Increasing accessibility and student readiness for statistical competitions

Nicole Dalzell and Ciaran Evans
Wake Forest University

What are statistical competitions?

- Students compete, often in teams, to analyze large, complex data and answer real research questions
- **ASA DataFest:**
  - Students spend a weekend analyzing data from a client, then present their findings to a panel of expert judges
  - Held at locations around the country
- Other competitions:
  - WiDS Datathon
  - NFL Big Data Bowl
  - Kaggle competitions

Goal

Our goal: make statistical competitions accessible to students with less background and experience

- Participation is valuable for a future career in statistics and data science
- Competitions, like ASA DataFest, often involve complex, open-ended data analyses which are much more challenging than what students work with in introductory courses
- Students with only one or two courses need additional preparation

Today

We will share

- Tips for helping all students feel welcome in competitions
- Skills students need to succeed in competitions
- Activities for helping students learn these skills for competition day
- The outcome of DataFest 2022

Step 1: Competition Levels

Step 1: Competition Levels

Our goal: make statistical competitions accessible to students with less background and experience

Challenge: Competing with students who have much more experience is intimidating

Our Approach: Create two competition levels.

- Levels are based on statistical background
- Students compete only with students in their level

Step 1: Competition Levels

- **Level 1**: Students who have taken only 1-2 stats courses
  - Typically first years and sophomores who have no more than regression (Stat2) knowledge
- **Level 2**: Students who have taken 3+ stats courses.
  - Typically juniors and seniors

Step 1: Competition Levels

Survey Question Post DataFest: *If you were a Level 1 team, did you feel more comfortable competing knowing you were competing with other students with similar level of statistics training as you? Choose yes or no*

Results: 100% of our Level 1 respondents replied “yes”

Step 2:
The Activities

Step 2: Activities

Our goal: make statistical competitions accessible to students with less background and experience

Challenge: Statistical competitions require skills these students do not have

Our Approach: Create activities to help students learn these skills over the course of a semester

Step 2: Activities

Our approach: create activities to help students gain the skills needed to succeed in statistical competitions

- Students engaged with these activities in a one credit Pass / Fail course
- The class met one hour a week
- The entire hour was spent working in their DataFest teams on these activities
- Both professors moved around the room answering questions

Discussion:
What skills do students need for statistical competitions?

Our List of Skills

- Working in teams
- Coding in teams (Collaborative Coding)
- Data Cleaning / Data Wrangling
- Working with large, messy data
- Choosing appropriate statistical methods based on the client request
- Presenting/explaining results
- Working on a short time frame

Data and Activities

- **Data:** predicting building energy use
  - Provided by Climate Change AI and Lawrence Berkeley National Laboratory
  - Used in WiDS March 2022 Datathon
- **Activities:** available at https://datafest-prep.github.io/calendar/
  - Designed for students with only a first course in regression
  - Reinforce existing skills and learn new tools for participating in DataFest
  - Used in an elective DataFest prep course

Image: https://www.widsconference.org/blog_archive/announcing-the-5th-annual-wids-datathon-2022-challenge-using-data-science-to-mitigate-climate-change
Data: Climate Change AI (CCAI) and Lawrence Berkeley National Laboratory (Berkeley Lab). (2022 January). WiDS Datathon 2022, Version 1. Retrieved January 10, 2022 from https://www.kaggle.com/competitions/widsdatathon2022/overview.

