

## **Using SPSS to perform Chi-Square tests:**

This handout explains how to perform the two types of Chi-Square test that were discussed in the lecture on Chi-Square last term: the Chi-Square Goodness of Fit test, and the Chi-Square test of association between two variables. (See the "Chi-Square test" on my website, [www.sussex.ac.uk/Users/grahamh/teaching06](http://www.sussex.ac.uk/Users/grahamh/teaching06), for more information on the Chi-Square test and how to calculate it by hand).

### **1. The Chi-Square Goodness of Fit test:**

The most common use of this test is to see whether or not instances of a number of categories have occurred equally frequently. The example used in the lecture last term was shoppers' preference for various soap-powder names. Suppose each shopper is given a list of four soap-powder names ("Kostik", "Smelloff", "Noscum" and "Grungefree") and asked to pick the one they like best. Our data consist of how many people pick each soap-powder; in other words, each person falls into one (and only one) of four categories. If names are chosen at random (i.e. shoppers show no consistent preference for one soap-powder over any other) then similar numbers of shoppers will choose each of the four names. However if there are any consistent preferences for soap-powder names, then one or more names will have a higher frequency of being chosen than the others. The Chi-Square Goodness of Fit test enables us to see whether the observed pattern of frequencies, obtained from our data, differs significantly from the frequencies we would expect to get by chance (i.e., all categories having roughly similar frequencies).

#### **Data entry:**

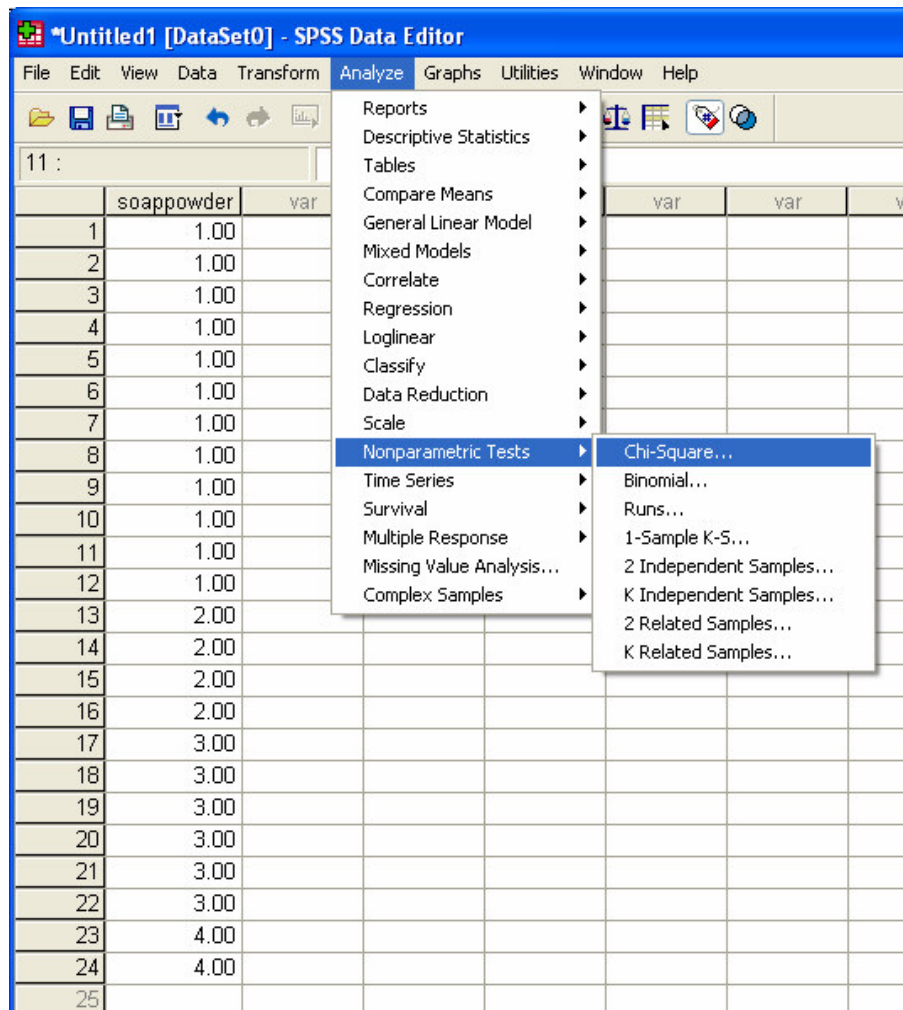
There are two quite different ways of entering data into SPSS in order to perform a Chi-Square test.

