3.2: Three (or more) variables 3.2.1 Elaboration

3.2.1.4 Elaboration 4 (Income differences BSA 2009-2014: CTABLES)

(1st, 2nd order and multi-way contingency tables)

This set of tutorials uses data from the British Social Attitudes series to explore the following research questions.

1: Is there a difference between the earnings (from paid work) of men and women?
2: What other variables might account for differences in earnings?
3: What effect do they have by themselves?
4: What happens to any differences in earnings between men and women when controlling for these other variables?

Data source: British Social Attitudes 1983 to 2014: Cumulative SPSS file

Dependent variable Y: Earnings from paid work

Independent variable X: Sex of respondent

Test variables T:

Work	Employee/Self-employed
	Public/private sector
	Hours worked
	Social class of job
Education	Age completed full time education
	Qualifications from school
	Qualifications post-school
	Highest qualification
Age	Age last birthday
	Age group
Location	Region
	Country

Previous tutorials:

3.2.1 Elaboration (Extract from Jim Ring's notes)
3.2.1.1 Elaboration 1 (Income differences BSA 1989)
3.2.1.2 Elaboration 2 (Income differences BSA 2009 - 2014)
3.2.1.3 Elaboration 3 (Income differences 2009 – 2014 CROSSTABS)

All tables in the elaboration series so far are from SPSS **Data Editors** with Names and Labels **ON**.

Because SPSS command **CROSSTABS** produces such cluttered output for two- and three-way tables, especially if you request both counts and row percents, it's better from now on to use **CTABLES.** The syntax is quite complex, but the tables are much clearer.

SPSS procedure **CTABLES** gives full control of output tables, but the syntax looks very complicated to the uninitiated (i.e. me!). I'm working with gurus from the SPSS list on a macro[1] that will do it for three variables using a new SPSS command **ELABORATE**:

elaborate Y = <dependent var> / RowVar = <independent var> / ColVar = <test var>.

[1] See **Appendix** for full syntax of beta version

CTABLES

The default output from **CTABLES** can still be a bit cluttered, but can be modified within the program.

Within the **CTABLES** command, tables have to be specified one at a time using:

```
/TABLE <variable>
```

1: Initial frequency counts.

For analysing one variable the default output can be very sparse, but at least for frequency distributions you don't get totally unnecessary cumulative totals for nominal variables.

1a: To display **counts** in **columns**:

```
ctables /table earngrp
        /table sex
        /table workmode.
```

		Count
earngrp Quartile group of R's earnings from paid work	1 Q1	2157
	2 Q2	2430
	3 Q3	2196
	4 Q4	2006

This table is much simpler than the default **FREQUENCIES** output which, in addition to **Frequency**, includes **Percent** (of all values) **Valid Percent** (of non-missing values) and (only really useful for very long ordinal and interval scale variables) **Cumulative Percent**.

earngrp Quartile group of R's earnings from paid work

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Q1	2157	11.1	24.5	24.5
	2 Q2	2430	12.5	27.6	52.2
	3 Q3	2196	11.3	25.0	77.2
	4 Q4	2006	10.3	22.8	100.0
	Total	8789	45.3	100.0	
Missing	System	10610	54.7		
Total		19399	100.0		

CTABLES output for **counts** in **columns** of **rsex** (independent) and **workmode** (test) variables:

		Count
Rsex Person 1 SEX	1 Male	8530
	2 Female	10869

		Count
workmode R working full- or part-time	1 Full time	6020
	2 Part time	2282

1b: To display **counts** in **rows**:

```
ctables /table by earngrp
        /table by sex
        /table by workmode.
```

earngrp Quartile group of R's earnings from paid work			
1 Q1	2 Q2	3 Q3	4 Q4
Count	Count	Count	Count
2157	2430	2196	2006

Rsex Person 1 SEX	
1 Male	2 Female
Count	Count
8530	10869

workmode R working full- or part-time	
1 Full time	2 Part time
Count	Count
6020	2282

1c: To display **percentages** in **rows** (dependent variable):

```
ctables
   /table by earngrp [ROWPCT.COUNT TOTALS [COUNT]].
```

earngrp Quartile group of R's earnings from paid work			
1 Q1	2 Q2	3 Q3	4 Q4
Row N %	Row N %	Row N %	Row N %
24.5%	27.6%	25.0%	22.8%

