Problem Set 1: Micro Analysis

1 Purpose

The purpose of this problem set is to assess your understanding of one key method of quantitative public opinion research: survey sampling and basic quantitative analysis of survey data.

2 Overview

You are asked to watch a short slidecast covering relevant material and then complete the following tasks. Questions regarding this material should be raised on the Moodle discussion board (`https://moodle.lse.ac.uk/mod/forum/view.php?id=521541`) or in instructor office hours.

3 Your Task

1. In your own words, what makes a sample representative of a population? What are some different ways of thinking of representativeness?

   Solution:

   - Representative sampling using a probability-based sample
   - Reweighting of a non-probability sample or a probability-based sample with any degree of unit or item nonresponse
   - Descriptive representativeness based upon demographics or other variables
   - Expert judgments or "fit for purpose" samples

2. In a few sentences, answer the following: What is coverage error? What consequences does it have for survey research? How can over-coverage and under-coverage errors be addressed?

   Solution:

   Coverage error is the degree to which a sampling framed used for recruiting survey respondents does not match the population. Overcoverage refers to a sampling frame that includes ineligible respondents (i.e., those who fall outside the specified population). Undercoverage refers to a sampling frame that excludes some population members, thereby making them ineligible for recruitment.

   Overcoverage can be addressed by trimming the sampling frame using other information (e.g., administrative data) or assessing eligibility at the point of recruitment.

   Undercoverage can be addressed by using a dual- or multi-frame sampling technique. In such designs, two or more sampling frames are constructed, their overlap is determined, and individuals are selected based on their membership in one or both (or more) lists. An alternative is simply to ignore undercoverage thereby accepting an unrepresentative sample, attempt to statistically correct for undercoverage (e.g., with weighting), or to redefine the target population (e.g., from the population to "those with landline telephones").

3. Define "element variance" and "sampling variance". How do these relate to "standard deviation" and "standard error"?

Solution:

"Element variance" is the variance of values for a variable. It can be defined for the population as:

$$\sigma^2 = Var(X) = E\left[(X - \mu_X)^2\right] \tag{1}$$

Or for the sample as:

$$s^2 = Var(x) = \sum_{i=1}^{n} \frac{x_i - \bar{x}}{n} \tag{2}$$

The population standard deviation σ is the square root of the population element variance, σ^2. The sample standard deviation s is the square root of the sample element variance, s^2.

Sampling variance is not a measure of variable elements but rather of sample statistics. If we were to repeat the sampling procedure an infinite number of times, calculating the mean (or another statistic) on each sample, then the "sampling variance" would be the variance of the resulting sample statistics. The "standard error" is the square root of this quantity, or equivalently the standard deviation of the sample statistics.

4. Imagine that 1000 respondents are recruited to participate in a survey and complete the interview, 1700 were invited to participate but chose not to participate, and 600 were invited but their eligibility for the study was unknown. What are the upper and lower bounds of the response rate for this study? Should we be concerned with this response rate? Defend your answer.

Solution:

If we ignore the eligibility of all unknowns, then we can calculate Response Rate 1:

$$RR1 = \frac{I}{(I+P) + (R+NC) + U} = \frac{1000}{1000 + 1700 + 600} = 0.30 \tag{3}$$

If we estimate the proportion of unknowns that are eligible, we might use Response Rate 3:

$$RR3 = \frac{I}{(I+P) + (R+NC) + (e*U)} \tag{4}$$

where e is estimated Pr(eligible) among unknowns. This allows us to place bounds on the response rate from:

$$max(RR3) = \frac{1000}{1000 + 1700 + (0*600)} = 0.37 min(RR3) = \frac{1000}{1000 + 1700 + (1*600)} = 0.30 \tag{5}$$

5. If one's goal is to estimate the proportion of the British population that supports the UK leaving the European Union, how large of a sample would be needed to estimate that percentage within +/- 2 percentage points? What about within +/- 0.5 percentage points? (Show your work.)

Solution:

Here we start from the equation for sampling variance of a proportion:

$$Var(p) = (1-f)\frac{p(1-p)}{n-1} = \frac{p(1-p)}{n-1} \tag{6}$$

(Note: Given the large population, we can ignore the finite population correction.)

We solve this for n and then plug in values recalling that margin of error is twice the standard error (SE). The maximum element variance of a proportion occurs when $p = 0.50$, so we will assume that here.

$$n = \frac{p(1-p)}{v(p)} \tag{7}$$

$$= \frac{p(1-p)}{SE^2} \tag{8}$$

$$= \frac{0.5(1-0.5)}{(0.01)^2} = \frac{0.25}{0.0001} = 2500 \tag{9}$$

$$= \frac{0.5(1-0.5)}{(0.0025)^2} = \frac{0.25}{0.00000625} = 40,000 \tag{10}$$

$$\tag{11}$$

6. In a simple random sample, all individuals in a sample are given equal *design* weights for the purposes of analyzing the resulting data. In a stratified sample (and any complex survey design), these weights differ across individuals. Consider a hypothetical survey of members of the population of England and Wales age 16 and over that is stratified by levels of education (i.e., three strata: those no qualifications, with some qualifications, and those with a university degree or greater). The census estimates of these population strata sizes are available from the Office for National Statistics. If the sample strata are equally sized (i.e., the same number of individuals are interviewed in each stratum) and individuals in the "university degree or above" stratum were assigned weights of 1, for which stratum are the design weights going to be larger than 1 and for which are the design weights going to be smaller than 1? Explain your answer.

