Unit 1: Data in Learners' Lives

Lesson 1.6: Building a Survey for Data Collection

In this lesson, students will create a Google Form to collect data. Additionally, they will learn about some basic data cleaning issues and solutions, and how to plan a sample to represent a larger population.

Duration: 90 minutes

Objective: By the end of this lesson, students will be familiar with different sampling designs and good data science questions.

Lesson Walkthrough Video: Unit 1 Lesson 6 - Teacher Walkthrough

CSTA Standards in this Lesson

Identifier	Concept	Subconcept	Standards
HS-DAA-23	Data & Analysis	Data Processing	Use a digital tool to clean and organize text-based data.

CSTA Data Science Specialty Standards in this Lesson

Identifier	Concept	Subconcept	Standards
S1-DSC-PP-12	Data Science	Professional Practice	Apply ethical principles to data collection, analysis, and communication to promote privacy, transparency, and accountability.
S2-DSC-CC-04	Data Science	Creation & Curation	Choose appropriate data to collect for a data science project based on available tools, skills, and project goals.

Lesson activities

Class Discussion - Local Issues Project (25 min)

(CSTA standards in this activity: 3A-AP-13, 3A-DA-12)

- The teacher presents slides on types of questions most appropriate for a data science investigation. Distinguish from **calculation questions** and **lookup questions!**
- Students discuss potential primary sources for their project. (This does not have to be the data they eventually analyze for the project, but they should prep this collection to have the option later.) They consider the following:
    - What question(s) do they want to answer?

This work was made possible through generous support from the National Science Foundation (Award # 2141655). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License

- Who would they like to gather data from?
- What kinds of responses do they want to get back?

Messy Data Activity (15 min)

(CSTA standards in this activity: 3A-AP-21, 3A-DA-10)

- The teacher asks students "What grade are you in?" and students respond to this question via a free-response survey.
  - Do all the responses come up the same?
  - If not, how could we fix this?
  - Plan out one strategy that you could use to fix this problem if you had already collected the data, and one strategy that you could use to fix this problem if you hadn't finished the survey yet.

Representative Samples (10 min)

(CSTA standards in this activity: 3A-DA-09, 3A-IC-25)

- Discuss how a sample is always drawn from a population. Ideally, this sample is representative of the population. Several sampling designs can accomplish this well:
  - **Simple Random Sample (SRS) -** members of the population are randomly selected to be part of the sample. In this sampling design, every set of *n* people from a population of size *N* are equally likely to be selected.
  - **Systematic sampling -** members of the population are put in an ordered list, and then every Xth person is selected to be part of the sample.
  - **Stratified Random Sample -** members of the population are first divided according to some meaningful characteristic, then randomly selected within these divisions. This can help ensure the representation of different groups, like racial groups, age groups, gender, or other factors that the researcher wants to make sure not to miss.
  - **Cluster Sample -** the entirety of a group (which is assumed to be representative of the larger population) is selected. For example, one might use an entire high school's student body as a cluster to represent all the high school students in a state. (The high school is assumed to represent all the different groups that might be represented across the state.)
- Some sampling designs do this poorly:
  - **Convenience sample -** selected based on ease of access, like only asking one's friends to respond to a survey, or giving a survey to the first 100 people to attend a school sports game. This might result in a sample that only represents a highly specific group.
  - **Voluntary response sample -** a study is advertised and participants can choose to participate if they want to share their opinions. This is likely to draw in very

This work was made possible through generous support from the National Science Foundation (Award # 2141655). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License

extreme opinions (ex: Yelp reviews are almost always 5 stars or 1 star – very few "Yeah, it was okay!" reviews!) and few mild opinions, because only those that feel strongly will volunteer to participate.

Impact of Sampling Design (5 min)

(CSTA standards in this activity: 3A-DA-09, 3A-IC-25)

- Samples get drawn for many reasons to attempt to represent a population.
  - When these samples do a **good** job of representing the population, we get an accurate and meaningful picture of what's really true in the population.
  - When these samples are **biased** (usually coming from a bad design!), the sample can misrepresent the population and lead to inaccurate conclusions.
- Food Deserts and Sampling Design:
  - Students consider what "bad sampling design" might look like if we were studying food deserts (we'll do this in the first lesson of Unit 2):
    - What kind of questions might you ask? What would make these questions better or worse?
    - How would you choose the people you were surveying? What could be a potential issue in the sample you choose?
- Racial Profiling as Bad Sampling Design
  - Racial profiling is a discriminatory act of suspecting, targeting, or discriminating against a person based on their ethnicity, nationality, or race rather than on individual suspicion or available evidence
  - How does this connect to the concept of a bad sampling design? What would the consequences of an investigated sample that focuses on people of a particular race or other group?

Google Form Survey Design (25 min)

(CSTA standards in this activity: 3A-AP-13)

- Students design a Google Form to collect data on their local issues.
- Students also design a sampling method for their data collection.

Exit Ticket (10 min)

(CSTA standards in this activity: 3A-DA-09)

- Students name each of the following sampling designs:
  - Every 10th name in the phone book is selected for a political poll phone survey.
  - A student measures the height of a patch of trees in his backyard to estimate the average height of all the trees in his town.
  - Every student at Roosevelt High School is numbered 001 - 797. A DCPS researcher randomly selects 30 numbers within this range and sends a satisfaction survey to the students corresponding with those numbers.

This work was made possible through generous support from the National Science Foundation (Award # 2141655). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License

    - A UMD campus administrator wants to gauge UMD's popularity in the local area, so they interview the first 200 arrivals to a UMD open house before the open house's official start time.
    - A local movie theater puts up a poster with a QR code to a survey that moviegoers can fill out to rate their satisfaction with food & drink prices at the movie theater.
- Key:
    - **Systematic.** Every 10th name is selected.
    - **Cluster.** A particular group, which is assumed to be representative of the broader population, is used.
    - **Simple random sample.** This process uses full random selection (probably with a random number generator or other tool).
    - **Convenience.** The first 200 arrivals are selected (and problematic because they are probably people who are VERY eager and excited to be touring UMD!)
    - **Voluntary response.** Only moviegoers who have strong feelings about the prices (probably negative!) are likely to fill out the survey, so this survey will probably be biased and show a more negative response than how the population actually feels.

Assessment:

Assess student understanding through participation in class discussions and class activities.

This work was made possible through generous support from the National Science Foundation (Award # 2141655). API Can Code is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License