#### **(a) Using Chi-Square on raw data:**

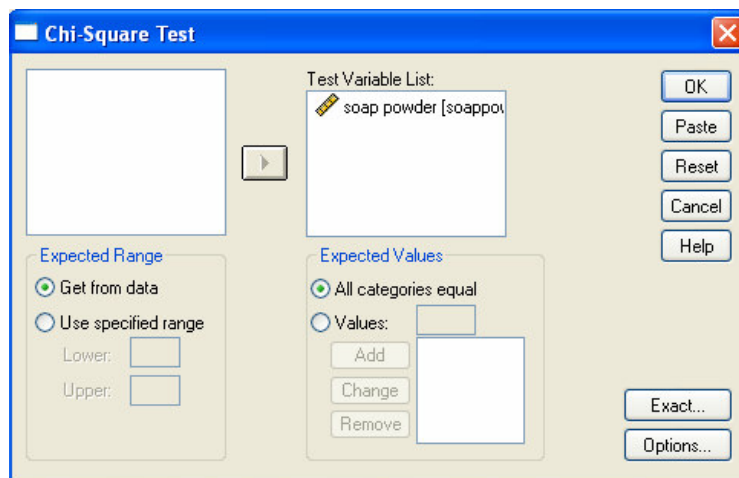
Using Chi-Square is most straightforward when you have all of the raw data. You produce one column that tells SPSS what each participant's choice was. In the example below, I've got 24 shoppers. I've used "1" as a code to represent "Kostik", "2" as a code for "Smelloff", "3" for "Noscum" and "4" for "Grungefree".

In line with the conventions for SPSS, each row represents one participant. Thus the first entry in the "soappowder" column is one participant's choice - "1", or "Kostik". The second entry is another participant's choice (also "Kostik") and so on. (I've not bothered to do it here, but you could go to "Variable view", and use "Values" to assign the name of each soap-powder to its particular code-number; this makes life easier when it comes to looking at the SPSS output).

To perform the Chi-Square Goodness of Fit test, go to "Analyze"; select "Nonparametric tests"; and then click on "Chi-Square".



The following dialog box appears. Click on the name of the variable containing the data, and then click on the arrow to move this name to the "Test variable list" .box. Then click on "OK" to run the Chi-Square test.



You should get the following output. First, you get a table that contains the observed and expected frequencies for each of the soap-powders. It also contains the "residuals", the difference between the observed and expected frequency or each category. Here, for example, you can see that more people picked soap-powder 1 ("Kostik") than would be expected by chance. Conversely, fewer people picked powders 2 ("Smelloff") and 4 ("Grungefree") than would be expected by chance. About as many people chose soap-powder 3 ("Noscum") as would be expected: there is no difference between the observed and expected frequencies.

	Observed N	Expected N	Residual
1.00	12	6.0	6.0
2.00	4	6.0	-2.0
3.00	6	6.0	.0
4.00	2	6.0	-4.0
Total	24		

**Test Statistics**

	soap powder
Chi-Square(a)	9.333
df	3
Asymp. Sig.	.025

a 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 6.0.

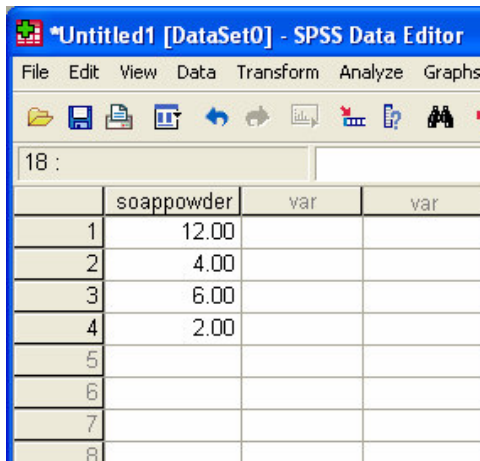
The other part of the output is the result of the Chi-Square test. The table shows the value of Chi-Square, the degrees of freedom, and the probability of obtaining this Chi-Square value merely by chance. Remember that in order for Chi-Square to be valid, no more than 20% of the expected frequencies should be less than 5. SPSS warns you if you violate this rule. Here, none of our expected frequencies is less than 5, so we are OK to use Chi-Square.

