IA SE
20 23

IASE 2023 Satellite Conference

Fostering Learning of Statistics and Data Science

Hybrid Conference

11 – 13 July 2023, Toronto, Canada

IASE

iasc

Infusion of Data Science and Computation into Introductory Statistics

Sayed Mostafa[1], Tamer Elbayoumi, Seongtae Kim

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

07/11/2023

[1] Assistant Professor & Coordinator of Introductory Statistics

Outline

- Introduction
  - Background
  - Study Objectives
  - Study Setting
- Infusion of Data Science and Computation into Intro Stats
  - Guiding Literature
  - Proposed DS/Computationally-Infused Intro Stats Design
- Evaluating the DS-Infused Intro Stats Design
  - DS awareness, readiness & aspirations
  - Statistical learning gains
  - Overall course performance
- Resources for Teaching a DS-Infused Intro Stats Course

Introduction

- Nolan and Temple Lang's (2010) 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 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.
- As a result, a significant body of literature on integrating computing in statistics courses emerged during the last decade
    - see the *JSDSE Special Issue* - Horton and Hardin (2021): "Integrating computing in the statistics and data science curriculum: Creative structures, novel skills and habits, and ways to teach computational thinking".

Study Objectives

In this study, we aim to

- introduce an Intro Stats course design that integrates computing as a core component of the course and
- 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.

Study Setting

North Carolina Agricultural
and Technical State University

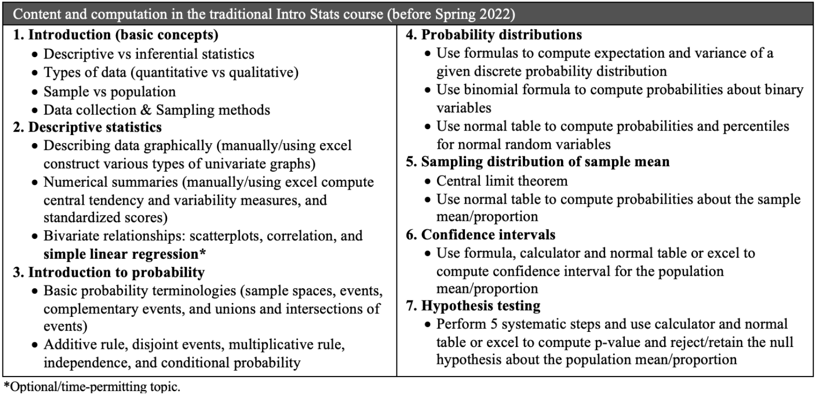
- The study took place at North Carolina A&T State University (NCA&T)
- NC A&T is a public, high-research activity land-grant university
- The largest Historically Black College and University (HBCU) in the United States
- Medium-sized university with Fall 2022 enrollment over 13,000
- Strong focus on STEM education

Study Setting

- “Introduction to Probability & Statistics” (MATH224)
- Algebra-based 3.00 credits course
- Serves as one of the Gen Ed courses for Mathematical, Logical, and Analytical Reasoning (MLAR)
- Serves STEM (∼46%) and non-STEM (∼54%) majors
- Most students in the course are from groups underrepresented in Statistics/DS
  - ∼82% are African Americans and ∼69% are females
- About 7 sections each semester (∼45 students in each section)

The "Traditional" Intro Stats Course Design

- "Consensus" course content (see below)
- 3 hours of lecture per week
- Use of a calculator and/or excel for course computation
- All inference rely on distribution tables (e.g., z/t tables)
- This design prevailed at NCA&T until before Spring 2022



Content and computation in the traditional Intro Stats course (before Spring 2022)

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.

Consequences of the “Traditional” Intro Stats Design

- Low statistical learning gains
  - In Fall 2019, we measured students’ learning gains from the course by the improvement in their scores on the Comprehensive Assessment of Outcomes in a First Statistics Course (CAOS) scale
  - the average learning gain (posttest - pretest) was 5.4% compared to a 9.1% national average (delMas et al., 2007)
- High overall course repeat rate
  - The course DFW rate was 24.4%

Opportunities for Intro Stats

- Intro Stats can (and should) help us attract and prepare a large diverse pool of undergraduates for further Statistics and Data Science (DS) education
- Intro Stats students are likely unaware of Statistics/DS educational/career opportunities
  - In Fall 2019, we survey our Intro Stats students ($n = 181$) about their awareness and interest in DS
  - Only 33.15% of students surveyed had heard about DS
  - Of those, only 27.12% knew NCA&T offers DS courses and only 18.64% knew NCA&T offers an undergrad DS certificate

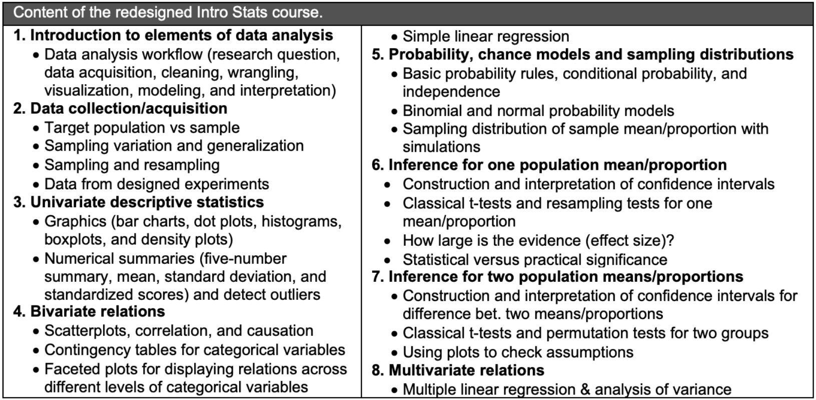
Infusion of Data Science and Computation into Intro Stats

Guiding Literature:

- The Intro Stats course should
  - introduce students to the entire data analysis cycle rather than pieces of it (**Cobb, 2015**),
  - expose students to multivariable thinking (**GAISE #1**),
  - 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**),
  - train students to think structurally with data and become data-savvy (**Horton et al., 2015**), and
  - expose students, early and frequently, to the elements of the DS workflow and the data scientist's toolbox (**Horton et al., 2015**)

Infusion of Data Science and Computation into Intro Stats

- Revised course content adopted in Spring 2022:



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

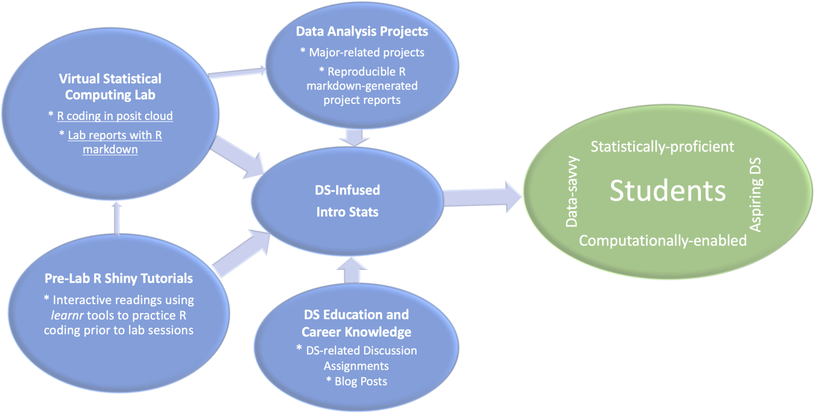
- Construction and interpretation of confidence intervals for difference bet. two means/proportions
- Classical t-tests and permutation tests for two groups
- Using plots to check assumptions

8. Multivariate relations

- Multiple linear regression & analysis of variance

DS/Computationally-Infused Intro Stats

- Redesigned Intro Stats Course – Phase I (Fall 2022)



Virtual Statistical Computing Lab

* R coding in posit cloud

* Lab reports with R markdown

Data Analysis Projects

* Major-related projects

* Reproducible R markdown-generated project reports

DS-Infused Intro Stats

Pre-Lab R Shiny Tutorials

* Interactive readings using *learnr* tools to practice R coding prior to lab sessions

