Preparing Students for the Data Science Era through Introductory Statistics: A Proposed Model

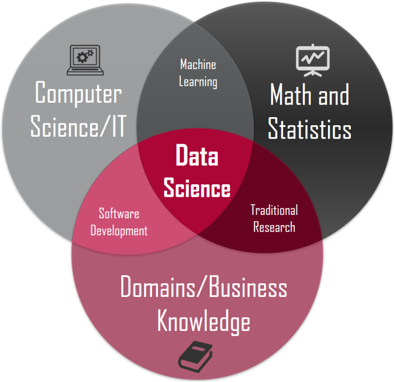
Sayed Mostafa[1]

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

[1] Assistant Professor & Coordinator of Introductory Statistics

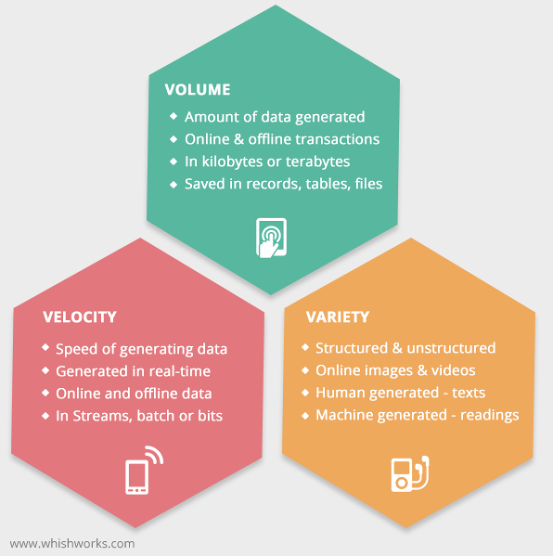
What is Data Science?

- A fast-growing *interdisciplinary* field which combines skills and concepts from
  - Statistics,
  - Mathematics, and
  - Computer Science



Source: Data Science Explained

Why Data Science?



VOLUME
- Amount of data generated
- Online & offline transactions
- In kilobytes or terabytes
- Saved in records, tables, files
VELOCITY
- Speed of generating data
- Generated in real-time
- Online and offline data
- In Streams, batch or bits
VARIETY
- Structured & unstructured
- Online images & videos
- Human generated - texts
- Machine generated - readings
www.whishworks.com

The three Vs of big data!!

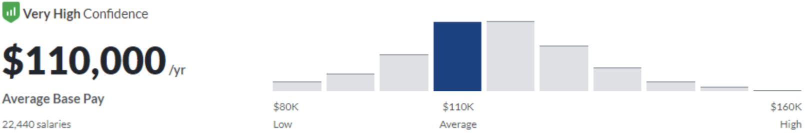
Why Data Science?

- A 2017 Business-Higher Education Forum (BHEF) and PwC joint report projected that by 2021,
  - 69% of employers will give preference to candidates with data science and analytics skills,
  - whereas only 23% of college and university leaders expect their graduates to have those skills

Why Data Science?

- There is a huge need for individuals with data science skills (10,071 job openings in 2022 and 5,971 in 2021 according to Glassdoor)

- "Data Scientist" was the best job in the US for 4 years (2016-2019) and is the third best job in the US in 2022 according to Glassdoor
  - **median base salary**: \$120,000
  - **job satisfaction**: 4.1/5



