Problem Set 2: Trends and Toplines

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 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:**
>
> Some possible criteria you might mention include:
>
> - Representative ("design-based") sampling using a probability-based sample
>
> - Descriptive representativeness based upon demographics or other variables
>
> - Reweighting of a non-probability sample or a probability-based sample (with nonresponse) to match the characteristics of the population
>
> - Expert judgments or "fit for purpose" samples (i.e., it's "good enough" for the present purposes)

2. A key challenge in the collection of original survey data is obtaining a representative sample. The process of doing so typically involves creating a "sampling frame", which is a list of all members of the population from which you will sample a small number of individuals to interview. In practice, it is not possible to obtain a complete list of all members of a population, so sampling frames are constructed based upon close approximations. For example, to sample from the entire population of UK citizens, a researcher may create a list of all telephone numbers and randomly sample numbers to call. Or they may create a list of all households in the UK (based on local council lists) and contact a random sample of households.

   Both of these sampling frames involve problems of *coverage error*. Overcoverage is when a sampling frame includes units that are not from the target population, meaning that some people from outside the population might be included the research unintentionally. Undercoverage is when a sampling frame does not include all members of the target population, meaning some members of the population are entirely excluded from eligibility to participate in the research. In a few sentences, explain why these two sampling frames (telephone listing and list of all UK households) have both undercoverage and overcoverage. What consequences might overcoverage and undercoverage have for claims of representativeness of survey data? And how might overcoverage and undercoverage 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.

The telephone sampling frame is likely to be subject to both overcoverage and undercoverage. If our goal is to sample UK citizens, then there will be many telephone numbers that belong to other types of individuals (e.g., non-citizens, businesses) and some non-working telephone numbers (i.e., numbers not tied to any individual). There will also be undercoverage because some UK citizens may not have telephones, may have telephone numbers from outside the UK (e.g., they live in another country) or, depending on the particular way that telephone numbers are selected, may have phone numbers that are not allowed to be called.

The household sampling frame also suffers from both problems, and a problem with the unit of analysis. Overcoverage — like with telephones — will emerge when an apparent housing unit is actually not a home. For example, an apartment might be unoccupied or may be used as a business or hotel suite. Overcoverage also occurs when a household is occupied but contains no members of the population (e.g., all household members are immigrants). Undercoverage emerges when individuals in the population do not live in what might be defined as a housing unit (e.g., they live in a hotel, in a vehicle, in a prison or other institution, or are homeless).

(Among those households that actually contain living individuals from the target population, there is also the issue that the household may contain multiple individuals while other households contain only one individual. In such cases, a secondary issue (unrelated to coverage) emerges: if we select only one individual in each household to study, the probability of being sampled in a multi-member household is lower than in a single-member household; if we interview all members of a household, then the probability of being sampled in a multi-member household is higher than in a single-member household. In either case, we need to decide how to proceed either by randomly sampling one or more individuals from each household and/or reweighting our respondents to account for the fact that multiple individuals live in a given household.)

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

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. A response rate is a measure of what proportion of individuals that have been asked to participate in a research study agree to do so and complete the study. Imagine for instance a telephone survey, *respondents* are those individuals who have been asked to participate and complete the survey. *Nonrespondents* are those individuals who have been asked to participate but decline. In many cases there are also individuals that were never able to be contacted (e.g., never answered their phone when called) and are referred to by survey researchers as *unknowns*. The response rate is simply the number of respondents divided by the total number of respondents and nonrespondents; sometimes the unknowns are included in the denominator and sometimes not (because it is not known of those individuals were eligible or if, for example, those were non-working telephone numbers that were not connected to any individual), thus providing a range of possible response rates. A response rate of 1.0 means all individuals asked to complete a study do so. A response rate of 0.0 means all individuals asked to participate refused to do so.

Now, 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? Explain why or why not.

Solution: The bound on the response rate are from $\frac{1000}{2700} = 0.37$ (all unknowns ineligible) to $\frac{1000}{3300} = 0.30$ (all unknowns eligible).

