[Commentary by **John F Hall**] [Last updated: 13 August 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 tutorial 5.1.5: Using **COMPUTE**[1] to create new variables. (4'10")

[**NB:** JM has all the syntax in the same file (messy) and is also working, not on a copy, but on the original data file (naughty!)]

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:	**COMPUTE** **RECODE**[2]
Technical terms:	**valid value**, **system missing**, **source variable**, **target variable** **numeric expression**

Task: Create an index of depression from a list comprising eight symptoms purporting to measure "depression", six negative and two positive.

Step 1: Reverse scoring of positively worded symptoms (page 2)
Step 2: Calculate an index of depression from all eight symptoms (page 7)

"I will now read out a list of the ways you might have felt or behaved during the past week. Using this card, please tell me how much of the time during the past week . . ."

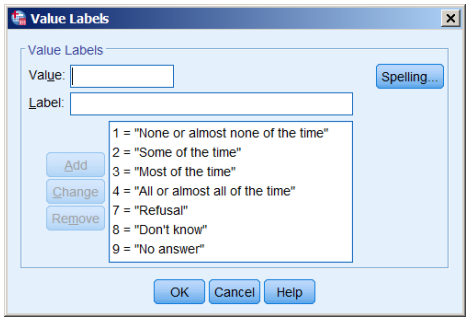
	None or almost none of the time	Some of the time	Most of the time	All or almost all of the time	(Don't know)	
D5 . . . .you felt depressed?	1	2	3	4	8	**[flteeff]s**
D6 . . . .you felt that everything you did was an effort?	1	2	3	4	8	**[slprl]**
D7 . . . .your sleep was restless?	1	2	3	4	8	**[flteeff]**
D8 . . . .you were happy?	1	2	3	4	8	**[wrhpp]**
D9 . . . .you felt lonely?	1	2	3	4	8	**[fltlnl]**
D10. . . .you enjoyed life?	1	2	3	4	8	**[enjlf]**
D11. . . .you felt sad?	1	2	3	4	8	**[fltsd]**
D12. . . .you could not get going?	1	2	3	4	8	**[cldgng]**

[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] For a brief introduction to the **RECODE** command, see 2.3.1.1 Data transformations (pp10,11) and 2.3.1.2a2 Recode into new variable

The associated variables are in rows 197 -204 of the **Data Editor**:

	Name	Label
197	fltdpr	Felt depressed, how often past week
198	flteeff	Felt everything did as effort, how often past week
199	slprl	Sleep was restless, how often past week
200	wrhpp	Were happy, how often past week
201	fltlnl	Felt lonely, how often past week
202	enjlf	Enjoyed life, how often past week
203	fltsd	Felt sad, how often past week
204	cldgng	Could not get going, how often past week

Values



JM does not give a detailed explanation of what he is trying to do, or why. Basically he wants to create an index of depression in which higher scores indicate higher levels of depression. However, he can't just add up all the item scores because the values of the two positive items **[wrhpp]** and **[enjlf]** run counter to the direction of coding for the six negative items and will offset the index downwards.

Step 1: Reverse scoring of positively worded symptoms

For the two positive symptoms of non-depression:

[wrhpp] **D8** " . . you were happy"
[enjlf] **D10** " . . you enjoyed life"

. . JM needs to reverse the scoring as follows:

	None or almost none of the time	Some of the time	Most of the time	All or almost all of the time	(Don't know)	
Original values						
D8 . . . .you were happy?	1	2	3	4	8	**[wrhpp]**
D10. . . .you enjoyed life?	1	2	3	4	8	**[enjlf]**
Reversed values:	↓	↓	↓	↓	↓	
D8 . . . .you were happy?	4	3	2	1	8	**[wrhpp2]**
D10. . . .you enjoyed life?	4	3	2	1	8	**[enjlf2]**