Very High Confidence

$110,000 /yr

Average Base Pay

22,440 salaries

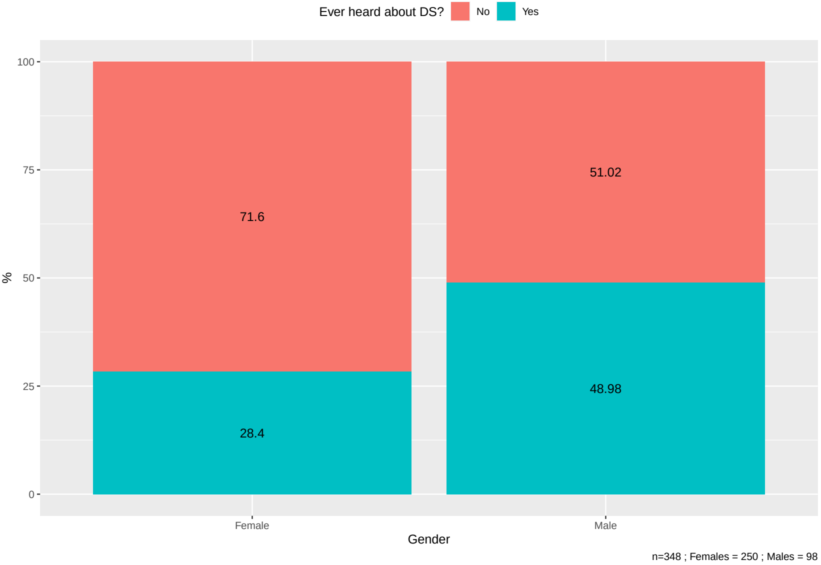
Data Science at NCA&T

- NCA&T offers several data science tracks to prepare students to become data scientists:
  - **Undergraduate Certificate in Data Science & Analytics**
  - **BS in Mathematics (Statistics & Data Science Concentration)**
  - **Post-Baccalaureate Certificate in Data Analytics**
  - **MS in Data Science and Engineering**
  - **PhD in Data Science & Analytics**

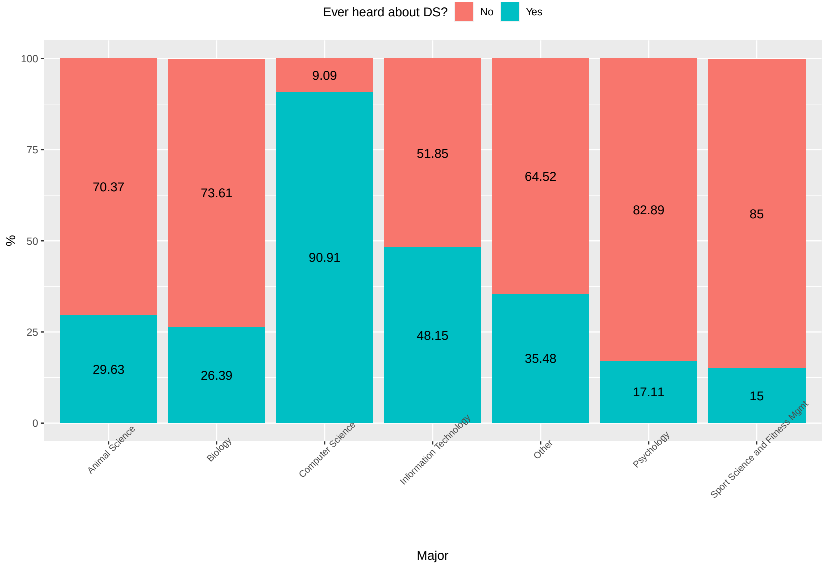
Students' Awareness of Data Science

- With DS being a relatively new field, most undergraduate students are unaware of DS and its opportunities!!
- We surveyed the NCA&T Intro Stats students about their awareness and aspirations of DS.
- Since Intro Stats is a very popular Gen. Ed. course at NCA&T, we think Intro Stats students give good representation of our UG students.

