Chi Square

Example: An insurance company was interested in finding out if there is a significant association between the age of a driver and whether or not s/he is likely to exceed the speed limit. The speed of 120 younger drivers (aged 18-30) was recorded: of these, 100 were found to exceed the speed limit. The speed of 85 older drivers (aged 40+) was also recorded. Of these, 5 were found to exceed the speed limit.

The above data take the form of 'frequencies' and can be represented thus in the following **contingency table**.

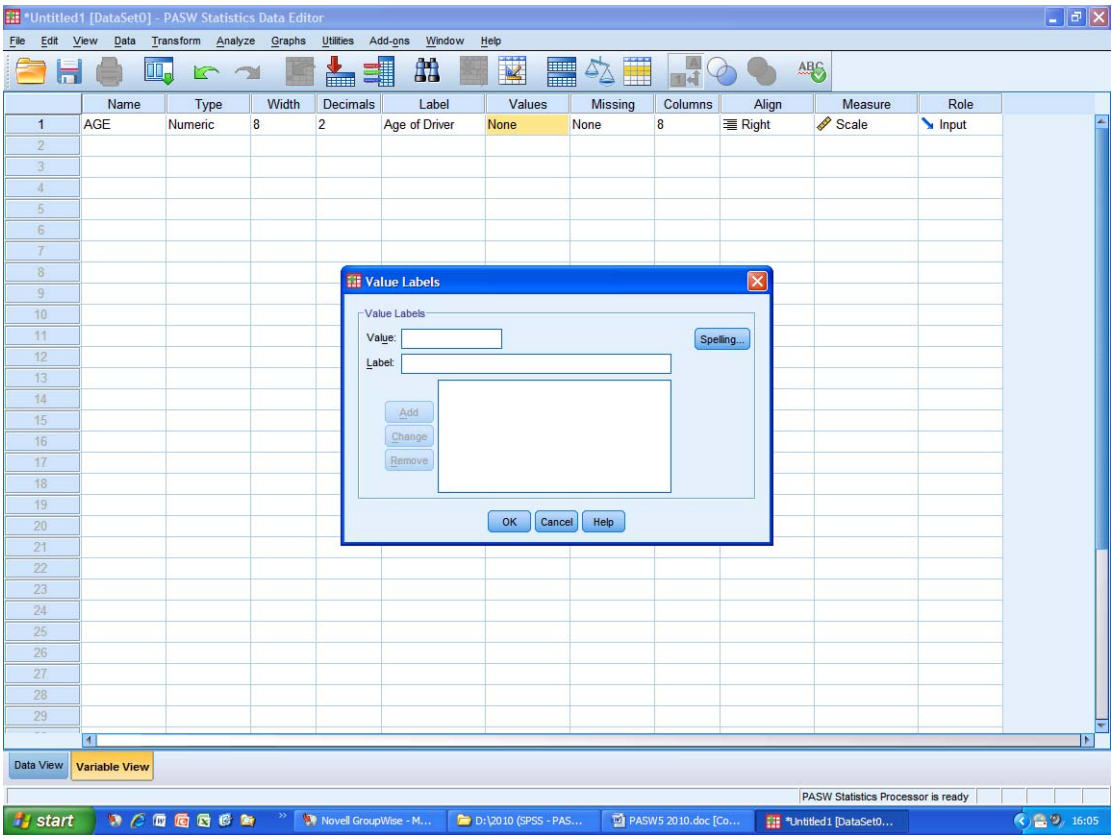
This section of the workbook uses Chi Square to examine whether or not there is a significant association between the age of a driver and whether or not the speed limit is likely to be exceeded by that driver.