DS Education and Career Knowledge

* DS-related Discussion Assignments

* Blog Posts

Statistically-proficient

Data-savvy

Students

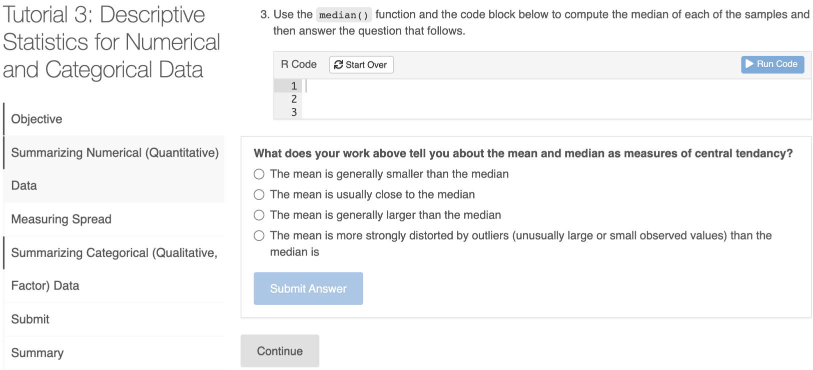
Aspiring DS

Computationally-enabled

- **Implementation**: 2 treatment sections and 2 control sections

DS/Computationally-Infused Intro Stats: Phase I-FA22

- Interactive Shiny Pre-Lab Tutorial



Tutorial 3: Descriptive Statistics for Numerical and Categorical Data
Objective
Summarizing Numerical (Quantitative) Data
Measuring Spread
Summarizing Categorical (Qualitative, Factor) Data
Submit
Summary
R Code
Start Over
Run Code
Submit Answer
Continue

3. Use the `median()` function and the code block below to compute the median of each of the samples and then answer the question that follows.

What does your work above tell you about the mean and median as measures of central tendancy?

○ The mean is generally smaller than the median

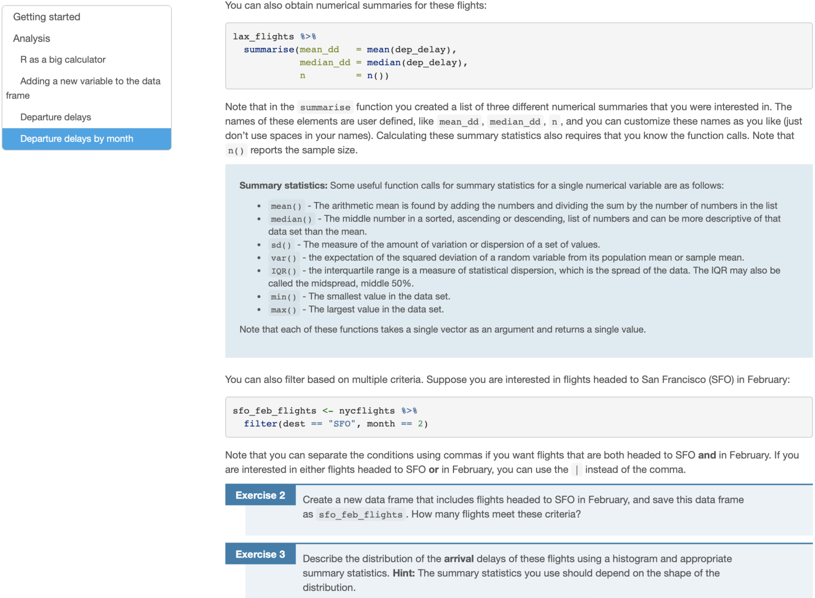
○ The mean is usually close to the median

○ The mean is generally larger than the median

○ The mean is more strongly distorted by outliers (unusually large or small observed values) than the median is

DS/Computationally-Infused Intro Stats: Phase I-FA22

▶ Computing Lab Description (Static)



You can also obtain numerical summaries for these flights:

```
lax_flights %>%
  summarise(mean_dd   = mean(dep_delay),
            median_dd = median(dep_delay),
            n         = n())
```

Note that in the `summarise` function you created a list of three different numerical summaries that you were interested in. The names of these elements are user defined, like `mean_dd`, `median_dd`, `n`, and you can customize these names as you like (just don't use spaces in your names). Calculating these summary statistics also requires that you know the function calls. Note that `n()` reports the sample size.

Summary statistics: Some useful function calls for summary statistics for a single numerical variable are as follows:

- `mean()` - The arithmetic mean is found by adding the numbers and dividing the sum by the number of numbers in the list
- `median()` - The middle number in a sorted, ascending or descending, list of numbers and can be more descriptive of that data set than the mean.
- `sd()` - The measure of the amount of variation or dispersion of a set of values.
- `var()` - the expectation of the squared deviation of a random variable from its population mean or sample mean.
- `IQR()` - the interquartile range is a measure of statistical dispersion, which is the spread of the data. The IQR may also be called the midspread, middle 50%.
- `min()` - The smallest value in the data set.
- `max()` - The largest value in the data set.

Note that each of these functions takes a single vector as an argument and returns a single value.

You can also filter based on multiple criteria. Suppose you are interested in flights headed to San Francisco (SFO) in February:

```
sfo_feb_flights <- nycflights %>%
  filter(dest == "SFO", month == 2)
```

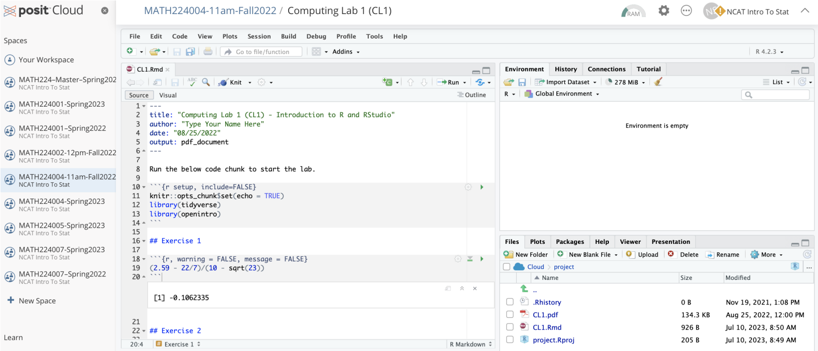
Note that you can separate the conditions using commas if you want flights that are both headed to SFO **and** in February. If you are interested in either flights headed to SFO **or** in February, you can use the `|` instead of the comma.

Exercise 2 Create a new data frame that includes flights headed to SFO in February, and save this data frame as `sfo_feb_flights`. How many flights meet these criteria?

Exercise 3 Describe the distribution of the **arrival** delays of these flights using a histogram and appropriate summary statistics. **Hint:** The summary statistics you use should depend on the shape of the distribution.

DS/Computationally-Infused Intro Stats: Phase I-FA22

- Computing Lab R Markdown Template



~~~
---
title: "Computing Lab 1 (CL1) - Introduction to R and RStudio"
author: "Type Your Name Here"
date: "08/25/2022"
output: pdf_document
---

Run the below code chunk to start the lab.

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(openintro)
~~~

Exercise 1

```{r, warning = FALSE, message = FALSE}
(2.59 - 22/7)/(10 - sqrt(23))
```

[1] -0.1062335

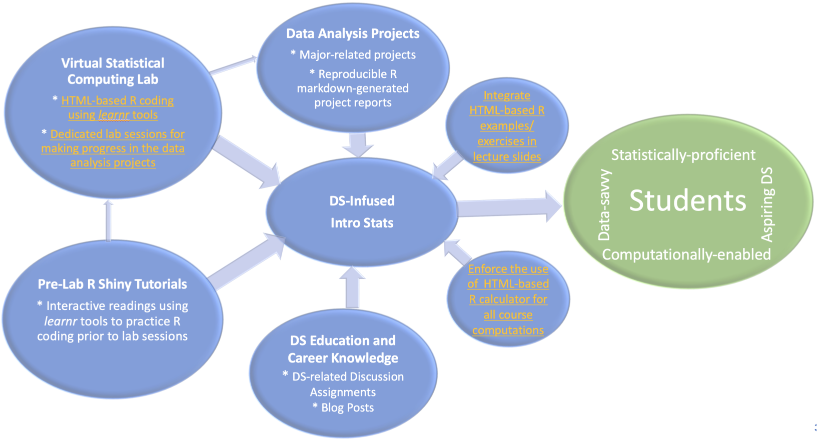
Exercise 2
```

Environment is empty

Name	Size	Modified
..		
.Rhistory	0 B	Nov 19, 2021, 1:08 PM
CL1.pdf	134.3 KB	Aug 25, 2022, 12:00 PM
CL1.Rmd	926 B	Jul 10, 2023, 8:50 AM
project.Rproj	205 B	Jul 10, 2023, 8:49 AM
```

