

[Commentary by [John F Hall](#)]

[**Draft only**: last updated 4 Novemebr 2017]

John MacInnes

[An Introduction to Secondary Data Analysis with IBM SPSS Statistics](#)

(Sage, Dec. 2017)

5.1 [Chapter 5 video tutorials](#) (direct link to companion website)


[NB: All video tutorials for chapter 5 are on the same web page and cannot (yet) be disaggregated]

Video 5.1.6: Using the **SUM** sub-command (6'49")

Exemplar:	European Social Survey 2012
SPSS file:	ESS6e02_1.sav
Variable to be derived:	Index of depression from the 8-item depression inventory
Source variables:	fltdpr flteeff slprl wrhpp fltlnl enjlf fltsd cldgng
SPSS commands:	COMPUTE¹ IF FORMATS MEANS VARIABLE LEVEL CORRELATIONS
Statistical function:	SUM
Technical terms:	argument, ratio scale, valid value, system missing, source variable, target variable, numeric expression

Task: Create a new variable: ". . the sum of the 8 values on the 8 variables that describe depression."

JM now gets round to doing what I think he should have done in 5.1.5 (and, out of research curiosity, I had already done). He should have started with a simple addition.

[NB: JM keeps running syntax by highlighting the whole command, but as long as the cursor is somewhere inside the command SPSS will run it with **Ctrl+R** or .

He points out that **[depress]** has range of values 7 to 32 and 1.5% missing cases, but **not** that **[depress]** has 2 superfluous decimal places. He doesn't even show **SUM** for the set of 8.

". . However, there's a small complication: not every respondent has given an answer to all eight of the variables. We want to take account of at least those respondents that have answered . . . at least 7 of the questions." He does not explain why and nothing is shown on the video, but says he is looking for people with at least 7 valid values across all 8 items. The commentary describes the syntax:

compute <new variable> = **SUM.7** (<var_1>², <var_2>,<var_n>)

He's already extracted the subset of variables: the ones he uses are on lines 200ff in the original file, but are now on lines 14 ff. so he's obviously using a different data set. He uses direct syntax "because it's easier". Using the six negative and the two recoded positive items he constructs the list of variables,

¹ For a brief introduction to the **COMPUTE** command, see [3.5.2.4 The COMPUTE command 1 - Attachment to status quo](#) and [3.5.2.7 The COMPUTE command 2 - Sexism](#)

² In SPSS these lists are known as logical **arguments**: each argument has to be separated by a comma)

but doesn't explain why you have to use **commas**, not spaces: he puts the first comma in and inserts the other commas afterwards,

Watch how **compute** remains **red**

compute depress = sum.7(fltdpr, flteeff, slprl, fltlnl, fltsd, cldgng, enjlf2, wrhpp2)

. . until the full stop goes on the end, when it turns **blue**.

compute depress = sum.7(fltdpr, flteeff, slprl, fltlnl, fltsd, cldgng, enjlf2, wrhpp2).
freq depress.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	7.00	24	.0	.0	.0
	8.00	2530	4.5	4.5	4.6
	9.00	3092	5.4	5.6	10.1
	10.00	5103	9.0	9.2	19.3
	11.00	5716	10.1	10.3	29.6
	12.00	5773	10.2	10.4	39.9
	13.00	5388	9.5	9.7	49.6
	14.00	5176	9.1	9.3	58.9
	15.00	4622	8.1	8.3	67.2
	16.00	3881	6.8	7.0	74.2
	17.00	3214	5.7	5.8	80.0
	18.00	2716	4.8	4.9	84.8
	19.00	1829	3.2	3.3	88.1
	20.00	1481	2.6	2.7	90.8
	21.00	1204	2.1	2.2	93.0
	22.00	1073	1.9	1.9	94.9
	23.00	721	1.3	1.3	96.2
	24.00	699	1.2	1.3	97.4
	25.00	418	.7	.8	98.2
	26.00	324	.6	.6	98.8
	27.00	198	.3	.4	99.1
	28.00	113	.2	.2	99.3
	29.00	154	.3	.3	99.6
	30.00	92	.2	.2	99.8
	31.00	48	.1	.1	99.9
	32.00	82	.1	.1	100.0
	Total	55671	98.0	100.0	
Missing	System	1164	2.0		
Total		56835	100.0		

There is no discussion of whether **sum.7** instead of **sum.8** distorts the depression score.

