A computer practical

Objectives

This computer practical is included to illustrate the potential for interactive sessions in which students work on the analysis of a prepared data set. In so doing, the lessons of the previous week's classroom session can be reinforced in a practical context, while the student simultaneously acquires technical proficiency with the statistical software. This practical focuses on the analysis and presentation of descriptive data from a cross-sectional survey. Students are also taught to carry out simple bivariate analyses using t-tests and chi-squared tests, and are introduced to the concept of the confidence interval. At the end of the session the students are taught how to test for the presence of confounding and how to adjust for confounding using a classical stratified analysis.

Notes for students

Work through this practical following all the instructions in italics. Answer the questions as you go.

Getting started

First call up the relevant file. It is called gospel.sav. This is an spss data file.
Click on file, and then on data, highlight gospel.sav with the cursor and then click on OK.

How was this data obtained?

This data was gathered in a cross-sectional catchment area survey. Gospel Oak is a geographically defined area of North London, UK, with a population of just over 6,000. The investigators attempted to identify and interview all those residents who were aged 65 and over. They developed their own register by carrying out a door-knock census. The 654 persons who consented to be interviewed (74% of all those who were eligible) were interviewed by a trained lay interviewer. Data was collected on physical and mental health status and social and demographic circumstances.

What data is contained in this file?

The data can be viewed as a spread sheet. Scrolling down the spread sheet you will see that there are 654 lines of data, one for each subject. Each subject has a unique identifier contained in the first variable CODENO. Scrolling across the spread sheet you will see that there are a total of 18 different variables. Double-clicking on the name of a variable gives you information about its contents. Thus if you double click on EDCAT you will see that this contains data on years of education. Click on labels and you will see that EDCAT is an ordered categorical variable with three levels

1. < 9 years of education
2. 9-11 years of education
3. >11 years of education

Click on cancel to clear these windows

SPSS will give you similar information on all variables in the dataset.
Click on utilities and then on file info.

Describing the sample

What are the characteristics of the sample by age, gender, marital status, educational status and handicap?
Starting with gender, look at the frequency distribution of the variable SEX.
Click on statistics - summarise - frequencies and select SEX. Then click on OK.
You should get the following output

SEX

sex

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
M	1	258	39.4	39.4	39.4
F	2	396	60.6	60.6	100.0
		-------		-------	-------
	Total	654		100.0	100.0

Valid cases 654 Missing cases 0

39% of the sample is male and 61% is female.

Q1.) Why do you think there is such a preponderance of women? Is this likely to cause any problems?

This data can be presented graphically, in a pie chart.
Click on graphs - pie - define. Select the variable SEX and paste into the box ‘define slices by’ by clicking the indicator arrow. Next click on OK.

Next age. Look at the frequency distribution of the variable AGE. We will not reproduce the output here, as it is rather long. Clearly this way of summarising the data on age is not particularly helpful. There are too many categories. Of course, age is a continuous rather than a categorical variable. Its values have real numeric significance. In this case values have been rounded up to integers (whole years). The data contained in such a variable is better summarised in the following way:
Click on statistics - summarise - descriptives. Select AGE and click OK. You should get the following output.

Variable	Mean	Std Dev	Minimum	Maximum	N Label
AGE	75.41	7.37	65	99	654

The 654 subjects are aged between 65 and 99 years. The mean age is 75.4 years, and the standard deviation (a measure of the dispersion of the distribution) is 7.4.

It is a good idea to look at the distribution of continuous variables using a histogram.
Click on graph - histogram. Select ‘display normal curve’. Select AGE and click on OK. The histogram consists of bars representing the numbers of subjects within consecutive 2.5 year age bands. Clearly this variable is not normally distributed. There are more younger than older subjects, and the distribution is truncated at the age of 65 years. Those under 65 were

not surveyed. With a skewed variable such as this, the mean and standard deviation do not provide a good summary of the central tendency and dispersion. The median (50th centile or halfway value) and interquartile range (25th and 75th centiles) are better. These can be displayed in a box plot.
Click on graph - boxplot. Select 'summaries of separate variables' and click on define. Select AGE and click on OK.
The upper and lower bars indicate the full range of values (65-99), the red box the interquartile range (25th to 75th centiles, 69-81) and the bold bar the median (50th centile, 74). The skewed nature of the distribution is immediately clear. All these values can be obtained in one procedure.
Click on statistics - summarise - frequencies , select the variable AGE then deselect 'display frequency values' by clicking on it. Then click on statistics and select quartiles, minimum, maximum, mean and median in the next window, then click on OK .

Now describe in appropriate summary form the distributions of the following variables - marital status (MARRY), educational status (EDCAT), and London Handicap Scale Score (HANDICAP). Write your answers in the space below.