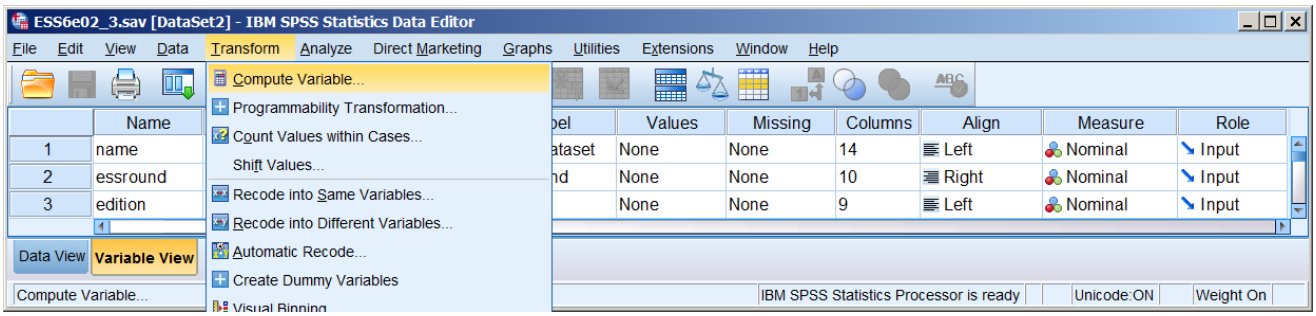
He states that the original values for these two items can be reversed by subtracting them from 5, but doesn't explain that this neat shortcut method applies to all Likert-type scales coded from 1 to n. In this case the range of **valid values** is 1 – 4, but for a range of 1 – 7 the values would be subtracted from 8, i.e. 1 higher than the highest point on the scale.

He also explains that if the **source variables** do not have valid values, the values for new variables will be set to the **system missing** value **SYSMIS**. This practice is not recommended by experienced users of SPSS: the original user-missing values should always be retained, but this requires more than a simple **COMPUTE** command.

When recoding the positive items it is also best practice to keep the original variables and create two new ones with the scores reversed, for instance using the original names with a prefix indicating the reversal, eg **[r_wrhpp]** and **[r_enjlf]**. JM does this by creating two new variables **[wrhpp2]** and **[enjlf2]**.

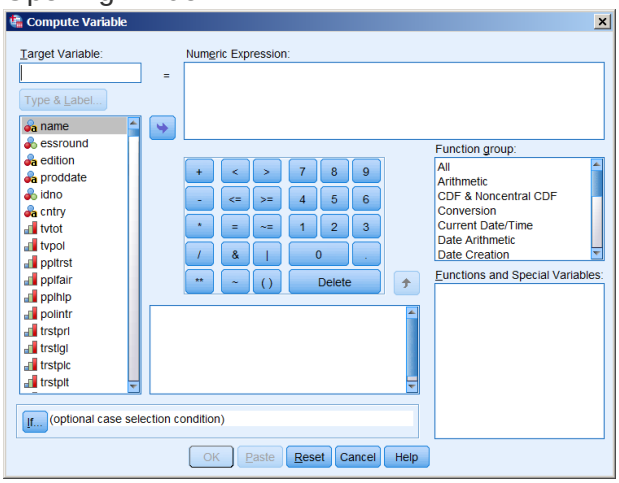
To create the first of these new variables, he uses the GUI

Transform >> Compute Variable

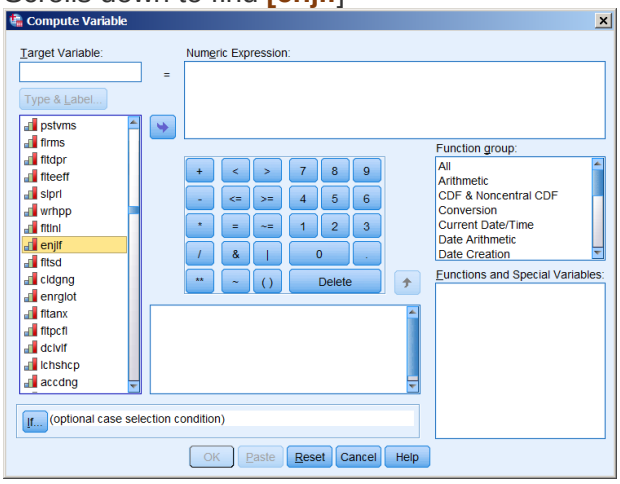


The opening **Dialog box** for **Compute Variable** displays a list of variable names in the left pane, starting at the top of the file. JM scrolls down the list looking for the first of his two positive symptoms **[enjlf]** (actually the second in the inventory). This could take users a long time as it is difficult to spot.