# DS/Computationally-Infused Intro Stats

- Redesigned Intro Stats Course – Phase II (Spring 2023)



**Virtual Statistical Computing Lab**

* HTML-based R coding using *learnr* tools

* Dedicated lab sessions for making progress in the data analysis projects

**Data Analysis Projects**

* Major-related projects

* Reproducible R markdown-generated project reports

Integrate HTML-based R examples/ exercises in lecture slides

**DS-Infused Intro Stats**

Statistically-proficient

Data-savvy

Students

Aspiring DS

Computationally-enabled

**Pre-Lab R Shiny Tutorials**

* Interactive readings using *learnr* tools to practice R coding prior to lab sessions

Enforce the use of HTML-based R calculator for all course computations

**DS Education and Career Knowledge**

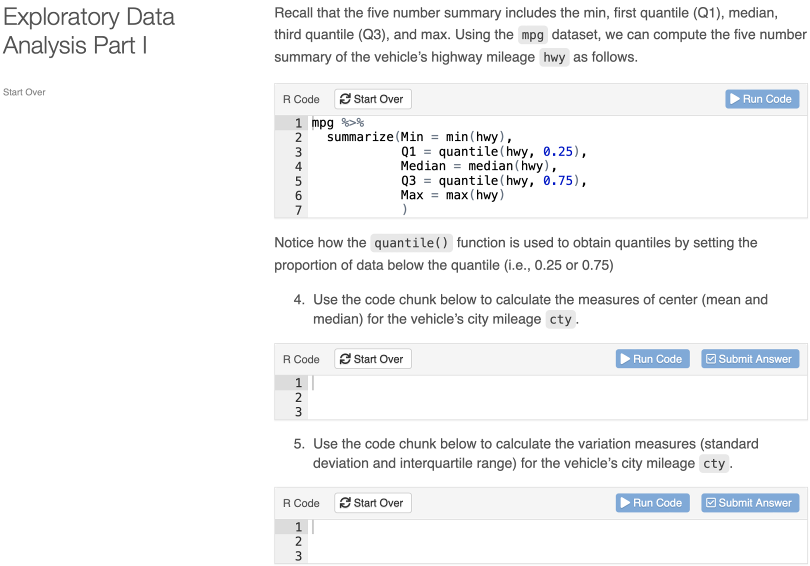
* DS-related Discussion Assignments

* Blog Posts

- **Implementation**: 4 treatment sections and 2 control sections

# DS/Computationally-Infused Intro Stats: Phase II-SP23

- Interactive Computing Lab (using the *learnr* package)



## Exploratory Data Analysis Part I

Recall that the five number summary includes the min, first quantile (Q1), median, third quantile (Q3), and max. Using the `mpg` dataset, we can compute the five number summary of the vehicle's highway mileage `hwy` as follows.

```r
mpg %>%
  summarize(Min = min(hwy),
            Q1 = quantile(hwy, 0.25),
            Median = median(hwy),
            Q3 = quantile(hwy, 0.75),
            Max = max(hwy)
            )
```

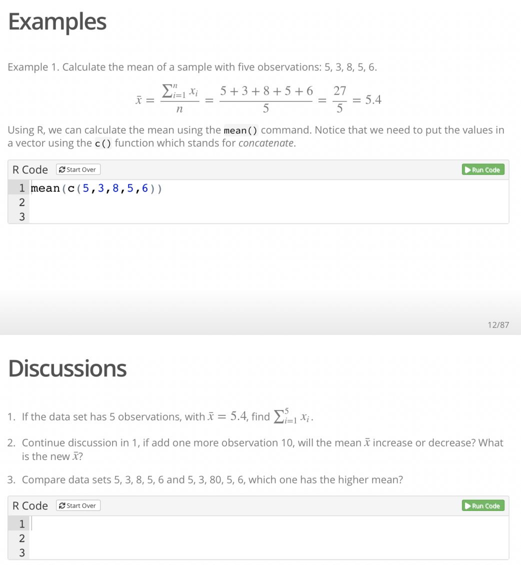
Notice how the `quantile()` function is used to obtain quantiles by setting the proportion of data below the quantile (i.e., 0.25 or 0.75)

4. Use the code chunk below to calculate the measures of center (mean and median) for the vehicle's city mileage `cty`.

5. Use the code chunk below to calculate the variation measures (standard deviation and interquartile range) for the vehicle's city mileage `cty`.

DS/Computationally-Infused Intro Stats: Phase II-SP23

- Slides with Interactive Coding



Examples

Example 1. Calculate the mean of a sample with five observations: 5, 3, 8, 5, 6.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{5+3+8+5+6}{5} = \frac{27}{5} = 5.4$$

Using R, we can calculate the mean using the `mean()` command. Notice that we need to put the values in a vector using the `c()` function which stands for *concatenate*.

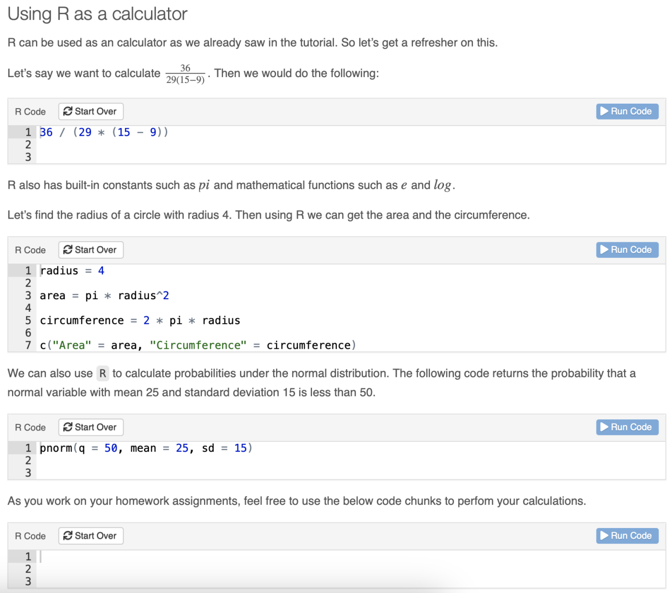
```r
mean(c(5,3,8,5,6))
```

Discussions

1. If the data set has 5 observations, with $\bar{x} = 5.4$, find $\sum_{i=1}^{5} x_i$.
2. Continue discussion in 1, if add one more observation 10, will the mean $\bar{x}$ increase or decrease? What is the new $\bar{x}$?
3. Compare data sets 5, 3, 8, 5, 6 and 5, 3, 80, 5, 6, which one has the higher mean?

DS/Computationally-Infused Intro Stats: Phase II-SP23

- Interactive R Calculator



Using R as a calculator

R can be used as an calculator as we already saw in the tutorial. So let's get a refresher on this.

Let's say we want to calculate $\frac{36}{29(15-9)}$. Then we would do the following:

```r
36 / (29 * (15 - 9))
```

R also has built-in constants such as pi and mathematical functions such as e and log.