Our obtained Chi-Square value has a  $p$  of .025: this is less than .05, and so we would conclude that our observed frequencies of soap-powder name choice are significantly different from what we would expect to get by chance - i.e., the soap-powder names are not equally likely to be chosen by shoppers.

**(b) Using Chi-Square on summary data:**

Sometimes, you already have the frequency for each category -i.e., in this case, you have the observed frequencies of 12, 4, 6 and 2. SPSS allows you to perform SPSS using these category summaries, rather than having to type in 24 rows of data! However, it is a little more complicated because you need to prevent SPSS from treating the summaries as raw data.

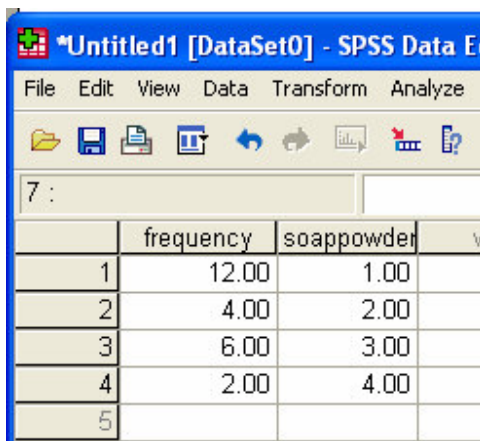
Here, if we simply type in the number of occurrences for each category (i.e. the number of shoppers picking each soap-powder name), SPSS will think we have four participants, scoring 12, 4, 6 and 2 respectively! We need to force SPSS to treat these values as frequencies for our four categories.



	soappowder	var	var
1	12.00		
2	4.00		
3	6.00		
4	2.00		
5			
6			
7			
8			

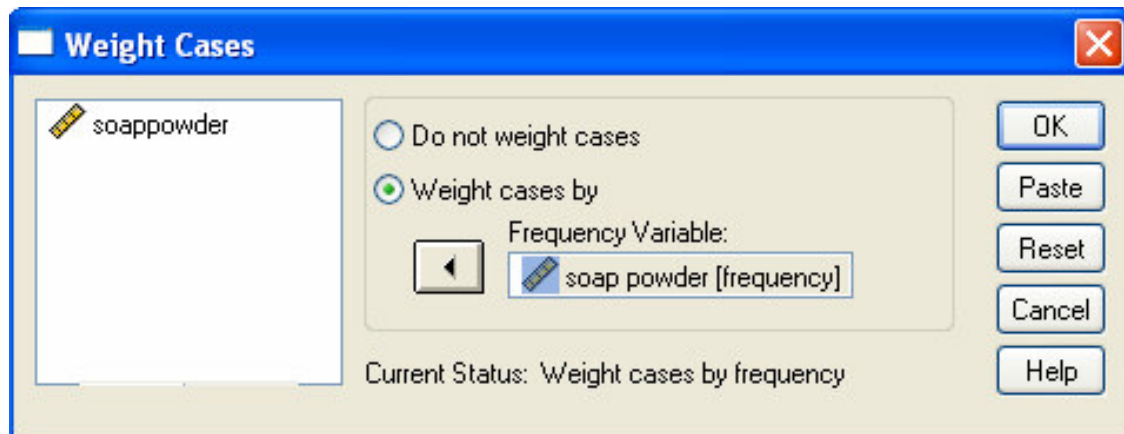
Here's how to do this.

1. You need two columns. One contains the frequency with which each category occurred. The other gives the category identifiers (1 to 4 as before, representing the names of the four soap-powders).



	frequency	soappowder	v
1	12.00	1.00	
2	4.00	2.00	
3	6.00	3.00	
4	2.00	4.00	
5			

2. Click on "Data", and then click on "Weight cases" at the bottom of the menu that appears. The following dialog box appears:



3. Click on "weight cases". Then put the variable that contains the frequency data (the number of occurrences of each category) into the "Frequency Variable" box, as above. Then click "OK". SPSS will now treat the numbers in the "frequency" column as the totals for the categories identified in the "soappowder" column. (In other words, by using the "weight cases" option, we have fooled SPSS into thinking that we have typed in "Kostik", "Smelloff", "Noscum" and "Grungefree" 24, 4, 6 and 2 times respectively).