1d: To display **percentages** in **columns** (independent and test variables):

```
ctables
   /table rsex [COLPCT.COUNT TOTALS [COUNT]]
   /table  workmode [COLPCT.COUNT TOTALS [COUNT]].
```

		Column N %
Rsex Person 1 SEX	1 Male	44.0%
	2 Female	56.0%

		Column N %
workmode R working full- or part-time	1 Full time	72.5%
	2 Part time	27.5%

2: Zero order tables (counts only).

```
ctables   /TABLE sex BY earngrp
          /TABLE workmode BY earngrp.
```

		earngrp Quartile group of R's earnings from paid work			
		1 Q1	2 Q2	3 Q3	4 Q4
		Count	Count	Count	Count
Rsex Person 1 SEX	1 Male	559	1113	1245	1382
	2 Female	1598	1317	951	624

		earngrp Quartile group of R's earnings from paid work			
		1 Q1	2 Q2	3 Q3	4 Q4
		Count	Count	Count	Count
workmode R working full- or part-time	1 Full time	575	1633	1705	1546
	2 Part time	1185	475	210	129

2b: Zero order tables (with row %).

Unlike **CROSSTABS**, **CTABLES** allows you to display the row totals on the same line as the row percentages (i.e. without the row counts) by requesting TOTALS [COUNT] as an additional element inside the square brackets:

```
ctables   /TABLE sex BY earngrp  [ROWPCT.COUNT TOTALS [COUNT]]
```

This will display the same table as the above. To display the row totals used as a base for percentages, you need an additional line for each /TABLE specification:

```
          /CATEGORIES VARIABLES=  earngrp TOTAL=YES .
```

```
ctables   /TABLE sex BY earngrp  [ROWPCT.COUNT TOTALS [COUNT]]
            /CATEGORIES VARIABLES=  earngrp TOTAL=YES
          /TABLE workmode BY earngrp  [ROWPCT.COUNT TOTALS [COUNT]]
            /CATEGORIES VARIABLES=  earngrp TOTAL=YES .
```

		earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total
		Row N %	Row N %	Row N %	Row N %	Count
Rsex Person 1 SEX	1 Male	13.0%	25.9%	29.0%	32.1%	4299
	2 Female	35.6%	29.3%	21.2%	13.9%	4490

		earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total
		Row N %	Row N %	Row N %	Row N %	Count
workmode R working full- or part-time	1 Full time	10.5%	29.9%	31.2%	28.3%	5459
	2 Part time	59.3%	23.8%	10.5%	6.5%	1999

For analytical purposes these tables are far more useful (and far less cluttered) than the output from **CROSSTABS** and it's much easier to calculate the epsilons.

Although the tables are now much easier to read and interpret, they can be further improved by getting rid of all the **%** signs in the body of the table and by changing the column headers. The elements in the /TABLES specification can be extended by adding labels enclosed in double primes eg: [ROWPCT.COUNT "%"] and [COUNT "n = 100%"]

The default formats are integer for cell counts and one decimal place for percentages, but if needed the latter can be changed by adding a format eg:

[ROWPCT.COUNT f4.2]

However two decimal places for percentages seems a bit pointless: we're trying to reduce the clutter, not add to it!

```
ctables
  /TABLE sex BY earngrp  [ROWPCT.COUNT f5.1 "%"  totals [count "n= 100%"]]
        /CATEGORIES VARIABLES=  earngrp TOTAL=YES
  /TABLE workmode BY earngrp  [ROWPCT.COUNT f5.1 "%" totals [count "n= 100%"]]
        /CATEGORIES VARIABLES=  earngrp TOTAL=YES .
```

		earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total
		%	%	%	%	n= 100%
Rsex Person 1 SEX	1 Male	13.0	25.9	29.0	32.1	4299
	2 Female	35.6	29.3	21.2	13.9	4490

Epsilon **-22.6** **-3.4** **+7.8** **+18.2**

		earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total
		%	%	%	%	n= 100%
workmode R working full- or part-time	1 Full time	10.5	29.9	31.2	28.3	5459
	2 Part time	59.3	23.8	10.5	6.5	1999

Epsilon **-48.8** **+6.1** **+20.7** **+21.8**

Variables **rsex** and **workmode** both have strong effects on earnings. The above tables do not have totals for the earnings groups as it's easier to compare the income groups of men/women and full-time/part-time workers without them (and to calculate the percentage point differences, epsilons).

