GAISE preK-12 II, The Introduction to Data Science Course, and Data Science Education

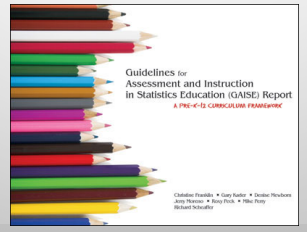
Rob Gould
UCLA

outline

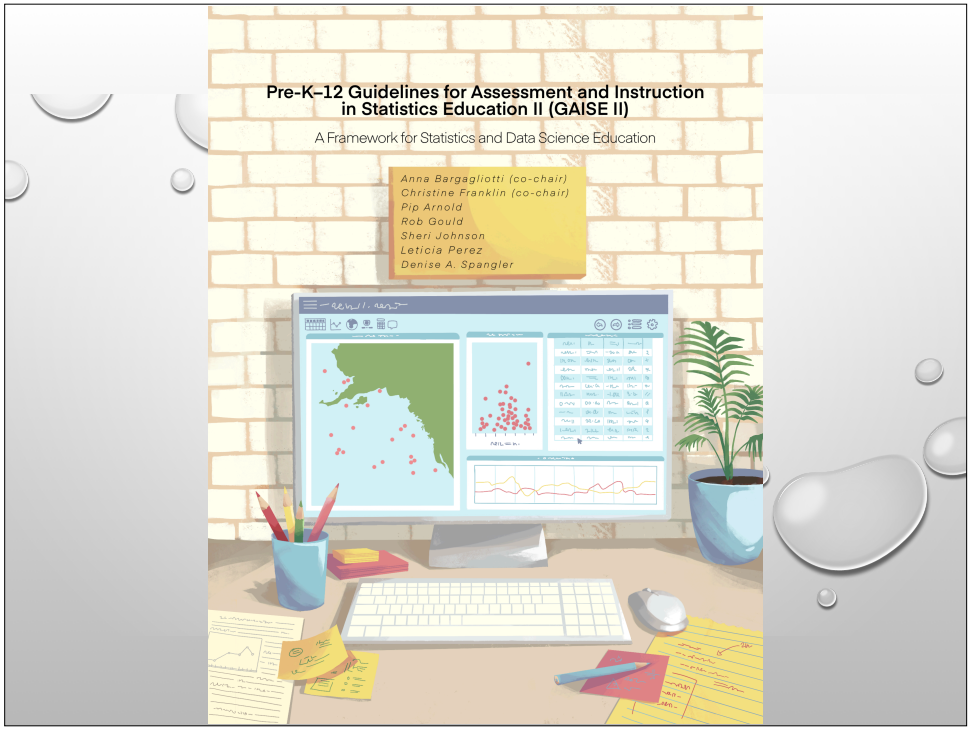
- Overview of GAISE II
- Overview of the Introduction to Data Science (IDS) high school course
- Examples of how IDS meets the GAISE II
- (If time remains, a demonstration of the participatory sensing dashboard.)

ORIGINAL VISION

- "Every high-school graduate should be able to use sound statistical reasoning to intelligently cope with the requirements of citizenship, employment, and family and to be prepared for a healthy, happy, and productive life." – Pre-k-12 GAISE, p.1



WE MUST HELP STUDENTS (STARTING AT YOUNG AGE) MAKE SENSE OF DATA THAT SURROUND THEM



MAJOR ENHANCEMENTS BEING ADDRESSED:

- Incorporating new Stats ed research
- Different data and variable types (images, words, etc.-
- Distinguishing between primary and secondary data
- Questioning throughout the statistical problem-solving process – question posing and question asking
- Multivariate thinking
- Science based examples
- Integrating technology and computational thinking