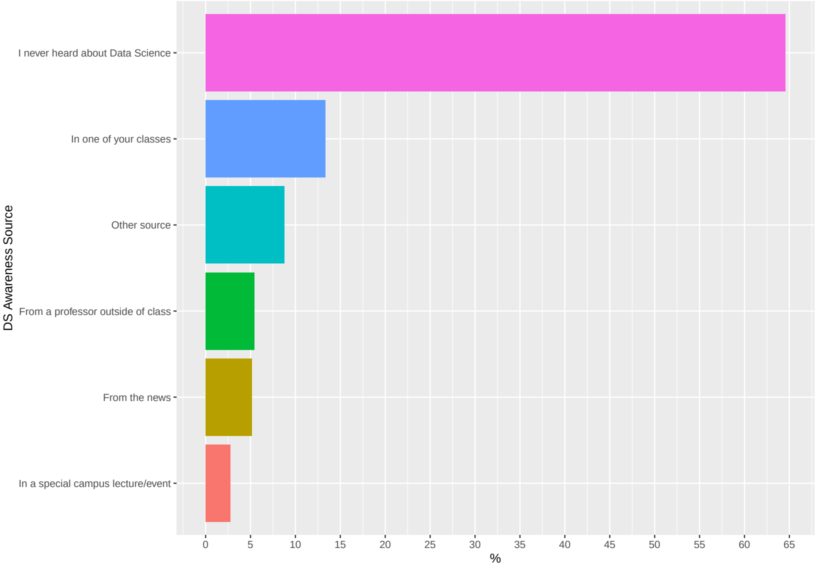
NCA&T Students' Awareness of Data Science



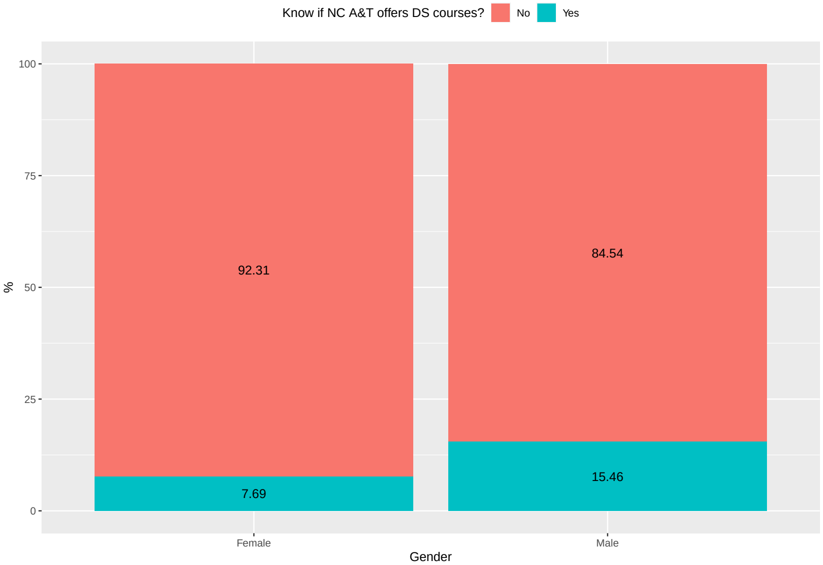
NCA&T Students' Awareness of Data Science



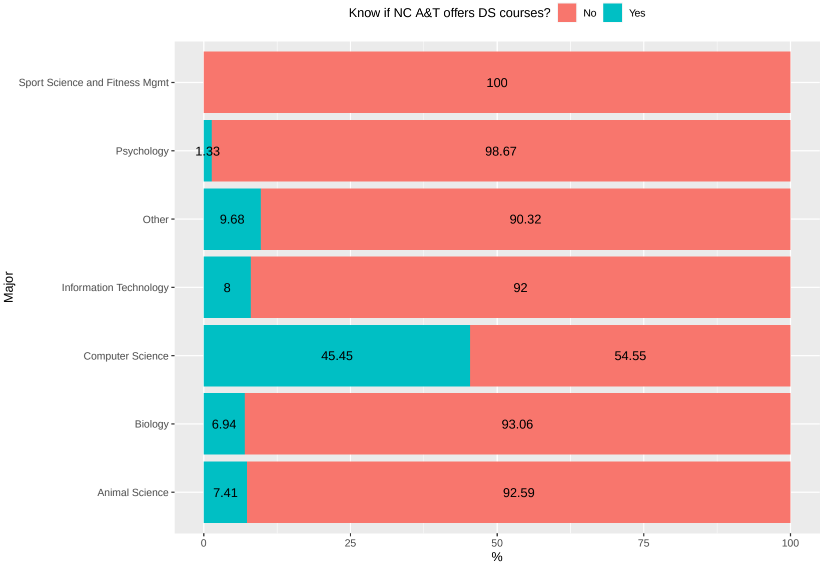
NCA&T Students' Awareness of Data Science