[**NB**: The epsilons were produced separately by copying the tables into Excel, performing the calculations, copying the epsilons back into Word, inserting the + signs and changing the colour of positive epsilons to **blue**] I wonder if **CTABLES** can produce tables with epsilons?

For elaboration purposes you need to compare these **conditional** distributions with the original **overall** distribution to see how they have been **partitioned** when controlling for test variables.

earngrp Quartile group of R's earnings from paid work				
1 Q1	2 Q2	3 Q3	4 Q4	
%	%	%	%	n = 100%
24.5	**27.6**	**25.0**	**22.8**	**8789**

Test variables can be added at any stage.

Model for elaboration exercise

Dependent variable	Y = Earnings	**rearngrp**	(Quartile groups)
Independent variable	X = Sex	**rsex**	(Men, Women)
Test variable 1	T_1 = Hours of work	**workmode**	(Part-time, Full-time)

Zero order tables

Model	**Table to display** (Row variable by Column variable)	
1: X → Y	<independent variable> by <dependent variable>	(<indvar> by <testvar>)
2: T_1 → Y	<test variable> by <dependent variable>	(<testvar> by <depvar>)

SPSS ctables syntax [General format for two-way tables]

```
ctables
  /TABLE <indvar> BY <depvar>  [ROWPCT.COUNT f5.1 "%"  totals [count "n= 100%"]]
  /CATEGORIES VARIABLES=  <depvar>   TOTAL=YES
```

The above syntax looks very complex, but all you have to worry about is copy/pasting it from here (or from the Appendix) substituting the variable names, repeating for each table required.

For the zero order tables:

Model	**<indvar> by <depvar>**	(**Row** variable by **Column** variable)
1: X → Y	**rsex** by **earngrp**	Sex of respondent by earnings group
2: T_1 → Y	**workmode** by **earngrp**	R working FT/PT by earnings group

```
ctables
  /TABLE rsex BY earngrp  [ROWPCT.COUNT f5.1 "%"  totals [count "n= 100%"]]
  /CATEGORIES VARIABLES=  earngrp TOTAL=YES
  /TABLE workmode BY earngrp  [ROWPCT.COUNT f5.1 "%" totals [count "n= 100%"]]
  /CATEGORIES VARIABLES=  earngrp TOTAL=YES .
```

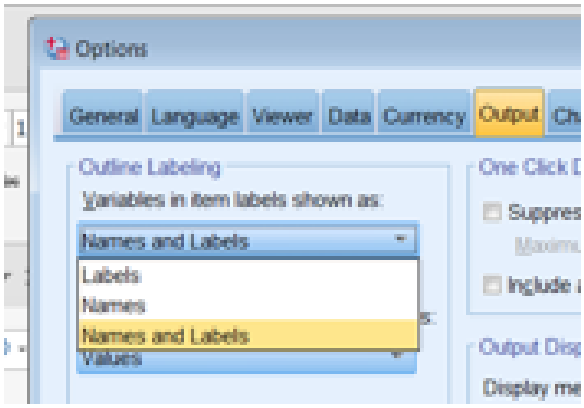
		earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total
		%	%	%	%	n= 100%
Rsex Person 1 SEX	1 Male	13.0	25.9	29.0	32.1	4299
	2 Female	35.6	29.3	21.2	13.9	4490

		earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total
		%	%	%	%	n= 100%
workmode R working full- or part-time	1 Full time	10.5	29.9	31.2	28.3	5459
	2 Part time	59.3	23.8	10.5	6.5	1999

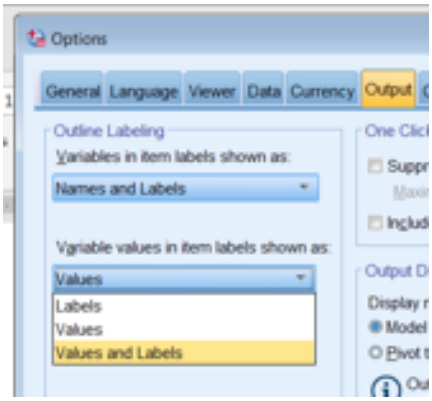
[NB: Unlike **CROSSTABS**, **CTABLES** does not display a title for the tables, so for variable labels SPSS settings are best left as Names and Labels. If you don't want to display the numeric values, change the settings for value labels to Labels only.]