FRAMEWORK

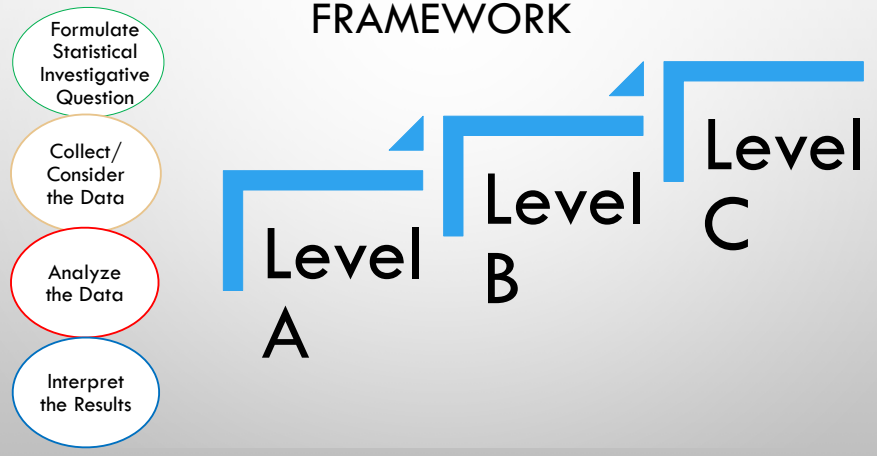


Table 1: The Framework

Process Component	Level A	Level B	Level C
II. Collect Data/ Consider Data	Understand that data are information; recognize that to answer a statistical investigative question, a person may collect data themselves specifically for that purpose, or a person may use data that have been collected by other people for another purpose Understand how to collect and record information from the group of interest using surveys and measurements collected from observations and simple experiments Understand that a variable measures the same characteristic on several individuals or objects and results in data values that may fluctuate Understand that within a data set there can be different types of variables (e.g., categorical or quantitative) Interrogate the data set to understand the context of the variables as they may relate to statistical investigative questions Understand that data are not always pristine but may contain errors, have missing values, etc., and that decisions have to be made about how to account for these issues	Understand that data are information collected and recorded with a purpose and can be organized and stored in a variety of structures (e.g., spreadsheets) Understand that a sample can be used to answer statistical investigative questions about a population. Recognize the limitations and scope of the data collected by describing the group or population from which the data are collected Understand that data can be used to make comparisons between different groups at one point in time and the same group over time Recognize that data can be collected using surveys and measurements, and develop a critical attitude in analyzing data collection methods Understand that quantitative variables may be either discrete or continuous Understand how to interrogate the data to determine how the data were collected, from whom they were collected, what types of variables are in the data, how the variables were measured (including units used), and the possible outcomes for the variables Understand that data can be collected (primary data) or existing data can be obtained from other sources (secondary data) Understand how random assignment in comparative experiments is used to control for characteristics that might affect responses	Word as: Apply an appropriate data collection plan when collecting primary data or selecting secondary data for the statistical investigative question of interest. Distinguish between surveys, observational studies, and experiments Understand what constitutes good practice in designing a sample survey, an experiment, and an observational study Understand the role of random selection in sample surveys and the effect of sample size on the variability of estimates Understand the role of random assignment in experiments and its implications for cause-and-effect interpretations Understand the issues of bias and confounding variables in observational studies and their implications for interpretation Understand practices for handling data that enhance reproducibility and ensure ethical use, including descriptions of alterations, and an understanding of when data may contain sensitive information Understand how concerns about privacy and human subjects may affect the collection and distribution of data Understand that in some circumstances, the data collected or considered may not generalize to the desired population, or this data may be the entire population

Introduction to Data Science

Vision

- All students, regardless of success in mathematics and regardless of whether they plan to attend college, should develop "data acumen" to be engaged citizens.
- All students can learn to analyze data to answer questions that interest them and to address problems that they feel are important.
- All students should understand how data analysis includes ethical responsibilities concerning privacy, confidentiality, and equity.

Introduction to Data Science

- Developed with funding from the National Science Foundation in partnership with
  - Los Angeles Unified School District (LAUSD),
  - UCLA Department of Statistics,
  - UCLA Department of Computer Science, and
  - UCLA Graduate School of Education and Information Science.

LAUSD Demographics

- LAUSD is second-largest school district in U.S., with about 750,000 students;
- 80% of students are below the poverty level
- IDS was designed for *all* students, regardless of career goals or past academic achievement

https://commons.wikimedia.org/wiki/File:Hamilton_High_School_LAUSD_Entrance.jpg

{{Information |Description= Hamilton High School Entrance in Los Angeles, CA |Source= Own Work |Date= 26 March 2006 |Author= User:Jorobeq }}

Introduction to Data Science

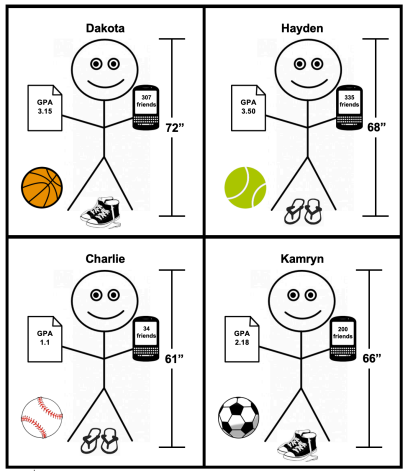
- A year-long course for high-school students (ages 14-18)
- Requirements: algebra
- Currently taught in 65 high schools in 4 states and has taught 12,500 students to date. (Currently, 5000 students are taking the course.)
- Open-source and freely available under Creative Commons license: http://introdatascience.org

Structure

- 4 units of daily lessons.
- Each unit ends in a "practicum": a long project that ties together the previous weeks and
  - each unit includes a *participatory sensing* data-gathering campaign.
- Lessons consist of
  - student-centered classroom activities to develop conceptual understanding and
  - data analysis labs using R (Rstudio) to apply concepts and learn basic coding.