Opening window

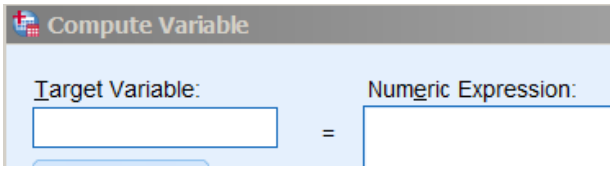


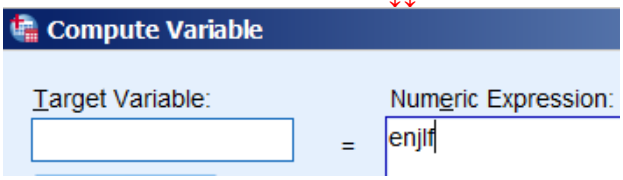
Scrolls down to find **[enjlf]**



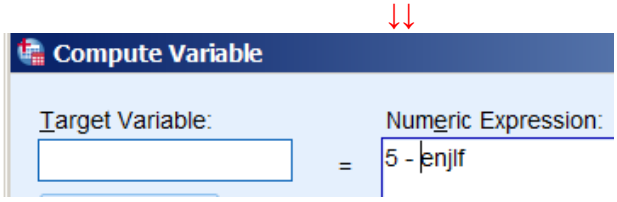
. .clicks on the blue arrow 

. . . to transfer **[enjlf]** to the Numeric Expression box:

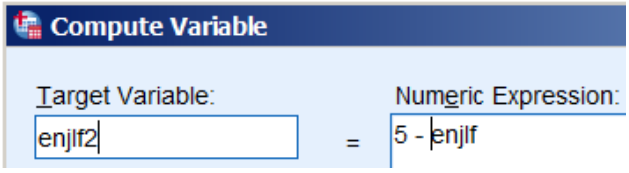




. . writes in **5 –** to make it **5 – enjlf**



Decides to call the new **target variable enjlf2** and writes it in the Target Variable box



Clicks on Paste to see the syntax generated by SPSS:

```
COMPUTE enjlf2 = 5-enjlf .
EXECUTE .
```

It would have been so much quicker to write, direct in the **Syntax Editor**:

```
compute enjlf2 = 5-enjlf.
```

To reverse the codes for the second variable **[wrhpp]** he does actually use direct syntax:

```
compute wrhpp2 = 5-wrhpp
```

Note the colour coding of **compute** until he types in the full stop.

```
compute wrhpp2 = 5-wrhpp .
```

He then runs the three commands:

```
COMPUTE enjlf2 = 5-enjlf .
EXECUTE .
compute wrhpp2 = 5-wrhpp .
```

If there are missing values in any of the **source variables**, JM's method sets the value of the new variables to **SYSMIS**. Recoding to **SYSMIS** is generally frowned on by researchers experienced in SPSS, but JM understandably excuses it for this exercise. In general it is better, and safer, to create new variables in which user-missing values are retained[3].

As a check, JM runs two separate crosstabs with direct (abbreviated) syntax:

```
cross enjlf2 by enjlf .
cross wrhpp2 by wrhpp .
```

enjlf2 * enjlf Enjoyed life, how often past week Crosstabulation

Count

		enjlf Enjoyed life, how often past week				
		1 None or almost none of the time	2 Some of the time	3 Most of the time	4 All or almost all of the time	Total
enjlf2	1.00	0	0	0	**13154**	13154
	2.00	0	0	**23661**	0	23661
	3.00	0	**15118**	0	0	15118
	4.00	**3845**	0	0	0	3845
Total		3845	15118	23661	13154	55778

[3] The reversing of values retaining user-missing values can be done with:

```
recode enjlf wrhpp (1=4)(2=3)(3=2)(4=1)(else = copy) into enjlf2 wrhpp2 .
missing values enjlf2 wrhpp2 (7,8,9) .
```

wrhpp2 * wrhpp Were happy, how often past week Crosstabulation

Count

		wrhpp Were happy, how often past week				
		1 None or almost none of the time	2 Some of the time	3 Most of the time	4 All or almost all of the time	Total
wrhpp2	1.00	0	0	0	**12560**	12560
	2.00	0	0	**26506**	0	26506
	3.00	0	**13898**	0	0	13898
	4.00	**2784**	0	0	0	2784
Total		2784	13898	26506	12560	55748