My settings for this tutorial are: Edit >> Options >> Output

Variables



Values



First order tables

1: X → Y . T_1
2: X → Y . T_2

To produce three-way contingency tables in **CTABLES**, the specification of variables is slightly different. One pair of variables has to linked by **>** (variable on the right of **>** is nested in categories of the variable on the left). There three ways of producing such tables:

X > T by Y **T > X by Y** **X > Y by T**

1: X > T_1 by Y

```
*nest workmode within sex.
ctables
        /TABLE rsex > workmode by earngrp
                [ROWPCT.COUNT f5.1 "%"  totals [count "n= 100%"]]
         /CATEGORIES VARIABLES=  earngrp TOTAL=YES.
```

				earngrp Quartile group of R's earnings from paid work				
				1 Q1	2 Q2	3 Q3	4 Q4	Total
				%	%	%	%	n= 100%
Rsex Person 1 SEX	1 Male	workmode R working full- or part-time	1 Full time	6.1	26.8	32.3	34.9	3028
			2 Part time	45.0	24.0	15.6	15.4	416
	2 Female	workmode R working full- or part-time	1 Full time	16.0	33.9	29.9	20.2	2431
			2 Part time	63.0	23.7	9.2	4.1	1583

2: T_1 > X by Y

***Nest sex within workmode.**

```
ctables   /TABLE  workmode > rsex by  earngrp            (nests sex within workmode):
                [ROWPCT.COUNT f5.1 "%"  totals [count "n= 100%"]]
          /CATEGORIES VARIABLES=  earngrp TOTAL=YES.
```

				earngrp Quartile group of R's earnings from paid work				
				1 Q1	2 Q2	3 Q3	4 Q4	Total
				%	%	%	%	n= 100%
workmode R working full- or part-time	1 Full time	Rsex Person 1 SEX	1 Male	6.1	26.8	32.3	34.9	3028
			2 Female	16.0	33.9	29.9	20.2	2431
	2 Part time	Rsex Person 1 SEX	1 Male	45.0	24.0	15.6	15.4	416
			2 Female	63.0	23.7	9.2	4.1	1583

However it's easier to compare men and women when the table is spread out using

3: **X by T_1 > Y**

```
ctables   /TABLE rsex by workmode > earngrp
                [ROWPCT.COUNT f5.1 "%"  totals [count "n= 100%"]]
          /CATEGORIES VARIABLES=  earngrp TOTAL=YES.
```

		workmode R working full- or part-time									
		1 Full time					2 Part time				
		earngrp Quartile group of R's earnings from paid work					earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total	1 Q1	2 Q2	3 Q3	4 Q4	Total
		%	%	%	%	n= 100%	%	%	%	%	n= 100%
Rsex Person 1 SEX	1 Male	6.1	26.8	32.3	34.9	3028	45.0	24.0	15.6	15.4	416
	2 Female	16.0	33.9	29.9	20.2	2431	63.0	23.7	9.2	4.1	1583

[NB: The tables above are manually edited from SPSS output with very wide columns]

This layout makes it easier to calculate epsilon (percentage point difference) between men and women in the different earnings groups. It's even easier if you copy the SPSS output into Excel and do the calculations (and the colours for negative values) there:

[Copied from Excel]

		workmode R working full- or part-time									
		1 Full time					2 Part time				
		earngrp Quartile group of R's earnings from paid work					earngrp Quartile group of R's earnings from paid work				
		1 Q1	2 Q2	3 Q3	4 Q4	Total	1 Q1	2 Q2	3 Q3	4 Q4	Total
		%	%	%	%	n= 100%	%	%	%	%	n= 100%
Rsex Person 1 SEX	1 Male	6.1	26.8	32.3	34.9	3028	45.0	24.0	15.6	15.4	416
	2 Female	16.0	33.9	29.9	20.2	2431	63.0	23.7	9.2	4.1	1583
	Epsilon	**-9.9**	**-7.1**	**+2.3**	**+14.7**		**-18.1**	**+0.3**	**+6.5**	**+11.3**	

