Final Assignment
Survey Methodology I
Due date: **January 25, 2023**

1. A political analyst reviews two studies conducted before the presidential election (two candidates, A and B). The first study, nationally representative of the population registered in the electoral registers and probabilistic in all its stages, yielded a point estimate of 54.5% (valid %) in favor of candidate A. Another study using sex and age quotas and a nationwide sample of registered voters obtained a 52.3% (valid %) point estimate for candidate A. The election result was 51.6% (valid votes). The analyst notices that the study conducted by quotas was closer to the final result, raising the question of the advantages of conducting a probabilistic study (usually always more expensive).

    a. What are the advantages and disadvantages of each type of sample? What is gained with one or the other?
    b. What background information is needed to know the precision of the estimates (MOE) of both studies?
    c. Why was the quota study closer to the final result?

2. A well-known researcher was assigned to study the workers' opinions on specific measures and decisions planned by the management. The management wanted to ensure that these decisions would be well-received and aimed to predict the outcome of a referendum scheduled for the end of the month. The researcher's sample design was as follows:

Group	People	%	Sample
A	2231	16,7	250
B	6450	48,3	250
C	1229	9,2	250
D	3440	25,8	250
Total	**13350**	**100**	**1000**

Once the sample was collected, the analyst conducted a frequency analysis to determine agreement on the most controversial decision. Out of the workers surveyed, 62% stated their agreement. The researcher anticipated a positive outcome, prompting the management to hold a referendum. The participation rate was high, with 96% of the workers expressing their opinion. However, the final result of the plebiscite revealed that only 44% agreed with the measure.

    a. What type of sampling does the researcher's design correspond to?
    b. What could explain the difference in the results of the prestigious researcher, given that it is a sample of 1000 cases?
    c. Based on the answer to the previous question, is it necessary to make any adjustments to the database before running the variable frequencies? If it is necessary to make the adjustment, define how it should be calculated.

3. Suppose an influential politician seeks your advice on a poll published in a local newspaper. The poll indicates that the prominent politician is leading by a significant margin (over 20 percentage points ahead of their primary opponent). However, considering the current street sentiment, it appears that the younger and more proactive opponent has been gaining supporters. Here are the technical details of the study:

   - A telephone survey of **600 cases**
   - Maximum error: ± 4 percentage points
   - Response rate with respect to the number of calls: 20%
   - Application date: April 1-7, 2023.
   - Design: A random sample of telephone numbers was taken from the three main provinces, out of a total of 21 provinces.

   a. Who is the target population of the study, and what specifications or recommendations should be given to the politician regarding this?
   b. If we assume that the error calculation is correct (mathematically speaking), is it appropriate to claim that the maximum sampling error is ± 4 percentage points? Why? Is there a lack of background information to answer this question? If so, what information is missing?
   c. What systematic bias could be identified based on the characteristics and performance of the study?
   d. Ultimately, should the prominent politician rely on the survey results or not?

4. A recent newspaper article reported that "sales of hand-held digital devices (e.g., tablets) are up by nearly 10% in the last quarter, while sales of laptops and desktop PCs have remained stagnant." This report was based on the results of an online survey in which 9.8% of the more than 126,000 respondents said that they had "purchased a hand-held digital device between January 1 and April 30 of 2023."

   Emails soliciting participation in this survey were sent to individuals using an email address frame from the 5 largest commercial Internet Service Providers (ISP) in the U.S. Data collection took place over 6 weeks beginning May 1, 2023. The overall response rate achieved in this survey was 13 percent.

   Assume that the authors of this study wanted to infer expected purchases of U.S. adults (18 years old +).

   a. What is the target population? What is the population in the sample frame?
   b. Briefly discuss how the design of this survey might affect the following sources of error (2 – 4 sentences each).

      ● Coverage error (specify the type of coverage error you are concerned with)
      ● Nonresponse error
      ● Measurement error

   c. Without changing the duration or the mode of this survey (i.e., computer-assisted, self-administration), what could be done to reduce the errors you outlined in 1b? For each source of error, suggest one change that could be made to reduce this error component, making sure to justify your answer based on readings and lecture material (1– 2 sentences each).

d. To lower the cost of this survey in the future, researchers are considering cutting the sample in half, using an email address frame from only the 2 largest ISPs. What effect (if any) will these changes have on sampling error and coverage error (1 – 3 sentences each)?

5. The following is a list of A = 10 blocks. Draw a PPS systematic sample, using units as the measure of size. Use a random start of 6 and an interval of 61.

Block	Units
1	32
2	18
3	48
4	15
5	37
6	26
7	12
8	45
9	46
10	21

6. After a five-year interval since the previous census, you carry out a household survey utilizing a telephone number database. When a chosen telephone number corresponds to a household, interviewers ask to speak with the person who has the most knowledge about the health of the household members. After the survey is completed, someone suggests assessing its effectiveness by comparing the demographic distributions (such as age, sex, race/ethnicity, and gender) of the "most knowledgeable" health informant with the demographic distributions of adults from the previous census. I would like to hear your thoughts on the wisdom of this suggestion.

7. You were asked to determine if there is a systematic difference in substance use between women and men using a dataset coming from sample of students (high schools). The sampling design was multistage:

   - First stage: simple random selection of schools
   - Second stage: simple random selection of students within schools

   The dataset `school-data.csv` in Github has the information you will need for the analysis:

   - Student responses to substance use in the past year. (`drug: 1 = yes, 0 = no`)
   - Gender (`female: 1 = yes, 0 = no`)
   - Identifiers of schools and students (`id_school`, `id_student`)
   - Total number of schools in the population (`total_schools`), and total students per school (`total_students`).

   Using this information:

   a. Estimate the difference substance use between women and men using the corresponding sampling design and weights.
   b. Estimate the 95% confidence interval of the difference and state a conclusion (Hint: You can use the `svyttest` command)