SCRP 2024
Survey Quality Research Lab
Specific Application

Information:
- Heather Kitada Smalley, PhD
  - Assistant Professor of Statistics and Data Science
- Office: Ford 216
- Email: hsmalley@willamette.edu

The Kitada-Smalley research group focuses on public opinion and survey statistics. The overarching goal of this group is to develop survey methodology to improve bias estimation, data quality, and the representation of historically undercounted populations. This summer, students in this group will focus on one of the most important surveys conducted in the United States, the American Community Survey. Students will be involved in all parts of the data science life cycle from data collection, cleaning, visualization, analysis, and reporting.

Project Description:
The American Community Survey (ACS) is conducted annually by the United States Census Bureau and is widely used by academicians and policy makers alike. The ACS tracks several metrics of data quality, such as unit non-response rates and allocation rates, which can be considered a proxy for item non-response rates. It has been shown that aggregated allocation rates vary across household language groups and lower levels of English speaking proficiency. This can have detrimental effects to data quality in those communities resulting in biased estimates. Given the applications and use of the data, it is imperative that they be representative of the population for an equitable allocation of resources. However, allocation rates are not the only point of concern, survey break-off may provide indications of significant respondent burden. Thus, we implement survival analysis methods to analyze the interaction between survey break-off across household language and level of English proficiency. The ACS employs multiple modes of data collection in sequence, which result in differential representation across modes, contributing an additional layer of complexity to the analysis.

Application Questions:
1. In one short paragraph, explain what most interests you about this project.

2. Briefly (one paragraph) describe your career interest or long-term goals post-Willamette.

3. In 1-2 paragraphs, describe how you see taking part in SCRP as benefiting your personal or academic development.

4. Academic preparation:
   a. **Statistics/Probability**:
      i. Have you taken any of the following statistics/probability courses?

DATA 152	MATH 138	IDS 138	AP Statistics
PSYC 252W	PSYC 253	SOC 341	DATA 280
DATA 252	MATH 239	MATH 266	MATH 376

   Please describe any of the course work or experience you have done in statistics.

      ii. **Content questions:** Please answer the following questions to the best of your ability in your own (non-technical) words
         1. What is a p-value?
         2. What type of errors can result from hypothesis testing?
         3. What are the assumptions of the linear regression model?
         4. What sources of error/bias are present in sampling and survey research?
         5. Explain the importance of representation when sampling.

   b. **Programming:**
      i. Have you taken any of the following courses?

DATA 151 R	CS 151 Python	DATA 370 Python	DATA 352 SQL

   Please describe any of the course work or experience you have with programming (in any language). What languages do you have experience programming in?

5. **Data Projects**
   a. Tell me about a time you did a data project. This could be for a class or just for your own fun!
   b. Tell me about the largest dataset you have worked with. How many rows did it have? How many variables?
   c. Tell me about a time you worked on a group project. How did that go? Please tell me about your communication and distribution of work. What went well? What could you have done better?
   d. In addition to working in a group, there will be time for you to work independently and unsupervised. Tell me about a time when you were self-motivated to reach a goal.

6. **Data Ethics**
    a. Since we are working with large datasets with observations from human beings, the ethical treatment of these data is imperative. When aggregating and modeling data we may forget the significance that it holds to an individual's lived experience.
    b. Please read the article **"*The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses*"** by William Seltzer and Margo Anderson. You can access this article from your Willamette library account on JSTOR (https://www.jstor.org/stable/40971467)
    c. After reading the article please write a one paragraph summary of the most salient points and how it resonates with you.