To get the same table with overall totals for earnings groups:

```
ctables
        /VLABELS VARIABLES=rsex earngrp workmode DISPLAY=NONE
        /TABLE rsex by workmode > earngrp
        [ROWPCT.COUNT f5.1 "%"  TOTALS [COUNT "n= 100%"]]
        /CATEGORIES VARIABLES=  rsex workmode earngrp  TOTAL=YES POSITION=after.
```

Produces a very wide table:

	1 Full time					2 Part time					Total				
	1 Q1	2 Q2	3 Q3	4 Q4	Total	1 Q1	2 Q2	3 Q3	4 Q4	Total	1 Q1	2 Q2	3 Q3	4 Q4	Total
	%	%	%	%	n= 100%	%	%	%	%	n= 100%	%	%	%	%	n= 100%
1 Male	6.1	26.8	32.3	34.9	3028	45.0	24.0	15.6	15.4	416	10.8	26.4	30.3	32.5	3444
2 Female	16.0	33.9	29.9	20.2	2431	63.0	23.7	9.2	4.1	1583	34.6	29.8	21.7	13.8	4014
Total	10.5	29.9	31.2	28.3	5459	59.3	23.8	10.5	6.5	1999	23.6	28.3	25.7	22.5	7458

Taking Q4 (the top 25% of earnings per annum) as the criterion category (for "high earnings") the figure of 22.5% for the whole sample of 7458 can be broken down into 32.5% of 3444 men and 13.8% of 4104 women. For hours worked the 22.5% of 7458 is broken down into 28.3% of 5459 working full time and 6.0% of 1999 working part-time.

For those working full time, the 28.3% of 5459 can be further partitioned into 34.0% of 3028 men and 20.2% of 2431 women. For those working part time the 6.5% of 1999 can be partitioned into 15.4% of men and 4.1% of 1583 women.

From these figures it is possible to construct a summary table to show what happens to differences in earnings of men and women when controlling for hours worked (full time is 30 or more hours per week).

Construct a blank table:

Difference in earnings of men and women controlling for hours worked (group Q4 only)

	All	Full time	Part time	Epsilon
All	% ( )	% ( )	% ( )	
Men	% ( )	% ( )	% ( )	
Women	% ( )	% ( )	% ( )	
Epsilon				

Unfortunately SPSS can't do this for you, but it's better for you to do it yourself as you will get a better fell for the data. Do it step by step. In the whole sample (of those with earnings from paid work) what percentage are in group Q4 and what is the base for the percentage?

	All
All	22.5% (7458)

Of those in this top earnings group, what is the percentage for men, what is the percentage for women, and what are the bases for these? Calculate epsilon (percentage point difference) between men and women.

	All
All	22.5% (7458)
Men	32.5% (3444)
Women	13.8% (4104)
Epsilon	+18.7

Of those in this top earnings group, what is the percentage for those working full time, what is the percentage for those working part time, and what are the bases for these? Calculate epsilon (percentage point difference) between those working full and those working part time.

	All	Full time	Part time	Epsilon
All	22.5% (7458)	28.3% (5459)	6.52% (1999)	**+13.8**

Now complete the rest of the table:

	All	Full time	Part time	Epsilon
All	22.5% (7458)	28.3% (5459)	6.52% (1999)	**+13.8**
Men	32.5% (3444)	34.9% (3028)	15.4% (416)	**+19.5**
Women	13.8% (4014)	24.2% (2431)	4.1% (1583)	**+20.1**
Epsilon	**+18.7**	**+10.7**	**+11.3**	

From this summary table, when controlling for hours of work, the original difference of **+18.7** points in earnings between men and women, decreases from to **+10.7** for full time and **+11.3** for p[art time.

