Subject: Why I am getting different total observations when using iweight for tabulating a variable
Posted by sujata on Tue, 07 Feb 2023 17:48:03 GMT
View Forum Message <> Reply to Message

I am trying to tabulate the hv025 variable in the PR file for the Indian state of Punjab. Applying svyset and tabulating this variable only gives the proportions and not the absolute values. I understand that svy: ta hv025 and ta hv025 [iw=shweight/1000000] will give the same results if we are not interested in standard errors. But the total observation (67913.878) differs from 67856 (the total number of observations). I am putting here both the results. 67913.878 is the population size. How is the population size different from the number of observations (68549)? If I apply aweight then I am getting the Total as equal to the total number of observations, which is 68549. But I should use iweight and not aweight.

```
gen weight_dis=shweight/1000000

ta hv025 [iw= weight_dis]

    type of |
   place of |
  residence |      Freq.     Percent        Cum.
------------+-----------------------------------
      urban |26,122.0845       38.46       38.46
      rural | 41,791.793       61.54      100.00
------------+-----------------------------------
      Total | 67,913.878      100.00


svyset [pw= weight_dis], psu( hv021) strata( hv022)

svy:ta hv025
(running tabulate on estimation sample)

Number of strata =  88                        Number of obs   =     68,549
Number of PSUs   = 915                          Population size = 67,913.878
                                    Design df       =        827


----------------------
type of   |
place of  |
residence | proportion
----------+-----------
    urban |      .3846
    rural |      .6154
          |
    Total |         1
----------------------
Key: proportion = Cell proportion
```

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by Bridgette-DHS on Thu, 09 Feb 2023 14:11:41 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

When Stata sees "pweight", which is the only type of weight you can use with svyset, it normalizes them to have a mean of 1. Stata does not automatically normalize iweights.

I opened the PR file and entered "tab hv024, summarize(shweight)". I see that the mean of shweight in Pujab (hv024=3) is 989497.01, which after division by 1000000 is .98948701. What's relevant is that this mean is NOT 1. Stata, with pweight, will re-scale to 1. With iweight it will NOT re-scale to 1.

So why does the mean of shweight differ from 1 (or 1000000) in each of the states? It's because DHS has normalized shweight in the HR file, not the PR file. I confirmed that by opening the HR file and entering "tab hv024, summarize(shweight)". Sure enough, the mean of shweight is 1000000 in the HR file.

Thus the discrepancy you observe is just due to the way that DHS normalized shweight for households rather than units, and you are using the PR file, with individuals as units, and Stata (with pweight) has re-normalized shweight. Hope this makes sense. Interesting question.

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by sujata on Thu, 09 Feb 2023 16:45:32 GMT
View Forum Message <> Reply to Message

Dear Tom,
Thank you very much for your reply.
I still have one query regarding the number of observations and population size. Even without applying svyset if I tabulate hv025 using iweight=shweight/1000000 the total number of observations is less.

gen weight_dis=shweight/1000000

ta hv025 [iw= weight_dis]
type of |
place of |

residence | Freq. Percent Cum.
------------+-----------------------------------
urban |26,122.0845 38.46 38.46
rural | 41,791.793 61.54 100.00
------------+-----------------------------------
Total | 67,913.878 100.00

Here Total is showing 67913.878, which is less than the actual number of observations, which is 68,549. Applying aweight gives a total of 68549. should I apply aweight or I am doing something wrong? Do I need to merge PR with HR and use shweight from HR file?

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by Bridgette-DHS on Thu, 09 Feb 2023 18:12:35 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

I am pretty sure that this is another manifestation of the fact that hv005 is normalized (so that total units = total weight, except for the factor of 1000000) for households, not individuals. That is, if you did the same thing in the HR file, you would get the consistency you expect, even though you do not get it in the PR file. You are ok--the discrepancy is small and we (the DHS research team) ignore it. If you want, you can re-normalize hv005 in the PR file by multiplying by a factor. The following lines will construct a normalized weight for the PR file.

```
gen unwtd=1000000
total unwtd hv005
matrix B=e(b)
matrix list B
scalar list sfactor
gen hv005_PR=round(sfactor*hv005)
```

I usually advise against using aweight. You can read what Stata says ("help weight") about aweight.

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by sujata on Fri, 10 Feb 2023 06:31:04 GMT
View Forum Message <> Reply to Message

Dear Tom,
I tried to normalize the weights in PR file using the command you mentioned here. But I am getting an error that the PR file doesn't contain sfactor. How do I get that?

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by Bridgette-DHS on Fri, 10 Feb 2023 12:58:18 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

Very sorry--somehow the line "scalar sfactor=B[1,1]/B[1,2]" was dropped. Please see below. Hope this will do what you want.

```
gen unwtd=1000000
total unwtd hv005
matrix B=e(b)
matrix list B
scalar sfactor=B[1,1]/B[1,2]
scalar list sfactor
gen hv005_PR=round(sfactor*hv005)
```

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by sujata on Fri, 10 Feb 2023 15:49:57 GMT
View Forum Message <> Reply to Message

Dear Tom,
Thank you very much.
Now the issue is resolved.