	Younger	Older
Exceeded	**100**	**5**
NOT exceeded	**20**	**80**

To enter the above data into *IBM SPSS Statistics*

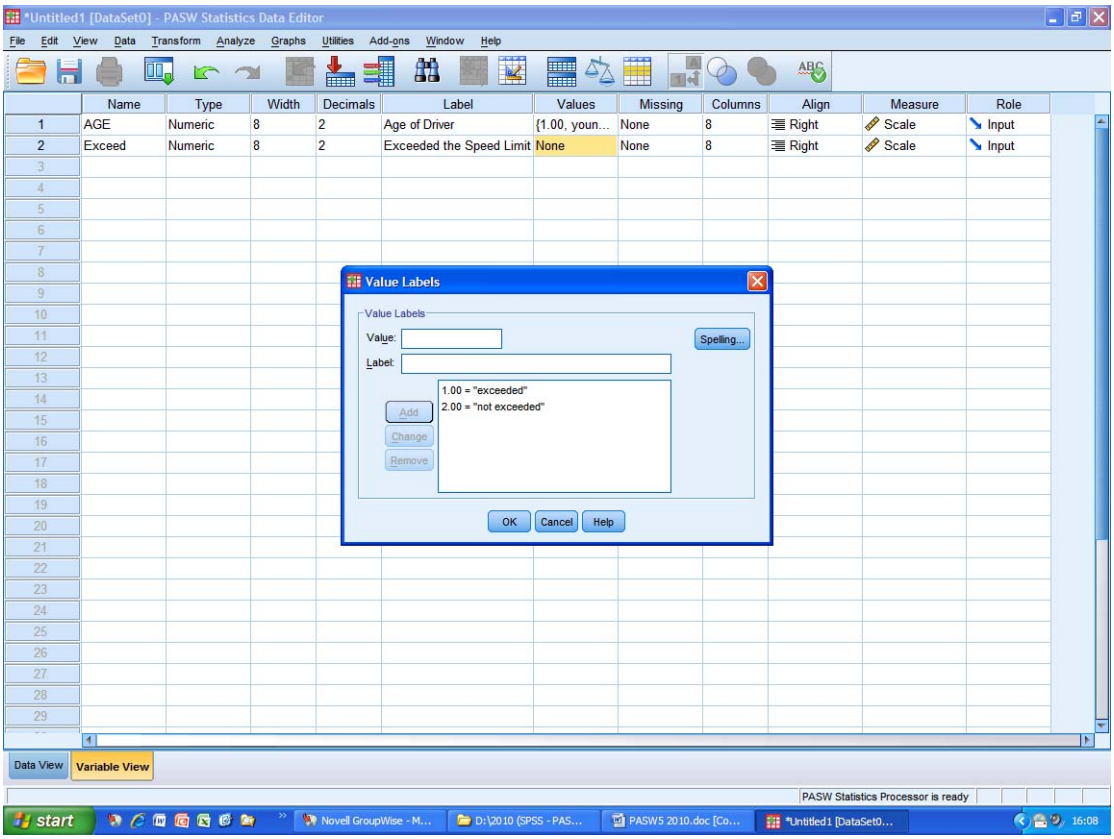
► Load *IBM SPSS Statistics*

► Click into the option **type in data**
► **OK**
► Click on **Variable View** at the bottom/left of screen
► In the **[Name]** box (top/left) enter the word **AGE**
► Enter a more full label such as **Age of Driver** in the [Label] box
► To the right click in the box under **[Values]**
► Click on the **[…]** area / box
You should now have on the screen two boxes named **Value:** and **Label:**



► Code younger drivers as **1** and older drivers as **2**
► To do this: in the **[Value:]** box type in the digit **1**
► Underneath this in the **[Label:]** box type in the word **younger**
► Click on **[Add]**
► Repeat this procedure to code **older** drivers as **2**
► Click on **[Add]**
► **OK**