[Draft only: 11 August 2016]

End of session: **3.2.1.4 Elaboration 4 (Income differences BSA 2009-2014: CTABLES)**

Back to: **3.2.1.3 Elaboration 3 (Income differences 2009 – 2014_CROSSTABS)**
3.2 Three (or more) variables

Appendix

Macro for ELABORATE

I am extremely grateful to Bruce Weaver who provided the initial code to replicate output from a procedure in earlier releases of SPSS.

```
means variables = depvar (x,y) indvar (a,b) testvar (p,q)
        /crossbreak <depvar> by <indvar>
```

I have tweaked it for elaboration, but I need to experiment with it to guarantee the same display formats as the tables in this tutorial.

```
* Encoding: UTF-8.
DEFINE elaborate
( Y = !CHAREND('/') /
   RowVar = !CHAREND('/') /
   ColVar = !CMDEND ).

CTABLES
  /TABLE !RowVar  [C] >  !ColVar [C] by !Y [c][ROWPCT.totaln, COUNT]
  /SLABELS POSITION=ROW VISIBLE=NO
  /CATEGORIES VARIABLES= !y  TOTAL=YES
  /TITLES CAPTION ='NOTE:  Cells show ROWPCT and COUNT'.
!ENDDEFINE.

* Read in some data to illustrate.
* Modify path on the FILE HANDLE command as needed.
FILE HANDLE TheDataFile /NAME='C:\Users\John\Desktop\ bsa83-14_elab4.sav'.
NEW FILE.
GET FILE = "TheDataFile".

* Call the macro.
elaborate Y = earngrp / RowVar = rsex / ColVar = workmode.
```

				earngrp Quartile group of R's earnings from paid work				
				1 Q1	2 Q2	3 Q3	4 Q4	Total
Rsex Person 1 SEX	1 Male	workmode R working full- or part-time	1 Full time	5.5%	24.3%	29.3%	31.6%	100.0%
				185	810	977	1056	3028
			2 Part time	39.0%	20.8%	13.5%	13.3%	100.0%
				187	100	65	64	416
	2 Female	workmode R working full- or part-time	1 Full time	14.6%	30.7%	27.2%	18.3%	100.0%
				390	823	728	490	2431
			2 Part time	55.4%	20.8%	8.0%	3.6%	100.0%
				998	375	145	65	1583

NOTE: Cells show ROWPCT and COUNT

The CROSSBREAK subcommand

This subcommand (no longer available) displayed means and (n) in tables. The example below is from the 1990 course and uses data from the 1981 Fifth Form Survey.

The only font available for both line-printer and VDU was Courier, but at least we had lower case.

```
MEANS VARIABLES = SEXISM(0,9) V348(1,2)ETHNIC(1,2)
                  /CROSSBREAK = SEXISM BY V348 BY ETHNIC
                 /CELLS = MEAN COUNT
```

The output looked like this:

```
                        ETHNIC
                 Mean :
                Count :  White      Black        Row
                      :                        Total
                      :       1   :       2   :
V348         --------:----------:----------:
                   1  :     4.63 :     3.43 :     3.98
  Boys             :       19  :       23  :       42
                   -:----------:----------:
                   2  :     1.89 :     1.80 :     1.84
  Girls            :       19  :       25  :       44
                   -:----------:----------:
     Column Total         3.26        2.58        2.88
                            38          48          86
```

A crafty use of CROSSBREAK in combination with RECODE allowed me to display percentages instead of means for the dependent variable. What I did was recode one or more criterion values of the dependent variable to 100 and everything else to 0. The 'means' displayed are then percentages!

Thus:

```
RECODE      SEXISM(2 THRU 7 = 100)(0,1 = 0)(ELSE = SYSMIS)
MEANS VARIABLES = SEXISM (0,100) V348 (1,2) ETHNIC (1,2)
      /CROSSBREAK = SEXISM BY V348 BY ETHNIC
          /CELLS = MEAN COUNT
```

The output looked something like this:

```
                    ETHNIC
              Mean :
             Count :  White      Other        Row
                   :                        Total
                   :       1   :       2   :
V348         --------:----------:----------:
                   1  :   100.00 :    82.61 :    90.48
  Boys             :       19  :       23  :       42
                   -:----------:----------:
                   2  :    47.37 :    44.00 :    45.45
  Girls            :       19  :       25  :       44
                   -:----------:----------:
     Column Total         73.68       62.50       67.44
                            38          48          86
```