Let's find the radius of a circle with radius 4. Then using R we can get the area and the circumference.

```r
radius = 4

area = pi * radius^2

circumference = 2 * pi * radius

c("Area" = area, "Circumference" = circumference)
```

We can also use `R` to calculate probabilities under the normal distribution. The following code returns the probability that a normal variable with mean 25 and standard deviation 15 is less than 50.

```r
pnorm(q = 50, mean = 25, sd = 15)
```

As you work on your homework assignments, feel free to use the below code chunks to perfom your calculations.

```r

```

Evaluating the DS-Infused Intro Stats Design

- DS awareness, readiness & aspirations
  - Students completed a DS awareness, readiness, and aspirations survey in Qualtrics
  - Pre-survey was completed during 1st week of semester; post-survey was completed during second-to-last week in semester
- Statistical learning gains
  - Students completed a revised version of the CAOS scale (e.g., Tintle et al., 2018)
  - Pre-test was completed during 1st week of semester; post-test was completed during second-to-last week in semester
- Overall performance
  - Measured by final course grade (focus on DFW rate)

Evaluating the DS-Infused Intro Stats Design

- DS awareness, readiness & aspirations: EFA results

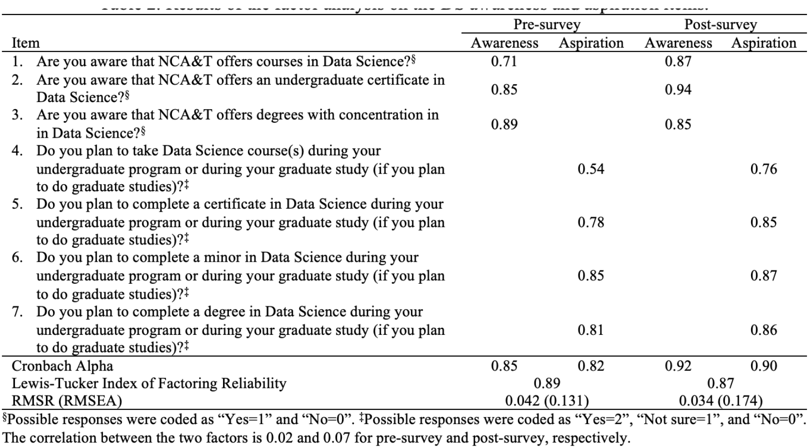


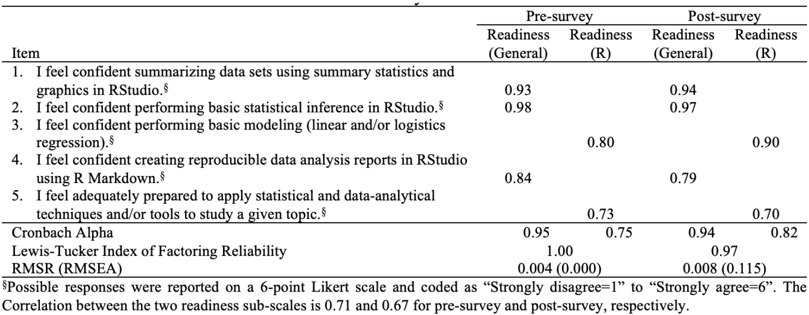
Table 2. Results of the factor analysis on the DS awareness and aspiration items.

	Pre-survey		Post-survey	
Item	Awareness	Aspiration	Awareness	Aspiration
1. Are you aware that NCA&T offers courses in Data Science?[§]	0.71		0.87	
2. Are you aware that NCA&T offers an undergraduate certificate in Data Science?[§]	0.85		0.94	
3. Are you aware that NCA&T offers degrees with concentration in in Data Science?[§]	0.89		0.85	
4. Do you plan to take Data Science course(s) during your undergraduate program or during your graduate study (if you plan to do graduate studies)?[‡]		0.54		0.76
5. Do you plan to complete a certificate in Data Science during your undergraduate program or during your graduate study (if you plan to do graduate studies)?[‡]		0.78		0.85
6. Do you plan to complete a minor in Data Science during your undergraduate program or during your graduate study (if you plan to do graduate studies)?[‡]		0.85		0.87
7. Do you plan to complete a degree in Data Science during your undergraduate program or during your graduate study (if you plan to do graduate studies)?[‡]		0.81		0.86
Cronbach Alpha	0.85	0.82	0.92	0.90
Lewis-Tucker Index of Factoring Reliability	0.89		0.87	
RMSR (RMSEA)	0.042 (0.131)		0.034 (0.174)	

[§] Possible responses were coded as "Yes=1" and "No=0". [‡] Possible responses were coded as "Yes=2", "Not sure=1", and "No=0". The correlation between the two factors is 0.02 and 0.07 for pre-survey and post-survey, respectively.

Evaluating the DS-Infused Intro Stats Design

- DS awareness, readiness & aspirations: EFA results



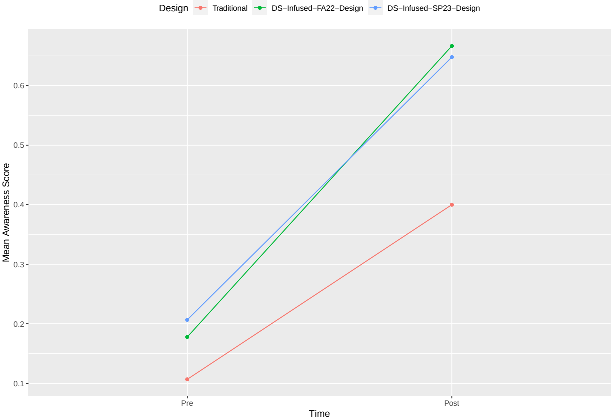
	Pre-survey		Post-survey	
Item	Readiness (General)	Readiness (R)	Readiness (General)	Readiness (R)
1. I feel confident summarizing data sets using summary statistics and graphics in RStudio.§	0.93		0.94	
2. I feel confident performing basic statistical inference in RStudio.§	0.98		0.97	
3. I feel confident performing basic modeling (linear and/or logistics regression).§		0.80		0.90
4. I feel confident creating reproducible data analysis reports in RStudio using R Markdown.§	0.84		0.79	
5. I feel adequately prepared to apply statistical and data-analytical techniques and/or tools to study a given topic.§		0.73		0.70
Cronbach Alpha	0.95	0.75	0.94	0.82
Lewis-Tucker Index of Factoring Reliability	1.00		0.97	
RMSR (RMSEA)	0.004 (0.000)		0.008 (0.115)	

§Possible responses were reported on a 6-point Likert scale and coded as "Strongly disagree=1" to "Strongly agree=6". The Correlation between the two readiness sub-scales is 0.71 and 0.67 for pre-survey and post-survey, respectively.

Awareness of Data Science

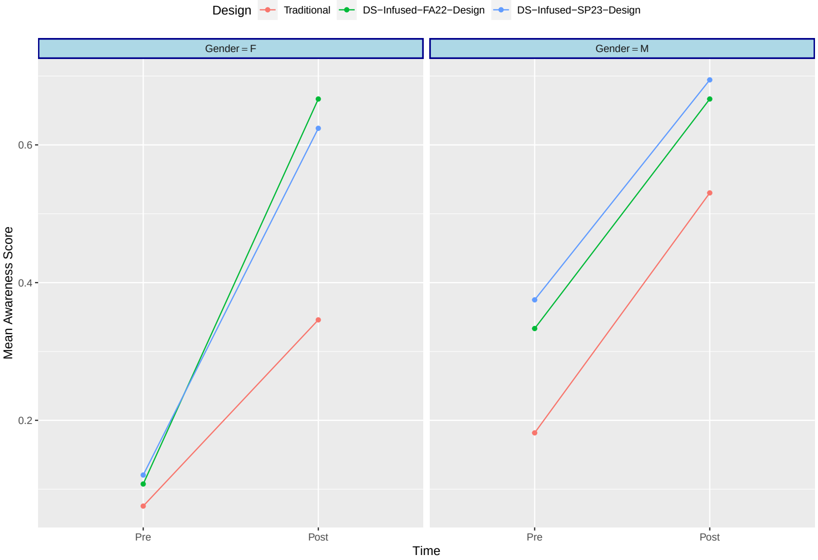
- DS awareness gains by course design

Course Type	n	Mean_Pre	SD_Pre	Mean_Post	SD_Post	Mean_Diff	SD_Diff
Traditional	75	0.11	0.27	0.40	0.46	0.29	0.43
DS-Infused-FA22-Design	45	0.18	0.33	0.67	0.46	0.49	0.44
DS-Infused-SP23-Design	71	0.21	0.37	0.65	0.46	0.44	0.46