Best regards,
Sujata

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by sujata on Thu, 30 Mar 2023 11:02:05 GMT
View Forum Message <> Reply to Message

Dear Tom,
In addition to the above queries, I am facing another issue while applying the normalized weight to

where j varies from 0 to i-1. for the same, I applied the following commands, but the mean of the fractional rank is not exactly 0.5. It is 0.4857. sv271 is the wealth index factor score for state-level studies. My study is on the Indian state of Punjab. I am following the world bank document "Analyzing health equity using household survey data" for your reference.

```
sort sv271s
```

```
egen raw_rank=rank(sv271s), unique
sort raw_rank

qui sum wgt_shweight_PR
gen wi = wgt_shweight_PR/r(sum)
gen cusum = sum(wi)
gen wj= cusum[_n-1]
replace wj=0 if wj==.
gen rank_CE=wj+0.5*wi
```

here wgt_shweight_PR is generated so that the mean is equal to 1. below are the commands used to normalize the weights in PR file:

```
gen unwtd=1000000
total unwtd shweight
matrix B=e(b)
matrix list B
scalar sfactor=B[1,1]/B[1,2]
scalar list sfactor
gen shweight_PR=round(sfactor*shweight)
gen wgt_shweight_PR= shweight_PR/1000000
```

Please let me know where I am going wrong.

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by Bridgette-DHS on Thu, 30 Mar 2023 15:04:06 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

I spent some time looking into your question but can't provide much help. Here are some thoughts.

First, wealth scores such as sv271 are household-specific and are constructed with the HR file. Then in the PR and other individual-level files they are exactly the same for everyone in the same household. When you calculate the fractional rank, using the PR file, you are basically dividing the household's rank by the number of people in the household. I don't know why you would do that. It would seem better to me to use the HR file and skip the calculation of the fractional rank.

Second, I don't know why you would expect the mean of the fractional rank to be 0.5. Is there a mathematical reason for this? Your formula for the fractional rank is not clear to me but I don't see a mathematical reason why the mean would be 0.5.

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by sujata on Fri, 31 Mar 2023 11:56:29 GMT
View Forum Message <> Reply to Message

Dear Tom,
Thank you very much for your response. Since my study is at the individual level, I am using a PR file. I am looking into wealth-related inequality in diabetes using a PR file. For the same, I am computing a fractional rank for individuals with individuals i=1,...,n ranked by the socio-economic status indicator in ascending order. Here, the socioeconomic indicator is the wealth index factor score.
I am attaching a screenshot of the content that I am referring to, which states that the mean of the fractional rank would be exactly 0.5 using the commands that I applied.
My analysis includes a semiparametric extension of the Wagstaff index comparing the values of indices across different districts.

File Attachments
1) Screenshot (177).png, downloaded 344 times

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by Bridgette-DHS on Fri, 31 Mar 2023 16:03:26 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

This is outside the scope of the forum but I'll suggest what MAY be going on. Two conventions regarding the normalization of weights may be in conflict. In the page you attached, the weights are normalized to add to one. That's one convention. However, with pweights (are you using pweights?) Stata automatically normalizes the weights to have a MEAN of 1. So far as I know, with pweights it is actually impossible to over-ride that. Stata does that so the weighted and unweighted totals will match, which is usually desirable from a sampling perspective.

I don't have time to do this, but if I did, I would go over the algebra of the fractional weights to confirm mathematically that the mean should be 0.5 under the first convention and then see what happens when a small set of maybe 10 numbers is analyzed with Stata.

Another option would be to ask one of the authors of the World Bank report. Sorry we can't be more helpful.

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by sujata on Sat, 01 Apr 2023 06:23:27 GMT

Dear Tom,
Thank you very much for looking into this.
I understand that this is outside the forum's scope, and I really appreciate that you spared some time for this.
However, I wanted to clarify further my understanding of how to treat the weights in my analysis. I want to ensure that I use them correctly and get accurate results.

Firstly, As per your suggestion, I normalized the data so that the mean of the shweight_PR in the PR file is equal to 1000000.

```
gen unwtd=1000000
total unwtd shweight
matrix B=e(b)
matrix list B
scalar sfactor=B[1,1]/B[1,2]
scalar list sfactor
gen shweight_PR=round(sfactor*shweight)
```

After that, I generated wgt_shweight_PR= shweight_PR/1000000. The mean of wgt_shweight_PR is 1.

```
svyset [pw= wgt_shweight_PR ], psu( hv021) strata( hv022)

sum wgt_shweight_PR

Variable              Obs     Mean     Std. dev.    Min      Max

wgt_shweig~R      52,682    1     .6350303    .05442   4.638086

egen raw_rank_CE=rank(sv271s), unique
sort raw_rank_CE
qui sum shweight_PR
gen wi = shweight_PR /r(sum)
gen cusum = sum(wi)
gen wj= cusum[_n-1]
replace wj=0 if wj==.
gen rank_CE=wj+0.5*wi

sum rank_CE

Variable    Obs     Mean       Std. dev.     Min         Max

rank_CE 52,682  .4857322   .2892787     6.00e-06 .9999868
```

I am getting the same mean (0.4857) with wgt_shweight_PR as well.

Is it the right way to use weights?

Thank you.

Subject: Re: Why I am getting different total observations when using iweight for tabulating a variable
Posted by Bridgette-DHS on Mon, 03 Apr 2023 12:35:58 GMT
View Forum Message <> Reply to Message

Following is a response from Senior DHS staff member, Tom Pullum:

With pweight, separately or within svyset, Stata automatically normalizes the weights to have a mean of 1. You did not have to do that with your construction of wgt_shweight_PR.

I have nothing to add to what I said earlier. I don't know why you are using "egen rank" or why you are calculating fractional weights or why you think you should get .5 instead of .4857. I hope someone else can help.
