Subject: Stratification
Posted by hlyons on Tue, 04 Feb 2014 01:09:56 GMT
View Forum Message <> Reply to Message

Hello -- I was wondering if there are some known "strata" variable errors in some of the surveys? My example is a random one, Zambia DHS IV:

Using the children's recode and R for computing, here is a somewhat fake example at the bottom -- fake in the sense the outcome (urban/rural split for children) is not something I'm really looking at. Basically, it looks like there are too many strata (v022). A less theoretical example would be DPT3 coverage nationally for urban and rural areas -- I think the standard errors for urban in the final report are more compatible with using v022. It's hard to say for sure because I do get some differences trying to duplicate the results in the final report. Maybe I'll write another post about that later...

Thoughts, anybody?

Thanks!

Hil

```
# R example
library(survey)
tmp.data = read.dta("ZMKR42FL.DTA")

# PSU's, checks out: 320
length(unique(tmp.data$v021))

# province, check out: 9
length(unique(tmp.data$v024))

# province and urban/rural combinations: 18
nrow(unique(tmp.data[,c("v023","v025")]))

# strata: 153 instead of 18
length(unique(tmp.data$v022))

# example of standard errors under two designs
# first, stratify on v022; second, on province and u/r combination
DHSdesign.v022 = svydesign(id = tmp.svy$v021, strata=~tmp.svy$v022, weights =
tmp.svy$v005/1000000, data=tmp.data)
DHSdesign.prov.ur = svydesign(id = tmp.svy$v021, strata=~tmp.svy$v023+tmp.svy$v025,
weights = tmp.svy$v005/1000000, data=tmp.data)

# proportion urban amongst children
svymean(~v025, design = DHSdesign.v022) #SE = 0.0099
svymean(~v025, design = DHSdesign.prov.ur) #SE = 0.0227
```

Subject: Re: Stratification
Posted by Bridgette-DHS on Wed, 26 Feb 2014 14:20:00 GMT
View Forum Message <> Reply to Message

We are working on a response to your posting. Thanks!

Subject: Re: Stratification
Posted by hlyons on Thu, 27 Feb 2014 19:52:55 GMT
View Forum Message <> Reply to Message

Thank you! I think one can fine other surveys/files in the whole DHS set in which this occurs so it will be helpful to understand the appropriate design. Whatever the answer is, can the team also advise on whether this answer should extend to other DHS surveys in which this occurs or whether this should be answered on a case-by-case basis?

Subject: Re: Stratification
Posted by Trevor-DHS on Fri, 28 Feb 2014 16:40:03 GMT
View Forum Message <> Reply to Message

DHS used to use an approach of producing implicit strata, based on combining clusters into pairs or small groups of 3 clusters. These implicit strata were used in the calculation of sampling errors. This approach was used in the World Fertility Surveys and in earlier DHS surveys were there was an implicit stratification based on an ordering of the clusters. The DHS sampling experts no longer recommend this approach for the stratification. v022 in the Zambia DHS IV survey contains the variable that has this pairing or grouping, as was used for the sampling error calculations reported in final report.

You will find this true in many DHS surveys, particularly the older ones.

In cases like this, check also v023 to see if this contains the strata. In many surveys this actually contains the regions rather than the strata. If the strata are not actually provided directly in the data set, then check appendix A of the final reports to see what strata were actually used. In many cases the strata will be urban and rural areas within region (in the case of the Zambia survey, urban and rural areas within province). However, in some surveys there is no stratification beyond the region, while in other surveys, there are 3 or 4 strata within each region. Appendix A of the final report should provide the information for this.

I hope this helps.

Subject: Re: Stratification
Posted by hlyons on Fri, 28 Feb 2014 19:35:11 GMT
View Forum Message <> Reply to Message

Thanks Trevor, this explanation does help. In terms of prescription, though, should one use the

strata as coded or the strata as described in the report? I interpret implicit strata to mean an organization and classification of PSU's undergone sometime after selection/sampling of PSU's (correct me if I'm wrong). I would want to use the strata that best reflects the true sampling design and my suspicion is that these should be generally region + urban/rural regardless of what strata are actually listed, which might have been an ex post facto creation.

