Data Collection & Sampling

Instructions: Please work with the same group that you worked with for the variable activity to tackle the three problems below! Please do not use your textbook or the internet.

1. A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments, and others a diverse mixture of housing structures. For each part below, identify the sampling methods described, and describe the statistical pros and cons of the method in the city's context.

   (a) Randomly sample 200 households from the city.

   This is an example of a *simple random sample*. This may be a reasonable approach: on balance, a simple random sample should produce a representative sample from the population of interest. However, there is always the chance that, just by (bad) random luck, we end up with a random sample that overrepresents some neighborhoods or underrepresents others.

   (b) Divide the city into 20 neighborhoods, and sample 10 households from each neighborhood.

   This is an example of a *stratified random sample*, where the stratification factor is the neighborhood. This approach will ensure that our sample provides better representation of all the neighborhoods in the suburban area of the city, which—depending on the purpose behind the household survey—may be important to us. However, this sort of approach is more complex and costly to implement, and so may not be ideal for the city council if time, cost, or logistics are factors.

   (c) Divide the city into 20 neighborhoods, randomly sample 8 neighborhoods, and then randomly sample 50 households from those neighborhoods.

   This is an example of a *multistage cluster sampling scheme*, where each neighborhood is now being treated as a "cluster". This sort of approach may be logistically simpler to implement than the stratified sampling approach in (b): the survey administrators would, for example, only need to travel to eight of the neighborhoods in order to conduct their survey instead of traveling to all 20. That being said, this particular context is not ideal for a cluster sampling scheme. We are told that the city's neighborhoods are all quite distinct from one another, with different zoning, different house sizes, and different housing structures. The proposed multistage cluster sampling approach would not guarantee that we end up with a subset of eight neighborhoods that reasonably reflect/represent this diversity. Furthermore, cluster sampling is best adapted to settings where all of our clusters look similar to one another (but there is large within-cluster variability), which does not appear to be the case for the neighborhoods in the suburban area of this particular city.

   (d) Sample the 200 households closest to the city council offices.

   This is an example of a *convenience sample*, wherein observations (i.e., households) that are easier to sample are more likely to be included in the final sample. This approach is definitely easy to implement—the city councilors administrating the survey certainly wouldn't need to walk very far!—but is not a random (probability) sample. As such, we would not expect the resulting convenience sample to be representative of our population of interest (all households in the suburban area of the city) and so would not want to rely on it to make statistical inferences.

2. In statistics, *missing data* refers to the scenario when—for whatever reason—data are not available for a particular variable or a particular observational unit. Consider, for example, a hypothetical study examining the relationship between self-reported marijuana use and blood cortisol levels (measured in mcg/dL) in a simple random sample of American college students. The dataframe below shows the available data for the first six study participants; variables with missing values are encoded as `NA`.

ID	marijuana_use	cortisol_levels
1	No	NA
2	No	21
3	NA	6
4	Yes	12
5	NA	23
6	Yes	NA

   (a) Blood cortisol levels are missing because the lab running the analysis accidentally dropped some of the test tubes (oops!). Do you think that the individuals with recorded (i.e., non-missing) cortisol levels still represent a random sample of all American college students? Why or why not?

   If the only reason that the blood cortisol levels are missing is due to an unrelated event (in this case, the lab dropping some of the test tubes), then yes, the individuals with recorded cortisol levels are likely still a random sample of all American college students: a random sample of a random sample is still itself a random sample! In other words, if the data are *missing completely at random*, the complete data can still be generalized to the broader target population.

   (b) Information on marijuana use was collected using an online survey; it's missing if participants either skipped or refused to answer that question. Do you think that the individuals with recorded (i.e., non-missing) information on marijuana use still represent a random sample of all American college students? Why or why not?

   No, the individuals with recorded information on marijuana use are likely not a random sample of all American college students. Marijuana use has not been legalized/decriminalized in all states, so it may be that—due to social pressures or a sense that certain responses are more "desirable" or safer than others—those who use marijuana are more likely to skip the question. So the data are *missing not at random*, and a sample of people with complete marijuana use information might underrepresent marijuana use relative to the broader population.

3. A school district is considering whether it will no longer allow high school students to park at school after two recent accidents where students were severely injured. As a first step, they survey parents by mail, asking them whether or not the parents would object to this policy change. Of 6,000 surveys that go out, 1,200 are returned. Of these 1,200 surveys that were completed, 960 agreed with the policy change and 240 disagreed. Which of the following statements are true? Why?

   (a) Some of the mailings may have never reached the parents.

   (b) The school district has strong support from parents to move forward with the policy approval.

   (c) It is possible that the majority of the parents of high school students disagree with the policy change.

   (d) The survey results are unlikely to be biased because all parents were mailed a survey.

   (a): True. Although 6,000 surveys were sent to parents/caregivers of students at the high school, only 1,200 surveys were returned. The fairly high non-response rate suggests that some caregivers elected not to (or forgot to) complete and return the survey, while others still may not have received the mailing at all.

(b): False. While it's entirely possible that the school district has strong support for their proposed policy, the information that we have at our disposal is not sufficient to make that claim. In particular, the sample we collected is likely not a random sample of all caregivers, as the observed levels of non-response were high. Caregivers who felt strongly about the proposed policy change (whether positive or negative) may have been more likely to fill out and return the survey. The "best case scenario" for caregiver support would be if all the survey non-responders were in favor of the proposed policy change; in that case, 5760 of the 6000 caregivers (96%) would support the policy and the claim of "strong support" would be reasonable. Under the "worst case scenario", however, all the survey non-responders would be *against* the proposed policy change, in which case the caregiver support would be at only 16%.

(c) True. See reasoning above.

(d) False. Although all the caregivers were *mailed* a survey, not all caregivers *returned* the survey. We can only collect and measure variables for those caregivers who actually returned the survey, so our data are only able to reflect the responders, not the non-responders.