Exercise Sheet: Foundations of Data Collection

Data, Data Storage, Data Collection

Exercise 1 - Quantitative vs Qualitative Data

For each of the following, indicate whether the data is quantitative or qualitative:

(**1**) Hours of sleep students gets the night before the examen.
(**2**) Transcripts of customer service calls.
(**3**) Number of steps recorded daily by a fitness tracker.
(**4**) Photos taken by wildlife cameras in a forest.
(**5**) Results of a multiple-choice survey about favorite foods.
(**6**) Scores from standardized math tests.
(**7**) Zip codes of shipment addresses.
(**8**) Ratings (1-5 stars) given to movies on a review site.

Quantitative:	1, 3, 6, 8.	⚠ Ordinals might still be quantitative
Qualitative:	2, 4, 5, 7.	⚠ Numbers might still be qualitative

Exercise 2 - Scope of Data

The California Environmental Protection Agency (CalEPA), the Office of Environmental Health Hazard Assessment (OEHHA) developed the CalEnviroScreen project to study how environmental hazards relate to individual health in California.

The project studies connections between population health and environmental pollution using:

- Demographic summaries from the US Census
- Health statistics from the California Department of Health Care Access and Information
- Pollution measurements from air monitoring stations maintained by the California Air Resources Board

Data are available at the level of census tracts (aggregation over geographic areas), not individuals. For example, one can examine rates of asthma hospitalizations vs. air quality across tracts, but not the impact on any specific individual.

(**1**) What is the target population for the original research question?
(**2**) What is the access frame used in this study?
(**3**) What is the sample in this context?
(**4**) Why does aggregation at the census tract level limit the conclusions that can be drawn?

(**1**) Remember that the target population is the group about which the study is trying to drawn conclusion. In this case, all individuals living in California. This means that ideally, the study would collect data on every individual in California to draw the most accurate conclusions. Also note that while the study uses pollution data to understand how environmental factors relate to the health of individuals, the pollution data itself is not part of the population being studied.

(**2**) The set of all census tracts in California. Each tract groups residents living in the same geographic area. While the U.S. Census collects data for the entire country, the CalEnviroScreen project only needs the tracts in California because these are the units (and geographic area) of interest. Since we know a priori that only a sub-part of the US census data is needed, we can filter them out before the collection. Additionally, the health statistics only over California, meaning that we cannot draw conclusions outside of California. The access frame is thus restricted to the overlap (the intersection) of both dataset.

(**3**) Here, the sample meets the access frame since data is available for the full access frame and there is no need to restrict to a smaller subset.

(**4**) Data are only available at the tract level, so individual-level effects cannot be studied. One can examine associations at the community level (e.g., asthma rates vs. air quality) but cannot make claims about specific individuals. For example, not everyone in a high-pollution tract may be equally exposed or affected. Tract-level data also masks the differences within each tract. For instance, a tract might have both wealthy and low-income neighborhoods, but the aggregated data will not show how pollution or health outcomes vary between these groups. This aggregation is necessary for privacy and practicality, but it means the study can only identify community-level patterns, not individual risks.