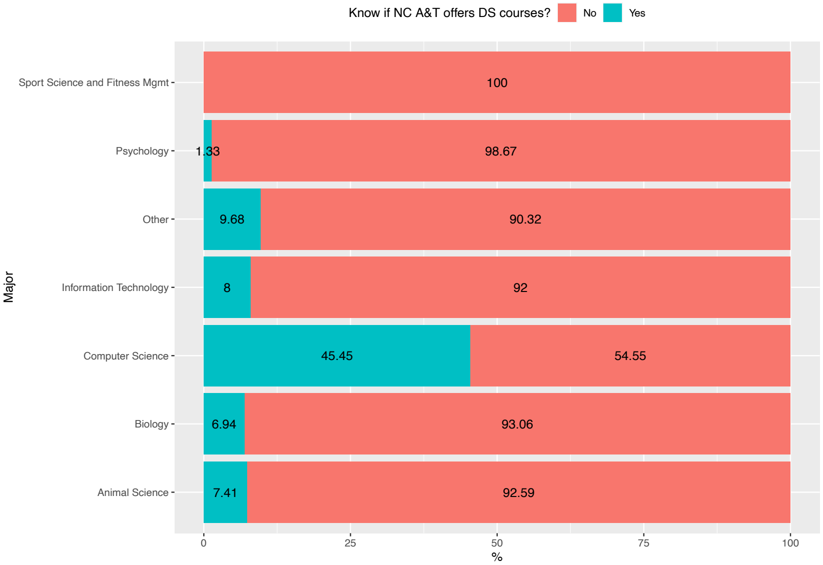
NCA&T Students' Awareness of Data Science by Source



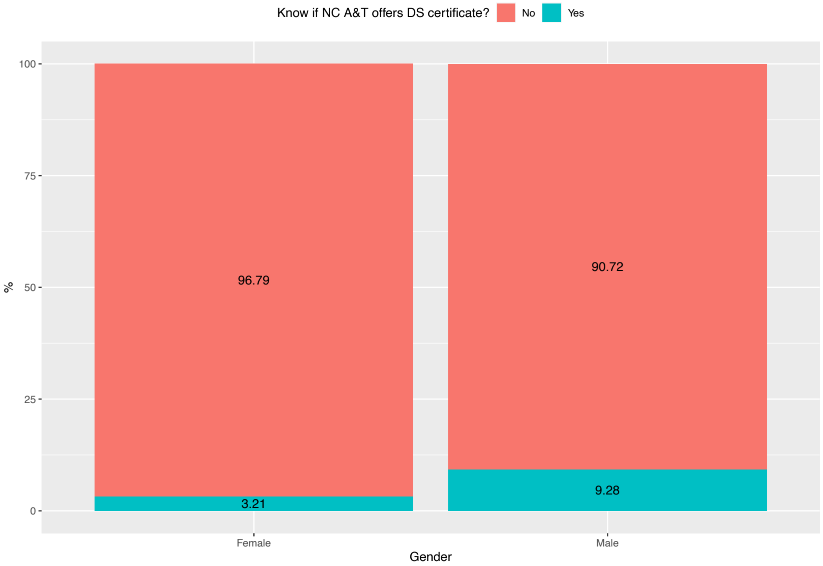
NCA&T Students’ Awareness of Data Science