Solution:

Design weights adjust the weighting given to each sample observation in order to correct for sampling design. In a simple random sample, all observations have a design weight of 1. In a stratified sample, strata that are "over-sampled" (i.e., the proportion of sample stratum observations is intentionally larger than the proportion in the population) are given design weights less than 1. Those strata that are "under-sampled" (i.e., the proportion of sample stratum observations is intentionally smaller than the proportion in the population) are given design weights greater than 1.

In this example, the sample consists of three equal sized strata (33% in each stratum). Those with no qualification (22.7%) are being oversampled and thus will have design weights less than 1. Those with some qualifications (50.1%) are being undersampled, so these observations will have design weights greater than 1.

Weighted sample statistics will then provide unbiased estimates of corresponding population values.

7. Consider the following hypothetical population and sample. For each of the following six groups, what post-stratification weights would make the sample data match the population distribution of sex and religion?

   - Male, Christian

  - Population: 20%
  - Sample: 12%
- Male, Muslim
  - Population: 10%
  - Sample: 12%
- Male, Any other religion
  - Population: 20%
  - Sample: 30%
- Female, Christian
  - Population: 24%
  - Sample: 26%
- Female, Muslim
  - Population: 8%
  - Sample: 4%
- Female, Any other religion
  - Population: 18%
  - Sample: 16%

Solution:

The post-stratification is the weight applied to each unit in a sample stratum that would make the proportion of sample observations in the stratum match the population distribution of strata. So sample strata that are large relative to the population will have post-stratification weights less than 1 and sample strata that are small relative to the population will have weights larger than 1. (Note: this is different from design weights because post-stratification weights are not used to adjust for a sample *design*, but rather for the sample that happens to result from a particular design after other weighting adjustments.)

The specific weights are just ratios of population sizes to sample sizes, $\frac{N_k}{n_k}$, so the weights are:

- Male, Christian: $\frac{0.20}{0.12} = 1.67$
- Male, Muslim: $\frac{0.10}{0.12} = 0.833$
- Male, Any other religion: $\frac{0.20}{0.30} = 0.33$
- Female, Christian: $\frac{0.24}{0.26} = 0.92$
- Female, Muslim: $\frac{0.08}{0.04} = 2$
- Female, Any other religion: $\frac{0.18}{0.16} = 1.125$

In this example the post-stratification weights are all quite small, though some groups (males, any other religion) are weighted quite low while other groups (female, Muslim) are weighted quite high. Often, in practice, these weights are trimmed to avoid massively overweight a single individual or two who are made to represent a population stratum.

8. Consider the following sample data:

Obs.	X	Y	Weight
1	1	92	1.30
2	1	85	0.99
3	1	88	1.21
4	1	69	1.01
5	0	71	0.73
6	0	54	0.55
7	0	68	1.05
8	0	59	0.60

(a) If these data were collected as a simple random sample, what would the sample mean of Y be? What would its standard error be (assuming an infinitely large population and simple random sampling?

Solution:
The sample mean is:

$$\bar{y} = \frac{\sum_{i=1}^{8} x_i}{n} = \frac{586}{8} = 73.25 \tag{12}$$

The standard error is:

$$SE = \sqrt{Var(p)} = \sqrt{\frac{s^2}{n-1}} = \sqrt{\frac{1331.5}{7}} = 13.79 \tag{13}$$

(b) What is the *weighted* sample mean of Y?

Solution:

$$\bar{y}_w = \frac{\sum_{i=1}^{8} x_i * w_i}{\sum w_i} = \frac{568.25}{7.44} = 76.38 \tag{14}$$

(c) What are the mean values of Y for groups X=0 and X=1? Assuming a simple random sample and, is the difference between these means statistically significant? Recall that the formula for a two-sample t-test is the mean difference between the groups divided by the pooled standard deviation; the pooled standard deviation is: $sqrt((s\_0/(n-1)) + (s\_1/(n-1)))$.

Solution:
The sample means should be straightforward to calculate:

$$\bar{x}_0 = 63\bar{x}_1 = 83.5 \tag{15}$$

To calculate the t-statistic, it is easiest to just use a statistical software package:

```
> x <- c(1,1,1,1,0,0,0,0)
> y <- c(92, 85, 88, 69, 71, 54, 68, 59)
> t.test(y ~ x)

        Welch Two Sample t-test

data:  y by x
t = -3.2048, df = 5.6671, p-value = 0.02002
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -36.377537  -4.622463
sample estimates:
```

```
mean in group 0 mean in group 1
           63.0            83.5
```

The interpretation here is that given the large t-statistic and small p-value, the mean-difference is statistically significantly different from zero.

4 Submission Instructions

You should submit your problem set as a Word (.docx) document or PDF via Moodle.

5 Feedback

You will receive feedback within two weeks.