Introduction to Survey Statistics – Day 2
Sampling and Weighting

Federico Vegetti
Central European University

University of Heidelberg

Sources of error in surveys

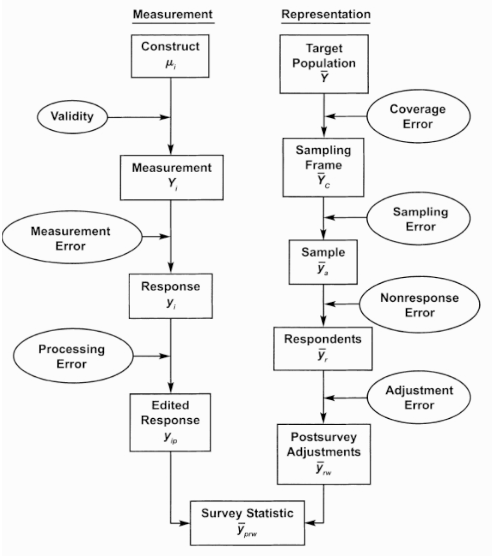

Figure 1: From Groves et al. (2009)

Representation error

- The difference between the values that we observe in the sample and the true values in the population
- It has many sources
  - Coverage, Sampling, Non-response
- **Sampling** is arguably the most relevant
- However, a similar logic applies to all of them

Two types of error

- **Bias**: when the deviation from the true value systematically goes in a specific direction
  - E.g. We want to know whether people liked the new Star Wars movie
  - We interview people leaving the Opera house after a Wagner's play
  - Our sample will probably show lower appreciation of the movie than the average moviegoer
- **Variability**: when the deviation from the true value is a random incidence
  - We sample 100 people from the phone list of Berlin, and ask them their attitude towards EU integration
  - The next day we draw other 100 people from the same list, and ask the same question
  - Most likely figures won't be identical

Sampling and variability

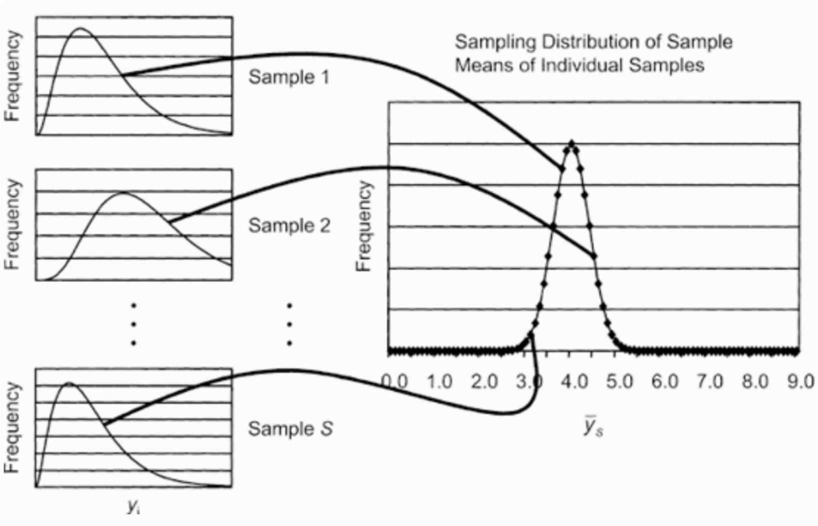

Figure 2: From Groves et al. (2009)

Standard error

- Variability *between* samples is reflected in the variability *within* the sample
- In fact, the **standard error** of an estimated parameter is interpreted as the standard deviation of such estimate across different independent samples
- It is calculated from the variance of the parameter in the sample
- It corrects by the number of observations
    - The more observations we have, the more information we have, and the more precise is our estimate

Two goals

1. Reduce the bias of the parameter estimates
2. Increase the precision of the parameter estimates

- We can do a lot to reach these goals when planning the data collection
- As a less optimal solution, we can also adjust the data after the collection, in order to make them more resemblant of the population

On inference, again

- We saw two inferences that we make when we work with survey data:
  1. From answers to questions to individual characteristics
  2. From samples to populations
- In statistics, there is a distinction between **model-based** and **design-based** inference
- To a certain extent, these two types mirror the two inferences we make with survey data

Model-based inference

- Inferences that require us to make **assumptions** regarding the process that generated the data
- Assumptions are **theories**
  - We assume/theorize that a dichotomic variable (e.g. voting/not voting) has been generated by a Bernoulli distribution
  - We assume/theorize that an outcome is a function of some predictors
- In fact we do not know what model generated the data, but we offer an approximation of reality with our theory
- As long as our assumptions are correct, our results can be generalized to other situations where the same process is at work

