Subject: Clubbing individual recode and mens recode file to calculate overall prevalence
Posted by sarizwan1986 on Fri, 12 Jan 2018 12:20:00 GMT
View Forum Message <> Reply to Message

Hi

I have been working on DHS datasets of India for quite some time now. The latest NFHS4 was released yesterday and i had crack at it.
I wanted to calculate the prevalence of any tobacco use among men, women and both combined.
I used the files IAMR71FL and IAIR71FL and mapped the common variables between them and created a separate combined file using the STATA append command.
The variables of interest were 463a thru 463g and sm609e sm609c (in the mens recode file) and s710c an s710e (in the individual recode file) and i created a single variable called anytobaccouse if the answer to any of the above was yes (1), else (0).

I calculated the proportions using the following STATA command:
proportion anytobaccouse [pweight=v005/1000000], over(sex)

The proportion was 44% in men, 6% in women (which are almost the same as reported in the India factsheet) but what i could not believe or understand is the proportion in both sexes combined which was 11% (which was not reported in the factsheet). I expected a proportion in the ballpark of half the sum of men and women values (more like (44+6)/2 ~ 25%).

I noticed that the sample size was about 700,000 for women and about 100,000 for men, which can explain the above 11% but i thought this was supposed to be corrected by applying the weights.
I have attached a file where i show the weighted and unwieghted numbers for your reference.

What am I missing?

1. Was combining the above mentioned files appropriate? if yes, does applying the weight correct for the sample size imbalances between men and women?
2. How do I get the combined sexes prevalence if combining the files is not appropriate?

Thanks for reading my query and really appreciate your time and effort.

File Attachments
1) dhs forum.docx, downloaded 406 times

Subject: Re: Clubbing individual recode and mens recode file to calculate overall prevalence
Posted by Bridgette-DHS on Tue, 23 Jan 2018 12:13:00 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS Stata Specialist, Tom Pullum:

There are a couple of issues. First, your measure of tobacco use is defective. There is a binary variable in the IR file (v463z) and in the MR file (mv463z) that takes the value 1 for "no tobacco use" and 0 otherwise. Therefore "anytobaccouse" should be defined as 1-v463z or 1-mv463z. Second, this is one of several surveys with only a subsample of men. Fifteen percent of the households have hv027=1, meaning "household selected for male interview". The other households have hv027=0. If the household was selected, then all men in the household (who satisfied other eligibility requirements such as age) would be interviewed. If hv027=0, no men would be interviewed. It can happen that a household has hv027=1, but there were no eligible men in the household. For that reason, you cannot identify all of the households with hv027=1 by just looking at the MR file. The HR and PR files are the only files that include hv027.

There are at least four ways to estimate the combined percentage of men+women who smoke. The first way is to restrict to men and women in the households with hv027=1. That requires merging the IR and MR files with the PR file and selecting the men and women who are in such households. Using weights, I then get anytobaccouse percentages of 6.8% for women, 45.5% for men, and 25.4% for women and men combined. This is a good estimate but it ignores most of the women, who were in households with hv027=0.

A second way is to multiply the weight for men in the households with hv027 by 6.67 or 1/.15, because their probability of selection was one-sixth 15% as high as that for women. I then get percentages of 6.8% for women, 45.5% for men, and 26.8% for women and men combined.

I will paste below the Stata code for the first two approaches.

A third way is to inflate the weights for men by a "post-stratification" factor rather than by a simple factor of 1/.15. We do not advise this because of the complexity of the sampling design for this survey.

Finally, a fourth way to approach this would be to calculate the percentage for women, using all the women, calculate the percentage for men, using all the men, and then estimate the pooled mean with a calculator or spreadsheet. You could use census data to estimate the numbers of men and women in the population in the specified age interval. If a fraction f of the population is female and a fraction m is male (f+m=1), and Pf is the percentage of women who use tobacco and Pm is the percentage of men who use tobacco, then calculate f*(Pf) + m*(Pm). That will be a good estimate of the pooled mean for adults (men and women combined) in the age interval.

These procedures would apply to any outcomes that are obtained from both men and women and surveys that involve a subsample of men. The question is actually very general.

```
* Prepare IR file for merge
use e:\DHS\DHS_data\IR_files\IAIR71FL.dta, clear
gen anytobaccouse_women=1-v463z
keep v001 v002 v003 v005 any
gen hv001=v001
gen hv002=v002
gen hvidx=v003
sort hv001 hv002 hvidx
```

```
save e:\DHS\DHS_data\scratch\IAIRtemp.dta, replace

* Prepare MR file for merge
use e:\DHS\DHS_data\MR_files\IAMR71FL.dta, clear
gen anytobaccouse_men=1-mv463z
keep mv001 mv002 mv003 mv005 any
gen hv001=mv001
gen hv002=mv002
gen hvidx=mv003
sort hv001 hv002 hvidx
save e:\DHS\DHS_data\scratch\IAMRtemp.dta, replace

* Prepare PR file for merge
use e:\DHS\DHS_data\PR_files\IAPR71FL.dta, clear
* hv027: household selected for male interview
keep hv001 hv002 hvidx hv005 hv104 hv027
sort hv001 hv002 hvidx

* Merge IR with PR
merge hv001 hv002 hvidx using  e:\DHS\DHS_data\scratch\IAIRtemp.dta
drop _merge
sort hv001 hv002 hvidx

* Merge MR with IR+PR
merge hv001 hv002 hvidx using  e:\DHS\DHS_data\scratch\IAMRtemp.dta
drop _merge

* This file is MR+IR+PR
gen anytobaccouse=. replace weight_adjusted= 6*weight if hv104==1 & hv027==1
summarize any* [iweight=weight_adjusted/1000000]
* The estimates are 6.8% (women), 45.5% (men), 25.8% (women+men)

replace anytobaccouse=anytobaccouse_men   if hv104==1
replace anytobaccouse=anytobaccouse_women if hv104==2

gen weight=.
replace weight=mv005 if hv104==1
replace weight= v005 if hv104==2

* Calculate estimate for women and men combined using only the cases with hv027=1
summarize any* [iweight=weight/1000000] if hv027==1
* The estimates are 6.8% (women), 45.5% (men), 25.4% (women+men) if limited to the
households with hv027=1

* Calculate estimate for women and men combined using all cases but re-weighting the men
summarize any* [iweight=weight/1000000]
gen weight_adjusted=weight
replace weight_adjusted= (1/.15)*weight if hv104==1 & hv027==1
```

summarize any* [iweight=weight_adjusted/1000000]
* The estimates are 6.8% (women), 45.5% (men), 26.8% (women+men)