4. Now run the Chi-Square analysis as before, using the "frequency" column in the "Test Variable List" box. You should get exactly the same results for the Chi-Square test as you did when using method (a).

## 2. The Chi-Square test of association between two independent variables:

This is the most common use of Chi-Square in psychology. We have categorical data for two independent variables and we want to see if there is some relationship between them. As with the Goodness of Fit test, there are two ways of entering the data, depending on whether you enter each participant separately, or want to use the summary frequency with which each category occurred.

### (a) Using Chi-Square on raw data:

Suppose we want to see if there is an association between brand of anti-dandruff shampoo ("Noflakes" and "Head and Shudders") and hair loss (totally bald versus no hair loss). In this case, we would have two columns. One would give the brand of shampoo that a participant used (coded 1 for "Noflakes" or "2" for "Head and Shudders") and the other would give the same participant's state of hairiness (coded with "1" for "bald" or "2" for "full head of hair").

If there is no association between hair loss and shampoo brand, we would expect to see as many slapheads using "Noflakes" as using "Head and

Shudders". On the other hand, if there is an association between the two variables, there should be a greater number of bald people using one shampoo rather than the other.

To perform the Chi-Square test of association on raw data, you need a row for each participant. One cell in that row tells SPSS which shampoo that participant used. The other cell in that row tells SPSS whether or not the same participant has hair. You thus end up with two columns of data, like this. (On the left, I've left the codes as numbers, while on the right I've used "variable label" to replace them with more meaningful labels):

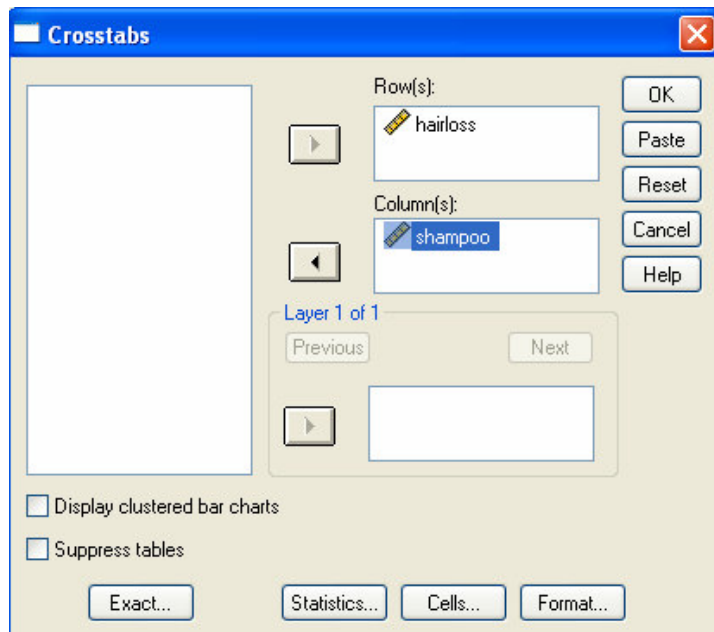
The screenshot shows the SPSS Data Editor window with a data table. The table has four columns: 'hairloss', 'shampoo', and 'var'. The 'hairloss' column contains values 1.00 and 2.00. The 'shampoo' column contains values 1.00 and 2.00. The 'var' column is empty. The rows are numbered 3 through 31.

	hairloss	shampoo	var
3	1.00	1.00	
4	1.00	1.00	
5	1.00	1.00	
6	1.00	1.00	
7	1.00	1.00	
8	1.00	1.00	
9	1.00	1.00	
10	1.00	1.00	
11	1.00	1.00	
12	1.00	1.00	
13	1.00	2.00	
14	1.00	2.00	
15	1.00	2.00	
16	2.00	1.00	
17	2.00	1.00	
18	2.00	1.00	
19	2.00	1.00	
20	2.00	1.00	
21	2.00	2.00	
22	2.00	2.00	
23	2.00	2.00	
24	2.00	2.00	
25	2.00	2.00	
26	2.00	2.00	
27	2.00	2.00	
28	2.00	2.00	
29	2.00	2.00	
30	2.00	2.00	
31			

The screenshot shows the SPSS Data Editor window with a data table. The table has three columns: 'hairloss', 'shampoo', and an unlabeled column. The 'hairloss' column contains values 'bald' and 'hairy'. The 'shampoo' column contains values 'Noflakes' and 'Head and Shudders'. The unlabeled column is empty. The rows are numbered 1 through 31.