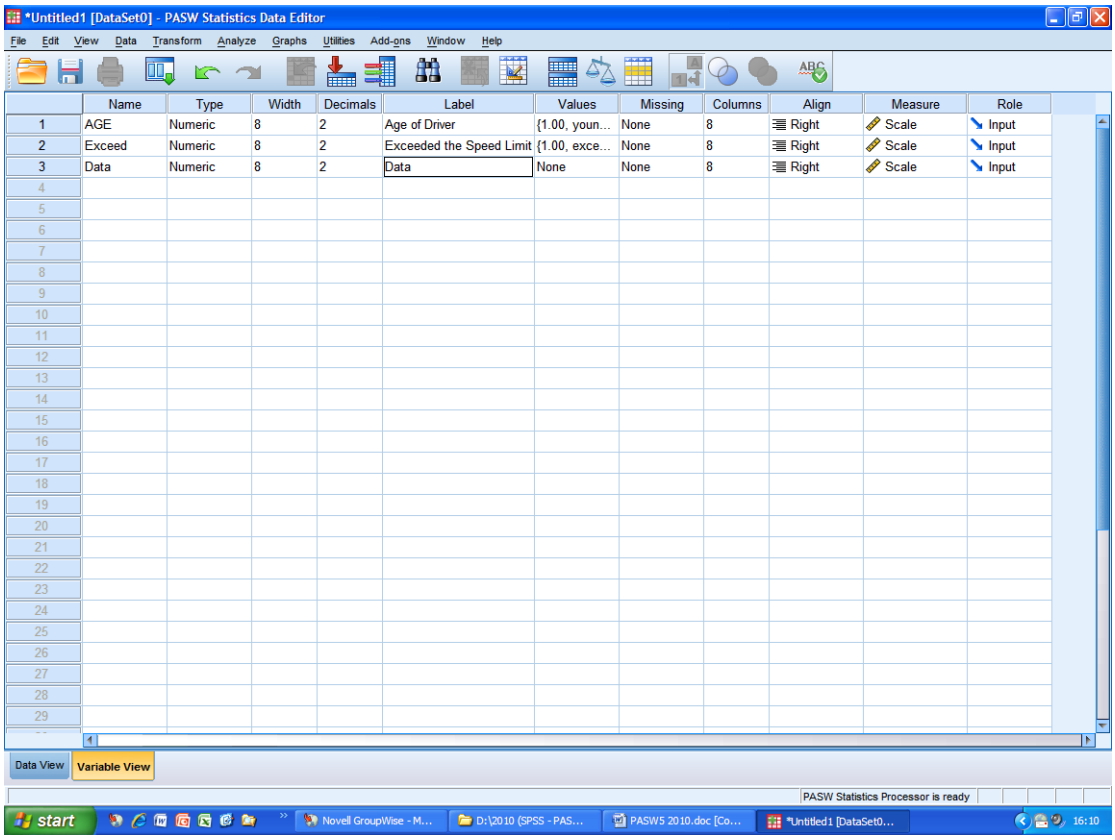
► On the next line below the above entries (i.e. under where the word **AGE** appears / top-left), repeat the above procedure to code for whether or not the speed limit was exceeded.
► Type in the word **Exceed** in the **[Name]** box
► In the **[Label]** box give this a full label such as **Exceeded the Speed Limit**
► Code **1 = exceeded**
► Code **2 = not exceed**
► Remember to click on **[Add]** each time