4 Lessons from IDS

Unit 1, Lesson 2: Stick Figures

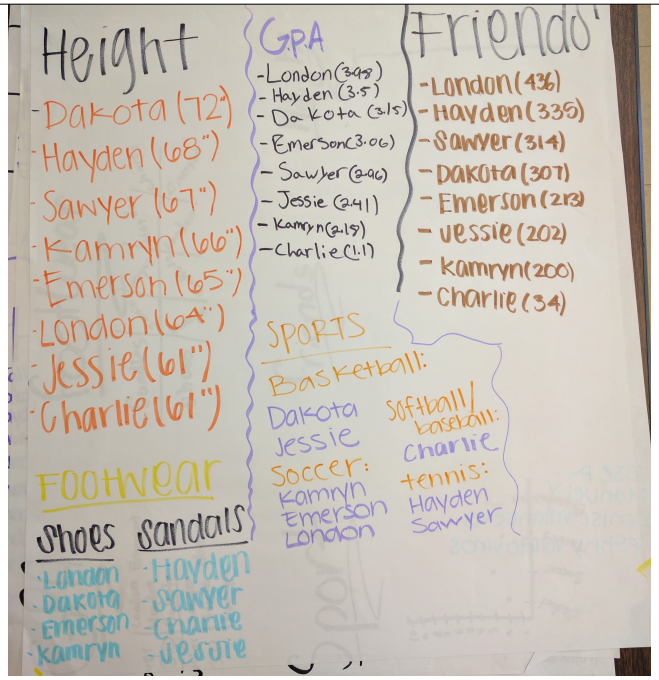


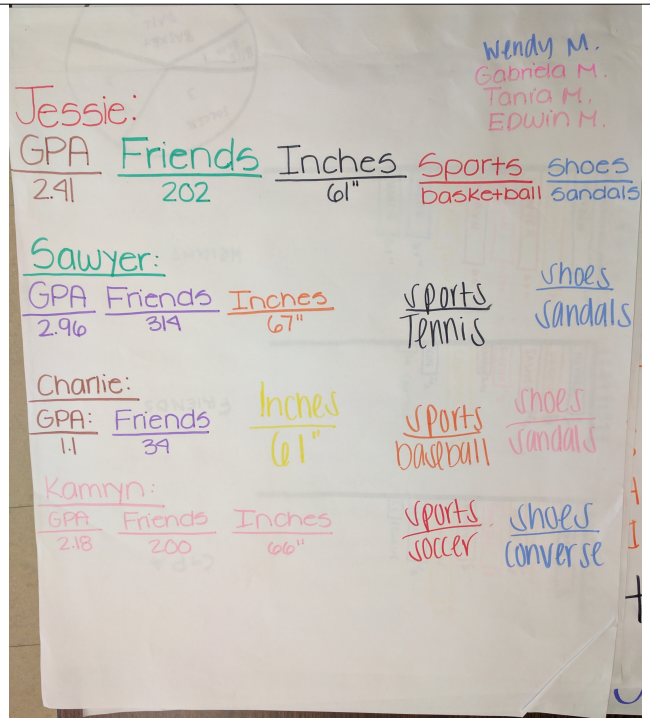
"Collect and record as much information as you can about these people"

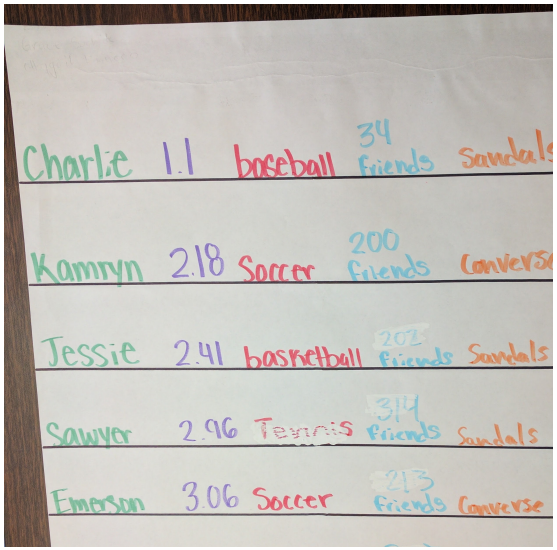
"Organize this information on a poster any way that you feel is helpful"

Posters are displayed, and students discuss:

- what are similarities and differences in the ways the data were organized
- what information ('variables') is available?
- which organizations made it easiest to see the variables?