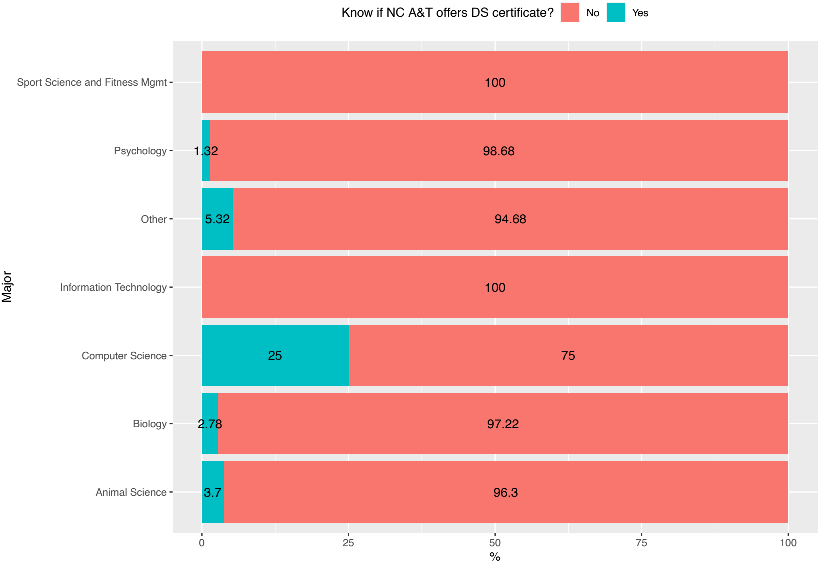
NCA&T Students' Awareness of Data Science



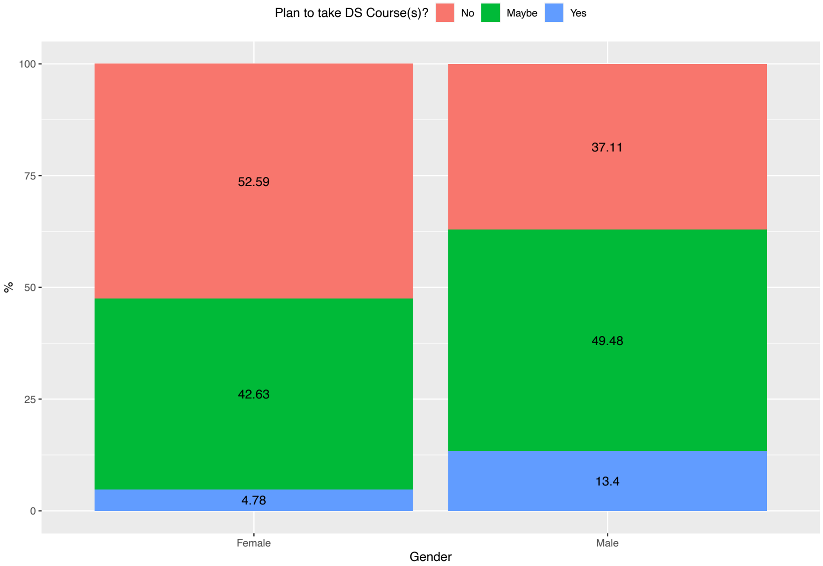
NCA&T Students' Awareness of Data Science



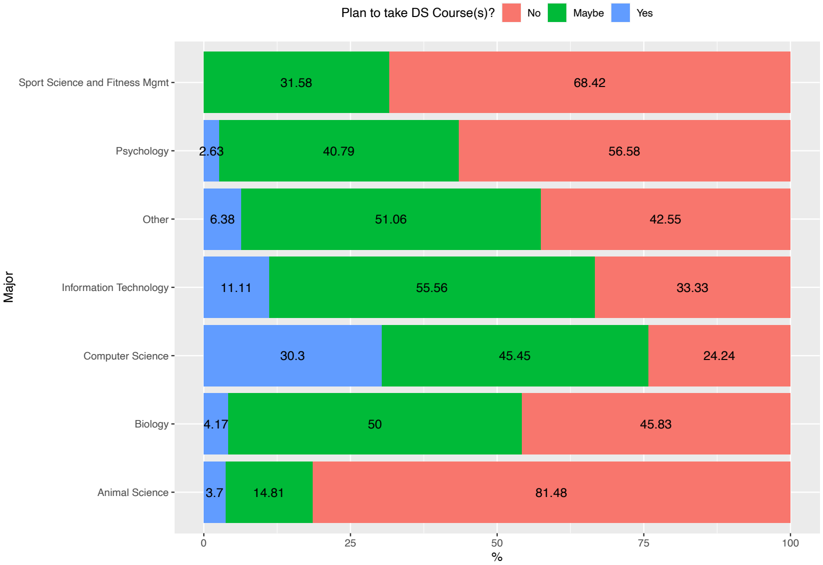
NCA&T Students' Awareness of Data Science



