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?

The source of data is

- **A**: The students who answer the questionnaire. Here the university acts as the organization collecting the data.
- **B**: The local transport authority, which originally collected and maintains the travel-card records.
- **C**: The students who post the comments. The social-media platform provides access to the data but the organization collecting the data is again the university.
- **D**: The campus mobility app and the users' devices, which record and transmit the GPS positions.

(**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?

A: Primary and designed. The questionnaire is created specifically for this investigation, and the questions are deliberately chosen to collect the required information.
B: Secondary and designed. The travel-card system was intentionally designed to record transport journeys, but the records were collected for another purpose and are now being reused.
C: Secondary and organic. The posts were produced as part of students' ordinary activity rather than specifically for the investigation, and the university is reusing existing posts.
D: Primary and organic. The GPS measurements are collected specifically for the study but as a byproduct of the app's navigation service rather than as a direct response to a question.

Note that the two classifications are independent.

(**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.

Possible answers include:

A: What version was used and when was it administered? This matters because it clarifies the context in which the data was actually measured and over what period.
B: Why were the travel-card records originally collected, and exactly which journeys do they contain? This matters because the records were created for operating the transport system and may not contain all the information needed for the investigation.
C: How were the posts identified and collected? This matters because the collection procedure may have filtered the available posts and therefore changed which observations are present in the dataset.
D: What exactly does the app record, and when does it record it? This matters because the recorded GPS positions depend on the app's recording protocol, for example when tracking starts and stops and how missing positions are handled.

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?

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.

(**Q2**) What is the access frame used in this study?

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.

(**Q3**) What is the sample in this context?

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.

(**Q4**) Why does aggregation at the census tract level limit the conclusions that can be drawn?

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.

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.

The total sum of the scores is the sum of the scores in the two groups and it suffices to weight the two averages by the number of students in each group:

$$\mu = \frac{1100 \times 12 + 100 \times 15}{1200} = 12.25.$$

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

The adjusted average is 12.25, which is higher than the average of the initial data (12). The misalignment arises because the sample excluded a higher-performing subgroup.

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.

Let $N = n_1 + n_2$. Then

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{n_1}(x_{1i} - \mu)^2 + \sum_{i=1}^{n_2}(x_{2i} - \mu)^2\right).$$

(**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

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

We obtain

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{n_1}\left[(x_{1i} - \bar{x}_1) + (\bar{x}_1 - \mu)\right]^2 + \sum_{i=1}^{n_2}\left[(x_{2i} - \bar{x}_2) + (\bar{x}_2 - \mu)\right]^2\right).$$

(**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 $(x_{2i} - \bar{x}_2)(\bar{x}_2 - \mu)$.

Expanding gives

$$\sigma^2 = \frac{1}{N}\left(\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + n_1(\bar{x}_1 - \mu)^2 + 2(\bar{x}_1 - \mu)\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) \right.$$
$$\left. + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2 + n_2(\bar{x}_2 - \mu)^2 + 2(\bar{x}_2 - \mu)\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2) \right).$$

The first and fourth terms describe the variation within the two subgroups. The second and fifth terms describe the differences between the subgroup means and the overall mean. The remaining terms are the cross terms.

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

By definition of the subgroup means,

$$\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) = 0.$$

Therefore,

$$2(\bar{x}_1 - \mu)\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1) = 0.$$

The same holds for the other term.

(**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.

Since

$$\sum_{i=1}^{n_j}(x_{ji} - \bar{x}_j)^2 = n_j s_j^2,$$

we obtain

$$\begin{aligned}\sigma^2 &= \frac{1}{N}\left(\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + n_1(\bar{x}_1 - \mu)^2 + \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2 + n_2(\bar{x}_2 - \mu)^2\right) \\ &= \frac{1}{N}\left(n_1 s_1^2 + n_1(\bar{x}_1 - \mu)^2 + n_2 s_2^2 + n_2(\bar{x}_2 - \mu)^2\right) \\ &= \frac{n_1 s_1^2 + n_2 s_2^2 + n_1(\bar{x}_1 - \mu)^2 + n_2(\bar{x}_2 - \mu)^2}{n_1 + n_2}.\end{aligned}$$

The first two terms describe the *within-subgroup variation*. The last two terms describe the *between-subgroup variation*: they measure how far the subgroup means are from the overall population mean.

(**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?

The between-subgroup terms are non-negative:

$$n_1(\bar{x}_1 - \mu)^2 \geq 0, \qquad n_2(\bar{x}_2 - \mu)^2 \geq 0.$$

Hence

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

Equality holds precisely when

$$\bar{x}_1 = \bar{x}_2 = \mu.$$

In this case, the subgroup means do not contribute to the overall variance. This inequality tells us that including the excluded group cannot decrease the overall variance below the weighted average of the subgroup variances. In other words, the excluded group introduces additional variability.

(**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?

We have

$$\sigma^2 \geq \frac{1100 \times 3^2 + 100 \times 2^2}{1200} = \frac{9900 + 400}{1200} \approx 8.58.$$

The variance of the complete population is at least approximately 8.58.

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.

The sampling fraction is

$$\frac{60}{1200} = 0.05.$$

Therefore:

Faculty	Sample size
Science	$520 \times 0.05 = 26$
Humanities	$40 \times 0.05 = 2$
Engineering	$280 \times 0.05 = 14$
Business	$360 \times 0.05 = 18$

The resulting sample contains $26 + 2 + 14 + 18 = 60$ students. In this exercise, the proportions give integer sample sizes, so no rounding is required.

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.

```python
import random

def stratified_sampling(faculties, sample_size):
    total_size = sum(len(students) for _, students in faculties)

    sample = []

    for faculty, students in faculties:
        strata_size = len(students) * sample_size // total_size
        faculty_sample = random.sample(students, strata_size)
        sample.extend(faculty_sample)

    return sample
```

Here, `random.sample` selects distinct students uniformly at random without replacement.

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

With simple random sampling, the number of students selected from each faculty is random. A small faculty could therefore be underrepresented, or even absent from a particular sample. If study habits differ between faculties, this helps ensure that the sample reflects the composition of the target population.

(**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).

There are 40 Humanities students out of 1200, so the probability that a randomly selected student is from Humanities is

$$p = \frac{40}{1200} = \frac{1}{30}.$$

The expected number of Humanities students in a sample of 60 is

$$\mathbb{E}[X] = 60 \times \frac{1}{30} = 2.$$

To obtain the probability that no Humanities students are selected, we use the hypergeometric distribution with $N = 1200$, $K = 40$, $n = 60$, and $k = 0$.
Therefore,

$$\Pr(X = 0) = \frac{\binom{40}{0}\binom{1160}{60}}{\binom{1200}{60}} = \frac{\binom{1160}{60}}{\binom{1200}{60}} \approx 0.124.$$

Thus, although the expected number of Humanities students is 2, a simple random sample does not guarantee that the faculty is represented.