SDS 220 - Lecture 3 Handout

1. 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.

2. **[IMS 1.7] Migraine and acupuncture.** A migraine is a particularly painful type of headache, which patients sometimes wish to treat with acupuncture. To determine whether acupuncture relieves migraine pain, researchers conducted a randomized controlled study where 89 individuals who identified as female diagnosed with migraine headaches were randomly assigned to one of two groups: treatment or control. Forty-three (43) patients in the treatment group received acupuncture that is specifically designed to treat migraines. Forty-six (46) patients in the control group received placebo acupuncture (needle insertion at non-acupoint locations). Twenty-four (24) hours after patients received acupuncture, they were asked if they were pain free. Results are summarized in the contingency table below. Also provided is a figure from the original paper displaying the appropriate area (M) versus the inappropriate area (S) used in the treatment of migraine attacks.

Group/Pain	**Yes**	**No**
Control	44	2
Treatment	33	10

   (a) What percent of patients in the treatment group were pain free 24 hours after receiving acupuncture?

   (b) What percent were pain free in the control group?

   (c) In which group did a higher percent of patients become pain free 24 hours after receiving acupuncture?

   (d) Your findings so far might suggest that acupuncture is an effective treatment for migraines for all people who suffer from migraines. However, this is not the only possible conclusion. What is one other possible explanation for the observed difference between the percentages of patients that are pain free 24 hours after receiving acupuncture in the two groups?

(e) What are the explanatory and response variables in this study?

(a) Treatment: 10/43=0.23 (b) Control: 2/46=0.04 (c) A higher percentage of patients in the treatment group were pain free 24 hours after receiving acupuncture. (d) It is possible that the observed difference between the two group percentages is due to chance. (e) Explanatory: acupuncture or not. Response: if the patient was pain free or not.

3. **[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.

4. A study is designed to test the effect of light level on exam performance of students. The researcher believes that light levels might have different effects on people who wear glasses and people who don't, so they want to make sure both groups of people are equally represented in each treatment. The treatments are fluorescent overhead lighting, yellow overhead lighting, and no overhead lighting (only desk lamps).

   (a) What is the response variable?

   (b) What is the explanatory variable? What are its levels?

   (c) What is the blocking variable? What are its levels?

   (a)The response variable is a student's exam performance. (b) The explanatory variable is the level of light in the room where students are taking the exam (aka our treatment). This is a nominal categorical variable taking on three levels: fluorescent overhead lighting, yellow overhead lighting, and no overhead lighting. (c) The blocking variable is whether or not a student wears glasses, which is a nominal categorical variable. It has two levels: a student either (1) wears glasses or (2) doesn't.