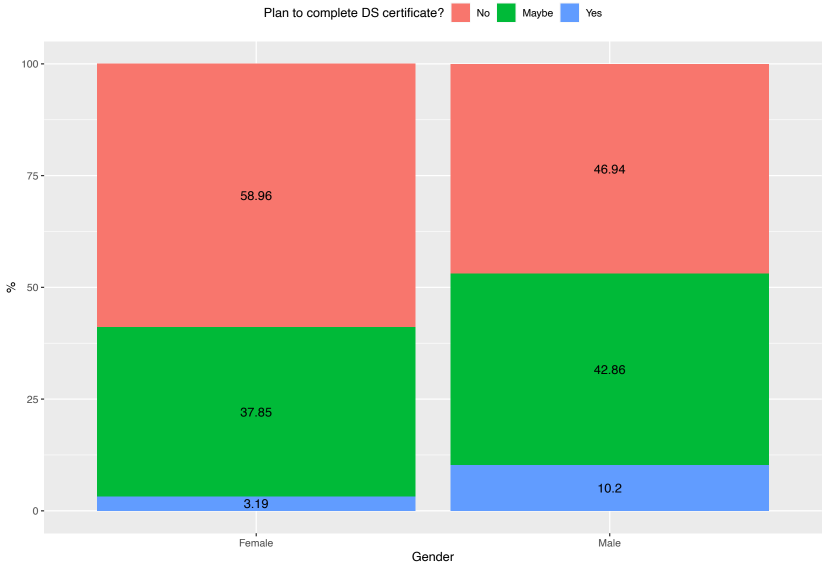
NCA&T Students' Aspirations of Data Science



NCA&T Students' Aspirations of Data Science



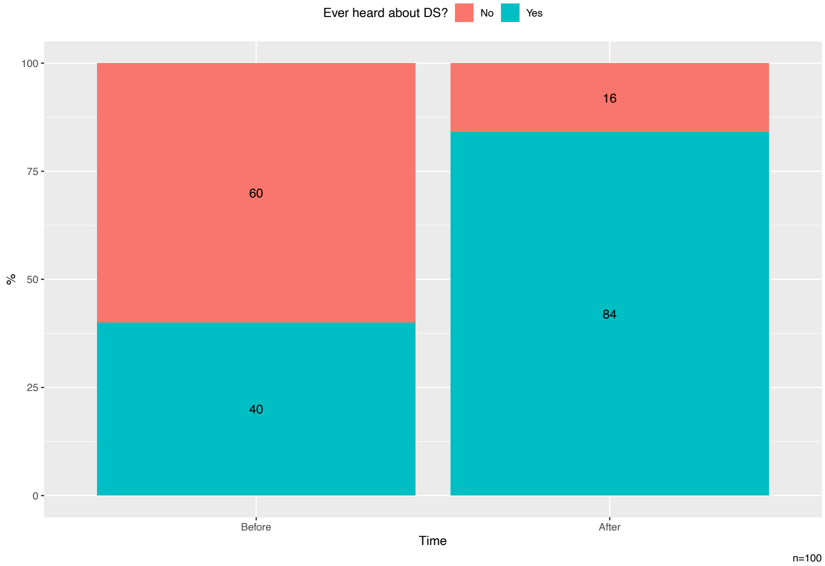
NCA&T Students' Aspirations of Data Science