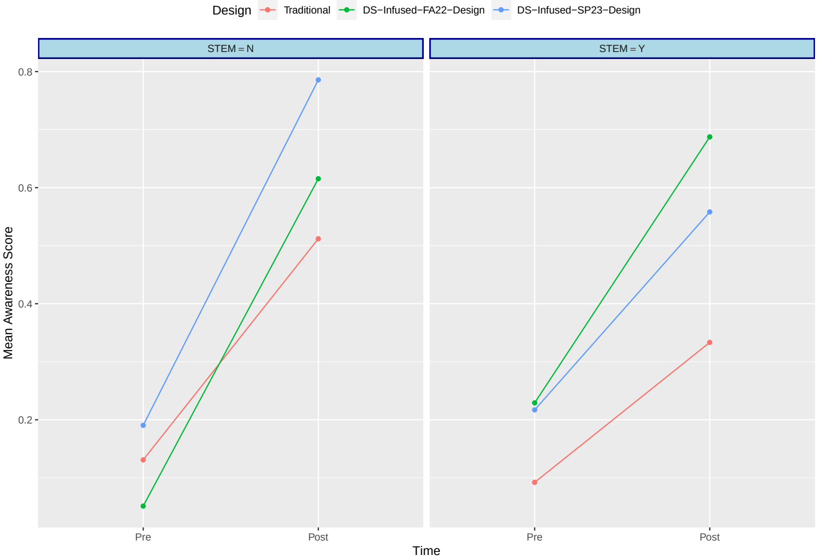
Awareness of Data Science

- DS awareness by gender



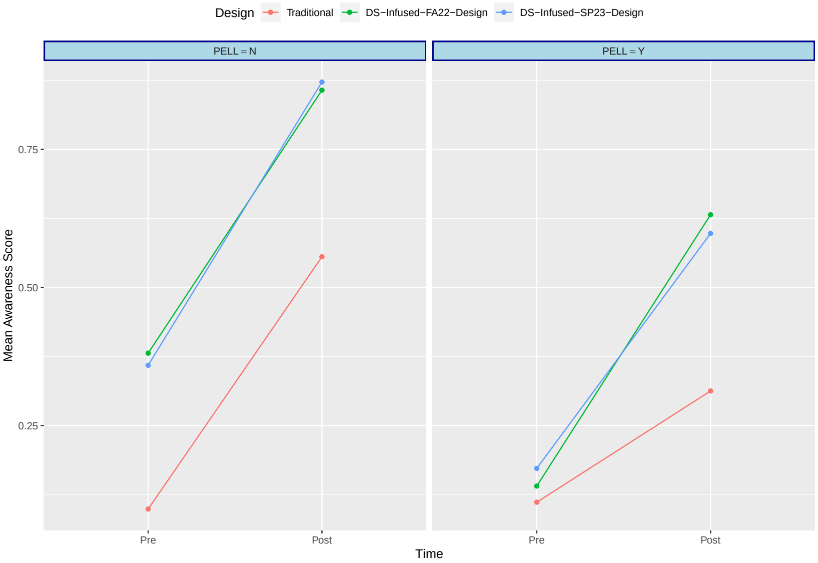
Awareness of Data Science

- DS awareness by STEM status



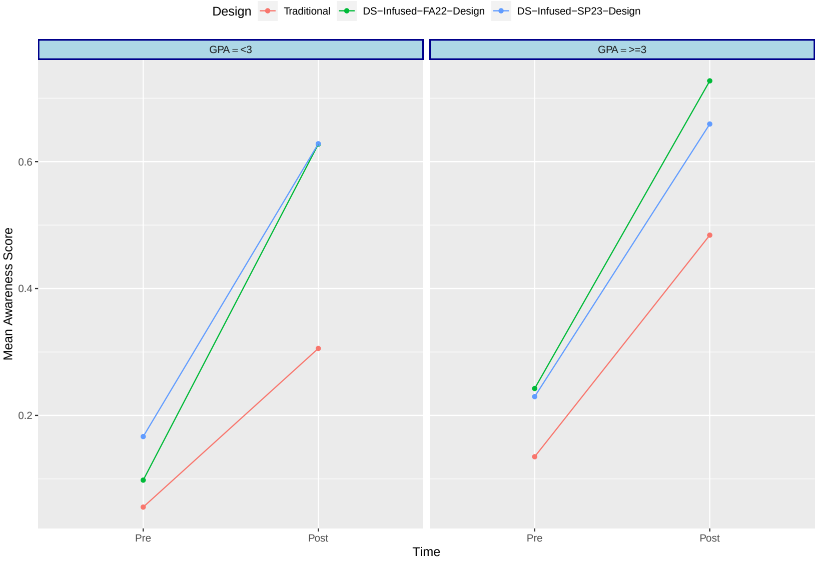
Awareness of Data Science

- DS awareness by PELL status



Awareness of Data Science

- DS awareness by pre-course GPA



Awareness of Data Science

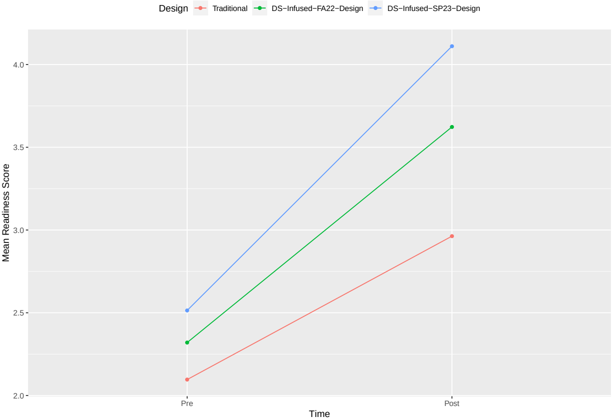
- Gains in DS awareness: regression on course design

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.

Readiness for Data Science

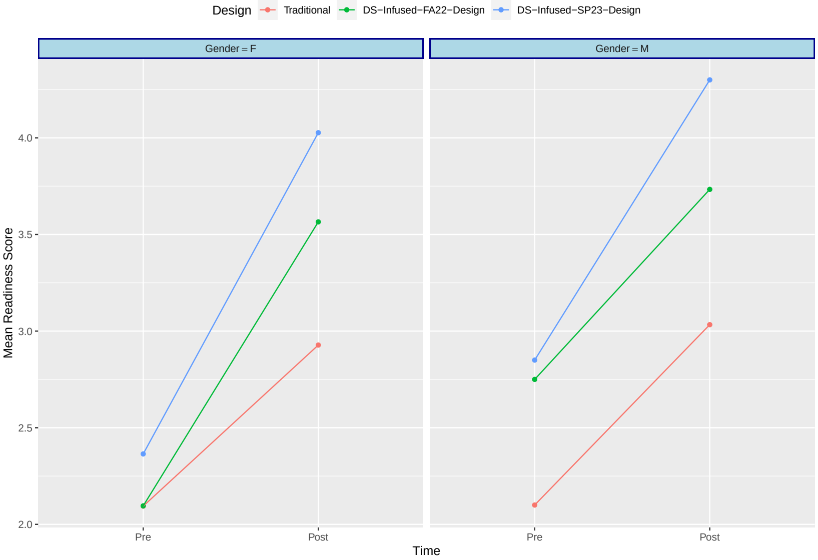
- Gains in DS readiness by course design

Course Type	n	Mean_Pre	SD_Pre	Mean_Post	SD_Post	Mean_Diff	SD_Diff
Traditional	54	2.10	0.88	2.96	1.10	0.87	0.92
DS-Infused-FA22-Design	35	2.32	0.80	3.62	1.08	1.30	1.06
DS-Infused-SP23-Design	65	2.51	1.05	4.11	0.87	1.60	1.11