Height
-Dakota (72")
-Hayden (68")
-Sawyer (67")
-Kamryn (66")
-Emerson (65")
-London (64")
-Jessie (61")
-Charlie (61")
FOOTWEAR
Shoes
-London
-Dakota
-Emerson
-Kamryn
Sandals
-Hayden
-Sawyer
-Charlie
-Jessie
GPA
-London (3.98)
-Hayden (3.5)
-Dakota (3.15)
-Emerson (3.06)
-Sawyer (2.96)
-Jessie (2.41)
-Kamryn (2.15)
-Charlie (1.1)
Friends
-London (436)
-Hayden (335)
-Sawyer (314)
-Dakota (307)
-Emerson (213)
-Jessie (202)
-Kamryn (200)
-Charlie (34)
SPORTS
Basketball:
Dakota
Jessie
Soccer:
Kamryn
Emerson
London
Softball/baseball:
Charlie
tennis:
Hayden
Sawyer


Wendy M.
Gabriela M.
Tania M.
EDWIN M.
Jessie:
GPA 2.41
Friends 202
Inches 61"
Sports basketball
Shoes sandals
Sawyer:
GPA 2.96
Friends 314
Inches 67"
Sports Tennis
Shoes Sandals
Charlie:
GPA: 1.1
Friends 34
Inches 61"
Sports baseball
Shoes Sandals
Kamryn:
GPA 2.18
Friends 200
Inches 66"
Sports soccer
Shoes converse



Konold, Finzer, Kreetong (2016)

- **"spreadsheet" format is not natural for many students (and their teachers)**
- **students need to develop the conception of "case"**
- **students have basically sound and solid notions of data**
- **students are comfortable and may even prefer hierarchical representations over spreadsheet representations**

Table 1: The Framework

Process Component	Level A	Level B	Level C
II. Collect Data/ Consider Data	Understand that data are information; recognize that to answer a statistical investigative question, a person may collect data themselves specifically for that purpose, or a person may use data that have been collected by other people for another purpose Understand how to collect and record information from the group of interest using surveys and measurements collected from observations and simple experiments Understand that a variable measures the same characteristic on several individuals or objects and results in data values that may fluctuate Understand that within a data set there can be different types of variables (e.g., categorical or quantitative) Interrogate the data set to understand the context of the variables as they may relate to statistical investigative questions Understand that data are not always pristine but may contain errors, have missing values, etc., and that decisions have to be made about how to account for these issues	Understand that data are information and recognize that to answer a statistical investigative question, a person may collect data themselves specifically for that purpose, or a person may use data that have been collected by other people for another purpose and can be organized and stored in a variety of structures (e.g., spreadsheets) Understand that a sample can be used to answer statistical investigative questions about a population. Recognize the limitations and scope of the data collected by describing the group or population from which the data are collected Understand that data can be used to make comparisons between different groups at one point in time and the same group over time Recognize that data can be collected using surveys and measurements, and develop a critical attitude toward data collection methods Understand that quantitative variables may be either discrete or continuous Understand how to interrogate the data to determine how the data were collected, from whom they were collected, what types of variables are in the data, how the variables were measured (including units used), and the possible outcomes for the variables Understand that data can be collected (primary data) or existing data can be obtained from other sources (secondary data) Understand how random assignment in comparative experiments is used to control for characteristics that might affect responses	Level B: Apply an appropriate data collection plan when collecting primary data or selecting secondary data for the statistical investigative question of interest. Distinguish between surveys, observational studies, and experiments Understand what constitutes good practice in designing a survey, an observational study, or an experiment, and an observational study Understand the role of random selection in sample surveys and the effect of sample size on the variability of estimates Understand the role of random assignment in experiments and its implications for cause-and-effect interpretations Understand the issues of bias and confounding in both observational studies and their implications for interpretation Understand practices for handling data that enhance reproducibility and ensure ethical use, including descriptions of alterations, and an understanding of when data may contain sensitive information Understand how concerns about privacy and human subjects may affect the collection and distribution of data Understand that in some circumstances, the data collected or considered may not generalize to the desired population, or this data may be the entire population

