EXAMPLES FOR PART 1: CONCEPTS AND LOGIC (CLOSED-NOTE)

Note: my example responses are longer than we'd expect from you. I just wanted to cover the possible responses thoroughly.

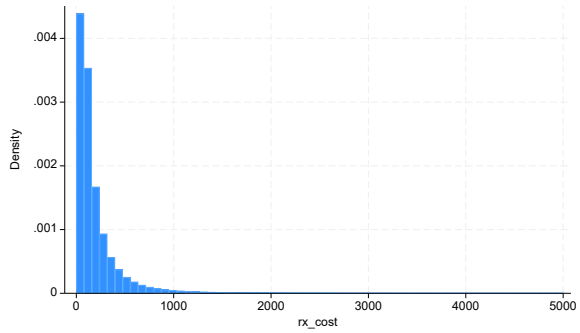
1. You are starting a research project on **whether having a baby leads to fewer hours of sleep**. There are two national data sets that ask questions about sleep habits and family size that you are considering using for this project:

> (1) National Longitudinal Survey of Youth 97: A longitudinal survey of a cohort of people who were teenagers in 1997, surveying those respondents annually from 1997 through 2025.
>
> (2) NHIS: A cross-sectional survey of a sample of Americans that has been conducted repeatedly from the 1950s to 2025.

Just based on this information, which of these two datasets is likely to be better suited for your research question and why?

The research question is trying to understand changes in sleep habits after a baby is born. Ideally, we'd find a data set that asks the same sleep habit questions repeatedly of the same people over time, some of whom have a baby during the follow-up period. Therefore, the National Longitudinal Survey of Youth 97 is better suited to our question, since we can do a "within person" or "individual fixed effects" analysis that estimates changes in sleep for the same person after having a baby. With cross-sectional data, all we can do is test whether people who have a baby get less sleep than a different group of people who don't have a baby, but there are probably many factors that can be confounders of that relationship.

2. We are analyzing data from a nationally representative large-scale survey of 100,000 people that asks them about their out-of-pocket health care spending. You notice that the distribution of out-of-pocket prescription medication spending is very right-skewed (see histogram below). Your supervisor suggests that you could try analyzing "logged spending". **What does your supervisor mean by this, and what steps would you take create a variable of "logged spending"?**



Spending data is often very right skewed, which means that the majority of the data is bunched at lower values with a very long tail on the right hand side (high-value outliers). We may want to convert it to behave like a normally distributed variable for more sensible interpretation of regression results. To do so, we can take the log of it, or log(rx_cost). This will transform it to a normal distribution. Practically

speaking, we will first add a small amount to anyone with $0 spending (because log(0) is undefined). Then we would generate a new variable that takes the log of that value. We can then use this logged value as we would any other continuous variable.

3. We want to use regression analyses to test whether there are differences in out-of-pocket prescription medication spending across different demographic groups. You decide to use survey weights in these analyses. **In a few sentences, explain to your colleagues why you included survey weights in your analysis.**

Survey weights are used to make analyses from survey data representative of the broader population (e.g., the whole US population). Because survey responses are not perfectly random due to design-based over-sampling of certain groups, low response rates in some groups, or just bad luck (a non-representative random sample was drawn), the results need to be re-weighted so that estimates reflect the broader population. A survey weight can be thought of as how many people each respondent counts as: a survey weight of 1.5 would suggest that the respondent counts as about 1.5 people.

4. You work at a pharmaceutical company researching treatments for infectious diseases. You have a data set of 7 infectious diseases your company works on, with one row per disease. It lists key facts about each disease in the columns. You obtain a second data set with thousands of rows listing all of the other pharmaceutical companies who also make treatments for hundreds of infectious diseases. (Examples below.) You want to combine these two data sets so you can identify the competing products. Briefly explain the steps you'd take to combine these data sets and your rationale.

YOUR COMPANY			
Disease	Cases per year (millions)	Fatality rate	Global priority level (1-5 scale)
Dengue fever	400	0.1%	4
Malaria	250	0.3%	5
Bacterial meningitis	2.5	15.0%	4
RSV	33	0.4%	3
Hepatitis B	1	18.0%	4
Chlamydia	150	0.0%	1
E. Coli	55	2%	3

COMPETITOR DATA (8 of 100,000 rows)			
Company	Disease	Market share	Price per dose
ABC Pharma	Bacterial meningitis	18%	$42
Boston Drugs	Hepatitis B	22%	$29
Boston Drugs	Bacterial meningitis	14%	$55
MedsUSA	COVID-19	76%	$90
RX123	E. Coli	12%	$7
RX123	Hepatitis B	15%	$30
RX123	Chlamydia	11%	$3
RX123	RSV	30%	$200

I would do a merge to combine these data sets – specifically, a 1:m merge – using the "disease" field as the merging variable. This is because each row in the first data set is unique by disease, while there are many rows per disease in the second data set (in the sample data, we can see this with Bacterial meningitis and Hepatitis B, which each have 2 rows). After merging, I would drop any records from the second data set that didn't match (e.g., COVID-19 treatment), because I don't care about those since my company doesn't work on treatments for that condition.