There is no objective criterion for assessing whether a particular response rate is "good" or "bad." Arguably one may want to obtain the highest response rate possible, but even a relatively high response rate says nothing about "response bias" (i.e., the extent to which the set of respondents are not representative of the target population). As soon as the response rate is below 1.0, there is a possibility for response bias. Yet, even if the response rate is very low, it may be the case that respondents and non-respondents do not differ from another and thus there is no response bias or reason to be concerned about the representativeness of the final sample.

The challenge of thinking about response bias, however, is that one needs to know characteristics of nonrespondents in order to assess how they differ from respondents. If we have no information about the nonrespondents (other than, for example, where they live or what their contact information is), then making these assessments is empirically difficult.

4. When sampling from a population, the goal is to make claims about population *parameters* (e.g., vote intention, support for a policy, etc.) based upon analysis of only the (limited) sample data. As such, estimates of population parameters generated from survey samples must be expressed with a measure of "uncertainty" that communicates how much the *sample estimate* might deviate from the true population parameter. This is often conveyed through a "margin of error", such as the number of individuals that intend to vote Labour in the next general election is "27% +/- 5% percentage points" where 27% is the sample estimate and 5% is the margin of error. In this case, the sample data suggest that the proportion of voters intending to vote Labour is between 0.22 (22%) and 0.32 (32%). For a "simple random sample" of a large population, the margin of error is primarily a function of sample size. For example, the margin of error of a proportion, p, is: $MoE = 2 * \sqrt{\frac{p * (1 - p)}{n}}$. As sample size increases, the margin of error decreases. Because we do not know the value of p, we can prospectively calculate this at the value that maximizes the margin of error, which is $p = 0.5$.

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 survey sample would be needed to estimate that proportion within +/- 0.02 (or 2 percentage points)? What about within +/- 0.005 (or 0.5 percentage points)? (Show your work.)

Solution: To obtain a margin of error for the proportion that is +/- 2 percentage points, plug the values of p and MoE into the below equation and solve for n:

$$MoE = 2 * \sqrt{\frac{p * (1 - p)}{n}}$$

$$0.02 = 2 * \sqrt{\frac{0.5 * (1 - 0.5)}{n}}$$

$$0.02 = 2 * \sqrt{\frac{0.5 * (0.5)}{n}}$$

$$0.02 = 2 * \sqrt{\frac{0.25}{n}}$$

$$0.01 = \sqrt{\frac{0.25}{n}}$$

$$(0.01)^2 = (\sqrt{\frac{0.25}{n}})^2$$

$$0.0001 = \frac{0.25}{n}$$

$$0.0001 * n = \frac{0.25}{n} * n$$

$$0.0001 * n = 0.25$$

$$\frac{0.0001 * n}{0.0001} = \frac{0.25}{0.0001}$$

$$n = 2500$$

Similarly, to obtain a margin of error for the proportion that is +/- 0.5 percentage points:

$$MoE = 2 * \sqrt{\frac{p * (1 - p)}{n}}$$

$$0.005 = 2 * \sqrt{\frac{0.5 * (1 - 0.5)}{n}}$$

$$0.0025 = \sqrt{\frac{0.25}{n}}$$

$$(0.0025)^2 = (\sqrt{\frac{0.25}{n}})^2$$

$$0.00000625 = \frac{0.25}{n}$$

$$n = 40,000$$

Thus, to produce a narrower margin of error (+/- 0.5 percentage points rather than +/- 2 percentage points) requires 16 times as many respondents.

5. Researchers are often not only interested in collecting data about public opinion at one point in time or only for the population as a whole. Instead, often comparisons are made between subgroups and over-time comparisons are made about how aggregated public opinions have changed over time.

    (a) If we are interested in comparing the vote intentions of two subgroups from a sample of 2000 respondents (where Group A has 1500 respondents and Group B has 500 respondents), how much larger is the margin of error for our estimate of the proportion of "leave" voters in Group B than Group A?

