# Illustrative examples to promote and encourage the appropriate use of survey data

Pamela Fellers

Grinnell College

# Goals and Outline

- Encourage intentional incorporation of survey data and survey concepts
- Provide a variety of specific examples and resources that can be easily incorporated into a class

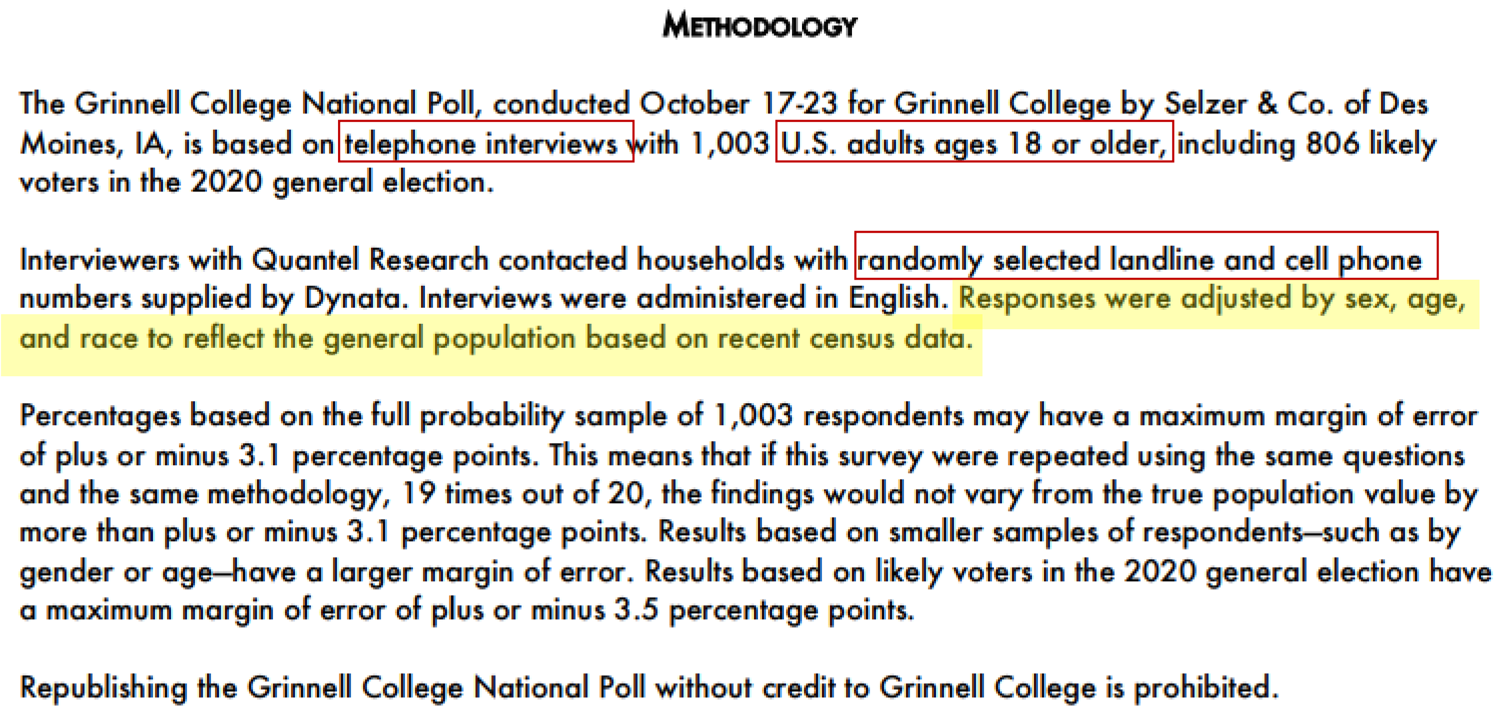
Moving beyond questions of the type:

A survey was conducted about this topic to address this question by collecting these variables. Next comes the analysis questions.

- Taking a look at survey data
  - Grinnell College National Poll
  - Survey Weights
- The Candy Activity
- Survey Data sets
- Utilizing Published Studies
- Advanced Examples

# Grinnell College National Poll

The Grinnell College National Poll (GCNP) pilot project was launched in August of 2018. It is a national public opinion poll designed by Grinnell College faculty and nationally-renowned pollster J. Ann Selzer that probes the political attitudes of the American public on current events, national trends and important policy questions.



**METHODOLOGY**

The Grinnell College National Poll, conducted October 17-23 for Grinnell College by Selzer & Co. of Des Moines, IA, is based on telephone interviews with 1,003 U.S. adults ages 18 or older, including 806 likely voters in the 2020 general election.