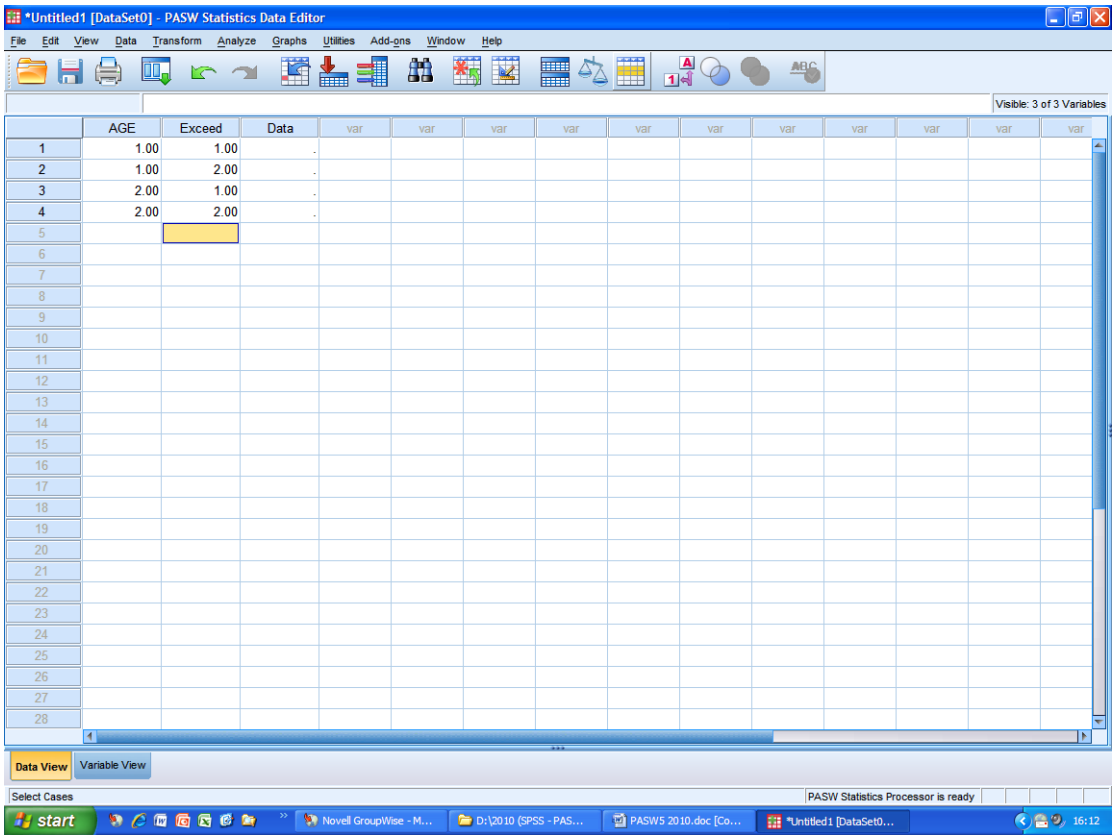
► **OK**

► Now we need to insert the **data** variable. On the next line (there are going to be 3 lines of coding in all) click into the box underneath the word **Exceed** / in the **[Name]** column
► Type in the word **Data**
► Also type in the word **Data** in the **[Label]** box

As we will be inserting the actual data, there is no need to go through any coding process for this line. Your screen should now look like this:



► With all **<u>three</u>** lines entered, click onto **Data View** (bottom/left of the screen). Type in the numbers thus:

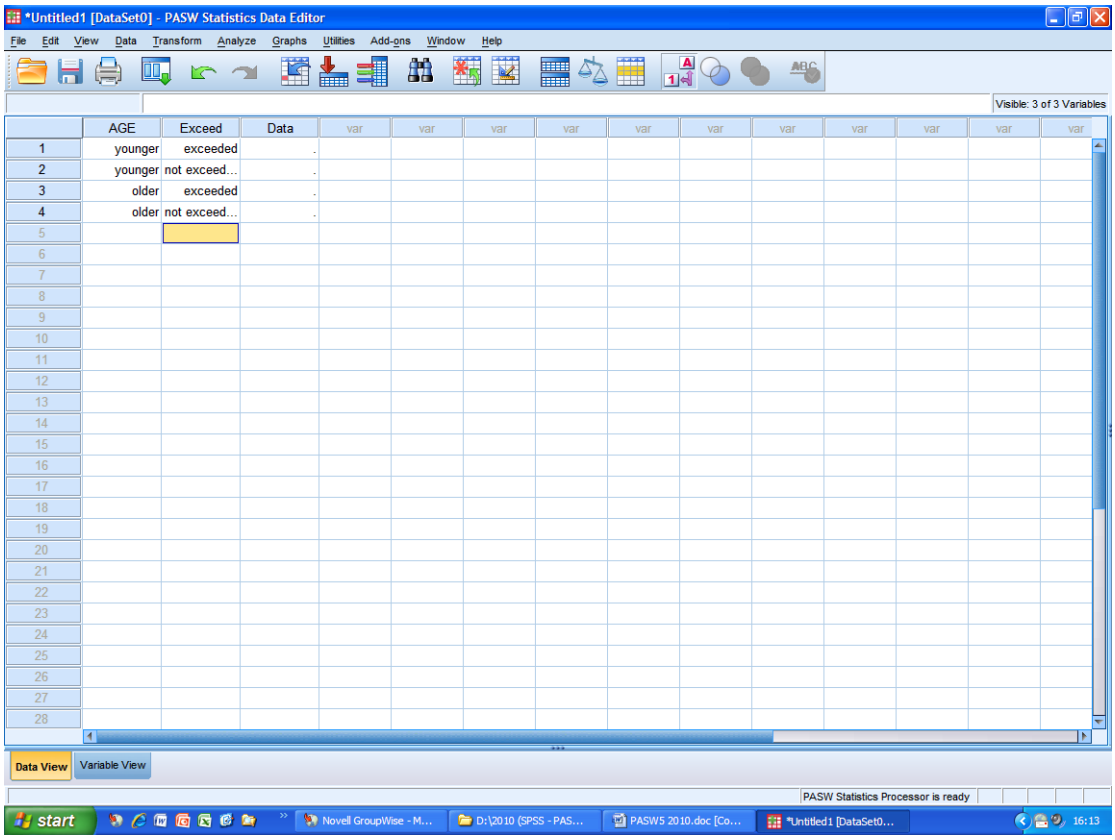


There is a series of icons at the top of the screen. Towards the right is an icon that has an illustration (in red) of a capital letter **A** and a digit **1**. This icon is fourth from the right in *IBM SPSS Statistics*. If you rest the cursor over this icon you will see that it is called **'Values Labels'**.