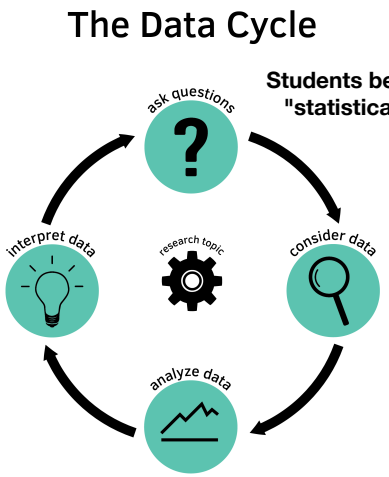
Understand that data are information

Understand how to collect and record..

Understand that a variable measures the same characteristic on several individuals

Understand that ...data can be organized and stored in a variety of structures

The Statistical Investigation Cycle is at the Foundation of IDS

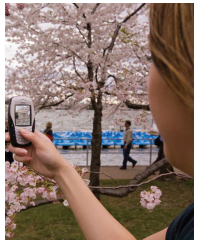


Students begin the class learn to pose "statistical investigative questions"

based on the GAISE Statistical Investigation Process and the PPDAC (Wild & Pfannkuch 1999)

participatory sensing

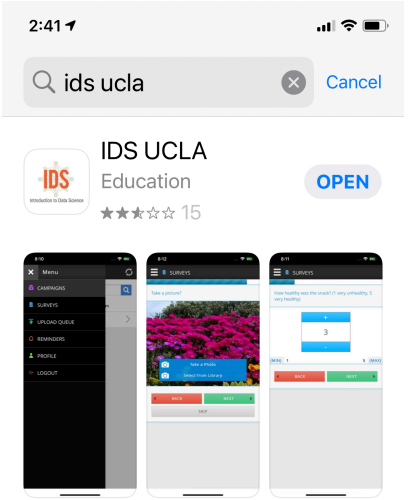
- A data-collection paradigm developed by Deborah Estrin's lab at UCLA (Center for Embedded Network Sensing)
- Students engage in participatory sensing campaigns.
- Mobile devices used to collect data to address various issues: Nutrition, recycling, stress, water conservation
- Students collect numbers, images, words, locations, times, dates.
- They are "human sensors", collecting a stream of data based on triggers, and not random samples.

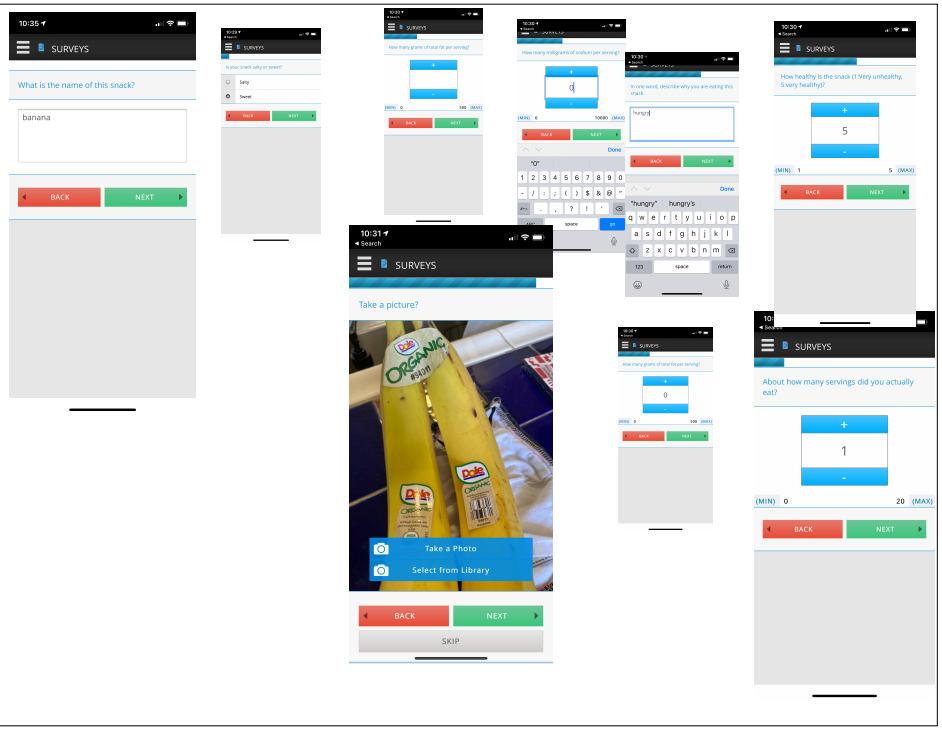


J. Burke, D. Estin, M. Hansen, A. Parker, N. Ramanathan, M. Reddy, M.B. Srivastava, Participatory Sensing. *Center for Embedded Network Sensing*. (2006).