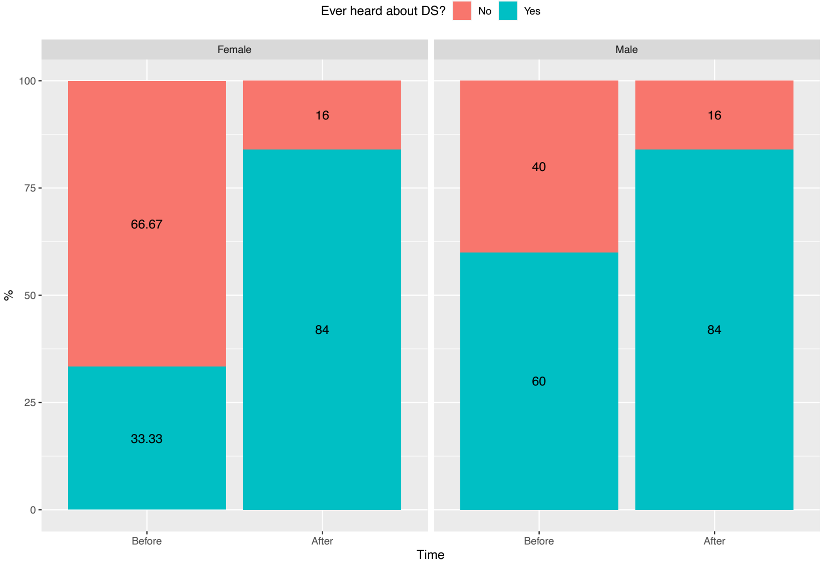
The Potential of Intro Stats to Promote Data Science

- **Intervention**:
  - Introductory lecture about the DS field and its opportunities
  - 45 minute informational presentation given during normal class session near middle of semester
  - Presentation is either given by the section instructor or course coordinator
  - Students completed the online DS awareness & aspirations survey before and after the lecture
  - 3 sections in Spring 2021 and 1 section in Summer 2021

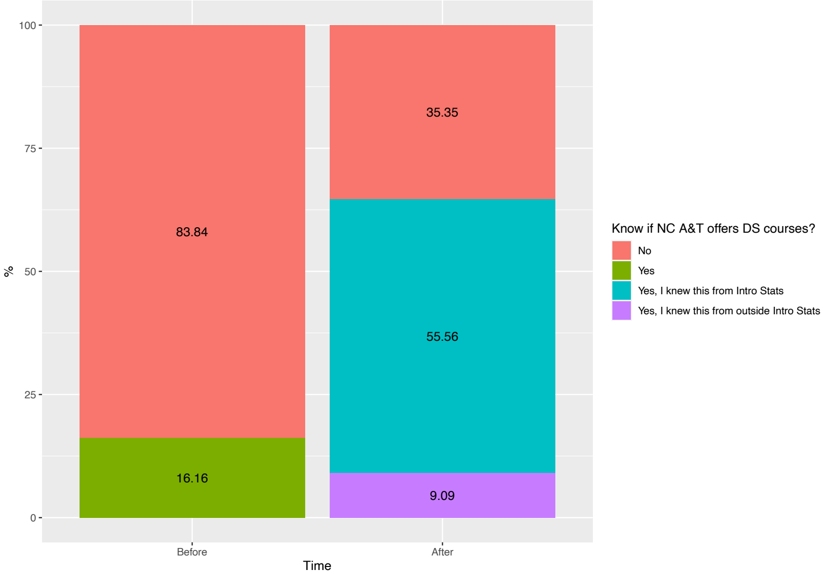
The Potential of Intro Stats to Promote Data Science



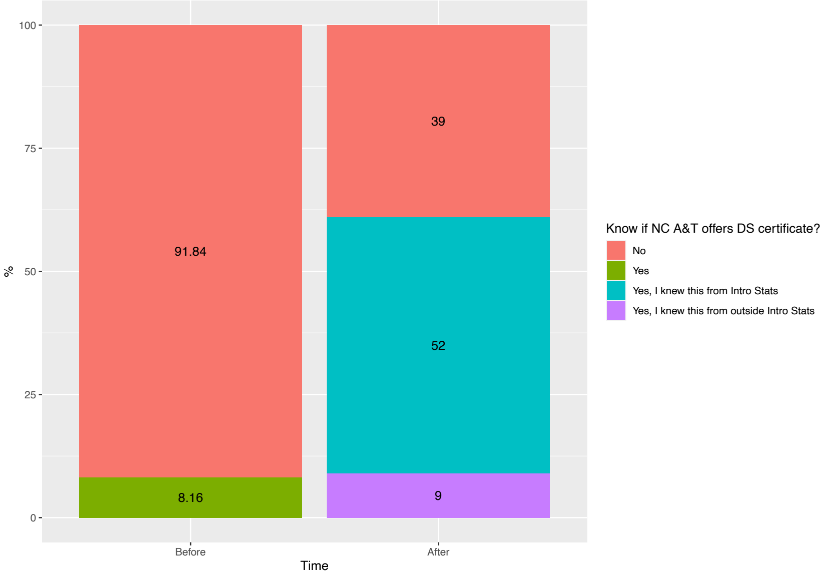
The Potential of Intro Stats to Promote Data Science