Activities

Skills	***Activity***
Fundamentals in R	Activity 1
Data Visualization	Activity 2
Data Cleaning / Data Wrangling Working with large, messy data	Activity 3 Part 1 and Part 2
Working in teams Coding in teams (Collaborative Coding)	Activity 4

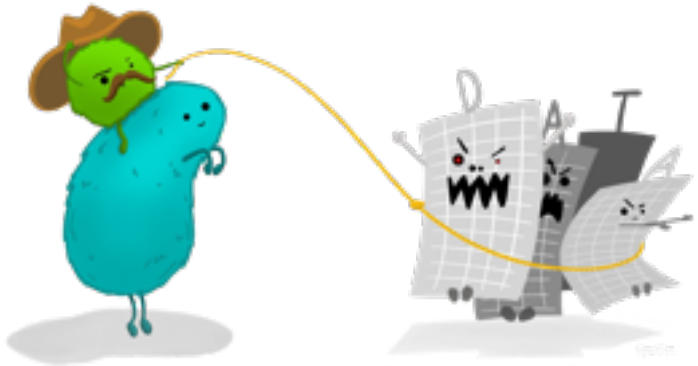
Activities

Skills	***Activity***
Choosing appropriate statistical methods based on the client request	Activity 5, 6 and 7
Getting started with an open-ended research question	Activity 8
Presenting/explaining results	Activity 9

Other Tips

Other Tips

- *Use complex data*
  - Data in competitions is often much larger, messier, and more complex than anything students see in class
  - Good data sources: datasets from previous competitions! (Lots of options on Kaggle)
- *Deliberately recruit students with less experience*
  - We emailed all students in introductory regression courses to advertise DataFest and our prep course
  - Advertising material specified the course was not for advanced students



Artwork by @allison_horst

Our Advertisement

If you are interested in DataFest, you can sign up for STA 175, a one credit pass/fail course that will meet once a week. The only pre-req is STA 112. No homework, no tests, but a chance to learn about R and learn about working with real data. This is especially useful for those who are relatively new to R or looking to learn more about it. If you participate in class and participate fully in DataFest, you pass!

We also emphasized the two levels, and that students would compete only against others in their level.

Other Tips

- *Give students a safe space to make mistakes*
  - Don’t grade work or course (except P/F)
  - **Do** have students submit work…
- *Competition prep helps students get familiar with the instructor*
  - Students get comfortable with the people running DataFest
  - Students feel more confident asking for help during the competition

Artwork by @allison_horst

The Results: DataFest 2022

The Results: Our Students

	2019	2020	2022
Level 1 Students	0	8	29
Level 2 Students	33	40	34
Total	33	48	63

Data Cleaning-Create A Variable

Level 0

Invitatio

id	safe	inviteText	peopleInvolve	acceptText	rejectText
0	1	Do you want to get pizza	0,3	The pizza is delicious	Too bad! [npc0] &
1	1	Want to come to the game	0	You have fun	Too bad! You
2	0	I'm ditching school to drink	2	The principal catches	Smart move! The
3	1	Hey, do you want to come	3	You do well on your	Too bad! You
4	0	My parents won't be home	1,2,4	You drink too much	Smart move! The
5	0	We're sneaking out to see	1,4	You get really banged	Smart move! In
6	1	My dance team is having	0,2,3	You have a lot of fun	Too bad! You
7	0	Hey, let me take you out so	1	During your date he	Smart move! You
8	1	Want to see a movie with	0,3	You see the latest	Too bad! You
9	0	Meet us behind the school	1,4	You get caught	Smart move! You
10	0	My brother said I should	2	You get drunk in her	Smart move! You

- **Invitation_id: The Invitation Message Number**
- **Safe: The activity is safe or not**

Significant factors for different PLAYERS

	Safe	Freindship
Player 1	×	×
Player 2	✓	✓
Player 3	×	✓

Reflections

Some student feedback:

- “Each activity was well thought out and prepared us for the competition later in the year”
- “Professors were both constantly engaged with the students and how they were progressing throughout the course”
- “It is a really exciting contest!” (from a Level 1 student)

Possible modifications:

- Additional R activities at the beginning?
- More time on data wrangling?

The Website link (again) and any questions?

Resources website: https://datafest-prep.github.io/

GitHub repository: https://github.com/datafest-prep/datafest-prep.github.io

- All files for the activities and web pages