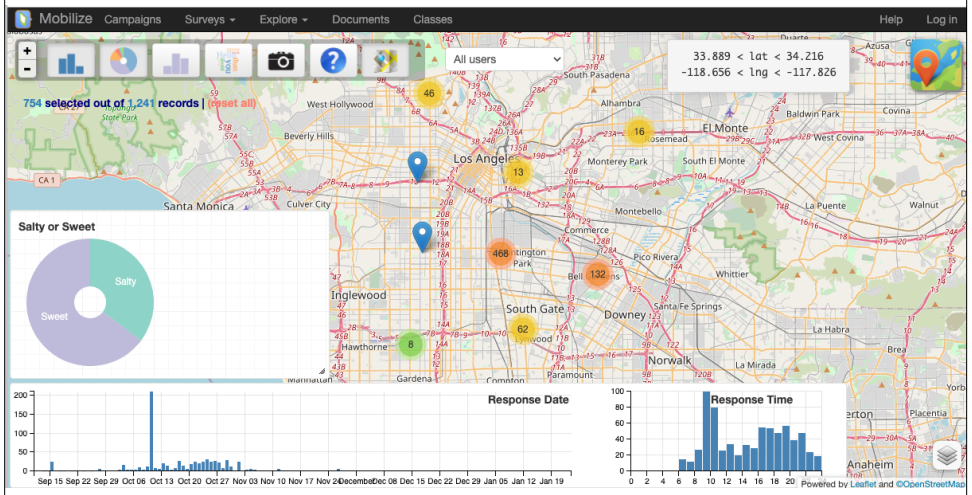
snack campaign

- Motivating Questions:
  - What is my snacking pattern?
  - How good am I at rating the healthiness of my snack?
  - Do I tend to eat healthy? How does this compare to the rest of my class?
  - Does knowing nutritional value change my habits?
- Data collection: Collect data every time you eat a snack for the next four days.





https://sandbox.mobilizingcs.org/#dashboard/#urn:public:nutrition



3. Markers

Practicum
The Data Cycle & My Food Habits

Instructions:

With a partner, you will engage in the Data Cycle to address the Research Topic:

How good are we at identifying healthy and unhealthy snacks?

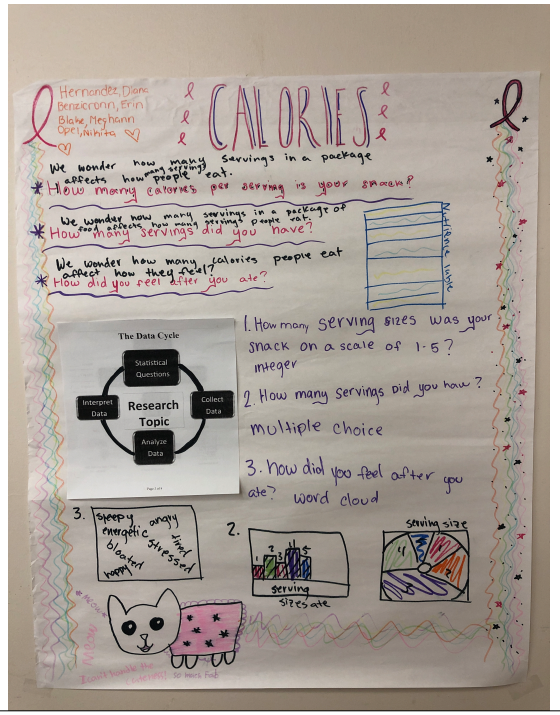
Task:

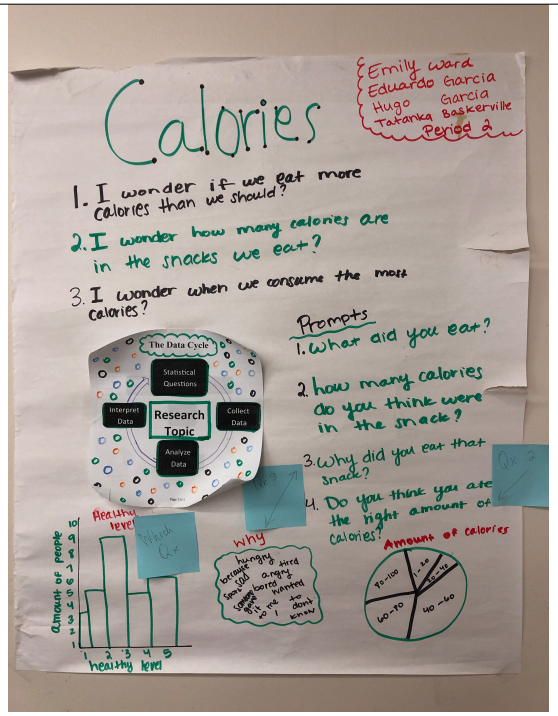
1. Create a Data Cycle poster.
2. The poster should illustrate how the Data Cycle is used to address the Research Topic.
3. Use RStudio to create at least one statistical graphic. The graphic MUST be included on the poster.
4. You and your partner will present your findings with appropriate evidence from the data.

Awards:

Your teacher will select the top posters in the following categories:

- Best Statistical Question
- Most Interesting Statistical Graphic
- Best Illustration of the Data Cycle



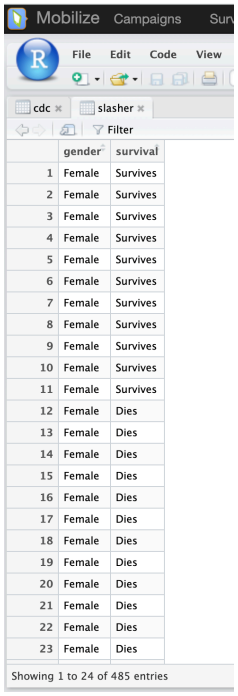


Framework (Participatory Sensing)

- Ask Questions: *Level C*
  - Formulate multivariable statistical investigative questions and determine how data can be collected and analyzed to provide an answer
- Collect and Consider Data: *Level A*
  - Understand that within a dataset that can be different types of variables
  - Interrogate the data to understand the context of the variables
- Collect and Consider Data: *Level B*
  - Interrogate to determine how data were collected, what types of variables, etc.
  - Understand that data can be collected or existing data obtained from other sources
- Collect/Consider Data: *Level C*
  - Understand how concerns about privacy and human subjects affect distribution and collection of data
  - Understand that in some situations the data may not generalize to the desired population
- Analyze: *Level C*
  - Use technology to filter and subset data
  - Summarize and describe relationships between multiple variables
- Interpret: *Level B*
  - State the limits of generalization
- Interpret: *Level C*
  - Use multivariate thinking to explain how variables impact one another

Horror Movie shuffle

Are women in slasher films more likely to survive until the end of the film than men?



	gender	survival
1	Female	Survives
2	Female	Survives
3	Female	Survives
4	Female	Survives
5	Female	Survives
6	Female	Survives
7	Female	Survives
8	Female	Survives
9	Female	Survives
10	Female	Survives
11	Female	Survives
12	Female	Dies
13	Female	Dies
14	Female	Dies
15	Female	Dies
16	Female	Dies
17	Female	Dies
18	Female	Dies
19	Female	Dies
20	Female	Dies
21	Female	Dies
22	Female	Dies
23	Female	Dies

Showing 1 to 24 of 485 entries

"What's the difference in proportion of survival rates for females and males?"

Tally whoa ... !

- Something you might have noticed is that these two lines of code aren't equivalent:

```
tally(gender ~ survival, data = slasher)

tally(survival ~ gender, data = slasher)
```

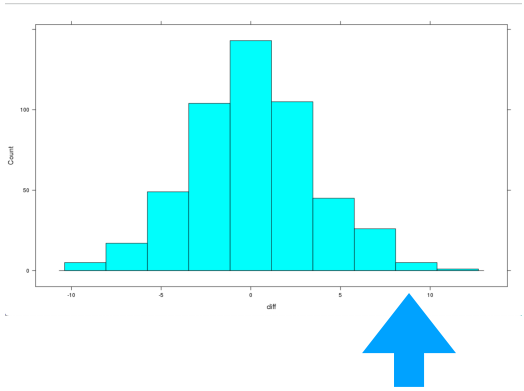
```
> tally(survival~gender, data=slasher, format="percent")
          gender
survival     Female     Male
  Dies     77.47748 86.69202
  Survives 22.52252 13.30798
```

```
> tally(gender~survival, data=slasher, format="percent")
         survival
gender       Dies Survives
  Female 43.00000 58.82353
  Male   57.00000 41.17647
```