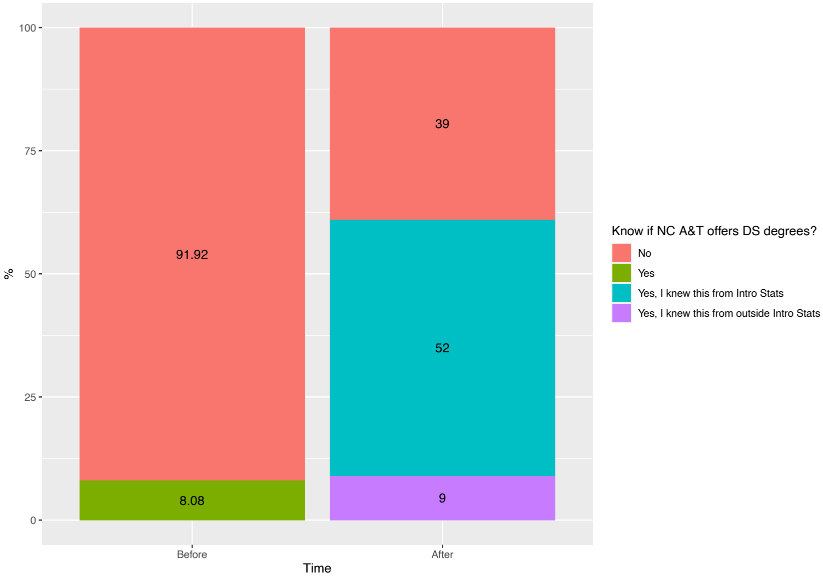
The Potential of Intro Stats to Promote Data Science



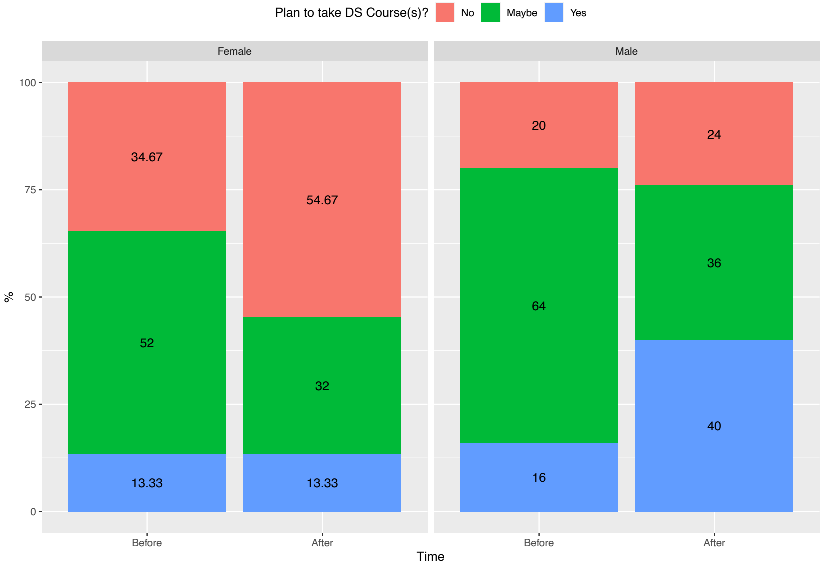
The Potential of Intro Stats to Promote Data Science



