Subject: Cluster identifiers and finite population correction
Posted by FanuelShem on Tue, 24 Nov 2015 10:19:18 GMT
View Forum Message <> Reply to Message

Hi,

I am using R to analyze DHS data (from Zimbabwe,Tanzania and others) on domestic violence, specifically spousal violence. To analyze such complex survey in R, one need to use the survey package. My analyses goes as follows:

#import and subset the data to get only respondents who received the Domestic Violence module
data=subset(zim,V044=="Woman selected and interviewed")
#initiating the suvydesign object
samplewt <- D005/1000000

```
(a) mydesign <-
    svydesign(
      id = ~V001+V002,
      strata = ~V022 ,
      data = data ,
      weight = ~samplewt,
   )
```

From this design, id is the cluster identifier, strata is the variable specifying strata, weight is the variable specifying sampling weights and data is the data frame.

Design (a) looks like this: Stratified 2 - level Cluster Sampling design (with replacement)
With (406, 6542) clusters. svydesign(id = ~V001 + V002, strata = ~V022, data = data, weight = ~samplewt)

Going through the Zimbabwe DHS report, I understand that DHS used stratified, two-stage cluster design implying that each stage of cluster has an identifier as V001 for first stage and V002 for second stage, which led to my choice of the cluster identifiers in the design stated above. However going through other posts on this forum, I realized that most analysts use STATA and their design looks like this if I am to do it in R:

```
(b) mydesign <-
    svydesign(
     id = ~V021,
     strata = ~V022 ,
     data = data ,
     weight = ~samplewt,
   )
```

Where V021 is the primary sampling unit. Which one of these designs is correct?

Design (b) looks like this : Stratified 1 - level Cluster Sampling design (with replacement)
With (406) clusters.
svydesign(id = ~V021, strata = ~V022, data = data, weight = ~samplewt)

About the finite population correction(fpc), which variable in the DHS data defines fpc? I have read that fpc is not often used when analyzing DHS data, is it okay to ignore fpc? ignoring fpc results into a design in which the sampling is with replacement.

I was able to replicate the values in table 16.10 on page 263 of the Zimbabwe DHS report using design (a). However, I was not able to replicate the values on table 16.9 on page 280 of the Tanzania DHS report. Any ideas about this as well would be great.

Many thanks,

Subject: Re: Cluster identifiers and finite population correction
Posted by Liz-DHS on Wed, 02 Dec 2015 19:48:50 GMT
View Forum Message <> Reply to Message

Dear User, Here is a response from Sampling Expert, Dr. Mahmoud Elkasabi:
Quote:Most of the DHS surveys are based on a 2-stage design, in which sampling clusters are selected as PSU's in the first stage and households are selected in the second stage. To properly calculate the sampling variance, only the sampling clusters should be declared in the survey package. Therefore, I would say design (b) is the correct one

```
mydesign <-
svydesign(
id = ~V021,
strata = ~V022 ,
data = data ,
weight = ~samplewt,
)
```

The sampling clusters can be id = ~V021 or id = ~V001. In most of the countries, both should be the same.

Regarding the fpc, for many reasons we do not provide the fpc value in the results report. Since we are dealing with national population surveys, the fpc value should be very small. Therefore, you can ignore the fpc in calculating the sampling errors and it should not affect the results.

Subject: Re: Cluster identifiers and finite population correction
Posted by FanuelShem on Thu, 10 Dec 2015 09:37:51 GMT
View Forum Message <> Reply to Message

Hi Liz-DHS,

Thank you so much. This is helpful.

Regards,