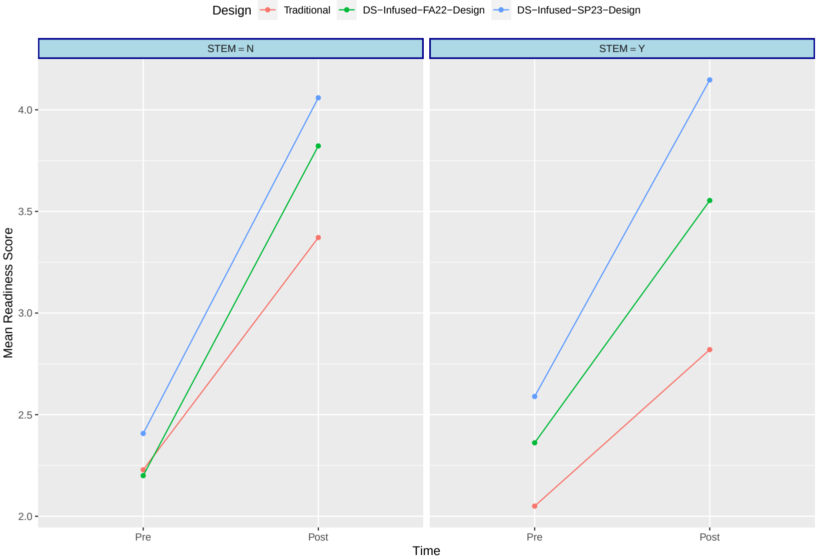
Readiness for Data Science

- DS readiness by course design & gender



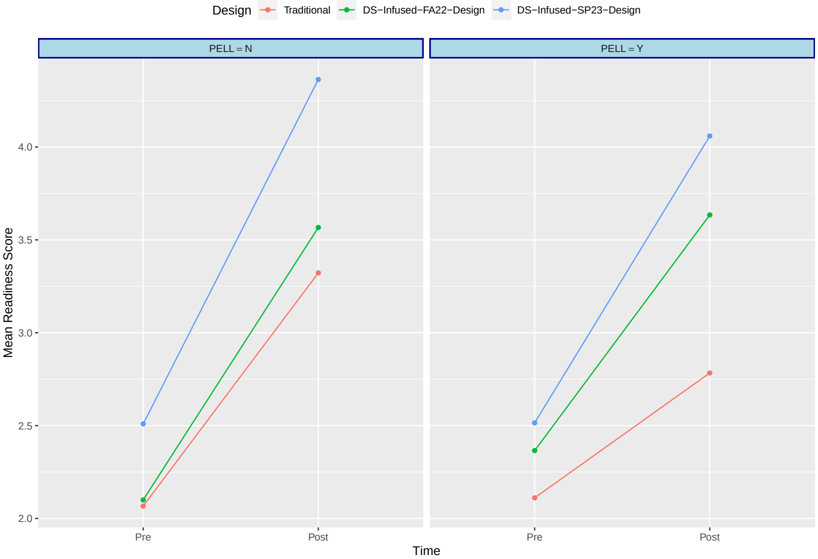
Readiness for Data Science

- DS readiness by course design & STEM status



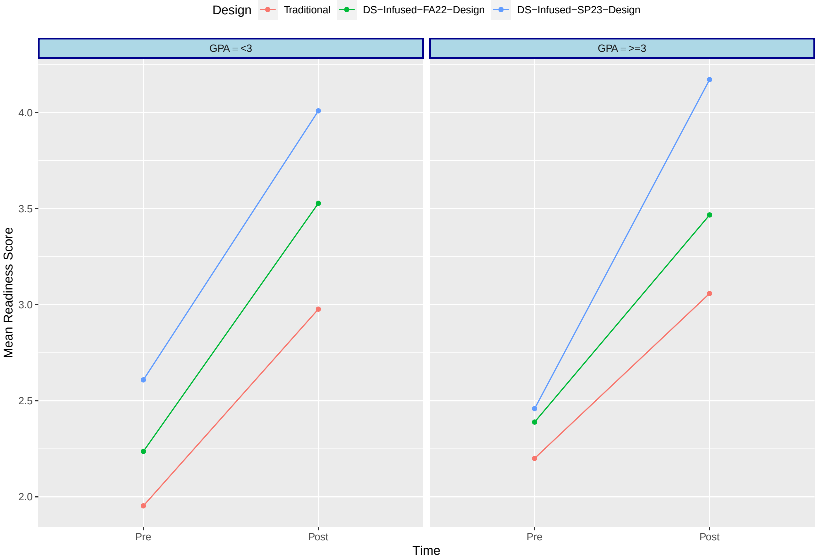
Readiness for Data Science

- DS readiness by course design & PELL status



Readiness for Data Science

- DS readiness by course design & pre-course GPA



Readiness for Data Science

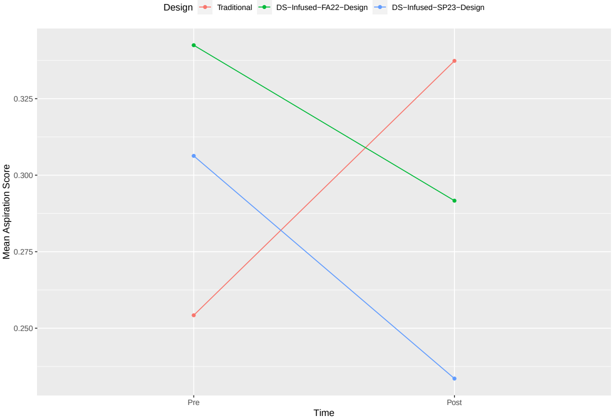
- Gains in DS readiness: regression on course design

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.

Data Science Aspirations

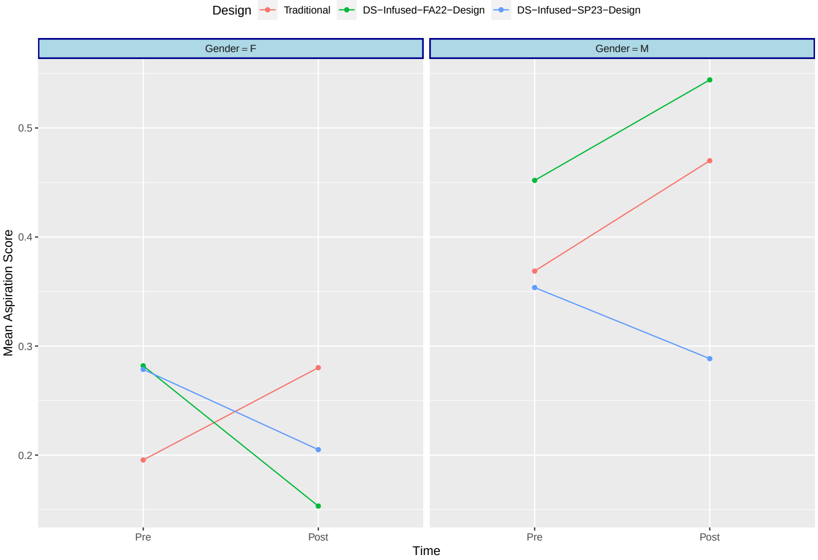
- Level of DS aspiration by course design

Course_Type	n	Mean_Pre	SD_Pre	Mean_Post	SD_Post	Mean_Diff	SD_Diff
Traditional	81	0.22	0.36	0.33	0.50	0.11	0.50
DS-Infused-FA22-Design	47	0.37	0.51	0.30	0.50	-0.07	0.55
DS-Infused-SP23-Design	72	0.22	0.36	0.23	0.39	0.01	0.37