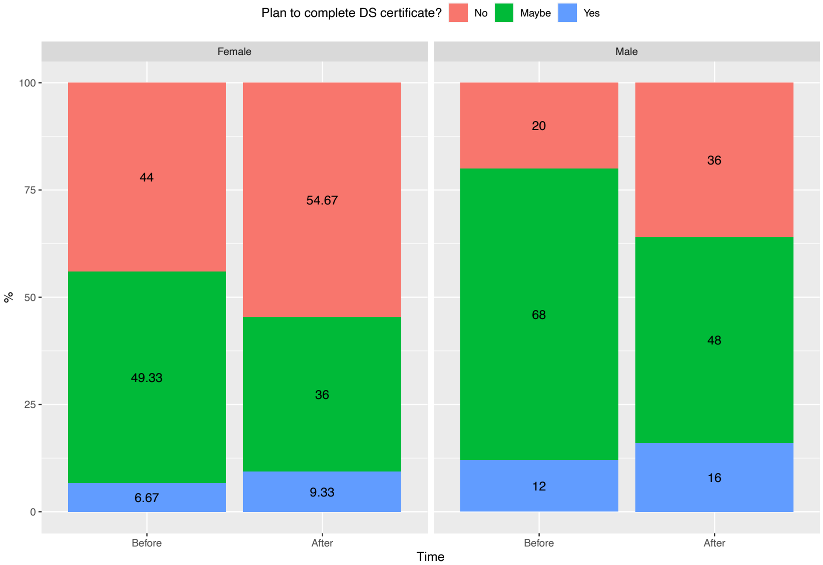
The Potential of Intro Stats to Promote Data Science



The Potential of Intro Stats to Promote Data Science



The Potential of Intro Stats to Promote Data Science



Discussion

- What DS programs do you offer at your institution?
- What is the level of DS awareness among Intro Stats students at your institution?
- What might be impactful practices for raising awareness and/or aspirations of DS among Intro Stats students?

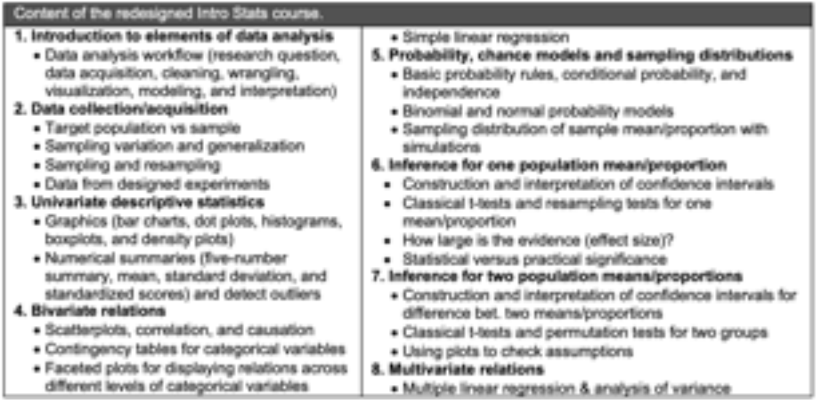
Redesigning Intro Stats to Promote DS at NCA&T

Goal: revolutionize Intro Stats at NC A&T to enhance the statistical and quantitative skills of and promote data science literacy among underrepresented minority (URM) students.

- The Intro Stats course should
  - introduce students to the entire data analysis cycle rather than pieces of it (**Cobb, 2015**)
  - leverage the use of technology for exploring concepts with simulations (**GAISE #2**)
  - help students learn statistics actively while analyzing real data using technology (**GAISE #3, 4 & 5**)
  - expose students to multivariable thinking (**GAISE #1**)
  - train students to think structurally with data, become data-savvy, and
  - expose students, early and frequently, to the elements of the DS workflow and the data scientist's toolbox

Redesigning Intro Stats to Promote DS at NCA&T

- Revised course content:



Content of the redesigned Intro Stats course.