Because **[cldgng]** is not available for Albanian respondents, JM gives them an **imputed** depression score, multiplying their score derived from the other seven items by a factor of $8 \div 7$, but makes no comparison of **sum.7** with **sum.8**. He finds that 2% missing cases for **SUM.7** rises to 6.2% for **SUM.8** and seems more intent on looking for a culprit country than analysing the structure of depression.

JM gives Albanian respondents the imputed depression score with:

if (missdep = 1) depress = (8/7)*depress.

That's a big assumption, that Albanian respondents, for whom only seven item scores are available, would have got the same depression score if they had answered all 8: it assumes that all items contribute equally to the index. His calculation is not necessarily comparing like with like. He needs to do a different calculation of a 7-item score **excluding [cldng]** and then compare the two. That way Albania stays in.

compute depress_7 = sum.7(fltldr, flteff, slprl, fltlnl, fltsd, enjlf2, wrhpp2)-7.
formats depress_7 (f2.0).
variable labels depress_7 "Depression score without cldngng" .
frequencies depress_7 /format notable /histogram normal .
means depress_7 by cntry.

depress_7 Depression score without cldngng
(Unweighted: unsorted)

cntry Country	Mea n	N	Std. Deviation
AL Albania	8.32	216	3.888
BE Belgium	4.71	916	3.582
BG Bulgaria	6.57	608	4.361
CH Switzerland	4.27	671	3.199
CY Cyprus	5.13	69	4.231
CZ Czech Republic	6.03	800	4.143
DE Germany	5.13	7041	3.432
DK Denmark	4.02	452	3.109
EE Estonia	5.90	110	3.714
ES Spain	5.56	3873	3.929
FI Finland	4.07	447	3.016
FR France	5.31	5305	3.886
GB United Kingdom	4.99	5180	3.677
HU Hungary	7.09	825	4.105
IE Ireland	4.20	356	3.643
IL Israel	5.26	538	3.656
IS Iceland	4.10	25	3.257
IT Italy	5.94	4990	3.811
LT Lithuania	6.46	236	3.418
NL Netherlands	4.34	1374	3.450
NO Norway	3.48	402	2.750
PL Poland	5.10	3143	4.135
PT Portugal	6.09	888	3.987
RU Russian Federation	6.51	10800	3.694
SE Sweden	4.02	784	3.267
SI Slovenia	4.11	173	3.366
SK Slovakia	6.03	443	3.418
UA Ukraine	6.98	3290	4.143
XK Kosovo	6.94	126	3.464
Total	5.64	54080	3.838

depress_7 Depression score without cldngng
(Unweighted: sorted in **descending** order of mean)

cntry Country	Mea n	N	Std. Deviation
AL Albania	8.32	216	3.888
HU Hungary	7.09	825	4.105
UA Ukraine	6.98	3290	4.143
XK Kosovo	6.94	126	3.464
BG Bulgaria	6.57	608	4.361
RU Russian Federation	6.51	10800	3.694
LT Lithuania	6.46	236	3.418
PT Portugal	6.09	888	3.987
CZ Czech Republic	6.03	800	4.143
SK Slovakia	6.03	443	3.418
IT Italy	5.94	4990	3.811
EE Estonia	5.90	110	3.714
ES Spain	5.56	3873	3.929
FR France	5.31	5305	3.886
IL Israel	5.26	538	3.656
DE Germany	5.13	7041	3.432
CY Cyprus	5.13	69	4.231
PL Poland	5.10	3143	4.135
GB United Kingdom	4.99	5180	3.677
BE Belgium	4.71	916	3.582
NL Netherlands	4.34	1374	3.450
CH Switzerland	4.27	671	3.199
IE Ireland	4.20	356	3.643
SI Slovenia	4.11	173	3.366
IS Iceland	4.10	25	3.257
FI Finland	4.07	447	3.016
DK Denmark	4.02	452	3.109
SE Sweden	4.02	784	3.267
NO Norway	3.48	402	2.750
Total	5.64	54080	3.838

The new variables **[depress_8]** and **[depress_7]** are appended to the file. Although **[depress_7]** and **[depress_8]** were both calculated using **COMPUTE**, SPSS has set the **Level** for **[depress_7]** to **Nominal** when it should really be **Scale**. This is something you have to watch out for if you leave everything to the SPSS 'heuristic' algorithm. When creating new variables it is better to set the measurement level yourself.

depress_8	Scale
depress_7	Nominal

Note that SPSS has still calculated **MEANS** on a **Nominal** variable !!