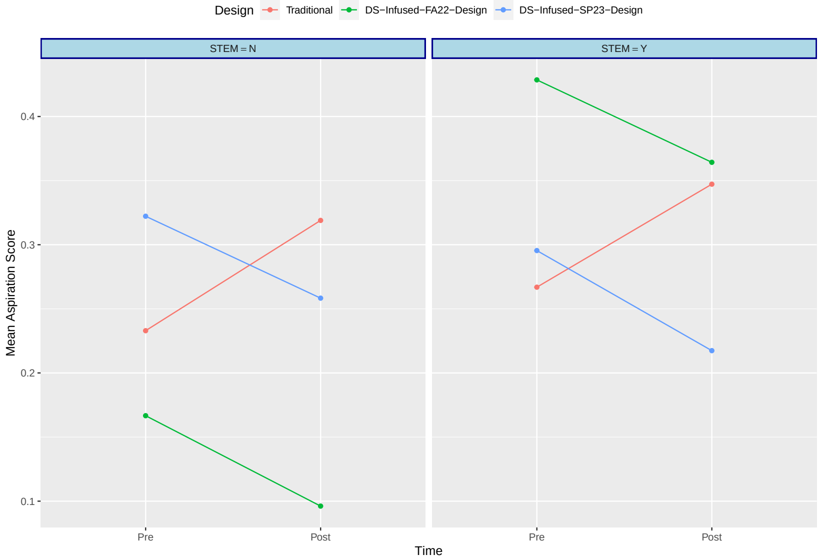
Data Science Aspirations

- Level of DS aspiration by course design & gender



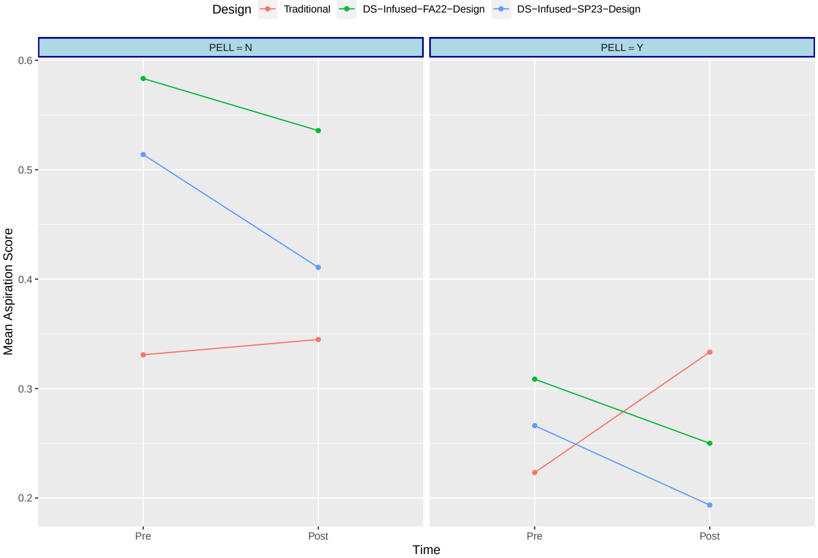
Data Science Aspirations

- Level of DS aspiration by course design & STEM status



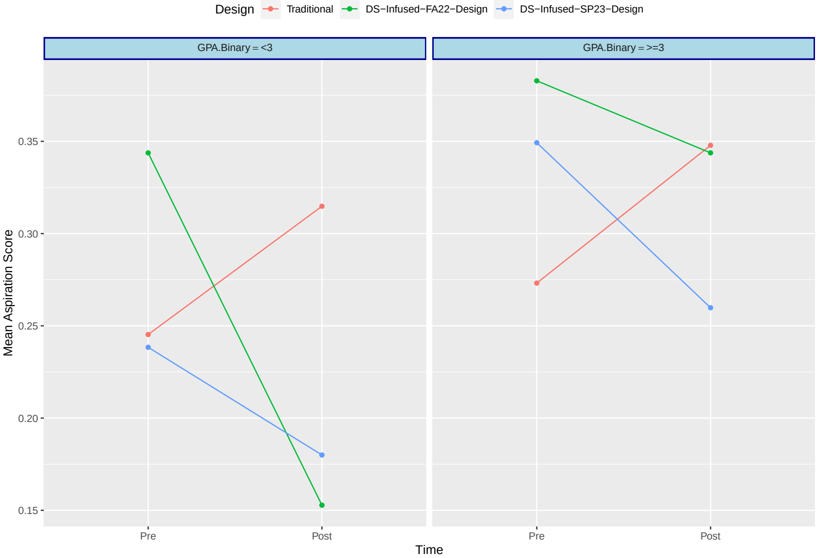
Data Science Aspirations

- Level of DS aspiration by course design & PELL status



Data Science Aspirations

- Level of DS aspiration by course design & pre-course GPA



Data Science Aspirations

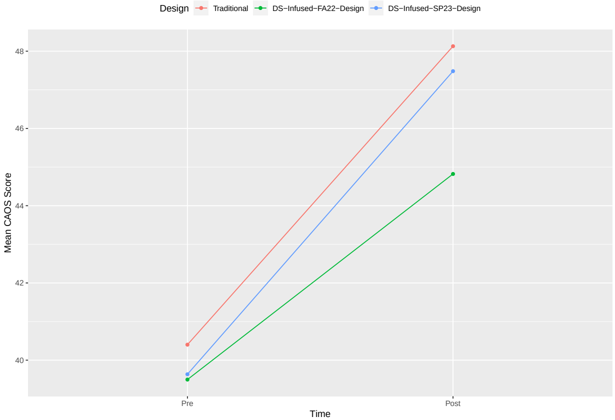
- Change in DS aspiration: regression on course design

Regression Term	**Estimate**	**LCL**	**UCL**	**p.value**	**Sig.**
Intercept	0.05	-0.53	0.63	0.8688	Not Sig.
Type: DS-Infused-FA22-Design	-0.25	-0.44	-0.07	0.0074	**
Type: DS-Infused-SP23-Design	-0.10	-0.26	0.05	0.2030	Not Sig.
Sex: Male	0.04	-0.11	0.19	0.5946	Not Sig.
Race: Not Black	-0.03	-0.21	0.16	0.7821	Not Sig.
PELL Recepient: Yes	0.14	-0.06	0.33	0.1645	Not Sig.
Rural: Yes	-0.01	-0.22	0.20	0.9514	Not Sig.
Residency: Out-of-State	-0.04	-0.20	0.13	0.6532	Not Sig.
STEM: Yes	-0.04	-0.19	0.11	0.5902	Not Sig.
AP Stat: Yes	0.17	0.01	0.33	0.0426	_*
Pre-Course Cum GPA	-0.07	-0.22	0.07	0.3156	Not Sig.
Attendance	0.00	0.00	0.01	0.4408	Not Sig.

Statistical Learning Gains

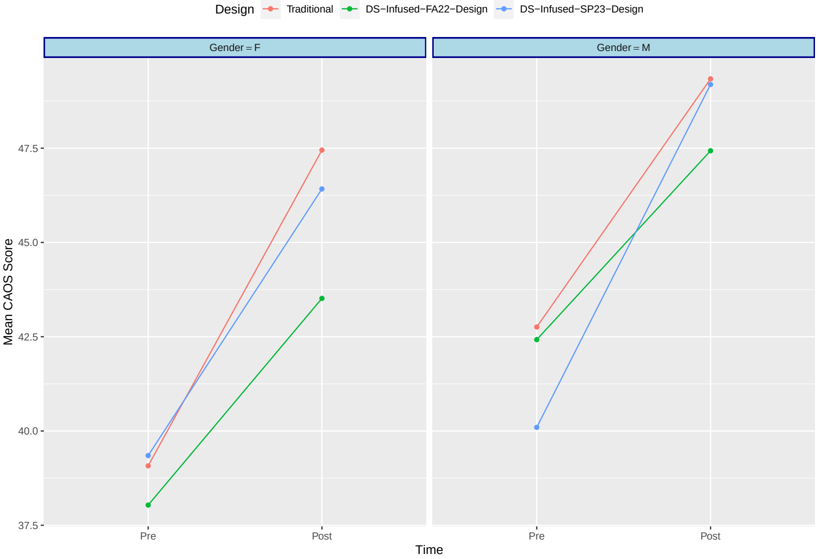
- Statistical learning gains by course design

Course Type	n	Mean_Pre	SD_Pre	Mean_Post	SD_Post	Mean_Diff	SD_Diff
Traditional	75	40.40	10.90	48.13	14.84	7.73	16.75
DS-Infused-FA22-Design	57	39.50	11.24	44.82	12.20	5.32	12.86
DS-Infused-SP23-Design	112	39.64	12.43	47.48	14.77	7.85	18.54



