Weekly Homework I Review
Survey Methodology I

1. For each of the following design decisions, identify which error source might be affected. Each design decision can affect at least two different error sources. Write short (2-4 sentences) answers to each point.

   a) The decision to use repeated calls persuading reluctant respondents in a customer satisfaction survey for a household product manufacturer.

   Using repeated calls might reduce non-response error as people who respond to a customer satisfaction survey could be happier with the manufacturer. However, it could affect the quality of measurement: respondents might declare less satisfaction due to getting repeated calls.

   b) The decision to increase the number of questions about assets and income in a survey of income dynamics, resulting in a lengthening of the interview.

   While increasing the number of questions might improve the validity of income dynamics measurement, it could also reduce the quality of information due to respondent fatigue. Longer questionaries would also increase item non-response.

   c) The decision to include prisons and hospitals in the sampling frame for a study of consumer expenditures.

   Assuming that people from prisons and hospitals are not active consumers, to include them correspondents to a specification error (units not suitable to the goals of the study), and coverage error (inclusion of ineligible units).

   d) The decision to change from a face-to-face interview design to a mailed questionnaire mode in a household survey of illegal drug usage.

   Using a mailed questionnaire might improve the report of illegal drug use (lower measurement error due to the reduction of social desirability bias), but it can also increase non-response error (e.g., problematic users will probably not respond to a mailed questionnaire).

2. A medical practice has records for N=830 patients. A simple random sample of n=273 was selected and 174 of the sample patients had private health insurance:

   a. Estimate the % of patients with private health insurance and the standard error.

   Assuming an infinite population:

   ```
   174/273 = 0.637
   SE = sqrt(174/273 * (1 - 174/273) / 273) = 0.029
   ```

   Multiple by 100 to get percentages.

   b. Calculate a 97% confidence interval for the population % and provide an interpretation. What are the advantages and disadvantages of using a 95% or 88% confidence interval instead?

   Assuming an infinite population:

   ```
   MOE = 2.17009 * sqrt(174/273 * (1 - 174/273) / 273) = 0.06314311
   Confidence interval = [0.5742195, 0.7005057]
   ```

Multiple by 100 to get percentages.

If we were to repeat the experiment (sample) over and over, then 97% of the time, the confidence intervals will contain the true mean.

Confidence (1-alpha) represents the likelihood of correctly not rejecting the null hypothesis (true negative), while alpha indicates the probability of Type I error (false positive). A smaller alpha (or higher confidence) corresponds to a lower probability of committing Type I error. Confidence intervals of 95% and 88% have a higher chance of Type I error compared to a 97% confidence interval, but they offer less precision (resulting in wider intervals). The convention is to use 95% confidence intervals, but that decision is rather arbitrary.

c. The study is to be repeated in another medical practice that has N=1150 patients. A standard error of 2.5 percentage points for the sample % of patients with private health insurance is required. What sample size is needed for a simple random sample to achieve this level of precision? For planning purposes, assume that the population percentage is 50%.

Assuming infinite population:

```
1.96^2 * 0.5*(1-0.5)/(1.96*0.025)^2 = 400
```

Using sampler:

```
library(sampler)
ssize(1.96*0.025)
```

Assuming finite population:

```
1150 * (0.5*(1-0.5))/((1150-1)*(0.025)^2 + 0.5*(1-0.5)) = 297
```

Using sampler:

```
library(sampler)
ssize(1.96*0.025, N=1150)
```

3. An organization has requested your assistance in designing a sample to estimate the proportion of engaged employees. The organization has provided data on the number of employees in four different types (A, B, C, D) based on the year 2023, as well as previous estimates of the proportion of engaged workers. These estimates were obtained from a survey conducted in 2021, which included 50 interviews per employee type, totaling 200 employees. The organization has allocated a budget to interview 300 employees. The executives are particularly concerned about the low engagement within group C (operation managers) and would like to prioritize maximizing the precision of the analysis for that specific group, to conduct additional analysis to understand factors influencing engagement within that group.

	Number of employees (2023)	Proportion Engaged (2021)
A	121	0.83
B	536	0.72
C	343	0.51
D	2642	0.32

Propose and justify a sample design for this problem specifying:

a. How many employees to interview by type.
b. Compute design weights if necessary.
c. MOE for the whole sample.
d. MOE only for group C.

See notebook homework-01-2023.ipynb in GitHub.