[**NB:** Neither of the variables **[enjlf2]** and **[wrhpp2]** has a variable or value label and both have 2 superfluous decimal places: they need changing to format (f2.0). Alternatively the SPSS settings can be changed to do this automatically, but some new numeric variables may actually need to be displayed with decimal places]

```
formats enjlf2 wrhpp2 (f2.0) .
crosstabs enjlf2 by enjlf
        / wrhpp2 by wrhpp .
```

enjlf2 * enjlf Enjoyed life, how often past week Crosstabulation

Count

		enjlf Enjoyed life, how often past week				
		1 None or almost none of the time	2 Some of the time	3 Most of the time	4 All or almost all of the time	Total
enjlf2	1	0	0	0	13154	13154
	2	0	0	23661	0	23661
	3	0	15118	0	0	15118
	4	3845	0	0	0	3845
Total		3845	15118	23661	13154	55778

wrhpp2 * wrhpp Were happy, how often past week Crosstabulation

Count

		wrhpp Were happy, how often past week				
		1 None or almost none of the time	2 Some of the time	3 Most of the time	4 All or almost all of the time	Total
wrhpp2	1	0	0	0	12560	12560
	2	0	0	26506	0	26506
	3	0	13898	0	0	13898
	4	2784	0	0	0	2784
Total		2784	13898	26506	12560	55748

The tables show that there are no cases with inconsistent pairings of values: all cases are on the main diagonal.

Users need to decide for themselves whether to add variable and value labels, keep the variables at the end of the file, move them to a new position or delete them once the new index of depression has been created. Good practice is to **keep the syntax in a separate file** for possible later use and as a record of what was done.

This section also warrants some explanation of the technical terms **Target Variable** and **Numeric Expression**

Syntax which retains the original user-missing values and specifies printing formats, missing values, measurement levels, variable labels and value labels is:

```
recode          enjlf wrhpp (1=4)(2=3)(3=2)(4=1)(else = copy) into enjlf2 wrhpp2 .
formats         enjlf2 wrhpp2 (f2.0) .
missing values  enjlf2 wrhpp2 (7,8,9) .
variable level  enjlf2 wrhpp2 (ordinal) .
variable labels enjlf2 "[enjlf] reverse coded" / wrhpp2 "[wrhpp] reverse coded".
value labels    enjlf2 wrhpp2
                1 "All or almost all of the time" 2 "Most of the time"
                3 "Some of the time" 4 "None or almost none of the time"
                7 "Refusal" 8"Don'tknow" 9 "No answer".
```

[**NB:** tabs inserted for clarity]

```
freq enjlf2 wrhpp2 .
```

enjlf2 [enjlf] reverse coded * enjlf Enjoyed life, how often past week Crosstabulation

Count

		enjlf Enjoyed life, how often past week				
		1 None or almost none of the time	2 Some of the time	3 Most of the time	4 All or almost all of the time	Total
enjlf2 [enjlf] reverse coded	1 All or almost all of the time	0	0	0	13154	13154
	2 Most of the time	0	0	23661	0	23661
	3 Some of the time	0	15118	0	0	15118
	4 None or almost none of the time	3845	0	0	0	3845
Total		3845	15118	23661	13154	55778

wrhpp2 [wrhpp] reverse coded * wrhpp Were happy, how often past week Crosstabulation

Count

		wrhpp Were happy, how often past week				
		1 None or almost none of the time	2 Some of the time	3 Most of the time	4 All or almost all of the time	Total
wrhpp2 [wrhpp] reverse coded	1 All or almost all of the time	0	0	0	12560	12560
	2 Most of the time	0	0	26506	0	26506
	3 Some of the time	0	13898	0	0	13898
	4 None or almost none of the time	2784	0	0	0	2784
Total		2784	13898	26506	12560	55748

The new variables will be appended to the **Data Editor** on lines 629 and 630.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
629	enjlf2	Numeric	2	0	[enjlf] reverse coded	{1, All or al...	7, 8, 9	10	Right	Ordinal
630	wrhpp2	Numeric	2	0	[wrhpp] reverse coded	{1, All or al...	7, 8, 9	10	Right	Ordinal

An alternative trick of the trade to check consistency is to use correlation:

```
correlations    wrhpp with wrhpp2
              / enjlf with enjlf2.
```

. . which produces perfect negative correlations of **-1.000** for both pairs.

Correlations for Analysis 1

		wrhpp2 wrhpp reverse coded
wrhpp Were happy, how often past week	Pearson Correlation	**-1.000**
	Sig. (2-tailed)	.000
	N	55748

Correlations for Analysis 2

		enjlf2 enjlf reverse coded
enjlf Enjoyed life, how often past week	Pearson Correlation	**-1.000**
	Sig. (2-tailed)	.000
	N	55779

Step 2: Calculate an index of depression from all eight symptoms.

