SDS 220 - Lecture 2 Handout

1. Take a small survey of your classmates and ask the following questions: *What is your name? What year are you? In which house do you live? What is your hometown? How many siblings do you have? What is the furthest away from Northampton (in miles) you were over the summer?*

   (a) Fill out the table below.

   (b) For each variable, describe the type of variable that it is (e.g., categorical/numerical, discrete/continuous, ordinal, etc.)

   (c) Describe your sampling strategy. What type of sample did you do (simple, convenient, stratified, cluster, etc.)

   (a) Varies. (b) Name - Categorical. Class Year - Orindal. House - Categorical. Siblings - discrete. Distance - Continuous. (c)Varies.

Name	**Class Year**	**House**	**Hometown**	**# of Siblings**	**Distance over break**

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'.

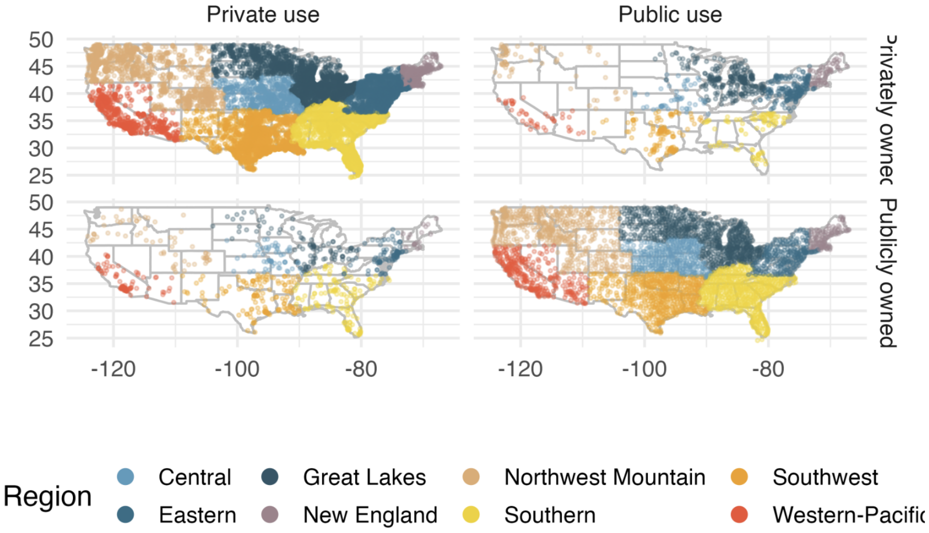
	marijuana_use	cortisol_levels
1	No	NA
2	No	21.00
3	NA	6.00
4	Yes	12.00
5	NA	23.00
6	Yes	NA

   (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?

   (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?

(a) 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) 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. **[IMS 1.13] US Airports.** The visualization below shows the geographical distribution of airports in the contiguous United States and Washington, DC. This visualization was constructed based on a dataset where each observation is an airport.



(a) List the variables you believe were necessary to create this visualization.

(b) Indicate whether each variable in the study is numerical or categorical. If numerical, identify as continuous or discrete. If categorical, indicate if the variable is ordinal.

(a) Airport ownership status (public/private), airport usage status (public/private), region (Central, Eastern, Great Lakes, New England, Northwest Mountain, Southern, Southwest, Western Pacific), latitude, and longitude. (b) Airport ownership status: categorical, not ordinal. Airport usage status: categorical, not ordinal. Region: categorical, not ordinal. Latitude: numerical, continuous. Longitude: numerical, continuous.

4. **[IMS 2.15] Haters are gonna hate, study confirms.** A study published in the Journal of Personality and Social Psychology asked a group of 200 randomly sampled participants recruited online using Amazon's Mechanical Turk to evaluate how they felt about various subjects, such as camping, health care, architecture, taxidermy, crossword puzzles, and Japan in order to measure their attitude towards mostly independent stimuli. Then, they presented the participants with information about a new product: a microwave oven. This microwave oven does not exist, but the participants didn't know this, and were given three positive and three negative fake reviews. People who reacted positively to the subjects on the dispositional attitude measurement also tended to react positively to the microwave oven, and those who reacted negatively tended to react negatively to it. Researchers concluded that "some people tend to like things, whereas others tend to dislike things, and a more thorough understanding of this tendency will lead to a more thorough understanding of the psychology of attitudes."

(a) What are the cases?

(b) What is (are) the response variable(s) in this study?

(c) What is (are) the explanatory variable(s) in this study?

(d) Does the study employ random sampling? Explain. How could they have obtained participants?

(e) Is this an observational study or an experiment? Explain your reasoning.

(f) Can we establish a causal link between the explanatory and response variables?

(g) Can the results of the study be generalized to the population at large?

(a) The cases are 200 randomly sampled men and women. (b) The response variable is attitude towards a fictional microwave oven. (c) The explanatory variable is dispositional attitude. (d) Yes, the cases are sampled randomly, recruited online using Amazon's Mechanical Turk. (e) This is an observational study since there is no random assignment to treatments. (f) No, we cannot establish a causal link between the explanatory and response variables since the study is observational. (g) Yes, the results of the study can be generalized to the population at large since the sample is random.

5. **[IMS 2.17] Sampling strategies.** A statistics student who is curious about the relationship between the amount of time students spend on social networking sites and their performance at school decides to conduct a survey. Various research strategies for collecting data are described below. In each, name the sampling method proposed and any bias you might expect.

   (a) They randomly sample 40 students from the study's population, give them the survey, ask them to fill it out and bring it back the next day.

   (b) They give out the survey only to their friends, making sure each one of them fills out the survey.

   (c) They post a link to an online survey on Facebook and ask their friends to fill out the survey.

   (a) Simple random sample. Non-response bias, if only those people who have strong opinions about the survey responds their sample may not be representative of the population. (b) Convenience sample. Under coverage bias, their sample may not be representative of the population since it consists only of their friends. It is also possible that the study will have non-response bias if some choose to not bring back the survey. (c) Convenience sample. This will have a similar issues to handing out surveys to friends.