Solution:

The calculations are the same as above. We can assume $p = 0.50$ unless we have a good reason to believe otherwise. Thus, for Group A ($n = 1500$), the margin of error is given by the same formula as above, but solving for MoE instead of n:

$$
\begin{aligned}
MoE_A &= 2 * \sqrt{\frac{p * (1 - p)}{n}} \\
&= 2 * \sqrt{\frac{0.5 * (1 - 0.5)}{1500}} \\
&= 2 * \sqrt{\frac{0.25}{1500}} \\
&= 0.0258
\end{aligned}
$$

For Group B, which is smaller ($n = 500$), the MoE will be larger given the fewer number of observations:

$$
\begin{aligned}
MoE_B &= 2 * \sqrt{\frac{p * (1 - p)}{n}} \\
&= 2 * \sqrt{\frac{0.5 * (1 - 0.5)}{500}} \\
&= 2 * \sqrt{\frac{0.25}{1500}} \\
&= 0.0447
\end{aligned}
$$

Thus, while the size of Group A is three times larger than the size of Group B, the margin of error for Group is only 73% larger ($0.0447/.0258 = 1.733$) than margin of error for Group A. This is because there is a "declining marginal return" to respondents; e.g., doubling the sample size does not halve the margin of error, as we saw in the previous question.

(Caveat: If the proportions p_A and p_B were different (i.e., there was actually a difference between the groups), then these margins of error might be different. The calculations here generate the maximum possible margin of error, given sample size.)

(b) If we find that the proportion of "leave" voters in Group A is 0.65 (or 65%) and the proportion of "leave" voters in Group B is 0.55 (or 55%), is this a *substantively* large or small difference? We would say this difference is *statistically significantly different from zero* if the difference between the proportion for group A, p_A, and the proportion for group B, p_B, is larger than the margin of error for the difference. The margin of error is given by: $2 * \sqrt{\frac{p_A * (1 - p_A)}{n_A} + \frac{p_B * (1 - p_B)}{n_B}}$, where n_A and n_B are the number of respondents in group A and group B, respectively. How large is the margin of error? And does that MoE imply the difference in proportions is statistically significant or not?

Solution:

The decision about whether this is substantively large depends to a large extent on what the basis for deciding that is. Is this difference large enough to swing the result of the election? Probably not, because both groups favour leave — it is only a question of by how much. Is this difference larger than other differences between groups (e.g., between the old and young or between Labour and Conservative voters)? Hard to say because we

don't have that information readily available. Without that information then we largely need to make a subjective judgement about whether this difference is large or small; it could be reasonable to argue either way provide some general justification is provided.

In terms of statistical significance, this is a pretty straightforward "plug in values" question:

$$
\begin{aligned}
MoE_{Diff} &= 2 * \sqrt{\frac{p_A * (1 - p_A)}{n_A} + \frac{p_B * (1 - p_B)}{n_B}} \\
&= 2 * \sqrt{\frac{0.65 * (1 - 0.65)}{1500} + \frac{0.55 * (1 - 0.55)}{500}} \\
&= 2 * \sqrt{\frac{0.2275}{1500} + \frac{0.2475}{500}} \\
&= 2 * \sqrt{0.00015 + 0.000495} \\
&= 0.05
\end{aligned}
$$

So the margin of error of the difference is 0.05 or 5 percentage points. Given the difference is $0.65 - 0.55 = 0.10$, then the difference is twice as large as the margin of error, which suggests that this difference is statistically distinguishable from no difference. Group A is thus statistically more supportive of leave than Group B.

If we were to report this difference, we should indicate our uncertainty by saying either that the difference is 10 percentage points +/- 5 percentage points, or that the range of possible differences is 5–15 percentage points.