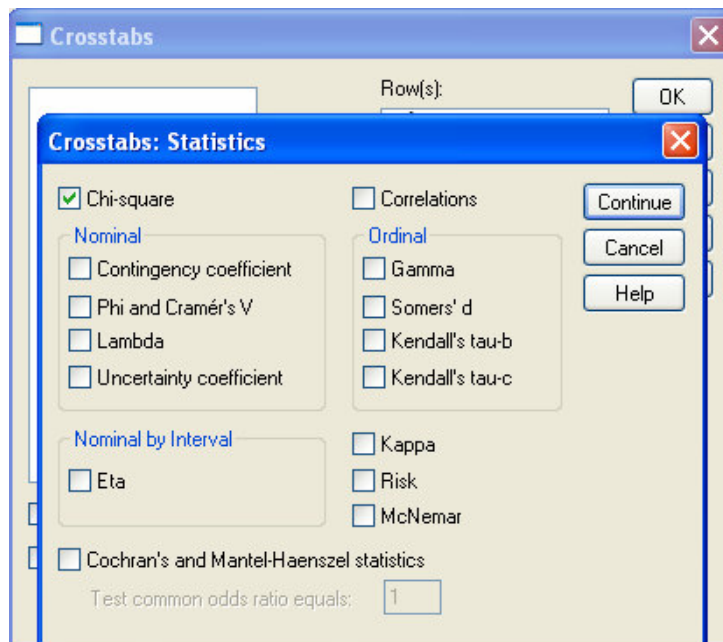
	hairloss	shampoo	
1	bald	Noflakes	
2	bald	Noflakes	
3	bald	Noflakes	
4	bald	Noflakes	
5	bald	Noflakes	
6	bald	Noflakes	
7	bald	Noflakes	
8	bald	Noflakes	
9	bald	Noflakes	
10	bald	Noflakes	
11	bald	Noflakes	
12	bald	Noflakes	
13	bald	Head and Shudders	
14	bald	Head and Shudders	
15	bald	Head and Shudders	
16	hairy	Noflakes	
17	hairy	Noflakes	
18	hairy	Noflakes	
19	hairy	Noflakes	
20	hairy	Noflakes	
21	hairy	Head and Shudders	
22	hairy	Head and Shudders	
23	hairy	Head and Shudders	
24	hairy	Head and Shudders	
25	hairy	Head and Shudders	
26	hairy	Head and Shudders	
27	hairy	Head and Shudders	
28	hairy	Head and Shudders	
29	hairy	Head and Shudders	
30	hairy	Head and Shudders	
31			

To perform the Chi-Square analysis, go to "Analyze" and pick "Descriptive Statistics" and then "Crosstabs" - do NOT use the "Chi-Square" command, strange as that may seem!



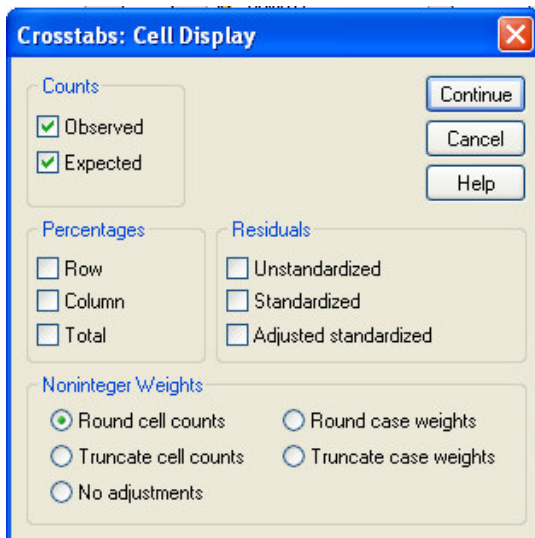
Move one of the variable names from the left-hand box to the box entitled "Row(s)" and the other variable name to the box named "Column(s)". Then click on "Statistics".