Model-based inference (2)

- Maximum Likelihood estimation is a classic example of model-based inference
- Our sample is assumed to be a realization of an infinite population that follows a given theoretical distribution
- Observations in the sample are linked to observations outside the sample by the assumption that they all come from the same distribution
- The parameters that we estimate from the sample are then our **best guess** about the values of the true parameters in the population *given the data*
- The sample does not need to be random, as long as we control by possible factors that make it different from the population

Model-based inference and measurement

- When we model a survey outcome (e.g. the response to a logic quiz) we assume that it has been produced by a random process that we theorize (e.g. intelligence)
- In this framework, both interpreting the output of a regression and the parametes of the distribution of a survey variable imply making a model-based inference
- The idea that measurement can be conceptualized as a statistical model where an observed outcome is a function of a hypothesized (latent) process is behind most psychometric methods

Design-based inference

- Example: a randomized experiment
  - We want to see if a drug cures depression
  - We take a pool of subjects with depression
  - We assign them randomly to either one of two groups
  - To the subjects in one group we give the actual drug, to the others we give a placebo
  - We keep them all in a clinic where they have the exact same treatment in all other respects

Design-based inference (2)

- In a randomized experiment:
  1. We know which subjects have been given the treatment
  2. We know that the only thing that differs between groups is the treatment itself
- What allows us to make a valid inference in experiments is **random assignment**
  - To make sure that the only systematic difference between the two groups is the occurrence of the treatment, we must assign units randomly to one group or the other
- In other words, we know that each unit has equal probability to end up in either one of the two groups
- This knowledge is the central point of design-based inference

Design-based inference in surveys

- Design-based inference allows us to draw conclusions about a variable in the the target population by looking at a sample and without assuming an underlying generative model
  - In other words, we can draw **descriptive** evidence directly from the sample to the population
- To be able to do so, we need to know the design that has been used to produce the sample
- This implies:
  - Knowing the sample frame (the finite population from which the sample is drawn)
  - Knowing the selection process for the observations (what rules drive the random sampling procedure)

Random samples

A **random sample** is a sample with the following characteristics (see Lumley 2010):

1. Every individual i in the sample frame has a non-zero probability π_i to end up in the sample
2. We can calculate this probability for every unit in the sample
3. Every pair of individuals i and j in the sample frame have a non-zero probability π_{ij} to end up together in the sample
4. We can calculate this probability for every pair of units in the sample

- Note that if individuals are sampled independently from each other, then $\pi_{ij} = \pi_i \pi_j$

Nonrandom samples

- When conditions 1 and 2 are not met, we have a **nonrandom sample**
- In nonrandom samples
  - We might not know the sampling frame
    - E.g. we take everyone who shows up in the lab
  - We might not be able to calculate the probabilities of selection
    - E.g. we use snowball sampling
- Nonrandom samples are very common in social science
- We can still use them to draw a model-based inference, under certain conditions (see Sterba 2009)

Simple random samples

- In a **simple random sample** we choose units at random from the entire population
- The probability of inclusion for all units is $\pi_i = n_i/N_i$
  - where n_i is the sample size and N_i the size of the sample frame
- Such probabilities serve as the basis to calculate **sampling weights**
- Weights are then calculated as $1/\pi_i$ for each unit i
- They reflect how many units in the sample frame each observation in the sample represents

Sampling weights in simple random samples (2)

- Example: we take a random sample of 1,000 respondents from a sample frame of 100,000 individuals
- For each individual, $\pi = 1000/100000 = 0.01$
- Then $1/0.01 = 100$
- Every respondent represents 100 people in the sample frame

Stratified samples

- We divide the population into groups that are
  - Internally homogeneous (with respect to specific characteristics)
  - Mutually exclusive
  - Collectively exhaustive
- We draw a random sample within each group
- This way we make sure that observations in each stratum end up in the sample
- Obviously, we need to know the stratum membership for each observation *before* we contact them

Stratified samples (2)

- Stratified samples increase the precision of the estimated parameters
  - They tend to have smaller standard errors than in simple random samples
  - **But only** when the variables for which we estimate the parameter are predicted by the variables used to stratify
- Why?
  - The precision of an estimate is always a function of the amount of information that we have
  - In stratified samples, the mere presence of an observation in the sample conveys information about some characteristics of that observation

Weights in stratified samples

- Stratified samples are simple random samples drawn within each stratum
- Hence, the probability of selection for an individual i in a stratum s is $\pi_{is} = n_{is}/N_{is}$
  - where n_{is} is the sample size and N_{is} the population size within the stratum s

