

Diversion into weights

Simple expression of the mean

$$\frac{\sum_{i=1}^n x_i}{n}$$

- Mean of this series:

– 17, 17, 17, 18, 18, 18, 19, 19, 19, 20, 21

$$\frac{17 + 17 + 17 + 18 + 18 + 18 + 19 + 19 + 19 + 20 + 21}{11} = 18.\overline{45}$$

- Rewrite series

17, 17, 17, 18, 18, 18, 19, 19, 19, 20, 21

as:

$$\frac{(3)17 + (3)18 + (3)19 + (1)20 + (1)21}{11} = 18.\overline{45}$$

17, 17, 17, 18, 18, 18, 19, 19, 19, 20, 21

Or as

$$\frac{(3)17 + (3)18 + (3)19 + (1)20 + (1)21}{3 + 3 + 3 + 1 + 1} = 18.\overline{45}$$

- Or as

– 17, 17, 17, 18, 18, 18, 19, 19, 19, 20, 21

Frequency weights

$$\frac{(3)17 + (3)18 + (3)19 + (1)20 + (1)21}{3 + 3 + 3 + 1 + 1} = 18.\overline{45}$$

- Or as

– 17, 17, 17, 18, 18, 18, 19, 19, 19, 20, 21

Frequency weights

$$\frac{(3)17 + (3)18 + (3)19 + (1)20 + (1)21}{\underbrace{3 + 3 + 3 + 1 + 1}_{\text{Sum of the frequency weights}}} = 18.\overline{45}$$

Sum of the frequency weights

Generalized expression of the mean

$$\frac{\sum_{i=1}^c f_c v_c}{\sum_{i=1}^c f_c}$$

$$\frac{\sum_{i=1}^n x_i}{n}$$

$$\sum_{i=1}^c f_c$$

Common case: weight by jurisdiction “size”

The screenshot shows the Stata software interface. On the left is the Command window with a list of commands. The main window is the Data Editor, showing a dataset with 31 observations (states) and 3 variables: state, tvotes, and obamapct. The status bar at the bottom indicates 3 variables, 51 observations, and the current mode is Edit.

Command Window:

```
37 replace hpct=100*hpct
38 replace hpct=hpct/100
39 twoway (lfit hpct ppct) (scat...
40 twoway (lfit hpct ppct) (scat... 198
41 twoway (lfit hpct ppct) (scat...
42 twoway (scatter hpct ppct, ...
43 twoway (lfit hpct ppct) (scat...
44 reg hpct ppct
45 reg hpct ppct [aw=tvotes]
46 twoway (scatter hpct ppct, ... 198
47 twoway (scatter hpct ppct, ...
48 twoway (lfit hpct ppct) (sca...
49 list if hpct==max(hpct) 198
50 summ hpct
51 list if hpct>90
52 list if stabbr=="MA"
53 predict py
54 reg hpct ppct
55 disp -23.25307+1.433516*64...
56 edit
57 edit if hpct~=.&ppct~=.
58 twoway (lfit hpct ppct) (scat...
59 reg hpct ppct
60 disp .2*5+.4*3
61 disp 2.2*20
62 disp 550/200
63 pwd
64 cd c:\dropbox\classes\17.87...
65 save hpct_ppct
66 clear
67 edit
68 clear
69 edit
70 edit
```

Data Editor (Edit) - [Untitled]

state[1]	state	tvotes	obamapct
1	Alabama	2.1e+06	.387838
2	Alaska	287316	.426847
3	Arizona	2.3e+06	.453866
4	Arkansas	1.0e+06	.378456
5	California	1.3e+07	.618728
6	Colorado	2.5e+06	.52748
7	Connecticut	1.5e+06	.58773
8	Delaware	408068	.59447
9	D. C.	288451	.925876
10	Florida	8.4e+06	.504423
11	Georgia	3.9e+06	.460434
12	Hawaii	427673	.717039
13	Idaho	633698	.335786
14	Illinois	5.2e+06	.585775
15	Indiana	2.6e+06	.447996
16	Iowa	1.6e+06	.529594
17	Kansas	1.1e+06	.388867
18	Kentucky	1.8e+06	.384572
19	Louisiana	2.0e+06	.412532
20	Maine	693582	.578599
21	Maryland	2.6e+06	.633217
22	Massachusetts	3.1e+06	.617857
23	Michigan	4.7e+06	.548005
24	Minnesota	2.9e+06	.539412
25	Mississippi	1.3e+06	.441981
26	Missouri	2.7e+06	.452213
27	Montana	469767	.429658
28	Nebraska	777145	.388706
29	Nevada	994940	.534075
30	New Hampshire	699479	.528338
31	New Jersey	3.6e+06	.589868

