Stat 222 Project 1: Randomized Response Survey on Cognition Enhancing Drugs

1 Introduction

In this project we'll analyze data obtained using a *randomized response* survey methodology. This strategy protects respondent's answers to sensitive questions by using a random mechanism to determine whether the respondent answers truthfully.

For example, suppose I'm wondering if you've ever cheated on an assignment, but I don't think you'd tell me if you had. I ask you to flip a coin and not show it to me, and then I tell you to answer me truthfully if the coin came up heads, but just to say "yes" if the coin came up tails. In this case, your particular "yes" response doesn't tell me whether you've cheated or not. But if I did the same thing with everyone in the class and substantially more than 50% of you said yes, I could reasonably infer that a non-zero percentage of you had cheated. Our focus will be on precisely *how* to make this kind of inference.

For more background and examples, see the paper by Bolstad et al. included with this project, or do your own research on the term "randomized response."

2 Data Description

Students in an upper-division undergraduate statistics course at UCB designed a randomized-response survey, included as `survey.pdf`. Students were instructed to answer the survey themselves, as well as administering it to at least two other UCB students. A handful of students administered more than two, including a member of a fraternity who turned in 23 surveys. The data are contained in the file `rrsurveydata.csv`, with `NA` indicating missing data.

3 Questions of Interest

The substantive questions we'll focus on with this data are:

1. What proportion of students have ever used cognition-enhancing drugs? For this question, group the "Yes, with a prescription" and "Yes, without a prescription" answers together.

2. Same as question 1, but reporting proportions for the three groups ("No", "Yes, with a prescription" and "Yes, without a prescription") separately.

3. What proportion of students with GPA ≥ 3.5 have ever used cognition-enhancing drugs? With GPA < 3.5?

Note 1: Some of you may end up diving deeply into the first question and not getting to the other questions, which is ok.

Note 2: What we mean by "students" needs to be clarified. What population are we talking about? More on this below.

4 Initial Guidance (to do for first data debrief)

Download the data and take a quick look, just so you understand the format and can spot any problems.

Before you do any work with the real data, I want you to *think about strategies* for answering the first question above. Here is some notation to get you started, which you can build upon for subsequent questions.

Define

$$X_i = \begin{cases} 1 & \text{Person i used cognition-enhancing drugs.} \\ 0 & \text{Person i did not use cognition-enhancing drugs,} \end{cases}$$

$$Z_{i1} = \begin{cases} 1 & \text{The red die for Q1 on the survey was 3,4,5, or 6 for person i.} \\ 0 & \text{The red die for Q1 on the survey was 1 or 2 for person i.} \end{cases}$$

$$Z_{i2} = \begin{cases} 1 & \text{The green die for Q1 on the survey was 3,4,5, or 6 for person i.} \\ 0 & \text{The green die for Q1 on the survey was 1 or 2 for person i.} \end{cases}$$

$$Y_i = \begin{cases} 1 & \text{Person i answered 2 or 3 to Q1 on the survey.} \\ 0 & \text{Person i answered 1 to Q1 on the survey.} \end{cases}$$

Note that we observe only the Y_i values; all the other variables are unobserved.

- Derive the distribution for Y_i conditional on X_i.

- Suppose the students in the survey had been drawn at random from a larger population in which the proportion of students who had used study drugs was θ. (This was not actually the case here, a point we'll come back to.) Then we could model $X_1, \ldots, X_n \overset{iid}{\sim} Ber(\theta)$. Under this model, what is the unconditional distribution for $Y_1, \ldots, Y_n$ (that is, conditional only on θ, not the X values)?

- Derive the MLE for θ. (Be careful to take note of the valid parameter space.)
- Think about *at least two* strategies for approximating the sampling distribution of the MLE. Write down either a mathematical expression or an algorithm. Consider whether n being small and/or θ being close to 0 or 1 presents a problem for each strategy you're considering. (To answer this question, it would be helpful to simulate some data and try it out.)
- Now find the MLE for θ using the real data, and implement your strategies for estimating the sampling distribution. Visualize and compare them.

5 Looking Ahead

Here are some more things to consider. (I don't expect you to answer all of these questions, nor do I think these are the only things to think about. I want you to engage critically with this data and am sharing some starting points that come to mind.)

- What would a Bayesian approach to the first question look like? How could we calculate/approximate the posterior distribution?
- What changes when we introduce a third category?
- What if we don't want to make the assumption that the survey participants are drawn from a larger population and instead want to make inference only about the proportion *among the survey participants*, e.g. $\bar{X}$ as defined for question 1?
- How could we use simulation to compare the performance of various estimators, confidence intervals, etc.?
- How would we set up a probability model for the third question, which combines two different survey questions and thus involves two sets of latent variables? (*Hint: The latent variables will be dependent across questions.*)
- Is there any other data you would like to have based on what I've told you? (I'm being a bit coy here and acting like a collaborator who might give you what he/she thinks is important, while letting slip something else about the data that you really should follow up on.)
- There's some missing data. How does it impact what you're doing?
- Have you intuited anything about designing these kinds of surveys? How should we reason about designing the random mechanism by which someone answers truthfully or not?