Final Assignment
Survey Methodology I
Due date: **January 21, 2022**

Note: You will need to wait until the last class to answer some of these questions.

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 prestigious researcher was asked to study workers' opinions regarding specific measures/decisions management wanted to implement. Management hoped decisions would be well-received and wanted to anticipate the results of a referendum planned at the end of the month. The researcher's sample design was:

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**

The researcher designed a survey and ran a frequency of agreement for the most controversial decision. 62% of the workers stated that they agreed with it. The researcher was sure of a favorable result, so the management conducted the referendum: the participation was high, and 96% of the workers registered their opinion. However, the final result of the plebiscite was 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: 1128/250 = 4.921
   D: 3440/250 = 13.76

3. Suppose an important politician asks you for advice regarding the results of a poll published in a regional newspaper. According to the poll, the prominent politician wins by far (more than 20 percentage points over his primary opponent). However, based on the atmosphere on the streets, the younger and more active opponent has been adding adherents. The technical details of the study are as follows:

   - 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, 2022.
   - Strategy: a random sample of telephone numbers of the three main provinces of a region (out of 21 provinces).

   a. What is the study's target population, and what specifications or recommendations would you give the leading politician in this regard?

      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. Assuming that the error calculation is correct (mathematically speaking), is it appropriate to state that the maximum sampling error is ± 4 percentage points? Why? Is there a lack of background information to answer this question? If yes, what 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:

   - *Did they use any stratification? What was the allocation per stratum?*
   - *Did they compute design or raking weights?*
   - *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 would it be possible to identify, given 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. Finally, should the prominent politician rely on the survey results or not?

   I would suggest to wait (or conduct) a survey with a better design:

   - *Get better representation of the 21 provinces of the region (e.g., PPS selection of provinces while keeping the main ones)*
   - *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*
   - *Be sure the final estimates adjust for non-response (e.g., weighing)*

4. You were provided a dataset from a school survey to estimate adolescent drug use (column `drug` in the file `final-assignment.csv` available in the homework folder on the Github repository). No sampling weights were provided, only the marginal distribution of key variables such as gender and age group.

   Marginal distribution of population
   Gender: Female (0.52), Male (0.48)
   Age: 9-12 years (0.40), 13-15 years (0.32), 16-18 years (0.28)

   Using the techniques reviewed in the course (write your answers and attach your code or computations):

   a. Estimate the proportion of drug use
   b. Estimate the DEFF of any survey weights you compute

   For this exercise, you need to use raking.
   Look at the notebook `final-assigment-review.ipynb`.

5. The following is a list of A = 10 blocks. Draw a PPS systematic sample, using Xa as the measure of size. Use a random start of 6 and an interval of 41.

Block	*Xa*	*Cumulative*	*Selection*
1	*32*	*32*	*Yes (6)*
2	*18*	*50*	*Yes (47)*
3	*48*	*98*	*Yes (88)*
4	*15*	*113*	*No*
5	*37*	*150*	*Yes (129)*
6	*26*	*176*	*Yes (170)*
7	*12*	*188*	*No*
8	*45*	*233*	*Yes (211)*
9	*46*	*279*	*Yes (252)*
10	*21*	*300*	*Yes (293)*