The following dialog box will appear. Click on the little box next to "Chi-Square" to select it. (There are another dozen tests here, but you can ignore them at present!) Then click on "continue" to get back to the previous dialog box.





Click on "Cells..." and make sure that there are ticks in the boxes next to "Observed" and "Expected", so that SPSS will show you both the observed and expected frequencies for each permutation of variables. Click on "Continue" to get back to the previous dialog box.



Finally click on "OK" to get the results of the analysis. The bits that you want are the table of observed and expected frequencies, and the results of the Chi-Square test.

The table shows us that 17 people used "Noflakes" and 13 used "Head and Shudders". It also shows us the observed frequencies (how many users of each shampoo actually were bald and how many were actually hairy) and the expected frequencies (how many bald and hairy users of each shampoo we would expect to get if baldness and shampoo choice had nothing to do with each other). Hopefully you can see that the observed and expected frequencies are rather different from each other.

**hairloss \* shampoo Crosstabulation**

			shampoo		Total
			Noflakes	Head and Shudders	
hairloss	bald	Count	12	3	15
		Expected Count	8.5	6.5	15.0
	hairy	Count	5	10	15
		Expected Count	8.5	6.5	15.0
Total	Count	17	13	30	
	Expected Count	17.0	13.0	30.0	



The second table shows the results of the Chi-Square test. The top line, entitled "Pearson Chi-Square", shows the results of the Chi-Square test: Chi-Square is 6.65, with 1 degree of freedom, and this is significant at  $p = .01$  (i.e. there is a significant association between shampoo choice and hair loss).

If you have a 2x2 table (so that you have only one degree of freedom), SPSS also calculates Chi-Square using Yates' correction for continuity. The results of this version of Chi-Square are given in the second row of the table. Notice that the value of Chi-Square is smaller than in the top row: Yates' correction makes the test more "conservative", meaning that it is harder to get a significant result (although in this case Chi-Square remains statistically significant with  $p$  of .027).

In the case of a 2x2 table, SPSS also works out some other tests and gives you the  $p$ -values associated with them, but you can ignore them at this stage.

**Chi-Square Tests**

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6.652 <sup>a</sup>	1	.010		
Continuity Correction <sup>b</sup>	4.887	1	.027		
Likelihood Ratio	6.946	1	.008		
Fisher's Exact Test				.025	.013
Linear-by-Linear Association	6.430	1	.011		
N of Valid Cases	30				

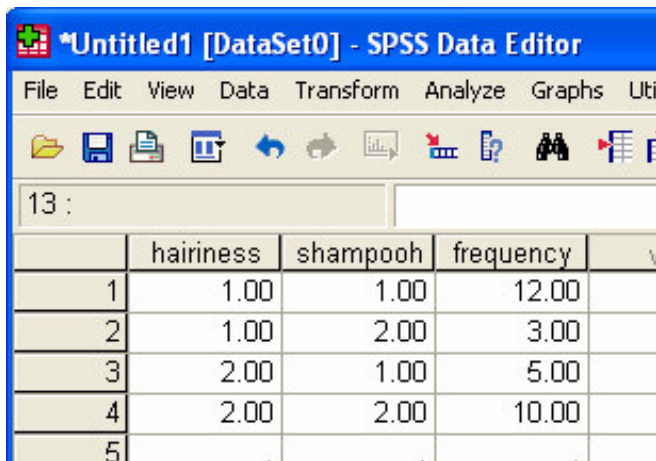
a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.50.

**(b) Using Chi-Square on summary data:**

You can use Chi-Square on summary data just as we did with the Goodness of Fit test.