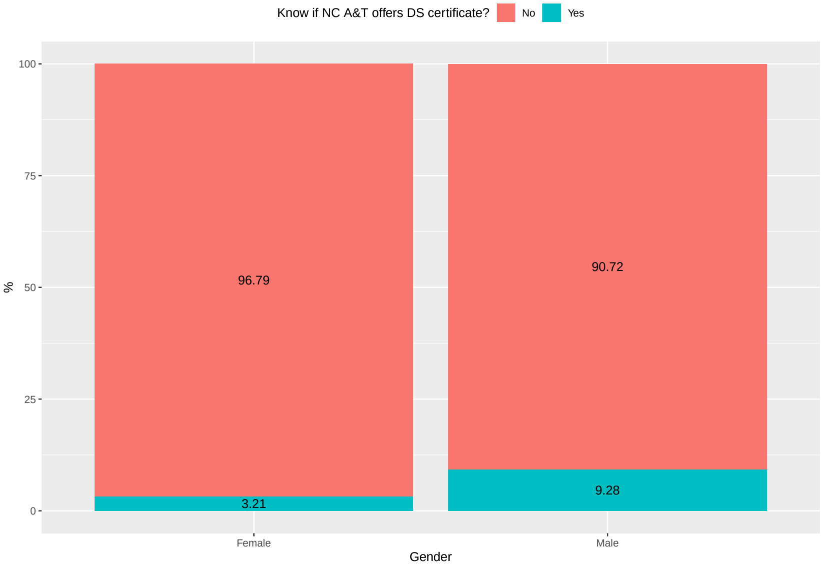
NCA&T Students' Awareness of Data Science



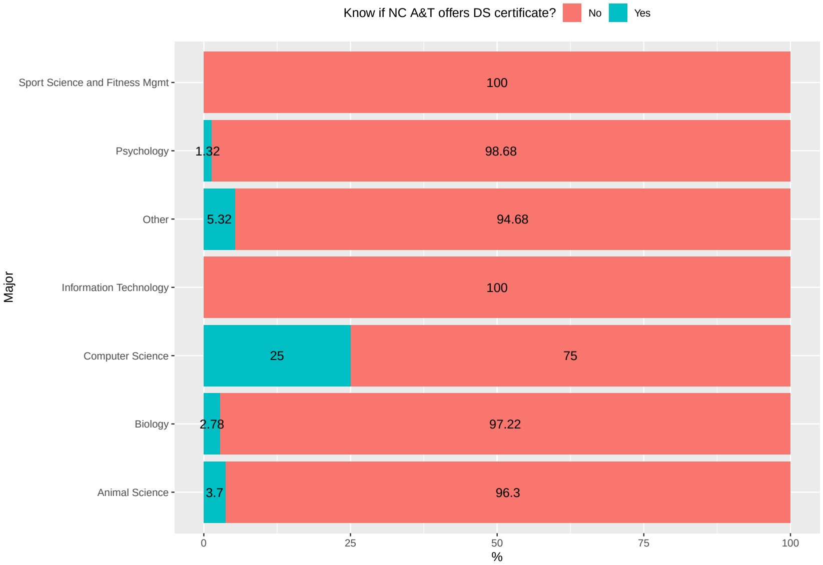
NCA&T Students’ Awareness of Data Science



