Scenario

Consider that the Cyprus public transport is evaluating whether to expand bus service. Before investing, administrators want an estimate of the proportion of UCY students who primarily use public transport to commute. Your team will design the study (no data collection), specify sampling plans, describe the data you would collect, and produce a data-description plan.

Tasks

A. Study design

1. Restate the research question and null hypothesis.
2. Describe the two sampling approaches (haphazard and random) and state whether sampling is with or without replacement and why.
3. Give a conceptual justification for chosen sample sizes for each plan (no arithmetic).

B. Data description & collection plan

1. **Data dictionary (required):** List each variable you would collect, giving:
    - Variable name
    - Type (categorical, ordinal, numeric, binary, date/time)
    - Possible values / coding (e.g., 0 = no, 1 = yes; commute modes: bus/train/car/bike/walk/other)
    - Short description and units (if applicable)
    - Whether the variable is required or optional
2. **Survey question wording:** Provide precise wording for the primary question(s) about commuting (to avoid measurement error). Include any definitions (e.g., used ≥3 days/week).
3. **Metadata to record:** For each observation, describe administrative fields you would capture (timestamp, sampling method label, interviewer ID, response mode, consent flag, location/time block if using stratified location sampling).
4. **Missing data plan:** Describe how you would record nonresponse and partial responses and a conceptual policy for dealing with them in descriptive summaries.

C. Data description tasks

1. **Variable-level summaries:** For each variable type, state the descriptive summaries you would present (no calculations required). For example:
    - Binary/categorical: frequency table, proportions, top categories.
    - Ordinal: distribution table, median/typical category.
    - Numeric: min, max, median, interquartile range, and a short description of spread.
2. **Data quality checks:** Conceptual checks you would run (e.g., impossible values, duplicate IDs, response time outliers, inconsistent answers).

3. **Interpretation of results**
    - Explain what a confidence interval would tell you about your estimate.
    - Explain how you would use a p-value to test your null hypothesis.

4. **Probability distribution question**
    - Based on your sampling plan, describe **what probability distribution** you would *expect* the number of "public transport users" in your sample to follow (conceptually) and why.
5. **Biases & limitations**
    - Identify potential biases for each sampling approach.
    - Suggest mitigation strategies.

A. Study design

Research question: What proportion of enrolled students primarily use public transport to commute to campus (defined as commuting by bus ≥3 days per week)?

Null hypothesis: H0: 40% of enrolled students primarily use public transport.

Sampling approaches:

- **Haphazard:** Ask first 120 students at the main campus entrance between 12:00–13:30 over two days. (Fast pilot; expected selection bias toward midday visitors).
- **Random (recommended):** Draw 600 unique student IDs from the list (without replacement), contact by email/text to complete a one-question survey. Without replacement avoids duplicate participants and matches finite population reality.

Sample size rationale (conceptual):

- Larger sample (600) for random sampling to improve subgroup coverage and reduce uncertainty. Haphazard (120) acts as pilot to refine question wording and logistics.

B. Data description & collection plan

1. Data dictionary (example variables)

- `student_id` — Type: identifier (string); Values: unique student code; Required: yes; Description: encrypted ID used only for linkage then removed.
- `response_id` — Type: identifier; Values: unique per response; Required: yes.
- `survey_timestamp` — Type: datetime; Values: ISO timestamp; Required: yes.
- `sampling_method` — Type: categorical; Values: `random`, `haphazard`; Required: yes.
- `response_mode` — Type: categorical; Values: `in-person`, `email`, `sms`; Required: yes.
- `primary_commute_mode` — Type: categorical (binary derived); Values: `public_transport`, `private_vehicle`, `walk_bike`, `other`; Required: yes; Description: mode used majority of commuting days.
- `public_transport_days_per_week` — Type: ordinal (0–7); Values: integers 0–7; Required: optional; Description: number of days per typical week the student takes bus/train.
- `residence_type` — Type: categorical; Values: `on_campus`, `off_campus_local`, `off_campus_faraway`; Required: yes.
- `faculty` — Type: categorical; Values: list of faculties/departments; Required: optional.
- `consent_given` — Type: binary; Values: `yes/no`; Required: yes.

- `interviewer_id` — Type: categorical; Values: code for interviewer; Required: conditional (if in-person).
- `response_duration_seconds` — Type: numeric; Required: optional; Description: proxy for inattentive responses.
- `missing_reason` — Type: categorical; Values: `declined`, `absent`, `other`; Required if nonresponse.

2. Survey question wording (precise)

Primary question (single-item, clear):

"In a typical week, on how many days do you commute to campus **primarily by public transport (bus)**? Please enter 0–7. (Definition: 'primarily' = you use bus or train for that commute on that day, not as part of a multimodal trip.)"

Derived binary for analysis: `primary_public_transport = yes` if `public_transport_days_per_week ≥ 3`.

Include consent text: "By answering, you consent to anonymized use of your responses for planning research."

3. Metadata to record

- `sampling_method`, `survey_timestamp`, `response_mode`, `interviewer_id` (if in-person), `contact_attempts` (for random sampling), `consent_given`, `missing_reason` if applicable.

4. Missing data plan

- Record all noncontacts and refusals with `missing_reason`.
- Report descriptive tables that include counts and percentages of missing responses by sampling method and strata.
- Avoid imputing for the primary binary outcome unless a documented missingness mechanism justifies it; report sensitivity interpretations (e.g., "If all nonresponders were users, p would be X; if none were users, p would be Y").

C. Data description tasks — what to produce

1. Variable-level summaries (no computations)

- `primary_commute_mode` / derived binary: frequency table (counts and percentages for each category), separate for random vs haphazard samples.
- `public_transport_days_per_week`: distributional description (e.g., majority at 0–2 vs 3–7), central tendency described in words (e.g., "most students report 0–2 days").
- `residence_type` and `faculty`: frequency counts to show subgroup composition.
- `response_mode` and `sampling_method`: metadata summaries to check how many responses came from each mode/method.

2. Data quality checks (conceptual)

- Check for duplicate `student_id` or `response_id`.
- Check `public_transport_days_per_week` values within 0–7.
- Examine `response_duration_seconds` for implausibly short times suggesting careless replies (flag for review).
- Cross-validate `primary_commute_mode` vs `public_transport_days_per_week` for consistency.

3. Interpretation of results

Confidence interval (conceptual):
A 95% CI would show a plausible range for the true campus proportion of public transport users. Narrower intervals mean more precise estimates.

p-value (conceptual):
We would test whether the true proportion differs from the 40% null. A small p-value would indicate evidence against H0; a large p-value indicates insufficient evidence. Validity depends on proper random sampling.

4. Probability distribution

Given our **random sample without replacement** from a large population and a binary outcome (yes/no public transport), the count of "yes" answers in the sample would approximately follow a **binomial distribution** with parameters n (sample size) and p (true proportion).

5. Biases & limitations

Haphazard sample:

- Selection bias (time and location).
- Volunteer bias (who agrees to respond).

Random sample:

- Frame bias (if list omits part-time or distance students).
- Nonresponse bias (non-responders differ systematically).
- Measurement error (definition of "primarily").

With vs. without replacement:

- Without replacement