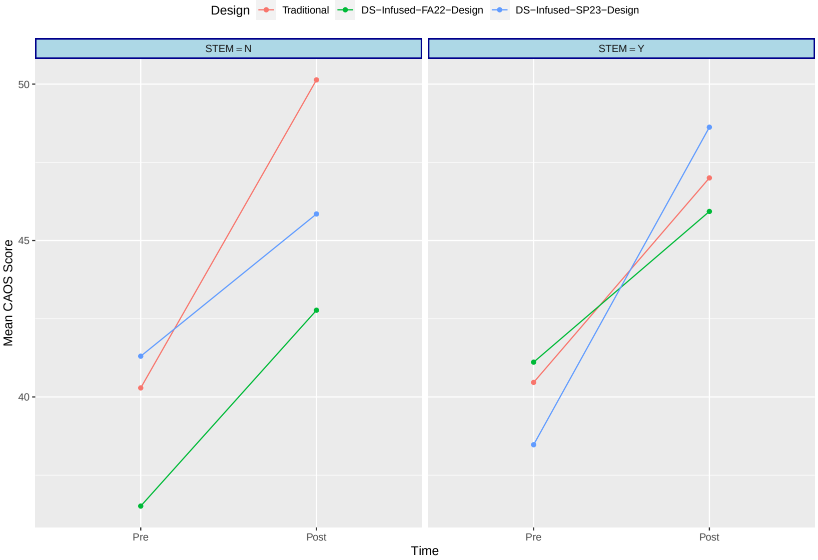
Statistical Learning Gains

- Statistical learning gains by course design & gender:



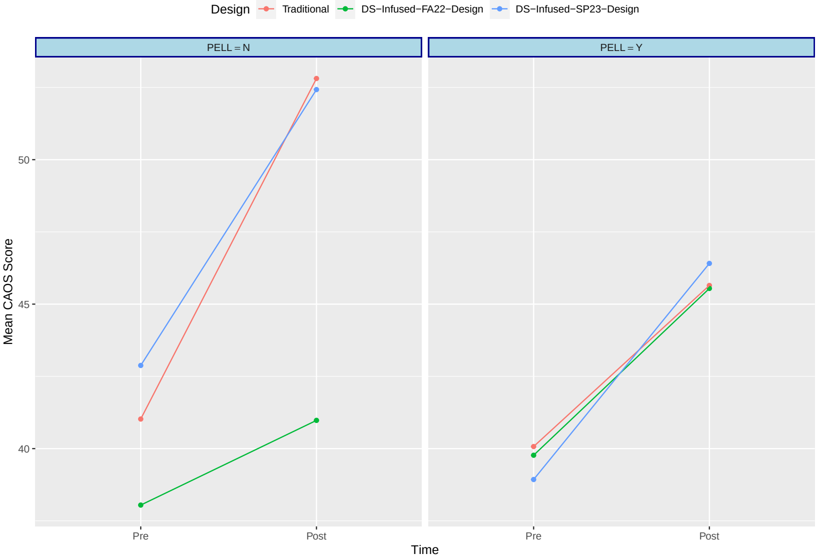
Statistical Learning Gains

- Statistical learning gains by course design & STEM status:



Statistical Learning Gains

- Statistical learning gains by course design & PELL status:



Statistical Learning Gains

- 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	-*

Overall Student Performance

- Overall student performance by course design:

Course Type	Grade	n	percent
Traditional	ABC	99	72.26
Traditional	***DFW***	***38***	***27.74***
DS-Infused-FA22-Design	ABC	56	65.12
DS-Infused-FA22-Design	***DFW***	***30***	***34.88***
DS-Infused-SP23-Design	ABC	120	83.92
DS-Infused-SP23-Design	***DFW***	***23***	***16.08***

Overall Student Performance

- Overall student performance by course design & gender:

Course Type	**Gender**	**Grade**	**n**	**percent**
Traditional	F	DFW	28	30.77
Traditional	*M*	*DFW*	*10*	*21.74*
DS-Infused-FA22-Design	F	DFW	13	24.07
DS-Infused-FA22-Design	*M*	*DFW*	*17*	*53.12*
DS-Infused-SP23-Design	F	DFW	13	14.77
DS-Infused-SP23-Design	*M*	*DFW*	*10*	*18.18*

Overall Student Performance

- Overall student performance by course design & STEM status:

Course Type	**STEM**	**Grade**	**n**	**percent**
Traditional	N	DFW	16	32.00
Traditional	*Y*	*DFW*	*22*	*25.29*
DS-Infused-FA22-Design	N	DFW	9	29.03
DS-Infused-FA22-Design	*Y*	*DFW*	*21*	*38.18*
DS-Infused-SP23-Design	N	DFW	12	20.34
DS-Infused-SP23-Design	*Y*	*DFW*	*11*	*13.10*

Overall Student Performance

- Overall student performance by course design & PELL status:

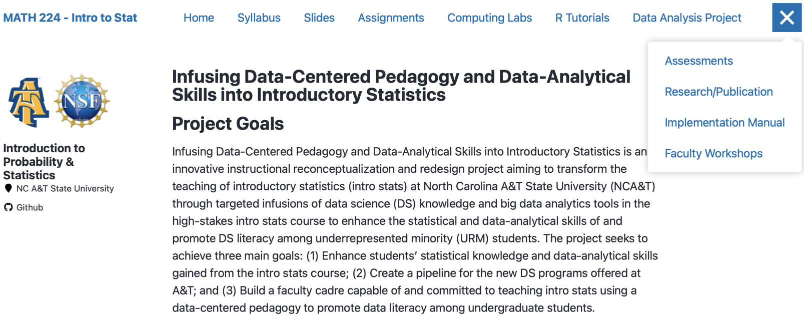
Course Type	**PELL**	**Grade**	**n**	**percent**
Traditional	N	DFW	6	15.38
Traditional	*Y*	*DFW*	*32*	*32.65*
DS-Infused-FA22-Design	N	DFW	4	36.36
DS-Infused-FA22-Design	*Y*	*DFW*	*26*	*34.67*
DS-Infused-SP23-Design	N	DFW	4	15.38
DS-Infused-SP23-Design	*Y*	*DFW*	*19*	*16.24*

Concluding Remarks

- Infusing computation and DS tools/knowledge into Intro Stats was associated with
  - significant gains in students' levels of awareness of and readiness for Statistics/DS education opportunities
  - modest statistical learning gains (to be confirmed by further data collection)
  - substantial improvement in the course success rate
- Infusing computation and DS tools/knowledge into Intro Stats seemed to drive some students away from aspiring for further DS education
  - This is somewhat in line with other findings in the literature that hinted at the complexity of computing and the challenges of integrating computing into introductory courses (e.g., Woodard and Lee, 2021)

Resources for Teaching a DS-Infused Intro Stats Course

- Project's Website on GitHub: https://introtostatncat.github.io



Infusing Data-Centered Pedagogy and Data-Analytical Skills into Introductory Statistics

Project Goals

Infusing Data-Centered Pedagogy and Data-Analytical Skills into Introductory Statistics is an innovative instructional reconceptualization and redesign project aiming to transform the teaching of introductory statistics (intro stats) at North Carolina A&T State University (NCA&T) through targeted infusions of data science (DS) knowledge and big data analytics tools in the high-stakes intro stats course to enhance the statistical and data-analytical skills of and promote DS literacy among underrepresented minority (URM) students. The project seeks to achieve three main goals: (1) Enhance students' statistical knowledge and data-analytical skills gained from the intro stats course; (2) Create a pipeline for the new DS programs offered at A&T; and (3) Build a faculty cadre capable of and committed to teaching intro stats using a data-centered pedagogy to promote data literacy among undergraduate students.

Future Work

- Continue to implement, evaluate, and refine the course design
- Evaluation of students' computational competency (e.g., competency in R coding)
  - We are not aware of an existing tool to assess computational competency

Acknowledgments

- This work is supported by NSF Grant #HRD2106945 (07/2021 – 06/2024)
  - Project Title: Infusing Data-Centered Pedagogy and Data-Analytical Skills into Introductory Statistics
- We are grateful to our fellow Intro Stats faculty who helped with the implementation and data collection: Mingxiang Chen and Kossi Edoh.

References I

- 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.
- 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.
- Nolan, D., and Temple Lang, D. (2010). Computing in the statistics curricula. *The American Statistician*, 64, 97–107.

References II

- 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.
- Woodard, V. and Lee, H. (2021). How students use statistical computing in problem solving. *Journal of Statistics and Data Science Education* 29(1), 1– 18.