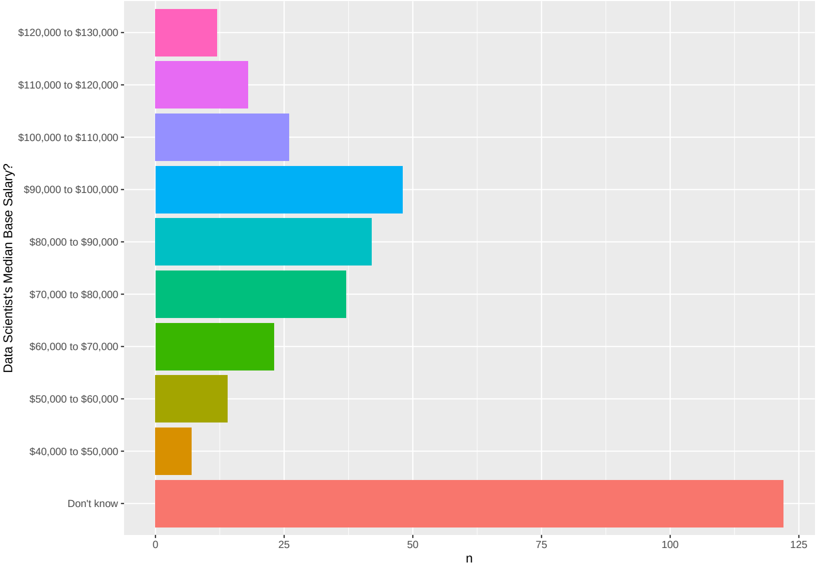
NCA&T Students' Awareness of Data Science



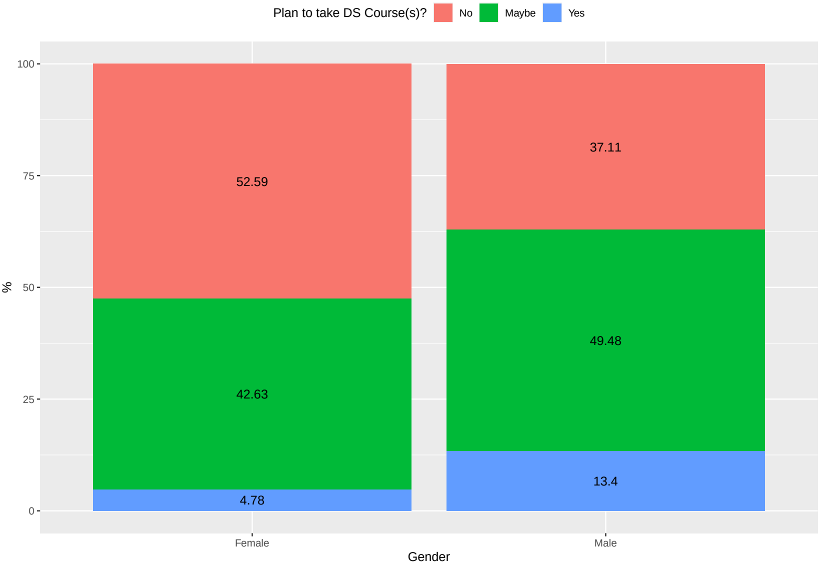
NCA&T Students' Awareness of Data Science



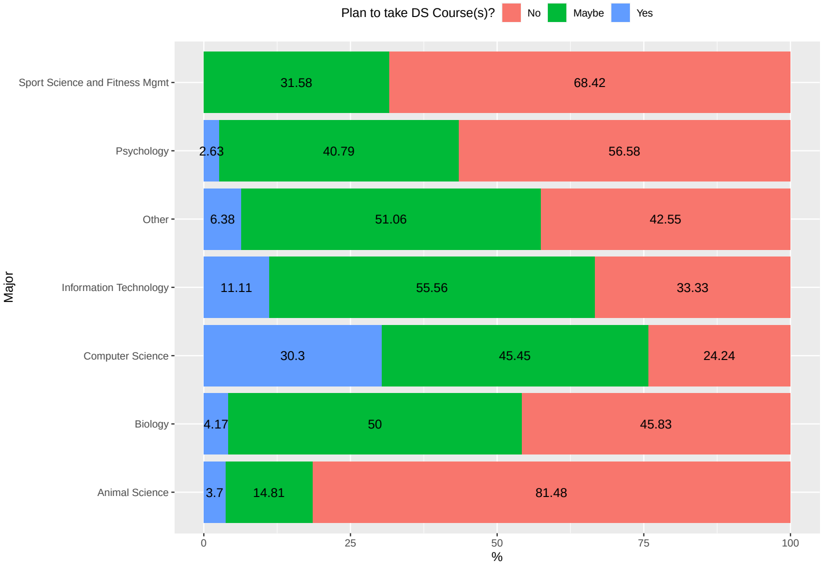
NCA&T Students’ Awareness of Data Science