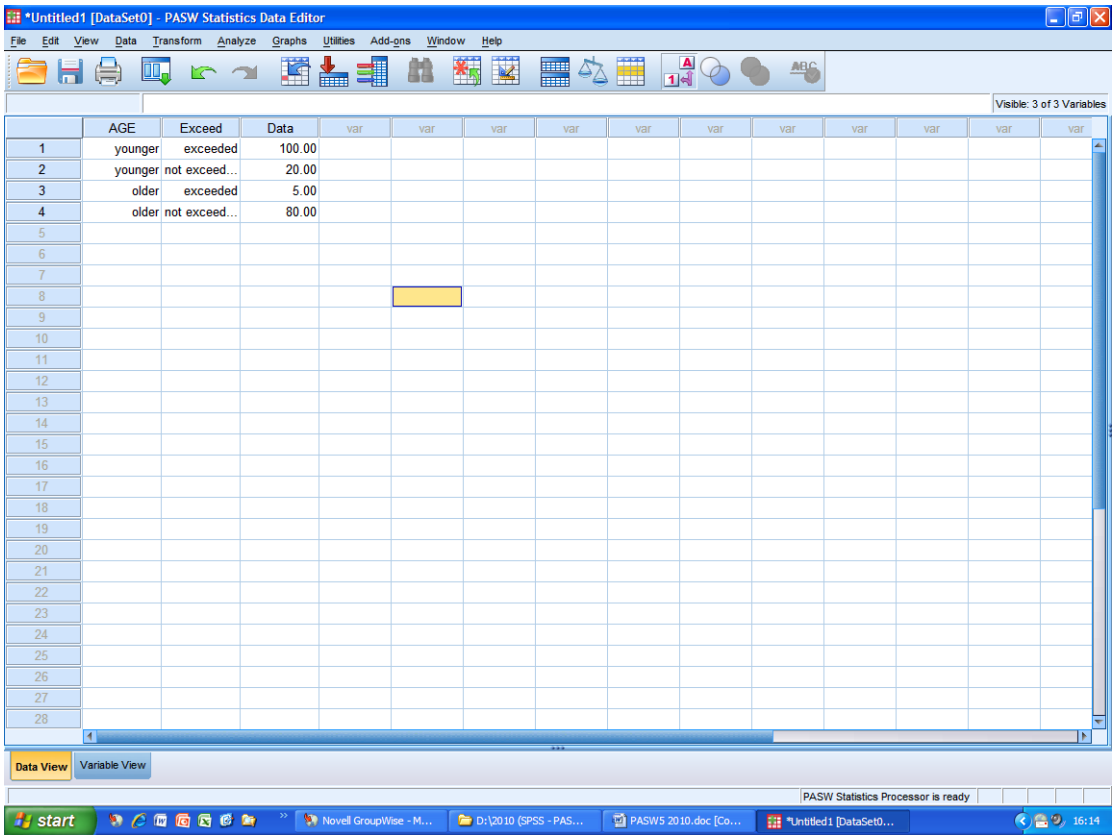
You will now see your coding in place. If you click on this icon you will see instantly all four possible categories in words. 1-1 = younger drivers who exceeded the speed limit; 1-2 = younger drivers who didn't; 2-1 = older drivers who exceeded; and 2-2 = older drivers who didn't.

(To 'toggle' between the labels (coding) and what they actually represent, simply keep clicking onto this icon).