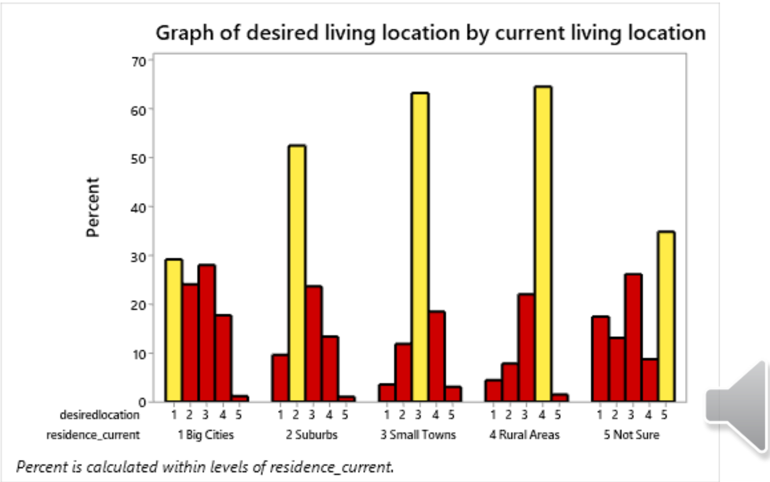
Interviewers with Quantel Research contacted households with randomly selected landline and cell phone numbers supplied by Dynata. Interviews were administered in English. Responses were adjusted by sex, age, and race to reflect the general population based on recent census data.

Percentages based on the full probability sample of 1,003 respondents may have a maximum margin of error of plus or minus 3.1 percentage points. This means that if this survey were repeated using the same questions and the same methodology, 19 times out of 20, the findings would not vary from the true population value by more than plus or minus 3.1 percentage points. Results based on smaller samples of respondents—such as by gender or age—have a larger margin of error. Results based on likely voters in the 2020 general election have a maximum margin of error of plus or minus 3.5 percentage points.

Republishing the Grinnell College National Poll without credit to Grinnell College is prohibited.

Need to look at the codebook

- Consider the variables desiredlocation and residence_current. What are the poll questions corresponding to these two questions? Are these two variables numerical or categorical? Explain.
- Explore the relationship between the two variables desiredlocation and residence_current by creating a two-way table and an appropriate graph. Do they appear to be associated or independent? Briefly explain.

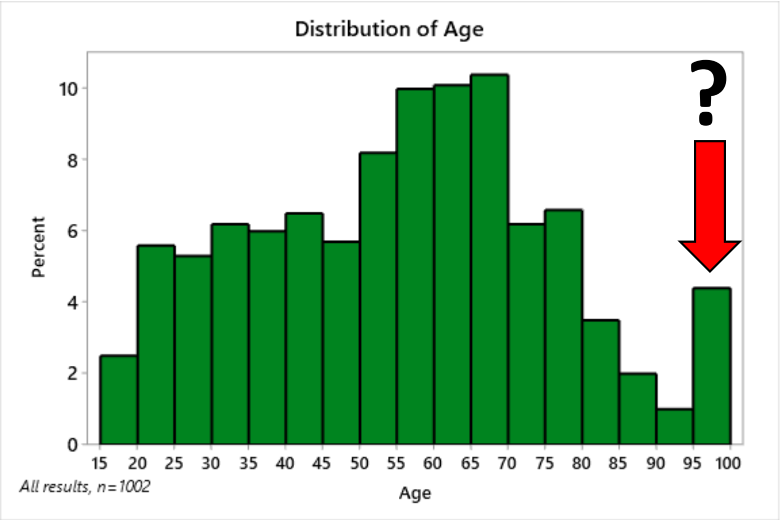


Graph of desired living location by current living location

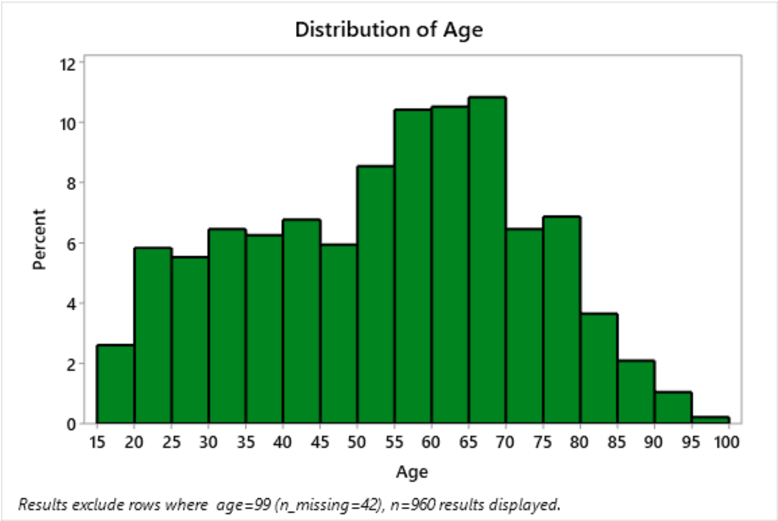
*Percent is calculated within levels of residence_current.*