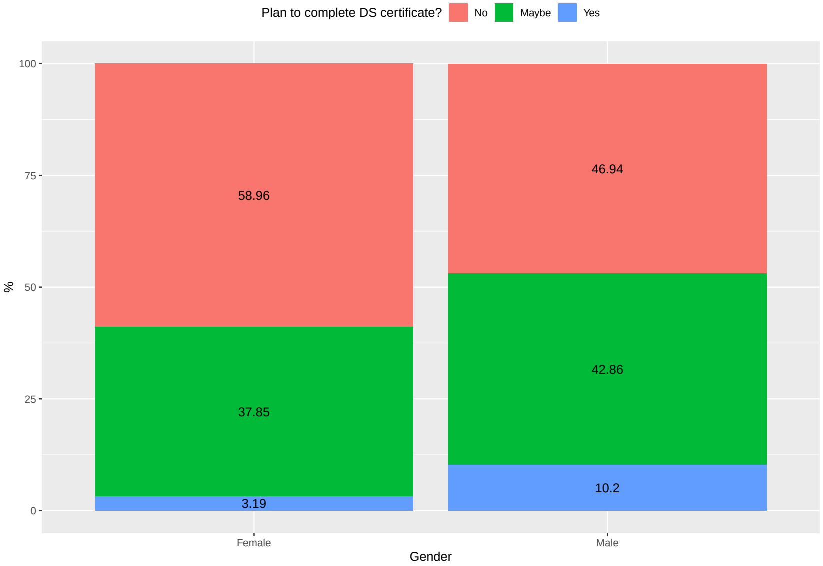
NCA&T Students' Aspirations of Data Science



NCA&T Students' Aspirations of Data Science



NCA&T Students' Aspirations of Data Science



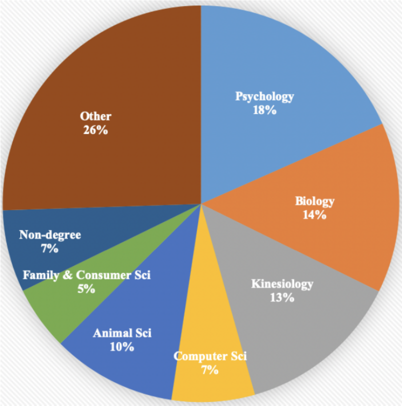
Promoting Data Science through Introductory Statistics

Why Intro Stats?

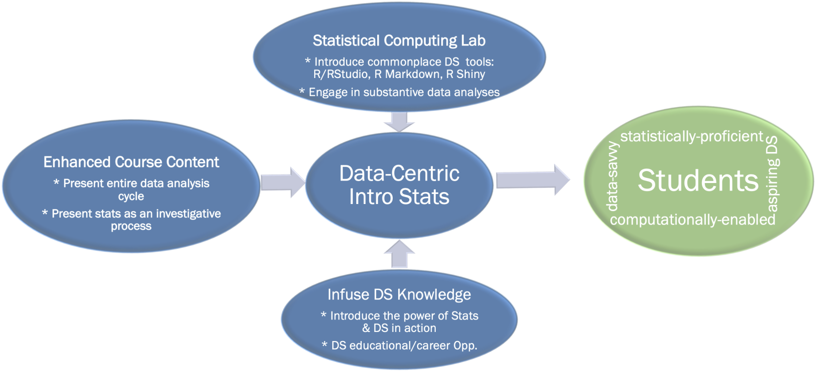
- Intro Stats is the main source for statistical training for UG students in the US and around the globe
- Being required for most STEM and many non-STEM majors, Intro Stats reaches a wide spectrum of students from varying backgrounds
- All students need to become data-literate to succeed in the data-driven world

