Survey data analysis – week 41: specifying your own survey design + cheat sheet

In the Take Home Exercise of week 40 you worked out the sampling design of your adopted survey. In some cases your survey is perhaps a simple SRS design, while in other cases you perhaps have a design with multistage sampling including both clusters and strata. The take home exercise of next week is to try to correctly specify the sampling in design. In order to help you do this, take the following steps

1. Is your survey SRS or using a design with unequal inclusion probabilities?
   ➔ If SRS: specify the svydesign object, you're done.
   ➔ If not SRS: go to step 2.
2. Did you find the stratification/cluster variables so you can specify the design in R?
   ➔ If yes:
      a. Use the cluster/ stratifying indicator in your sampling design
      b. Specifying the ~fpc can be complicated, because you will need to know the total number of clusters in the population and/or the population size for each stratum. In some cases, the survey documentation does specify these numbers, but often it will not. In these cases, you either have to some research about the specific population you study. In some cases (e.g. when your population is the adult population of the U.S.A. you can treat the population size to of "infinite size" (because fpc doesn't matter). In other cases, you may be able to find register data on your strata (e.g. a survey stratified on 'province' or 'council' in the Netherlands.
   ➔ If no: step 3.
3. Possibly, your survey does not include the stratification and cluster variables. In those cases, information should be available in the form of a "design weight" or "sampling weight". Some surveys also include multiple weights correcting for nonresponse (also called "poststratification weight") or population size equivalence weights. Read the documentation carefully to understand which weight you will need to use in your case.
      a. Look at how to use the Horvitz-Thompson estimator, and include the weight as inclusion probabilities in the survey design object instead of clusters and strata.
      b. In some other instances, there is only a "total weight" available, that combines design and nonresponse weights, and that usually sums up to either the sample size or population size. In this case you typically specify the weight under the `weight` command in the `svydesign()` function (see also next page).

Bring your survey design (in R) (and any problems you encounter) to class next week. If completed, you are all set to go and work on assignment 2 after we have discussed deadling with missing data due to nonresponse.

Please see the cheat sheet on the next page or read the following article for tips on what to look for in survey documentation:

Kolenikov, S., West, B.T., and Lugtig, P. (2020). A checklist for assessing the analysis documentation for public-use complex sample survey data sets. *The Survey Statistician, 81*, 50-62. Or https://statstas.shinyapps.io/svysettings/

Cheat-sheet for specifying your survey design object in R

In order to draw correct inferences in R, you need to always first specify a survey design object using the `svydesign()` function. After this, you can compute the correct statistic under your (complex) sampling design.

```
Svydesign(ids=~id, strata=~strata, weights=~weights, fpc=~fpc,
data=data)
```

ids:

- in an SRS this is simply specified as `id=~1`
- in a one-stage cluster sample, you specify the design as `id=~[clustervar]`
- In a two-stage (or multistage sample involving clusters), you include multiple ids, as `id=~[clustervar1]+[id]`.

strata:

- when your dataset includes a stratification variable, include that here as `strata=~[stratavar]`

fpc:

- in an SRS, ~fpc is equal to the population size. You typically need to add the fpc as a variable to the survey dataframe you are going to analyze.
- In a stratified sample, the fpc is equal to the population size within every stratum.
- In a cluster sample, you need as many fpc variables as you have levels in your study. The first fpc specifies the number of clusters in your population at the highest level.

Weights:

Often, datasets do no include all the information on strata or clustering variables. This may be for various (often not very good) reasons, but one possible good reason is that including a cluster id may lead to disclosure of certain individuals in the dataset. If this is the case, the dataset should include a "weight" variable to prevent disclosure. The weight is usually either a sampling weight (a standardized weight that is equal to the inverse of the inclusion probability of every case), or a population weight (a weight that directly tells you many elements in the population a case represents (also based on inverse inclusion probabilities). Using a sampling weight instead of `clustervar` and/or `stratavar` will automatically result in the use of the Horvitz-Thompson estimator.

- Specifying a sampling weight is done with `weights=~[weightvar]`.

Data:

This is your dataset as a data.frame object.

- `data=data`

Note that many datasets also include a nonresponse weight (see week 44) or a "totalweight", being the product of both the design and nonresponse weight. So be careful what weight you are using.

Finally please note that there are various other options, that would be relevant in specific sampling designs. For example, the `nest` option allows clusters to be nested in strata, the `variables` command can be used to only include certain variables from the data (handy if you have a very big dataset). Another option that is important in cluster sampling is the `probs` argument. Here, you can specify that the sampling probability of clusters (if not an SRS of clusters).