# Grinnell College National Poll

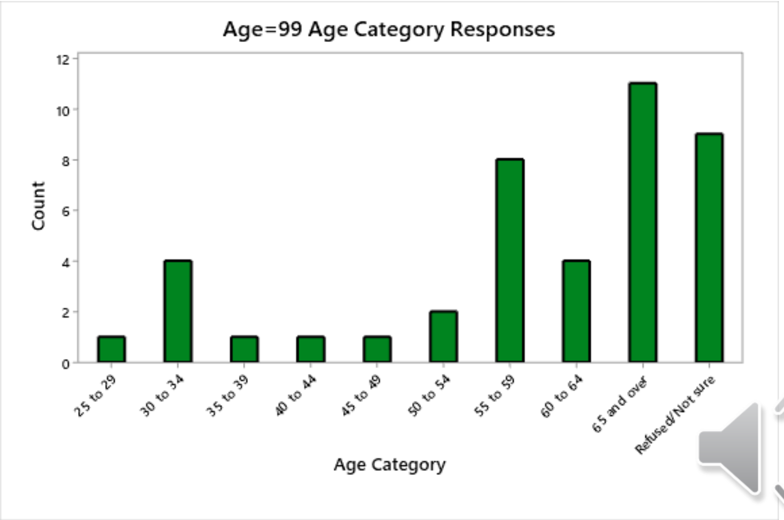
- Create a histogram of the variable age. A feature of this distribution should appear questionable. What feature is of concern?
    - Examine the codebook to explain this feature, how many observations have this response? How should you treat these peculiar observations when creating a histogram of age?
- Create a new histogram of the age variable, properly addressing the concern highlighted in the previous question. Describe this distribution, calculate any numerical summaries that may be helpful in further describing this distribution.


