Subject number : XXXXXXXXX

Sujet corrigé - 3 – Data Collection

This is the **question paper**. It is **NOT** the answer sheet.
Please check that the number on your question paper matches the number on your answer sheet. To complete the answer sheet correctly, you must:

- use a **black** ink pen
- shade in the boxes **completely without going over the edges**
- if you make a mistake, erase the box with a whiteout ("Tipp-Ex"), **but do not redraw it**
- every question has a **unique correct answer**.

Box correctly ticked

1 Ⓐ ● Ⓒ Ⓓ Ⓔ

Box incorrectly ticked

1 Ⓐ ☒ Ⓒ Ⓓ Ⓔ

1 **What is the primary purpose of the data collection stage in the data lifecycle?**
1 Point - Only one correct choice

☐ **A.** To visualize initial trends before full analysis
☒ **B.** To gather relevant data for analysis
☐ **C.** To preprocess raw data to remove errors and inconsistencies
☐ **D.** To ensure data is stored in a structured format for easy retrieval

> Data collection is the first step, where raw data is acquired to answer a research question.

2 **What is the difference between primary data and secondary data?**
1 Point - Only one correct choice

☐ **A.** Primary data is collected by researchers, while secondary data is collected by machines
☒ **B.** Primary data is collected for the study, and secondary data is reused from other sources
☐ **C.** Primary data is quantitative, secondary data is qualitative
☐ **D.** Primary data is more reliable because it is collected firsthand, unlike secondary data

> Primary data is tailored to the research question, while secondary data is repurposed.

3 **In the CalEnviroScreen example, what is the target population?**
1 Point - Only one correct choice

☐ **A.** The census tracts in California
☐ **B.** The air monitoring stations
☒ **C.** All individuals living in California
☐ **D.** The health statistics from hospitals

> The target population is the group about which conclusions are drawn.

4 **What is a confounding variable in an experiment?**
1 Point - Only one correct choice

☐ **A.** A variable that is directly manipulated by the researcher
☒ **B.** A variable that affects both the independent and dependent variables
☐ **C.** A variable that should have been collected but was forgotten
☐ **D.** A variable that is irrelevant to the study

> Confounding variables can distort the relationship between the variables of interest.

5 **What is the main advantage of using simulations in data collection?**
1 Point - Only one correct choice

☐ **A.** They provide exact real-world measurements
☒ **B.** They allow exploration of scenarios that are impractical or unethical to test experimentally
☐ **C.** They are always faster to obtain than data from real experiments
☐ **D.** They eliminate the need for data cleaning, as we can simulate exactly the required data

> Simulations are useful for testing hypotheses in controlled, virtual environments.

6 **What is the purpose of stratified sampling?**
1 Point - Only one correct choice

- ☐ **A.** To eliminate the need for randomization of the entire population
- ☒ **B.** To ensure proportional representation of subgroups in the sample
- ☐ **C.** To collect data from the less accessible participants
- ☐ **D.** To reduce the sample size as much as possible without introducing bias

> Stratified sampling divides the population into subgroups and samples proportionally from each.

7 **Which of the following is a potential bias in survey data collection?**
1 Point - Only one correct choice

- ☐ **A.** Using a huge sample size
- ☒ **B.** Having participants drop out
- ☐ **C.** Overlapping data collected from multiple sources
- ☐ **D.** Forgetting to clean the data

> Non-response bias occurs when certain groups are underrepresented in the responses.

8 **In the OpenWeatherMap API lab, what is the primary role of the API key?**
1 Point - Only one correct choice

- ☐ **A.** To encrypt the collected weather data
- ☒ **B.** To authenticate and authorize access to the API
- ☐ **C.** To ensure a real-time connection with the API, avoiding delay
- ☐ **D.** To clean the weather data automatically

> API keys ensure secure and authorized access to data providers.

9 **In Monte Carlo simulations, what is the role of randomness?**
1 Point - Only one correct choice

- ☐ **A.** To mitigate variability in the input data
- ☒ **B.** To model uncertainty and explore a range of possible outcomes
- ☐ **C.** To mimic real-world experiments where events can never be fully understood
- ☐ **D.** To parametrize the noise, which simplifies cleaning

> Randomness in Monte Carlo simulations helps estimate probabilities and outcomes under uncertainty.

10 **A data scientist realizes that the dataset cannot answer the research question even after excellent cleaning and preprocessing. What is the most likely reason?**
1 Point - Only one correct choice

- ☐ **A.** The machine learning algorithm was not trained long enough; they should request more computing power
- ☐ **B.** The collection did not capture enough data; they should rerun the collection procedure at a larger scale
- ☐ **C.** The metadata was poorly documented; they should ask for further details from the data publisher
- ☒ **D.** The collection did not capture the proper variables; they should redesign the collection protocol

> Even with high-quality data, if the collected variables do not address the research question, the analysis will fail. Redesigning the collection protocol to capture relevant variables is essential.