Before entering the actual data, it is a good idea to have clicked on the **'Value Labels'** icon so that your screen appears thus:



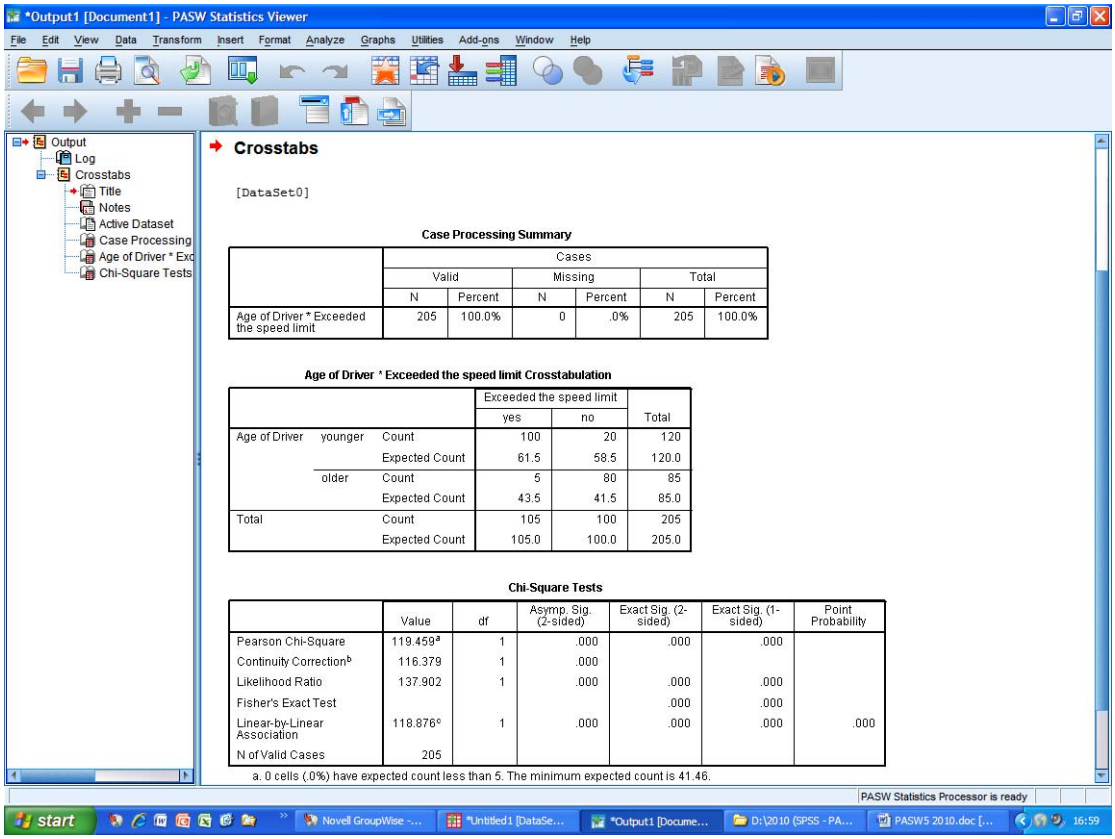
Now enter the data in the appropriate boxes above. Thus:



To perform the ***chi-square*** test of association

► Click on **[Data]** towards the top-left of the screen
► **Weight Cases…**
► Click into the **[Weight cases by…]** circle to select this option
► Click onto **Data [Data]** – just to the left in the box where it reads **weight cases by** - in order to highlight this
► Click onto the forward-facing arrow to insert the **Data** variable into the **[Frequency Variable]** box
► Click on **[OK]**
► **Analyse**
► **Descriptive Statistics**
► **Crosstabs…**
► Click to place **Age of Driver** into **[Rows]**
► Click to place **Exceeded the Speed Limit** into **[Columns]**
► **Statistics ► Chi-Square ► Continue ►**
► **Exact ►** <u>select</u> **exact ► Continue**
► **Cells ►** click to select **Observed** <u>and</u> **Expected**
► **Continue**
► **OK**

You should now be provided on screen with the **data output** page which follows below:



Crosstabs

[DataSet0]

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age of Driver * Exceeded the speed limit	205	100.0%	0	.0%	205	100.0%

Age of Driver * Exceeded the speed limit Crosstabulation

			Exceeded the speed limit		
			yes	no	Total
Age of Driver	younger	Count	100	20	120
		Expected Count	61.5	58.5	120.0
	older	Count	5	80	85
		Expected Count	43.5	41.5	85.0
Total		Count	105	100	205
		Expected Count	105.0	100.0	205.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	119.459[a]	1	.000	.000	.000	
Continuity Correction[b]	116.379	1	.000			
Likelihood Ratio	137.902	1	.000	.000	.000	
Fisher's Exact Test				.000	.000	
Linear-by-Linear Association	118.876[c]	1	.000	.000	.000	.000
N of Valid Cases	205					

