Understanding xtsum output

Consider a simple dataset with three students, for whom we observe two test scores each:

Table 1

id	score
1	70
1	70
2	60
2	80
3	90
3	50

Hand enter these data in the Data Editor and get summary statistics for *score*:

```
. sum score
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       score |         6          70    14.14214         50         90
```

Without taking into account the panel structure, we have 6 observations, with a mean of 70 and $s = 14.14214$.

Now *xtset* the dataset using *id*, and execute the *xtsum* command. You should get this output:

```
. xtset id
       panel variable:  id (balanced)
. xtsum score
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
score    overall |        70   14.14214         50         90 |     N =       6
         between |                    0         70         70 |     n =       3
         within  |             14.14214         50         90 |     T =       2
```

Note that the "overall" line of output matches the results from the *sum* command; Stata is simply calculating summary statistics for the entire (overall) dataset.

The "between" output is easy to understand. Looking at Table 2, we can see that the mean test score for each student equals 70. The mean of 70, 70, and 70 equals 70, and because these do not vary, $s = 0$. Note the $n = 3$; Stata is telling you it is calculating this line of numbers from 3 observations (the student-level means), not all 6.

The within output is a bit more tricky. T=2 tells us that Stata is calculating this within each student, because we have two observations per student. The s of 14.14214 in the output is conceptually the average of the standard deviations for the 3 students.

Table 2

id	score	$\bar{X}$	s
1	70	70	0
1	70		
2	60	70	14.14214
2	80		
3	90	70	28.28427
3	50		
Mean	70	0	14.14214

However, it is not exactly the average, as the following data in Table 3 and output illustrate. Similar setup, but now the mean student test score varies between students. The overall mean is still 70, but the overall s is larger, because there is increased variance in test scores. Note that the min and max values reflect the full range of student test scores.

Table 3

id	score	$\bar{X}$	s
1	70	70	0
1	70		
2	70	80	14.14214
2	90		
3	90	60	42.42641
3	30		
Mean	70	70	18.85618

```
. xtsum score
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
score    overall |        70    21.9089         30         90 |     N =       6
         between |                   10         60         80 |     n =       3
         within  |                   20         40        100 |     T =       2
```

The s for the between data now equals 10, because s is calculated on the three test score means of 70, 80, and 60 (confirm this for yourself). Note that the min and max values refer to the minimum and maximum values for the mean student test scores of 70, 80, and 60.

Main takeaway: the “between” output first estimates unit-level averages for every unit used in the *xtset* command, and then calculates s for these means.

Now compare the min and max values for the “within” output for the 6 test scores: Stata is reporting a max test score value of 100, which does not exist in the dataset! The average s

from Table 3 is close, but does not match the Stata output of 20.

If you read the *xtsum* documentation, you will learn that the within calculation does not use

$$x_{it} - \bar{x}_i$$

in its calculations (that is, the individual score minus the mean score for the unit). Instead *xtsum* adds back the global (overall) mean to "make results comparable":

$$x_{it} - \bar{x}_i + \bar{\bar{x}}$$

where $\bar{\bar{x}}$ is the overall mean for x in the dataset.

Table 4 relabels Table 3 with notation, and adds columns showing how the calculations are done. For example, for the first observation for Student 2 (third row in the table), Stata does not use -10 (70-80) when calculating s, but instead 70-80+70, yielding a final difference value of 60. So the last column explains why the min and max values for the within output can appear a bit odd.

Table 4

i	x_{it}	$\bar{x}_i$	$x_{it} - \bar{x}_i$	$x_{it} - \bar{x}_i + \bar{\bar{x}}$
1	70	70	0	70
1	70		0	70
2	70	80	-10	60
2	90		10	80
3	90	60	30	100
3	30		-30	40
Mean	70	70	0	70

What about s? For the overall s, Stata uses $x_{it} - \bar{\bar{x}}$; that is, it calculates variance around the overall mean in the dataset. But for the within s, we only want variation within units, and we want to ignore variation between units. Thre are two ways to do this. First, we could calculate the within variance using Stata's approach:

$$x_{it} - \bar{x}_i + \bar{\bar{x}}$$

Take the six numbers in the last column, and calculate s for these numbers: the result is 20, which matches the *xtsum* output.

Alternatively, we could calculate variance around each unit's mean:

$$x_{it} - \bar{x}_i$$

If you look at this column in Table 4, take the values 0, 0, -10, 10, 30, -30, square them, sum them, and divide by 5 ($n - 1$); the result is 20. Personally, I find this easier to understand than the method of adding back the global mean.

These $x_{it} - \bar{x}_i$ differences help us understand the min and max values for the within output. The "true" min value for the within calculation is -30, and the max is +30. Stata thinks putting these on the output would be confusing for some people, so they decided to add back the global mean. So the printed min value is $-30 + 70 = 40$, and the max is $+30 + 70 = 100$, whihc matches the *xtsum* output above.

Main takeaway: the "within" s tells you how much a variable varies within units, while ignoring all variation between units. Conceptually, you can think of this as calculating the standard deviation of x for each unit separately, and then averaging these values to get the typical s. To make the between and within numbers comparable, Stata adds back the overall mean to each observation in its calculations. This makes the min and max values a bit hard to interpet, unlike the min and max values for the between output. To get the original differences, just subtract the overall mean from the min and max values.

Summary

The *xtsum* command provides a quick way to understand the between and within variance for your covariates. The min and max between values can help you understand the range of your unit-level means, but are less helpful for the within data. The between and within standard deviations should be examined closely. Do the results match your intuitions? If not, you might have a data management issue.

Consider the Michigan school district panel from Wooldridge:

```
. xtset distid year
       panel variable:  distid (strongly balanced)
        time variable:  year, 1992 to 1998
                delta:  1 unit

. xtsum lunch enrol expp

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
lunch    overall |  27.20165   15.40591          0      91.27 |     N =    3850
         between |              15.0169          0   83.96143 |     n =     550
         within  |             3.490854  -6.806924   69.62879 |     T =       7
                 |                                            |
enrol    overall |  3043.734   8153.137         26     183151 |     N =    3850
         between |             8149.347   38.28571   177606.9 |     n =     550
         within  |             406.5732  -3708.123   8587.877 |     T =       7
                 |                                            |
expp     overall |  5237.257   1224.984        946      13982 |     N =    3850
         between |             910.4692   2500.857   10121.29 |     n =     550
         within  |             820.3198   2123.828     9903.4 |     T =       7
```

For the proportion of students on free and reduced lunch and student enrollments, we would expect big differences between districts, but relatively small differences over time within districts. This is exactly what we see when we compare the between and within s for these two variables.

Expenditures, on the other hand, may be more volatile over time within a district. Note that the between and within s are almost equal for this variable.