Thanks again!
Hil

P.S. What this also implies to me is that reports that use strata not originally in the design may also have incorrect or inadvisable standard errors (though maybe not far off)? I can't recall off the top of my head whether with-replacement standard error estimation will be somewhat robust to overdoing the strata for typical estimators (totals, ratios).

Subject: Re: Stratification
Posted by Trevor-DHS on Mon, 03 Mar 2014 17:12:08 GMT
View Forum Message <> Reply to Message

The implicit stratification is an ordering that has been made prior to selection of the clusters, and it is that ordering that permits the implicit stratification. However, I would suggest using the explicit strata defined at the time of the sample design - that is what DHS does now. The strata are generally the urban and rural areas in each region, but not always. Read appendix A of the final reports to confirm what is used for each survey.

I would not say that the sampling errors are incorrect or inadvisable. With using the implicit stratification, the assumption is that there is some improvement over simply randomly selecting the sample in a strata, thus the sampling errors tend to be slightly smaller and the confidence intervals slightly narrower. However, DHS has decided to be more conservative in its estimates of sampling error and confidence intervals and using the original design strata, you will get slightly wider confidence intervals. Of course, the difference is typically tiny and frequently negligible.

Subject: Re: Stratification
Posted by hlyons on Mon, 03 Mar 2014 18:12:42 GMT
View Forum Message <> Reply to Message

Thank you Trevor, that is the specific guidance I was hoping for. I can't say I completely understand the implicit strata, though I can imagine that an ordering along some field might allow grouping that reduces between cluster variability within implicit strata, thus reducing SE. Is there an available document describing and justifying the old practice? If there isn't, that is fine -- just asking for curiosity's sake. It may reside in one of the standard references, or just be a variation on standard methods applicable when no fpc is used.

Thanks again Trevor and DHS team, very helpful instruction.

Hil

Subject: Re: Stratification
Posted by Trevor-DHS on Mon, 03 Mar 2014 19:49:49 GMT
View Forum Message <> Reply to Message

I can't find anything that I can provide that really gives justification for this. The approach was used throughout the World Fertility Surveys (WFS) with systematic sampling, and was continued for a considerable time in DHS. You can find reference to implicit stratification in http://unstats.un.org/unsd/demographic/meetings/egm/Sampling _1203/docs/no_2.pdf. you may also find more if you search for "implicit stratification in sampling" on Google.

The documentation for the "Clusters" software used by the WFS to calculate sampling errors contains the following statements:
"It should be noted that the strata used for computation of sampling errors are not necessarily identical to the original explicit strata used in the sample selection. The difference between the two may arise for two main reasons:
1) Whenever PSUs are selected by systematic sampling* from an ordered list, adjacent units should be paired or grouped to form new smaller strata which are used for sampling error computations.
2) Sampling error computations require that there be at least 2 PSUs per stratum. Any strata for which only one PSU has been selected must be 'collapsed' together to form pairs (or other groups) of PUSs. These constitute new strata to be used for sampling error computations. Such grouping is done on the basis of characteristics of the whole strata population (pairing most similar strata), and not on the characteristics of the selected PSUs. Collapsing of the strata in this way leads to slight over estimation of the sampling error.
* 'systematic sampling' means selection, at a fixed interval, from a list starting from a randomly determined point."

This doesn't provide a justification, and unfortunately I only have a very old paper version of the documentation (that is falling apart), but does describe the procedure that was used in WFS and the early part of DHS.

Subject: Re: Stratification
Posted by Trevor-DHS on Mon, 03 Mar 2014 20:13:52 GMT
View Forum Message <> Reply to Message

I realized that the following older DHS document also proivides some information:
http://dhsprogram.com/pubs/pdf/AISM5/DHS_III_Sampling_Manual .pdf

Subject: Re: Stratification
Posted by hlyons on Mon, 03 Mar 2014 23:08:00 GMT
View Forum Message <> Reply to Message

Thanks for the additional info. I had never encountered it before, but it seems to be (or have been) a pretty common (even standard?) method with systematic sampling where within an explicit stratum the frame is usually sorted geographically. Doing a search, there might be further

description of it in Kish's Survey Sampling which I don't have. Anyway, thanks for the tip and docs.

Hil