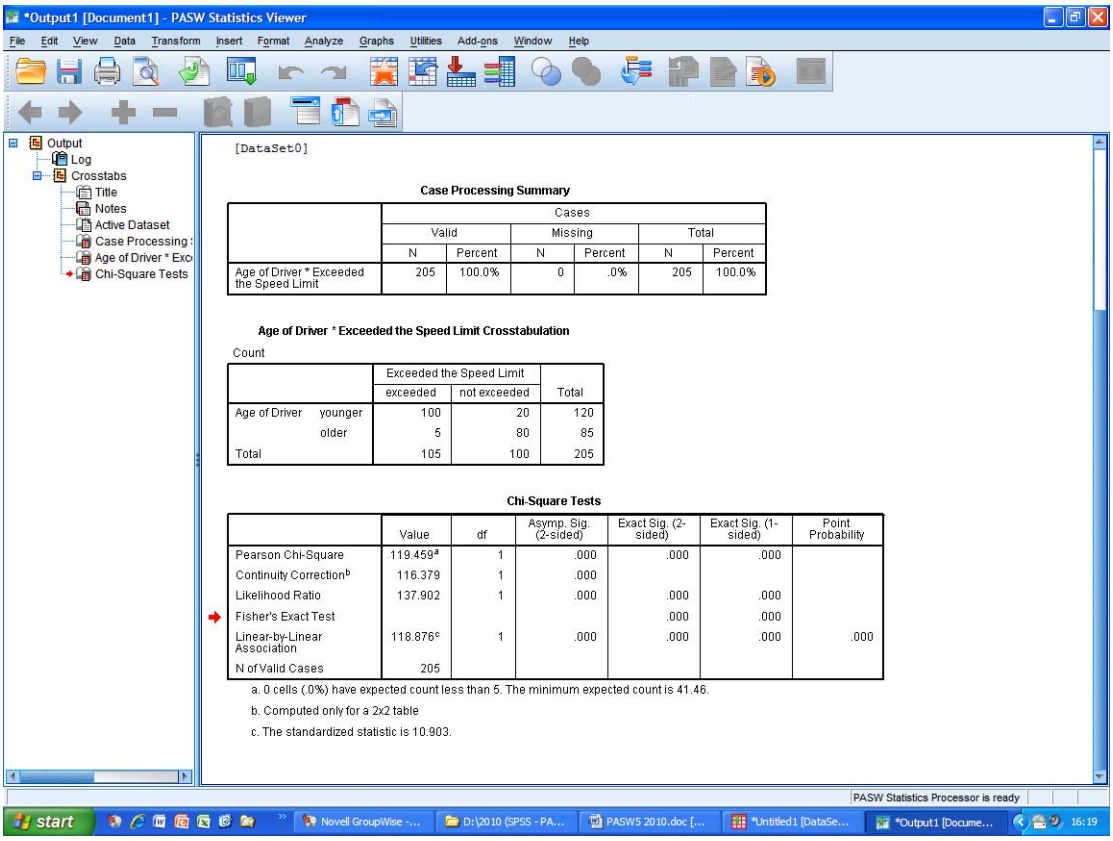
a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 41.46.

Explanation of the above *IBM SPSS Statistics* output

(1) The middle of the output page presents what is called a 'contingency table'. Clearly this is a **2 x 2** *Chi-Square*. You will notice that all **4** of the *Expected Frequencies* are clearly indicated. The large differences between the *Expected Frequencies* **(E)** and the *Observed Frequencies* **(O)** are what contribute towards the very high *Chi-Square* value that this data analysis has yielded. Have a look at the formula for calculating *Chi-Square* and you will understand that the bigger the difference between the values of **O** and **E**, then the bigger the value of *Chi-Square* that will be generated.

(2) The final table provided here shows the value of *Chi-Square* obtained. The one that we are interested in is the top line: **Pearson Chi-Square**. Here the value of 1.195E2 has been achieved. (Alternatively, this may be presented as 119.46. Refer to (3) below). This is massively significant. This should come as no surprise: just look at the contingency table to see how the data - for this fictitious example - have been extremely contrived to suggest that young drivers really do tend to break the speed limit whilst older drivers are particularly law-abiding.

