Final Assignment
Survey Methodology I

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?

The probabilistic sample allows you to know the probability of selection of each respondent and use statistical theory to compute the sampling error. However, full probabilistic samples are generally more expensive than non-probabilistic methods. Also, depending on the level of non-response, the probabilistic sample could still have some biases that need to be fixed with the data available.

With a quota sample, we don't know the probability of each respondent's selection. Thus, we cannot compute the standard error based on the sampling design without making strong assumptions about the sample (random selection of respondents). Quota samples are usually cheaper and faster than probabilistic samples, as interviewers do not have to make an extra effort to get a response from a select respondent.

    b. What background information is needed to know the precision of the estimates (MOE) of both studies?

In the case of the probabilistic design, we need to know how the sample was designed. What is the sample size? Does it include stratification? How was that allocation? Does the design include clustering? Can we estimate the design weights, or are they provided? Are there any post-stratification weights used to get estimates?

In the case of a quota sample, we cannot compute the precision of the estimates (MOE) as we don't know the probability of selecting respondents. Assuming the sample was randomly drawn, we can compute the MOE, but that would be a problematic and strong assumption.

    c. Why was the quota study closer to the final result?

Some possible reasons:

The estimate of the probabilistic sample has a given precision. Is the 54.5% estimate systematically different from 52.3% and 51.6%? We can run those tests and check if, given the sample size and design of our sample, we can conclude that the probabilistic estimate is different from the final result. By chance, the probabilistic estimate can be different from the target statistic. The critical point is that we can estimate the precision of that estimation. In the case of the quota study, we can only estimate that precision by making very strong assumptions (e.g., that we know the probability of selection).

Another possibility is that the probabilistic sample may have a systematic non-response bias that could bias the results. For instance, younger respondents (that tend to answer surveys less often) can be prone to vote for candidate B instead of A. Adjusting for non-response might correct the bias.

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?

Stratified sampling, assuming the selection of respondents within each stratum is random.

b. What could explain the difference in the results of the prestigious researcher, given that it is a sample of 1000 cases?

Participation was rather high, so it doesn't seem to be an issue of non-response bias. The issue seems to be related to the design allocation. Allocation is fixed (same sample size for each stratum), even though strata don't have even populations. Therefore, the probability of selection by stratum isn't the same. It could be that one of the groups underrepresented by the sample design (e.g., B) have a critical opinion of the measure management wanted to implement. Thus, when estimating the proportion agreeing with the controversial measure, we need to use weights to correct the uneven probability of selection, otherwise, we will get the wrong estimate.

It's also possible that a systematic bias could alter the results. For instance, an event between the survey and referendum that change the support for the controversial decision, or some a social desirability effect (e.g., if the surveys were not self-administrated). In any case, I would check first the allocation issue and once it is dismissed, I would explore other possible explanations.

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.

We need to compute design weights for each stratum. That is, each united sample should be multiplied by the factor `population_strata_i / 250` *when estimating the proportion:*

A: 2231/250 = 8.924
B: 6450/250 = 25.8
C: 1129/250 = 4.516
D: 3440/250 = 13.76

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?

The target population is the people eligible to vote in the election. However, the sampling frame of telephone numbers comes only from the three main provinces (out of 21). The 18 remaining provinces might have a different voting behavior: they might be more likely to vote for the younger candidate. It's hard to say without having additional data, for example, previous elections' results by province. It's also not clear what proportion of the total population eligible to vote is represented by the three central provinces. It that proportion is pretty big (e.g., 95%), it shouldn't be a big problem, but if it is 60%, the remaining 40% can make a big difference in the final result.

The characteristics of the sampling frame need to be clarified. Do they include mobile numbers or only landlines? We will need more information on the quality and coverage of the sampling frame of telephone numbers to conclude: Are they including phones from all companies? What is the level of phone coverage in the region being studied?

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?

A MOE of 0.04 with a sample of 600 respondents is what you get when using the proportion formula of error for a simple random sample and p=0.5 and 95% of confidence. If that is right will depend on the actual design of the sample that we don't know:

i. *Did they use any stratification? What was the allocation per stratum?*
ii. *Did they compute design or raking weights?*
iii. *How were respondents selected when calling a household (landline)? Do they select who first answered, use quotas, or select them randomly?*

c. What systematic bias could be identified based on the characteristics and performance of the study?

The response rate is pretty low, and we need to find out if they use any non-response adjustment (e.g., weighting). If there was no adjustment, responses will likely be biased towards older respondents, where the candidate seems to be doing better, as younger voters have lower survey participation rates.

d. Ultimately, should the prominent politician rely on the survey results or not?

I would suggest to conduct a survey with a better design:

i. *Get better representation of the 21 provinces of the region (e.g., PPS selection of provinces while keeping the main ones)*
ii. *Be sure that the sampling frame gets as closely as possible to the target population (those eligible to vote, e.g., using landlines and mobile numbers)*
iii. *Be sure the final estimates adjust for non-response (e.g., weighing)*

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 sample frame?

The target population is US adults, 18 years older or over. The sample frame is clients of the five largest commercial ISP, assuming that ISPs have emails of all their clients.

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)

The sample frame doesn't include all elements of the target population. People with other internet providers or who don't have internet access or email are excluded from the study, and could have different behavior on using technology devices.

● Nonresponse error

The response rate is quite low (13%). People who responded to the survey might differ from those who did not respond (e.g., younger and more involved in the use of technology). This might bias the estimation of purchases of hand-held devices.

● Measurement error

It's possible respondents don't remember exactly when they bought a device. Memory problems could overestimate purchases (e.g., including purchases that occurred a year ago, not the period of interest).

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).

Regarding coverage error, I would suggest extending the sample frame to include other directories of emails (small ISP providers, institutional providers). For non-response error, I would provide better incentives (e.g., prizes) to respondents to increase participation rates. Regarding measurement error, I would change the time reference of the question: Have you bought some hand-held devices (e.g., tables, kindles, etc.) during the last four months?

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)?

The precision of purchase estimates will decrease (more sampling error), assuming the same variance, statistician confidence, and we know the probability of selection of units. Coverage error, on the other hand, might increase because email addresses from the two largest ISP can have a different profile of users (e.g., older, from big cities or specific areas of the US).

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

Block	*Units*	*Cumulative*	*Selection*
1	*32*	*32*	*Yes (6)*
2	*18*	*50*	*No*
3	*48*	*98*	*Yes (67)*
4	*15*	*113*	*No*
5	*37*	*150*	*Yes (128)*
6	*26*	*176*	*No*
7	*12*	*188*	*No*
8	*45*	*233*	*Yes (189)*
9	*46*	*279*	*Yes (250)*
10	*21*	*300*	*No*

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.

   It is a good idea to compare the distribution of sociodemographic variables that comes from a telephone survey with census data, assuming that we are contrasting the same geographic areas and demographic groups. However, we would need to compare the distribution of all household members to the census distribution, not only the most knowledgeable, as we know they would probably be women or older. Comparing distributions from all household members would give us a better idea of the potential demographic biases of our telephone survey. This won't resolve measurement issues related to the health report.

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)

```
See notebook final-2024.ipynb in GitHub.
```