Properties Panel:

- Variables: state, tvotes, obamapct
- Variable: state
- Type: str14
- Format: %14s
- Value Labels: (empty)

```
. summ obamapct
```

Variable	Obs	Mean	Std. Dev.	Min	Max
obamapct	51	.500045	.1201509	.2537381	.9258765

```
. summ obamapct
```

Variable	Obs	Mean	Std. Dev.	Min	Max
obamapct	51	.500045	.1201509	.2537381	.9258765

But...

Obama received 65,909,451 votes

Romney received 60,932,176 votes

Therefore, Obama's national pct. is 51.96%

```
. summ obamapct
```

Variable	Obs	Mean	Std. Dev.	Min	Max
obamapct	51	.500045	.1201509	.2537381	.9258765

```
. summ obamapct [fweight=tvotes]
```

Variable	Obs	Mean	Std. Dev.	Min	Max
obamapct	1.268e+08	.51962	.0869835	.2537381	.9258765

But...

Obama received 65,909,451 votes
Romney received 60,932,176 votes

Therefore, Obama's national pct. is 51.96%

A special case, when the sum of the
weights = 1

A special case, when the sum of the weights = 1

Undergraduate class year	Avg. age (in years)	Fraction
1	18.4	.25
2	19.1	.25
3	20.2	.25
4	21.6	.25
Total	19.8	1.00

$$f_1 \frac{\sum_{i=1}^{n_1} x_{i,1}}{n_1} + f_2 \frac{\sum_{i=1}^{n_2} x_{i,2}}{n_2} + f_3 \frac{\sum_{i=1}^{n_3} x_{i,3}}{n_3} + f_4 \frac{\sum_{i=1}^{n_4} x_{i,4}}{n_4}$$

A special case, when the sum of the weights = 1

Undergraduate class year	Avg. age (in years)	Fraction
1	18.4	.15
2	19.1	.20
3	20.2	.25
4	21.6	.40
Total	20.3	1.00

$$f_1 \frac{\sum_{i=1}^{n_1} x_{i,1}}{n_1} + f_2 \frac{\sum_{i=1}^{n_2} x_{i,2}}{n_2} + f_3 \frac{\sum_{i=1}^{n_3} x_{i,3}}{n_3} + f_4 \frac{\sum_{i=1}^{n_4} x_{i,4}}{n_4}$$

More typical case: when the *sample* fractions don't equal the *population* fractions

Undergraduate class year	Avg. age (in years)	Sample fraction	Actual fraction
1	18.4	.2217	.2549
2	19.1	.2913	.2528
3	20.2	.2043	.2415
4	21.6	.2826	.2508
Total		1.00	1.00
Equal weighting	19.83		
Using sample fraction	19.88		
Using actual fraction	19.81		

More typical case: when the *sample* fractions don't equal the *population* fractions

Undergraduate class year	Avg. age (in years)	Sample fraction	Actual fraction	Actual fraction/ Sample fraction	Analytical weight
1	18.4	.2217	.2549	1.15	.28
2	19.1	.2913	.2528	0.87	.21
3	20.2	.2043	.2415	1.18	.29
4	21.6	.2826	.2508	.89	.22
Total		1.00	1.00		
Equal weighting	19.83				
Using sample fraction	19.88				
Using actual fraction	19.81				

```
. summ obamapct
```

Variable	Obs	Mean	Std. Dev.	Min	Max
obamapct	51	.500045	.1201509	.2537381	.9258765

```
. summ obamapct [aweight=tvotes]
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
obamapct	51	126841627	.51962	.087849	.2537381	.9258765

But...

Obama received 65,909,451 votes

Romney received 60,932,176 votes

Therefore, Obama's national pct. is 51.96%

Weighting summary

- Be aware of cases where weights are called for
 - Grouped data (now infrequent)
 - Most survey data
 - Other cases where samples are drawn (e.g., the previous example)
 - Estimating grand means when underlying units are unequal in size
 - Average size of militaries
 - Incarceration rates
 - Election returns
- Use frequency weights only when you are *really* analyzing a dataset in which the data are in summary form
- Use analytical weights in most cases (especially important in statistical tests)