(3) What the value of 1.195E2 actually means can be found by clicking onto the box presenting the Chi Square values (i.e. the bottom box). You will need to click once into the box; then twice onto the value presented of 1.195E2; then twice onto the value again. You will then be presented with the following display:



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age of Driver * Exceeded the Speed Limit	205	100.0%	0	.0%	205	100.0%

Age of Driver * Exceeded the Speed Limit Crosstabulation

Count

		Exceeded the Speed Limit		Total
		exceeded	not exceeded	
Age of Driver	younger	100	20	120
	older	5	80	85
Total		105	100	205

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	119.459[a]	1	.000	.000	.000	
Continuity Correction[b]	116.379	1	.000			
Likelihood Ratio	137.902	1	.000	.000	.000	
Fisher's Exact Test				.000	.000	
Linear-by-Linear Association	118.876[c]	1	.000	.000	.000	.000
N of Valid Cases	205					

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 41.46.

b. Computed only for a 2x2 table

c. The standardized statistic is 10.903.

Hence the value of Chi obtained is **119.46**

IBM SPSS Statistics sometimes will initially provide the value of Chi (as in this case) in the format such as (as in this case) 1.195E2. This style of presentation is known as 'scientific (floating point) notation'. It simply means that 1.195E2 is the equivalent of moving the decimal point (of the value 1.195) two places to the right. Moving one place to the right would give a value of 11.95. Moving two places gives a value of 119.5.

Hence 1.195E2 is another way of presenting 119.5

Clicking on the value (as indicated at the start of this note) 'translates' this value of Chi (119.5) into the more exact actual value obtained in this analysis of 119.458917.

In the presentation of values of Chi the convention is to present to two decimal places – hence 119.46

(4) **Note**: The significance level has 'gone off the scale', so to speak. Remember that (as mentioned before) whenever you are provided with a significance level of .000 by *IBM SPSS Statistics* this means that the results are significant at p = .0009 <u>at least</u>. Thus, this should be reported as p < .001

(5) On the data output page you will note that it reads: *Exact Sig. (2-sided)* ***OR*** *Exact Sig. (1-sided)*. Well, you need to decide whether you would be expecting this to be one-tailed or two-tailed in terms of the **<u>direction</u>** of any possible association between age and driving speed. When you have any doubts at all, it is advisable to go for two-sided (2-tailed). In this example, perhaps young people could be expected to exceed the speed limit more frequently than do older people because younger people are known to take more risks in life than older people. One the other hand, perhaps one could argue that older people (being more experienced drivers – generally speaking) might drive more quickly than younger people because they feel more competent and thus feel 'safer' doing so than younger people.

(6) **Note**: *Chi-Square* is **<u>not</u>** a valid test to use if any of the expected[1] counts (or 'frequencies') falls below **5**. This explains the footnote (a) on the data output

[1] The actual rule is: if 20% or more of the cells has an expected frequency of less than **5**, then *Chi-Square* cannot be employed. With a 2 x 2 *Chi-Square* (i.e. 4 cells) obviously none of the expected frequencies should fall below 5.

page. This footnote just reassures the user of *IBM SPSS Statistics* that the test is valid in this case.

APA style reporting on the results of the above statistical analysis

A chi-square analysis was performed to examine a possible significant association between the age of a driver and whether or not s/he was likely to exceed the speed limit. A significant association was observed, $\chi^2(1, N = 205) = 119.46$, $p < .001$. Younger drivers were more likely to exceed the speed limit than older drivers.

Workshop Exercise (5): Chi Square

It has been suggested that the media generally portray males who smoke cigarettes in a more positive light than females who smoke. It was hypothesized that this might result in (or, at least contribute to) more males smoking than females. It was also thought that this might be particularly true of school children. A survey was conducted amongst a group of **100** people aged between 13 and 16 years of age. The following data were obtained:

- **20** people reported that they **did** smoke – **18** of these were male.
- **80** people reported that they did **not** smoke – **75** of these were female.

1. Conduct a 2 x 2 Chi-Square analysis of these data using *IBM SPSS Statistics.*
2. Does the value of Chi obtained indicate a significant association between gender and smoking?
3. If so, at what level is the association significant?