Exercise Sheet: Foundations of Data Collection

Data, Data Storage, Data Collection

Exercise 1 – Understanding a Data Source

A university wants to understand how students experience commuting to campus. Consider the following datasets.

- **A**: A questionnaire created by the university asks students about their travel time, number of transfers, and difficulties encountered during their commute.
- **B**: The university obtains anonymized travel-card records from the local transport authority. The records contain timestamps and stations used by students carrying a university transport pass.
- **C**: Students post comments about their commute on a public social-media platform as part of their ordinary activity. The university collects these existing posts to study students' experiences.
- **D**: As part of the study, a group of students are asked to use a campus mobility app for navigation during their commute. The app records their GPS position every minute.

For each dataset, answer the following questions:

(**Q1**) Who or what is the source of the data?
(**Q2**) Was the data collected specifically for this investigation (primary) or was it collected previously for another purpose (secondary)? Was it designed to collect data, or is it an organic byproduct of another activity?
(**Q3**) Identify one piece of information about the dataset's provenance that you would want to check before using it, and explain why it matters.

The exercise shows that each source has its own strengths and weaknesses. They all provide particular observations about commuting, but none of them is the phenomenon itself. The university may need to combine all these sources to better understand students' commuting experiences.

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.

(**Q1**) What is the target population for the original research question?
(**Q2**) What is the access frame used in this study?
(**Q3**) What is the sample in this context?
(**Q4**) Why does aggregation at the census tract level limit the conclusions that can be drawn?

Exercise 3 – Sampling: Misalignment and Design

A university wants to understand the performance of its students in a compulsory course. There are 1200 students in total, belonging to four faculties: Computer Science, Mathematics, Engineering, and Business. At the end of the semester, the university obtains the following information:

- Student ID (unique identifier)
- Faculty (Science, Humanities, Engineering, Business)
- Exam Score (score at the exam: 0–20)

Due to some late grading and delay in transcription, the initial dataset is incomplete and excludes 100 students. The university would still like to use the available data to obtain an initial estimate of the students' performance. Later, the university adds the missing students to the dataset and wants to understand how the initial estimate was affected by the missing data.

A – Sampling and Scope Misalignment

The average exam score for the 1100 students present in the dataset is $\bar{x}_1 = 12$ with standard deviation $s_1 = 3$. For the lately graded 100 students, the average score is $\bar{x}_2 = 15$ with standard deviation $s_2 = 2$.

(**Q1**) Compute the average exam score μ of all 1200 students.
(**Q2**) The original dataset suggested an average score of 12. What does the adjusted value tell us about the effect of the missing students?

Computing the adjusted average from the initial one is quite straightforward, but what about the standard deviation? The result is an application of the law of total variance, which describes how the variance of a population can be decomposed into the variance within subgroups and the variance between subgroup means. Let us prove it for a discrete, finite population divided into two subgroups. Consider a population of N divided into two subgroups:

- Subgroup 1: n_1 individuals with average $\bar{x}_1$ and variance s_1^2.
- Subgroup 2: n_2 individuals with average $\bar{x}_2$ and variance s_2^2.

The overall population mean is μ, and the overall population variance is σ^2.

(**Q3**) Recall the formula for the variance of a finite population of N observations:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

Write the expression for the overall variance by separating the observations belonging to the two subgroups.
(**Q4**) For each subgroup, add and subtract its subgroup average ($\bar{x}_1$ or $\bar{x}_2$) inside the squared term. Rewrite the expression as

$$x_{ji} - \mu = (x_{ji} - \bar{x}_j) + (\bar{x}_j - \mu).$$

(**Q5**) Expand the two squared terms. Identify the terms corresponding to:

(a) the variance within each subgroup (i.e., $(x_i - \bar{x}_1)^2$ and $(x_i - \bar{x}_2)^2$);
(b) the difference between each subgroup mean and the overall mean;
(c) the cross terms involving $(x_{1i} - \bar{x}_1)(\bar{x}_1 - \mu)$ and $(x_{2i} - \bar{x}_2)(\bar{x}_2 - \mu)$.

(**Q6**) Show that the cross terms are zero.

(**Q7**) Deduce that the overall variance can be written as

$$\sigma^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{x}_1 - \mu)^2 + n_2(\bar{x}_2 - \mu)^2}{n_1 + n_2}.$$

This is the law of total variance for two subgroups. Interpret each term in this expression.

(**Q8**) Conclude that

$$\sigma^2 \geq \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}.$$

Is the inequality tight? When does it become an equality? What do you learn from that?

(**Q9**) Apply this inequality to the exam-score data. What lower bound can you obtain for the variance of the scores of all 1200 students?

B – Designing a Stratified Sample

To further understand the performance of the students, the university decides to survey a sample of 60 students about their study habits. Knowing that the study habits may differ between faculties, the university wants to ensure that the sample is representative of the target population. The 1200 students are distributed across the four faculties as follows:

Faculty	Number of Students
Science	520
Humanities	40
Engineering	280
Business	360

(**Q10**) Calculate the number of students that should be selected from each faculty to obtain a proportional sample of 60 students.

Stratified sampling is a technique where the population is divided into subgroups called *strata*, which share a common characteristic (here, the faculty). Instead of drawing a simple random sample from the entire population, we sample from each stratum proportionally to its size in the population.

(**Q11**) Write a Python function that generates a stratified random sample.
The input is a list of students grouped by faculty:

```
faculties = [(faculty,students), ...]
```

and the desired sample size `sample_size`. The function should select students uniformly at random within each faculty and preserving the proportions within each faculty.

(**Q12**) Why might stratified sampling be preferable to simple random sampling if the study habits of students differ between faculties?

(**Q13**) Suppose that instead of stratifying, we select 60 students uniformly at random from all 1200 students. What is the expected number of Humanities students in the sample? What is the probability that the sample contains no Humanities students? You may use

the following formula. If X is the number of students from a particular group in a sample drawn uniformly without replacement, then

$$\Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where N is the population size, K is the number of students in the group, and n is the sample size. You can evaluate the probability with `scipy.stats.hypergeom.pmf(k,M,n,N)` (see documentation) where the parameter are the following:

- `k` is the number of "success" items (here 0);
- `M` is the total population size;
- `n` is the total number of ""success" objects inside the population;
- `N` is the sample size (the number of objects drawn without replacement).