IDS uses Rstudio Cloud but relies on the package "mobilizR" which is based on the "mosaic" package (Pruim, Horton, Kaplan)

Could this difference be due to chance?
Or is it too large?

```
> set.seed(1)
> shuffles <-do(500)*tally(shuffle(survival)~gender, data=slasher, format="percent", margins=TRUE)
> shuffles <- mutate(shuffles, diff = Survives.Female - Survives.Male)
> histogram(~diff, data=shuffles)
```



outcome in data

Analyze Data Framework:

Level A: Observe whether there appears to be a difference in two groups
Level B: Explore patterns of association between two categorical variables
Level C: Describe associations between two categorical variables
Level C: use simulations to investigate associations between two categorical variables

Interpret Results Framework:

Level B: Use statistical evidence from analyses to answer ...questions through structured answers with some teacher guidance.
Level B: Generalize beyond the sample...including a statement of uncertainty
Level C: Understand what it means for an outcome...to be plausible or not compared to chance variation

informal inference

- Makar & Rubin (2009): generalizing beyond the data at hand and expressing uncertainty.
- IDS emphasizes the "informal" aspects. Does not teach p-values or confidence intervals or formal hypothesis tests.
- Instead, develops understanding and intuition to assist when students take Statistics (which does cover 'formal' inference')
- Informal inference is mostly levels A and B in GAISE II

Why teach coding?

- Learning to "code" using R has many advantages:
  - Students use code to communicate models and ideas
  - Students more easily understand code than mathematical notation
  - Teaches reproducible research habits and communication
  - Some coding is needed for students to access data.
  - Heinzman (2020):
    - Students find that coding is "helpful" and "productive" for solving problems. (Heinzman, 2020)
    - Students find it "efficient" and "empowering" (Heinzman, 2020)

Summary

- The GAISE II revision provides learning outcomes that can guide the development of a data science education curriculum.
- The GAISE II assumes a "flavor" of data science that has the primary focus on teaching students to develop "data acumen"--learning to reason with data.
- Many different groups are creating data science curriculum. It is extremely important that statisticians play a primary role in shaping these curricula

Thank you!

rgould@stat.ucla.edu
amstat.org/education/gaise
introdatascience.org

GAISE II team
Anna Bargagliotti
Christine Franklin
Pip Arnold
Rob Gould
Sheri Johnson
Leticia Perez
Denise A. Spangler

IDS Team
Rob Gould
Suyen Machado
Monica Casilla
Brendan Chang
Leeanne Trusela
Shubahm Kayastha
James Molyneux
Amelia McNamara
Terri Johnson
Hongsuda Tangmungarunkit
Steve Nolan

references

A. Bargagliotti, C. Franklin, P. Arnold, R. Gould, S. Johnson, L. Perez and D. Spangler, Pre-K-12 Guidelines for assessment and instruction in statistics education (GAISE) report II. Alexandria, VA: American Statistical Association and Reston, VA: National Council of Teachers of Mathematics. (2020)

J. Burke, D. Estin, M. Hansen, A. Parker, N. Ramanathan, M. Reddy, M.B. Srivastava, Participatory Sensing. *Center for Embedded Network Sensing*. (2006).

R. Gould, A. Bargagliotti, T. Johnson, An Analysis of Secondary Techers' Reasoning with Participatory Sensing Data, Statistics Education Research Journal, 16(2) November 2017.

E. Heinzman, Math is No Longer a Four-Letter Word: A Mixed Methods Study of Two Non-Traditional Fourth-Year Mathematics Classes. PhD dissertation, University of California, San Diego, 2020. https://escholarship.org/uc/item/25d4s7vq

C. Konold, W. Finzer, K. Kreeton, Modeling as a Core Component of Structuring Data, Proceedings International Conference on Statistical Reasoning, Teaching and Learning, 2016

K. Makar and A. Rubin, A Framework For Thinking About Informal Statistical Inference, Statistics Education Research Journal 8(1)(2009): 82-105.

C. Wild and M. Pfannkuch, Statistical Thinking in Empirical Enquiry, International Statistical Review, 67(1999), 223-265.

Traditional Mathematics Pathway to University

Algebra 1 | Geometry | Algebra II | Pre-Calculus

Advanced Mathematics Pathway

Geometry | Algebra II | Pre-Calculus | Calculus

Proposed Data Science Pathway

Algebra | Computer Science | Into Data Science | Statistics

Independent study found that IDS increases "college readiness"

https://sandbox.mobilizingcs.org/#dashboard/#urn:public:snack