1. Introduction to elements of data analysis

- Data analysis workflow (research question, data acquisition, cleaning, wrangling, visualization, modeling, and interpretation)

2. Data collection/acquisition

- Target population vs sample
- Sampling variation and generalization
- Sampling and resampling
- Data from designed experiments

3. Univariate descriptive statistics

- Graphics (bar charts, dot plots, histograms, boxplots, and density plots)
- Numerical summaries (five-number summary, mean, standard deviation, and standardized scores) and detect outliers

4. Bivariate relations

- Scatterplots, correlation, and causation
- Contingency tables for categorical variables
- Faceted plots for displaying relations across different levels of categorical variables
- Simple linear regression

5. Probability, chance models and sampling distributions

- Basic probability rules, conditional probability, and independence
- Binomial and normal probability models
- Sampling distribution of sample mean/proportion with simulations

6. Inference for one population mean/proportion

- Construction and interpretation of confidence intervals
- Classical t-tests and resampling tests for one mean/proportion
- How large is the evidence (effect size)?
- Statistical versus practical significance

7. Inference for two population means/proportions

- Construction and interpretation of confidence intervals for difference bet. two means/proportions
- Classical t-tests and permutation tests for two groups
- Using plots to check assumptions

8. Multivariate relations

- Multiple linear regression & analysis of variance

Redesigning Intro Stats to Promote DS at NCA&T

- Adding **Virtual Statistical Computing Lab**:
  - **virtual lab** using RStudio Cloud
  - provides free and effortless access to computing in R/RStudio
  - reduces the faculty and students effort to deal with device-specific issues with the R/RStudio software
  - removes the logistic restrictions associated with physical computer labs
  - 1-hour-long weekly virtual lab sessions
  - R will be used during both class and lab sessions
  - In the lab sessions, students will be guided to
    - further explore concepts via simulations,
    - practice using R commands introduced in class, and
    - analyze real datasets and make data-driven decisions
- Well-aligned with the principles of the data-centered pedagogy

Redesigning Intro Stats to Promote DS at NCA&T

- **Integration of DS knowledge and tools in the course:**
  - Horton et al. (2015) argue that *"by introducing students to commonplace tools for data management, visualization, and reproducible analysis in DS, and applying these to real-world scenarios, we prepare them to think statistically"*
  - The DS precursors integrated into the course will include:
    - **R & RStudio** to engage students in substantive data analyses and allow them to practice answering questions with data
    - **R Markdown** to train students to perform reproducible analysis
    - **Datasets that satisfy the 3 R's** of Kim et al. (2018) (Rich: to answer meaningful questions, Real: has context, and Realistic: needs wrangling; e.g., `gapminder` and `fivethirtyeight`)

Redesigning Intro Stats to Promote DS at NCA&T

- **Integration of DS knowledge and tools in the course:**
  - Reading assignments on DS projects at famous data scientist employers (Google, Amazon, Facebook, etc.)
  - Major-related data analysis projects (e.g., Kinesiology majors are assigned projects related to sports analytics)
  - Posts about current trends in the DS job market
  - Posts about DS educational opportunities

Redesigning Intro Stats to Promote DS at NCA&T

- NSF Grant #2106945 (07/2021 – 06/2024)
  - PI: Sayed Mostafa
  - Co-PIs: Seongtae Kim, Guoqing Tang, Tamer Elbayoumi, Mingxian Chen
- Project Title: Infusing Data-Centered Pedagogy and Data-Analytical Skills into Introductory Statistics
- Project Goals:
  - **Enhance** the students' statistical knowledge and data-analytical skills gained from the Intro Stats course;
  - **Create** a pipeline for the new DS programs offered at NC A&T;
  - **Build** a faculty cadre capable of and committed to teaching Intro Stats using a data-centered pedagogy to promote DS literacy among undergraduate students

Discussion

- Challenges with redesigning Intro Stats to promote DS??
- Challenges with integrating coding in Intro Stats??

References

- 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., and Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2), 28-58.
- Horton, N.J., Baumer, B.S. and 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.
- 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.