A comparison is needed of the alternative methods of calculating depression scores:

JM's method:

```
compute depress = sum.7(flt dpr, flteeff, slprl, fltlnl, fltsd, cldgng, enjlf2, wrhpp2).
if (missdep = 1) depress = (8/7)*depress.
```

The **IF** command over-writes the new variable: better to create another new variable:

```
if (missdep = 1) depress2 = (8/7)*depress.
```

Alternative method 1 (automatically eliminates Albania):

```
compute depress_8 = sum.8(flt dpr, flteeff, slprl, fltlnl, fltsd, cldgng, enjlf2, wrhpp2)-8.
```

Alternative method 2 (includes Albania):

```
compute depress_7 = sum.7(flt dpr, flteeff, slprl, fltlnl, fltsd, enjlf2, wrhpp2)-7.
```

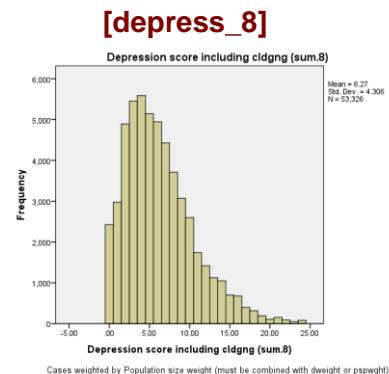
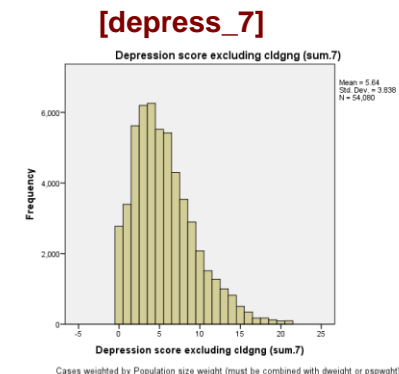
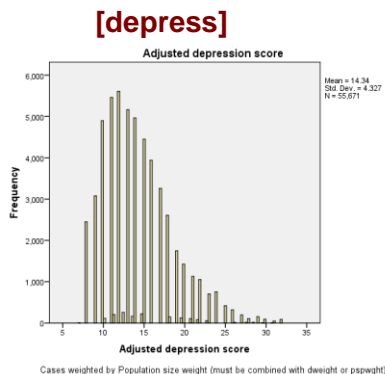
All methods:

```
formats depress depress2 depress_7 depress_8 (f2.0).
```

```
variable labels depress2 "Adjusted depression score"
/depress_7 "Depression score excluding cldgng (sum.7)"
/depress_8 "Depression score including cldgng (sum.8)".
```

```
frequencies depress depress_7 depress_8 /format notable /histogram .
```

		Statistics		
		Adjusted depression score	Depression score excluding cldgng (sum.7)	Depression score including cldgng (sum.8)
N	Valid	55671	54080	53326
	Missing	1164	2754	3508



correlations depress depress_7 depress_8.

Correlations		Adjusted depression score	Depression score excluding cldgng (sum.7)	Depression score including cldgng (sum.8)
Adjusted depression score	Pearson Correlation	1	.989	1.000
	Sig. (2-tailed)		.000	.000
	N	55671	54080	53326
Depression score excluding cldgng (sum.7)	Pearson Correlation	.989	1	.989
	Sig. (2-tailed)	.000		.000
	N	54080	54080	53326
Depression score including cldgng (sum.8)	Pearson Correlation	1.000	.989	1
	Sig. (2-tailed)	.000	.000	
	N	53326	53326	53326

[depress_7] and **[depress_8]** are perfectly correlated at **1.000** and correlate **0.989** with JM's tortuously derived and adjusted **[depress]**.

What happens if depression is imputed for all cases, based on depress_7? For countries other than Albania, how does the imputed score relate to the actual score?

End of: **5.1.6:** Using the **SUM** sub-command

Back to: [MacInnes \(2017\)](#)