Now that the scores on items **[wrhpp]** and **[enjlf]** have been reversed and stored in the new variables **[wrhpp2]** and **[enjlf2]** we can add the reversed scores to the scores on the other six variables to create an index of depression.

The SPSS command **COMPUTE**. has the general format:

COMPUTE <newvar> = <numerical expression> . For example:

compute depress = fltdpr + flteeff + slprl + wrhpp2 + fltlnl + enjlf2 + fltsd + cldgng .

That's the beauty of SPSS. Much of it is just like writing English, but if you mis-type variable names SPSS cannot run a spell-check on them: you'll get an error message. You need a bit of grammar as well, so don't forget to type a stop at the end of each command.

There is no need for an **EXECUTE** command as the calculation will automatically be performed when the next statistical procedure is performed:

frequencies depress .

depress

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8.00	54	.1	.1	.1
	9.00	73	.1	.1	.2
	10.00	462	.8	.9	1.1
	11.00	1090	1.9	2.0	3.1
	12.00	3904	6.9	7.3	10.5
	13.00	5874	10.3	11.0	21.5
	14.00	9226	16.2	17.3	38.8
	15.00	8143	14.3	15.3	54.1
	16.00	7034	12.4	13.2	67.2
	17.00	5342	9.4	10.0	77.3
	18.00	3737	6.6	7.0	84.3
	19.00	2715	4.8	5.1	89.4
	20.00	1921	3.4	3.6	93.0
	21.00	1308	2.3	2.5	95.4
	22.00	904	1.6	1.7	97.1
	23.00	539	.9	1.0	98.1
	24.00	361	.6	.7	98.8
	25.00	228	.4	.4	99.2
	26.00	224	.4	.4	99.7
	27.00	104	.2	.2	99.8
	28.00	49	.1	.1	99.9
	29.00	14	.0	.0	100.0
	30.00	5	.0	.0	100.0
	32.00	14	.0	.0	100.0
	Total	53326	93.8	100.0	
Missing	System	3508	6.2		
Total		56835	100.0		

Users still need to add a label to **[depress]** and get rid of the superfluous decimals:

```
variable labels depress "Score on 8-item depression scale" .
formats depress (f2.0) .
frequencies depress .
```

depress Score on 8-item depression scale

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	54	.1	.1	.1
	9	73	.1	.1	.2
	10	462	.8	.9	1.1
	11	1090	1.9	2.0	3.1
	12	3904	6.9	7.3	10.5
	13	5874	10.3	11.0	21.5
	14	9226	16.2	17.3	38.8
	15	8143	14.3	15.3	54.1
	16	7034	12.4	13.2	67.2
	17	5342	9.4	10.0	77.3
	18	3737	6.6	7.0	84.3
	19	2715	4.8	5.1	89.4
	20	1921	3.4	3.6	93.0
	21	1308	2.3	2.5	95.4
	22	904	1.6	1.7	97.1
	23	539	.9	1.0	98.1
	24	361	.6	.7	98.8
	25	228	.4	.4	99.2
	26	224	.4	.4	99.7
	27	104	.2	.2	99.8
	28	49	.1	.1	99.9
	29	14	.0	.0	100.0
	30	5	.0	.0	100.0
	32	14	.0	.0	100.0
	Total	53326	93.8	100.0	
Missing	System	3508	6.2		
Total		56835	100.0		

Finally, a measure with a range of 8 to 32 is not particularly easy to interpret: it would be more understandable if it were converted to a **ratio scale**[4] with a true 0 and a range of 0 – 24.

Standard procedure is to subtract the number of items in the scale (in this case 8) from the total score: JM does it keeping the same variable name (which over-writes any earlier values)

```
compute depress = depress - 8 .
freq depress.
```

A safer practice would be to create a new variable **depress2**[5]

[4] JM actually changes the range to 0 -24 in a later video, but then, as it's "easier for a lay audience to understand," divides it by 2.4 to change the range to 0-10 (which creates values with two decimal places)

[5] **compute** depress2 = depress - 8 .