(c) Another common comparison is to assess whether a sample estimate of a population parameter has changed over time. For example, we may want to assess whether support for "leave" or "remain" has increased or decreased since the last poll. Because we are only sampling from the population, some small changes will simply reflect "sampling error" (the chance variation in estimate due to sampling rather than conducting a population census). Imagine that we conduct two polls, each with a sample size of ($n = 2000$), about one month apart from each other. In the first poll, we find that the proportion intending to vote "leave" is 0.43 and in the second poll we find that the proportion intending to vote "leave" is 0.52. Is this a substantively large change? Using the formula from (b) for comparing two proportions, is this change statistically significant?

Solution:

Whether this is a substantively large change depends on the factors discussed in (b). We might consider this substantively large because it involves a switch in the majority preferred position (from the majority favouring remain to the majority favouring leave), but we might also consider it substantively large because it is similar in size to the between-group differences from (b). There may be other criteria for deciding this, as well.

In terms of statistical significance, we use the same formula as above to determine the size of the margin of error for the difference, this time comparing the two points in time rather than two subgroups:

$$MoE_{Change} = 2 * \sqrt{\frac{p_1 * (1 - p_1)}{n_1} + \frac{p_B * (1 - p_2)}{n_2}}$$

$$= 2 * \sqrt{\frac{0.43 * (1 - 0.43)}{2000} + \frac{0.52 * (1 - 0.52)}{2000}}$$

$$= 2 * \sqrt{\frac{(0.43 * (1 - 0.43)) + (0.52 * (1 - 0.52))}{2000}}$$

$$= 2 * \sqrt{\frac{0.2451 + 0.2496}{2000}}$$

$$= 2 * \sqrt{0.00024735}$$

$$= 0.0315$$

The MoE for the over-time change is 0.0315 (or about 3 percentage points). This is smaller than the margin of error from (b) because the two groups being compared are larger (both are $n = 2000$) but it is not dramatically smaller given the declining marginal return on respondents.

This means the change is nearly three times larger than the margin of error, so it is statistically significant. We should report this change with uncertainty as a 9 percentage point change (+/- 3 percentage points) in favour of "leave", or a change of between 6 and 12 percentage points in favour of leave.

(d) Just a note: If we conducted the two polls from (c) using the same set of respondents (i.e., we conducted the study as a panel), the margin of error would likely be smaller because the two sets of observations are not "independent". For this reason, it can be easier to statistically distinguish changes over-time from sampling noise in a panel design than from a "repeated cross-section design" like the one described in (c).

6. Not all samples are "simple random samples". In a simple random sample all individuals in the sampling frame have an equal probability of being sampled. In *stratified sampling designs*, some individuals are more likely to be sampled ("oversampling") and others are less likely to be sampled ("undersampling"). This can be useful for obtaining large (and therefore precise estimates) of the characteristics of small population subgroups, but then requires *reweighting* the resulting data in order for it to be representative. (E.g., calculating the unweighted mean opinion of a stratified sample that includes an oversample of young people will suggest the population is closer to the views of younger people than it actually is; younger people would need to be weighted less when calculating a weighted mean in order for that estimate to be representative of the population).

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 sampled for interviewing in each stratum), which strata are being oversampled and which strata are being undersampled? As such, if the respondents from the "no qualifications" group were weighted 1, would the weights given to respondents in the "some qualifications" and "university degree or greater" categories be smaller or larger than 1?

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.

If we analyse the data by ignoring these differential sampling probabilities, the sample estimates we generate will not reflect the population as a whole but rather a skewed population consisting of more of the oversampled individuals and fewer of the undersampled individuals. However, the correct analysis, using weighted sample statistics (e.g., calculating a weighted mean, etc.) will provide unbiased estimates of corresponding population values.

3 Submission Instructions

Please submit your answers as a PDF document via Moodle. It should be single-spaced, in Times New Roman font size 12, on A4 paper with standard 2.54cm margins. This problem set is self-assessed. A solution set will be provided on the course website and the activity will be discussed in class.

4 Feedback

Group feedback will be provided during class. If you would like more specific individual feedback on your work, please ask the instructor during office hours.