Distribution of Age

*All results, n=1002*


Distribution of Age

*Results exclude rows where age=99 (n_missing=42), n=960 results displayed.*

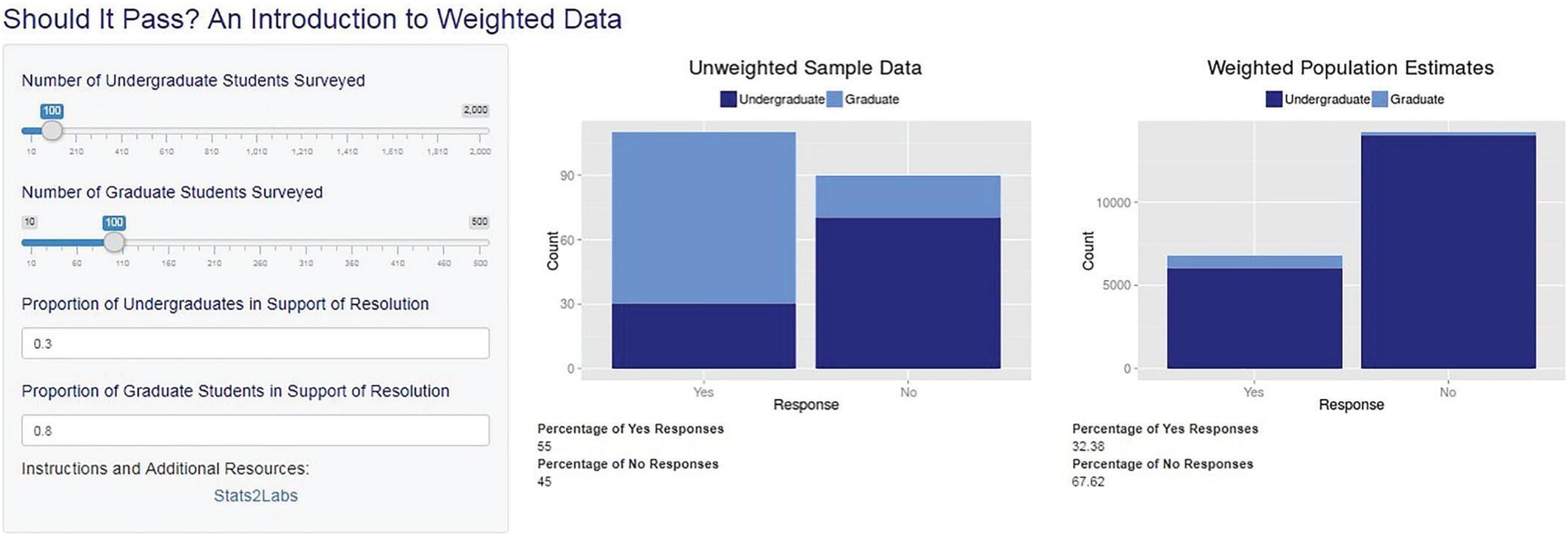

Age=99 Age Category Responses

# Survey Weights

Weighted Data webpage with all handouts, data sets and ShinyApps

- Introductory: Should it pass? and Political Preferences 1 activities
- Introductory or Intermediate: Political Preferences 2, CAM, and NHANES activities

Introducing Undergraduates to Concepts of Survey Data Analysis - JSE paper with more details



Should It Pass? An Introduction to Weighted Data

Number of Undergraduate Students Surveyed: 100

Number of Graduate Students Surveyed: 100

Proportion of Undergraduates in Support of Resolution: 0.3

Proportion of Graduate Students in Support of Resolution: 0.8

Instructions and Additional Resources:

Stats2Labs

Unweighted Sample Data

Percentage of Yes Responses
55
Percentage of No Responses
45

Weighted Population Estimates

Percentage of Yes Responses
32.38
Percentage of No Responses
67.62

# The Candy Activity

**Identify a question or problem:**
How much does this entire bag of candy weigh? Estimate the total weight of all 125 pieces.

**Collect relevant data on the topic:**
As a group select 5 pieces of candy, weigh and record the total weight of your 5 pieces.

**Analyze the data:**
Multiply the total weight of your 5 pieces of candy by 25 to get an estimate of the total weight of all the candy.

**OVER ESTIMATED**

Add the weight of your 5 pieces and your total weight estimate to the class data sheet and board graphs.

Uncertain of origin of this activity but see Chapter 9: Gelman, Andrew, and Deborah Nolan. *Teaching statistics: A bag of tricks*. Oxford University Press, 2017.

## Extending the Activity

**small group discussion**

Consider the situation where instead of 1 bag of candy we had 10 bags - not entirely the same, but of similar content. For each of the three sampling strategies listed, come up with how you could use the sampling strategy to obtain a sample of candy pieces to estimate the total weight. Also give how many samples you would collect (n=#).

- Simple Random Sample
- Cluster Sampling
- Stratified Sampling

What characteristics of our observational units might be important to consider?

# Survey Datasets

- Health Information National trends Survey (HINTS)
- General Social Survey
- National Health and Nutrition Examination Survey (NHANES
  - Also check out the NHANES package in R
- The CDC, Pew Research center, and FiveThirtyEight (as well as many others not listed) all have data available to download.
- Recreated xkdc survey dataset – shared with me by a student as an idea for their final project

Check out their Flying Etiquette Survey

(SurveMonkey Audience Poll)

- Article
- Data

# Utilizing published studies – looking at methodology

**The Scenario:**

Consider the following sampling design quote from the Methods and Materials section of the article “Association between Spouse/Child Separation and Migration-Related Stress among a Random Sample of Rural-to-urban Migrants in Wuhan, China”

- “The participants of this study were recruited in Wuhan, the capital city of Hubei Province. Wuhan has a total population of 10 million and per capita GDP of $12,708 [40]. Rural-to-urban migrant in this study was defined as possessing a legal rural Hukou, 18–45 years old, working or living in city for at least one month by time of the survey day. Participants were selected using a GIS/GPS-assisted stratified random sampling design. First, residential areas in urban Wuhan were divided, on computer, into mutually exclusive geounits of 100 meters by 100 meters as the primary sampling frame (PSF). Altogether 60 geounits were randomly selected from the PSF. Considering cost-effectiveness of the study, relatively more geounits were sampled in districts with a larger number of migrants using the optimal design method[41]. Second, Approximately 20 households were then random sampled from each geounit. Last, one person per household per gender was selected from the sampled households and recruited to participate in the study. To ensure independence, one participant by gender in one household was selected using the random digits method. To ensure adequate sample size, 20% extra geounits were added, considering of absence and refusal. A total of 1414 rural migrants were invited to participate, and 1293 (91.44%) agreed to participate and completed the survey.”