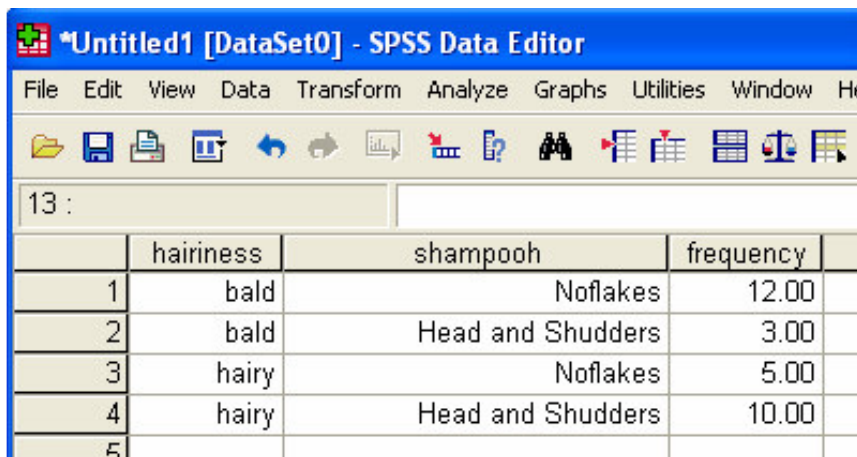
In the column entitled "Frequency", I've entered the totals for each permutation of my two independent variables. Thus there were 12 users of "Noflakes" who were bald, and 3 who had hair. There were 5 users of "Head and Shudders" who were bald, and 10 who had hair.



The screenshot shows the SPSS Data Editor window titled '\*Untitled1 [DataSet0] - SPSS Data Editor'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, and Utilities. The toolbar contains icons for file operations, navigation, and analysis. The data grid shows the following data:

	hairiness	shampoo	frequency
1	1.00	1.00	12.00
2	1.00	2.00	3.00
3	2.00	1.00	5.00
4	2.00	2.00	10.00
5	.	.	.

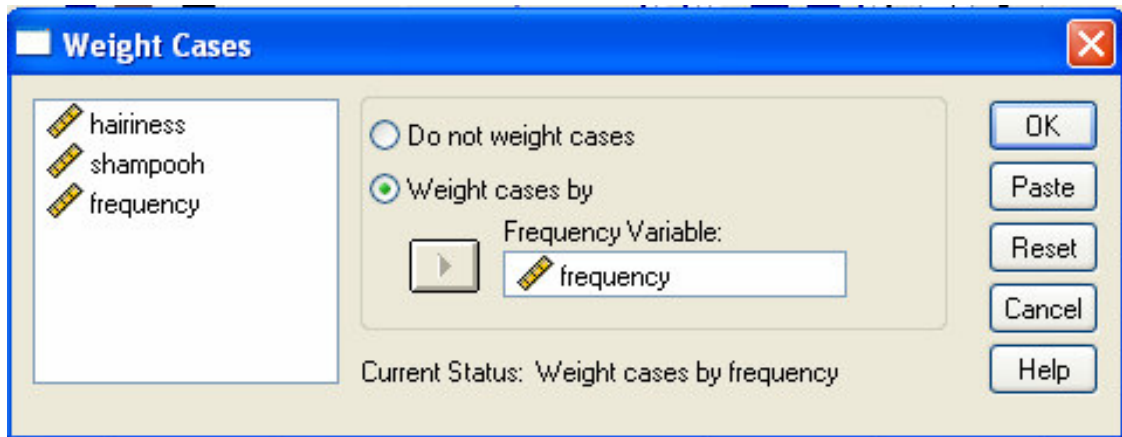
In the other two columns, I've used "1" and "2" to tell SPSS which permutation of conditions each frequency refers to. They are the same codes as I used before: "1" for "Noflakes" or "2" for "Head and Shudders", and "1" for "bald" or "2" for "full head of hair". Thus a "1" for "hairiness" and "1" for "shampoo" means "bald Noflakes users". "1" for "hairiness" and "2" for "shampoo" means "bald Head and Shudders users"; and so on. Replacing the variable labels with words hopefully makes this clearer:



The screenshot shows the same SPSS Data Editor window, but the data grid has been updated with descriptive labels for the 'hairiness' and 'shampoo' columns:

	hairiness	shampoo	frequency
1	bald	Noflakes	12.00
2	bald	Head and Shudders	3.00
3	hairy	Noflakes	5.00
4	hairy	Head and Shudders	10.00
5	.	.	.

The next step is to weight the cases. Click on "Data", and then on "Weight cases". In the dialog box that appears, move "frequency" into the "frequency variable" box. Then click on "OK".



Now perform the Chi-Square analysis as before - you should get the same results as when you used all the raw data.

**Reassurance for the cognitively challenged:**

If you find the "weight cases" business confusing, don't worry - so do many other people (including me)! It's not exactly intuitive, but you can always resort to using the more straightforward method of data entry, as long as you don't mind typing in the data.