Introductory Statistics at NCA&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



Designing Intro Stats to Promote DS



Statistical Computing Lab

* Introduce commonplace DS tools: R/RStudio, R Markdown, R Shiny

* Engage in substantive data analyses

Enhanced Course Content

* Present entire data analysis cycle

* Present stats as an investigative process

Data-Centric Intro Stats

Infuse DS Knowledge

* Introduce the power of Stats & DS in action

* DS educational/career Opp.

statistically-proficient

data-savvy

Students

aspiring DS

computationally-enabled

Proposed Model for Intro Stats

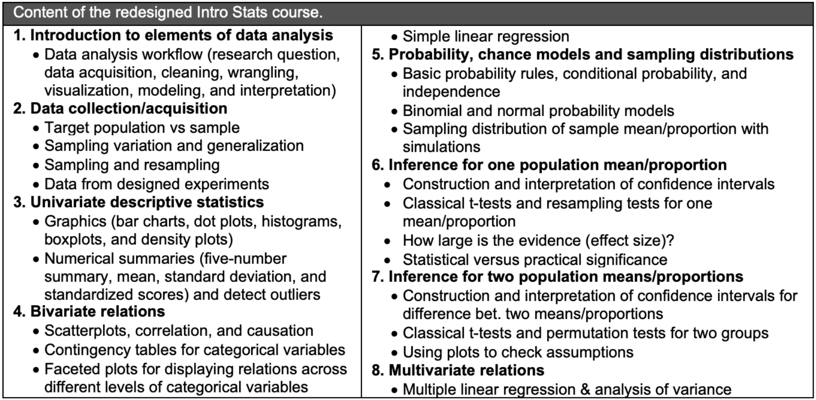
Designing Intro Stats to Promote DS

Supporting Literature:

- 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, 2016, Recommendation #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 and become data-savvy (**Horton et al., 2015**)
  - expose students, early and frequently, to the elements of the DS workflow and the data scientist's toolbox (**Horton et al., 2015**)

Designing Intro Stats to Promote DS

1. **Enhanced 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

Designing Intro Stats to Promote DS

2. **Virtual Statistical Computing Lab**:

- 1-hour-long weekly **virtual lab** using RStudio Cloud
- **Before lab sessions**, students complete assigned interactive **R shiny** tutorials involving reviewing concepts from lecture, examples and running R codes
- **During lab sessions**, students are guided to write and run R codes in **RStudio** Cloud
- **At the end of each lab session**, students submit a lab report written using **R Markdown**
- Exposes students, early and frequently, to the elements of the DS workflow
- Infuses DS precursors [Horton et al. (2015)]:
  - R & RStudio to engage students in substantive data analyses
  - R Markdown to train students to perform reproducible analysis

Designing Intro Stats to Promote DS

3. **Integration of DS Knowledge within the Course**:

- Discussion board assignments promoting the power of stats and DS for solving real-world problems
- Posts about DS educational opportunities and current trends in the DS job market
- Major-related data analysis projects (e.g., Kinesiology majors are assigned projects related to sports analytics)

Model Implementation at NCA&T

- NSF Grant #HRD2106945 (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 DS programs offered at NCA&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

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.
- 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. 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.

Acknowledgment

- I am grateful to the Intro Stats faculty at NCA&T who helped with the data collection and/or discussion of results: Giles Warrack; Mingxiang Chen; Tamer Elbayoumi; Seongtae Kim; and Suzanne O'Regan (currently at UGA).

Thank you