Cluster sampling

- Using a random sample of the entire population may be difficult in case surveys are conducted face-to-face
- An alternative is to divide the population into clusters (e.g. districts) and take a random sample of clusters
- Then we can either:
  - Take all units inside of the cluster (single-stage sampling)
  - Sample further (multistage sampling)

Cluster sampling (2)

- Unlike stratified sampling, cluster sampling decreases the precision of the estimated parameters
- Why?
  - People in the same cluster tend to be more similar to one another (more so than people from different clusters)
  - Formally, values of respondents from the same cluster tend to be more correlated
  - With a clustered sample, the correlation between units will be on average higher
  - Hence, the information that we get from each respondent will be a bit less than with a random sample of the full population
- This is less of a problem the more the clusters are similar to one another

Weights in clustered samples

- In single-stage cluster sampling, the probability π_i that an individual i is sampled is equivalent to the probability π_c that the cluster c to which the individual belongs is sampled
  - Where $\pi_c = n_c/N_c$
  - n_c is the number of sampled clusters
  - N_c is the total number of clusters in the sample frame
- In multistage sampling, π_i is also a function of the probability π_{ic} that i is sampled within the cluster c so that $\pi_i = \pi_c \pi_{ic}$
  - Where $\pi_{ic} = n_{ic}/N_{ic}$
  - n_{ic} is the sample size
  - N_{ic} is the population size within the cluster c

What do we do with weights?

- We may need weights to calculate sample statistics, especially if we want to obtain descriptive statistics about the sample
  - For instance, if we have a stratified sample, weights allow us to compute unbiased and efficient (i.e. with high precision) parameter estimates
- We can adjust the sample weights to correct for deviations of the sample from some (known) parameters of the population

Horvitz-Thompson estimator

- Estimates of the **population total** are the basis for most other more complex statistics
- The Horvitz-Thompson estimator is a method used to calculate the population total (and its standard error)

$$\hat{T}_X = \sum_{i=1}^{n} \frac{1}{\pi_i} X_i$$

- Where:
  - X_i is the measurement of variable X for respondent i
  - π_i is the probability of inclusion for respondent i

Horvitz-Thompson estimator (2)

- From here we can obtain, for instance, the estimated **population mean** of X by dividing $\hat{T}_X$ by the population size N

$$\hat{\mu_X} = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}X_i$$

- Which *in a simple random sample*, is equivalent to the sample average

$$\hat{\mu_X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$

- In a stratified sample, the formula for $\hat{\mu_X}$ produces what is often called the **weighted mean** of X, which is an unbiased and efficient estimator of the population mean

Post-stratification

- Suppose we have a sample where females are 48% and males are 52%, but we know that in the population females are 52% and males are 48%
- If our sample was stratified on sex, this difference in proportion would be reflected in the weights
- However
  - The sample can not be stratified on everything
  - Nonresponse patterns may be different between groups
  - Group proportions in the sample may end up being different from the ones in the population by chance
- Even in these cases, we can adjust the weights so that groups have the same proportion that they would have in a stratified sample
- This adjustment is called post-stratification

Post-stratification (2)

- When we apply post-stratification, we substitute the sampling weights $1/\pi_i$ with g_i/π_i
  - Where $g_i = N_k/\hat{N}_k$
  - N_k is the population size in the group (or stratum) k
  - $\hat{N}_k$ is the Horvitz-Thompson estimator of the population size in the group k
- In other words, we change the values of the weights so that the group size in the sample matches the group size in the population

Raking

- We may need post-stratification to be performed for more than one variable
- This is more often the rule than the exception
- Ideally we would need a complete cross-classification of the variables
  - E.g. Males of age 18-24 and low education, males of age 18-24 and high education, etc.
- However, some resulting combinations may be so untypical that nobody ends up sampled in those categories
- Raking is an iterative procedure that allows to post-stratify on multiple grouping factors without the need for a full cross-classification

Final remarks

- Note that the use of weights and of post-stratification adjustments is necessary to have unbiased estimates of population parameters under a design-based inference paradigm
- When we make a model-based inference, what counts is that our model is correctly specified
- This usually implies
  - Assuming the correct data generating process for the outcome variable
  - Assuming a correct specification for the function predicting the outcome variable
- In regression models, we often include as predictors the variables that in design-based inference we use to post-stratify

References

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2 edition. Hoboken, N.J: Wiley.

Lumley, Thomas. 2010. *Complex Surveys: A Guide to Analysis Using R*. 1 edition. Hoboken, N.J: Wiley.

Sterba, Sonya K. 2009. "Alternative Model-Based and Design-Based Frameworks for Inference from Samples to Populations: From Polarization to Integration." *Multivariate Behavioral Research* 44 (6): 711–40.