**Possible Questions for in-class, quiz, or homework:**

- Sketch/diagram the sampling process outlined in the information provided from the Methods and Materials Section of the paper.
- The sampling design presented has multiple stages, identify key terms from the information that help in understanding the sampling design that may be the same as we have discussed in class or they may use different terminology.
- The first stage would in general seem to best be described by which sampling method discussed in class (stratified RS, Cluster RS, or SRS)? As appropriate, also describe what represented clusters or strata.
  - About how many of these units were selected?
- The second stage would in general seem to best be described by which sampling method discussed in class (stratified RS, Cluster RS, or SRS)? As appropriate, also describe what represented clusters or strata.
  - About how many of these units were selected?
- How would you describe the final stage of sampling based on the information provided?
- We have discussed non-response and the potential for non-response bias in surveys. How many of the invited participants did not respond, what is the non-response rate?

Link to paper: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154252#sec012 Guo, Yan, Xinguang Chen, Jie Gong, Fang Li, Chaoyang Zhu, Yaqiong Yan, and Liang Wang. "Association between spouse/child separation and migration-related stress among a random sample of rural-to-urban migrants in Wuhan, China." PloS one 11, no. 4 (2016).

# Other methodology statements

USDA APHIS Honey Bee Pests and Diseases Survey Project Plan

- "When sampling an apiary, it is critical to select colonies at random, which is different than haphazard or regularly spaced. Colonies should under no circumstances be preferentially selected because they seem "healthy" or "sickly". To help select colonies as random, we will provide sheets of randomly generated numbers. Instructions on the use of this will be provided with the sampling kit sent to each state."

Survey Scavenger Hunt

Health Information National Trends Survey 5 (HINTS 5)

- The sampling frame of addresses was grouped into two explicit sampling strata:

  1 Addresses in areas with high concentrations of minority population

  2. Addresses in areas with low concentrations of minority population.
- The high and low minority strata were formed using the census tract-level characteristics from the 2013–2017 American Community Survey data file. Addresses in census tracts that had a population proportion of Hispanics or African Americans that equaled or exceeded 34 percent were assigned to the high-minority stratum. All the remaining addresses were assigned to the low-minority stratum.
- The purpose of creating high- and low-minority strata and then oversampling the high-minority stratum is to increase the precision of estimates for minority subpopulations. The gains in precision stem from the increase in sample sizes for the minority subpopulations produced by the oversampling.

# Utilizing published studies – looking at limitations

## Music festival attendees' illicit drug use, knowledge and practices regarding drug content and purity: a cross-sectional survey

Limitations

- The results of our research are based on a convenience sample of music festival attendees, and as such, survey respondents are not likely to be representative of the general population. The predominance of female respondents does not correspond with national data demonstrating that males typically consume more illicit drugs than females [1], potentially limiting the generalizability of this study. The high proportion of females surveyed (60.5%) may also skew the interpretation of the results, given the known differences in drug use prevalence between genders [1].

Day, Niamh, Joshua Criss, Benjamin Griffiths, Shireen Kaur Gujral, Franklin John-Leader, Jennifer Johnston, and Sabrina Pit. "Music festival attendees' illicit drug use, knowledge and practices regarding drug content and purity: a cross-sectional survey." *Harm reduction journal* 15, no. 1 (2018): 1.

## Diversifying the Sports Department and Covering Women's Sports: A Survey of Sports Editors

Interpreting the Findings

- First, these findings must be read within their limitations; most notably, the number of editors who participated…
- For instance, the miniscule number of female sports editors (6) in the sample made it impossible to consider important questions about the differences that may exist in the ways male and female sports editors may understand issues around coverage and hiring.
- The cyclical, seasonal nature of the sports calendar is also a factor in any survey that asks editors to think about content, as they likely think about what they are dealing with in the moment to answer larger questions, which could skew their responses. This survey was administered in the late spring and summer 2012, just before the Olympics.

Laucella, Pamela C., Marie Hardin, Steve Bien-Aimé, and Dunja Antunovic. "Diversifying the sports department and covering women's sports: A survey of sports editors." *Journalism & Mass Communication Quarterly* 94, no. 3 (2017): 772-792.

# Advanced Examples

- Estimation of a Proportion with Survey Data
    - Logistic regression estimator, utilizing auxiliary information
- Weighted Data webpage with all handouts, data sets and ShinyApps
    - Introductory or Intermediate: Political Preferences 2, CAM, and NHANES activities
    - Advanced Supplements
    - Introducing Undergraduates to Concepts of Survey Data Analysis - JSE paper with more details
- Incorporate the use of simulations to explore survey concepts

# Thank You!

Pamela Fellers
fellerspam@grinnell.edu
Grinnell College

Just to clarify: I don't only use survey examples in my classes, I also use a variety of study types and have started to think about the continued discussion of the data collection methods in general.