depress Score on 8-item depression scale

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	54	.1	.1	.1
	1	73	.1	.1	.2
	2	462	.8	.9	1.1
	3	1090	1.9	2.0	3.1
	4	3904	6.9	7.3	10.5
	5	5874	10.3	11.0	21.5
	6	9226	16.2	17.3	38.8
	7	8143	14.3	15.3	54.1
	8	7034	12.4	13.2	67.2
	9	5342	9.4	10.0	77.3
	10	3737	6.6	7.0	84.3
	11	2715	4.8	5.1	89.4
	12	1921	3.4	3.6	93.0
	13	1308	2.3	2.5	95.4
	14	904	1.6	1.7	97.1
	15	539	.9	1.0	98.1
	16	361	.6	.7	98.8
	17	228	.4	.4	99.2
	18	224	.4	.4	99.7
	19	104	.2	.2	99.8
	20	49	.1	.1	99.9
	21	14	.0	.0	100.0
	22	5	.0	.0	100.0
	24	14	.0	.0	100.0
	Total	53326	93.8	100.0	
Missing	System	3508	6.2		
Total		56835	100.0		

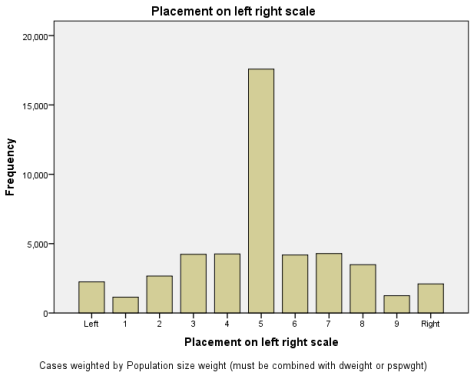
Because **[depress]** is effectively an **interval level** measure (ie **Scale** in SPSS parlance) it is legitimate to use statistics which are not permissible on **ordinal level** measures such as the eight individual symptoms, nor (strictly) on the (0-10) **[lrscale]** " Placement on left right scale".

For this reason the correct chart to use for **[lrscale]** is a **barchart**, (with spaces between the bars because there is no known fixed interval between the points on the scale). The only permissible measures for centrality are **median** or **mode**, and for spread, **range** or a **percentile** based measure such as **interquartile range**.

formats depress2 (f3.0).
variable labels depress2 'Depression score modified to 0-24'.

SPSS allows you to **suppress**[6] the table and display only the barchart, median, mode and percentiles:

freq lrscale /for not /bar /sta med mod /per 25 75.



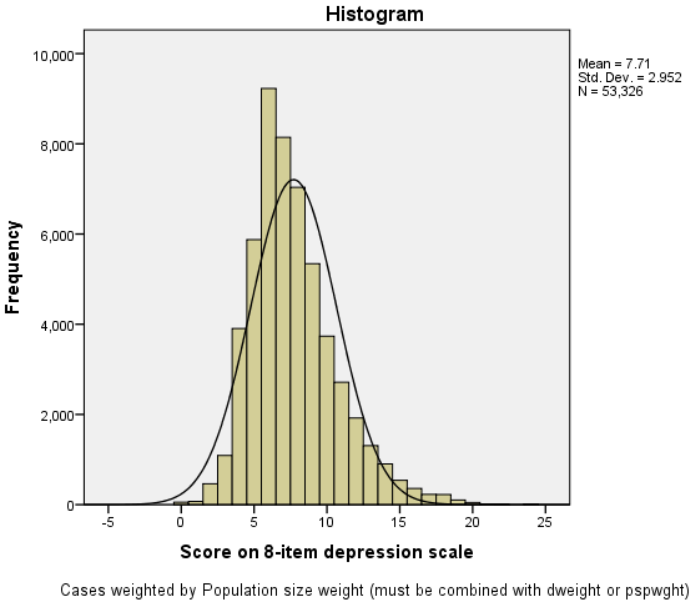
Statistics

enjlf Enjoyed life, how often past week

N	Valid	55779
	Missing	1056
Median		3.00
Mode		3
Percentiles	25	2.00
	75	3.00

For **[depress]** we can use a **histogram** (in which the bars are touching because there is a known fixed interval between the points) and for which it is legitimate to calculate descriptive statistics such as **mean** and **standard deviation**. SPSS allows you to suppress the table and display only the histogram[7]. Mean and standard deviation are automatically displayed and the histogram can optionally be overlaid with a **normal distribution curve**:

freq depress /for not /his nor .



End of: 5.1.5 Using **COMPUTE** to create new variables

Back to: MacInnes (2017)

[6] **frequencies** lrscale /format notable /barchart /statistics median mode /percentiles 25 75 .
[7] **frequencies** depress /format notable /histogram normal.