Q2.) Describe the following variables
Marital status Educational status London Handicap Scale Score

Q.3) What is the prevalence of depression?
Consulting your list of variables, you will see that the variable DPDSCASE signifies whether or not subjects were identified as cases of 'pervasive depression' according to the algorithm applied by the SHORT-CARE instrument used in the survey.

Prevalence of depression =

How does the prevalence of depression vary with gender?
So far we have described the distribution of the data only in terms of one variable at a time. This question requires that we look at two variables DPDSCASE and SEX simultaneously. The first step is to cross-tabulate the two variables.
Click on statistics - summarise - crosstabs. Paste SEX into 'rows' and DPDSCASE into 'columns'. Click on OK.
What output will you see?

35/258 men and 81/396 women were identified as having pervasive depression.

We can get more information on the association between gender and depression in the following way.
Click again on statistics - summarise - crosstabs. Click on Cells and select ‘row percentages’. Click on Statistics and select ‘Chi-squared’.

What output will you see?

The prevalence of depression is 13.6% among men and 20.5% among women.

Q4.) Is this difference statistically significant?

For any association it is also good practice to report an effect size with 95% confidence intervals. In this case, one way to summarise the size of the effect would be to report the difference between the proportion of depressed between the two genders. This is 20.5-13.6 = 6.9%. SPSS does not calculate confidence intervals for the difference between two proportions, but we can calculate these values using another program, CIA (Confidence Interval Analysis). Open this program in another window by double clicking on cia.exe. You will need chapter 4, section 2 (two samples, unpaired case). You should get the following answer

Difference in proportions depressed (between genders) = 6.9%, 95% confidence intervals 1.1%-12.7%.

Q5.) What additional information is conveyed here?

Another useful way of summarising the size of this effect is by calculating an odds ratio. You will see that odds ratios have been widely used in the two publications reporting the results of this survey (Prince et al 1997a, Prince et al 1997b). The odds ratio in this case is the odds of being female if depressed, divided by the odds of being female if not depressed. You should be able to calculate the odds ratio by hand using the above cross-tabulation. Do this now.

However, SPSS will carry out this task for you, and will give 95% confidence intervals.
Repeat the crosstabs commands, but this time select ‘risk’ in the statistics subcommands instead of chi-squared.
What output will you see?

The odds ratio (OR) is given in the first line of the output. It is 1.64 with 95% confidence intervals of 1.06 to 2.52. The odds of being depressed if female is 1.64 times that of being depressed if male. The 95% confidence intervals do not include the null hypothesis value of 1.0. Therefore this effect reaches conventional levels of statistical significance. Note that you could also have obtained the OR with 95% confidence intervals by entering in the data from the cross tabulation using Chapter 6 of the CIA program.

How does the prevalence of depression vary with age?
AGE is a continuous variable, whereas SEX was a categorical variable. We could assess the relationship between depression and age by comparing the mean age of those who are depressed and those who are not depressed. However, this would tell us little about the form of the association between the two variables. In the early stages of an exploratory analysis it is good practice to divide continuous variables into categories which can then be cross-tabulated. This has already been done for age. There is a variable NAGE which is labelled as follows

NAGE Quartiles of AGE

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
65-69	1	178	27.2	27.2	27.2
70-74	2	165	25.2	25.2	52.4
75-81	3	159	24.3	24.3	76.8
82+	4	152	23.2	23.2	100.0
		-------		-------	-------
	Total	654		100.0	100.0

Valid cases 654 Missing cases 0

The sample has been ranked by age in such a way that as near as possible to one quarter of the sample falls into each of the four age groups.

Now, cross-tabulate NAGE (rows) with DPDSCASE (columns) and request row percentages and chi-squared tests as before.
What output will you see?

Q6.) Describe the form of the association between age and depression. Is this association statistically significant?
Is there an association between frequency of contact with friends and pervasive depression?
Examine the effect of frequency of contact with friends (variable = FRDFRQ) on depression. Again describe the form of the association. Is it linear across the three categories?
When, as in this case, the association is not linear, it can be helpful in an exploratory descriptive analysis of this type to dichotomise the exposure variable into high risk and low risk categories. This has already been done for you. Using the variable FRDCAT calculate the difference in proportions depressed (with 95% confidence intervals) between those never seeing friends and those with at least some contact. Write your answer below.

Q7.) What is the difference in proportions depressed between those never seeing friends and those with some contact?

Now calculate the odds ratio (with 95% confidence intervals) for depression associated with never seeing friends, and write your answer below.

Q8.) What is the odds ratio (with 95% confidence intervals) for depression associated with never seeing friends

Is there an association between handicap and depression?
The first of the two papers reporting the results of this survey (Prince et al. 1997a) considered the effects of impairment, disability and handicap on depression. In your dataset subjects have been divided into four quarters according to their scores on the London Handicap Scale (NHANDICA). For convenience, these four categories have been collapsed into two in another variable HANDICAT.

Q9.) What is the relationship between handicap and pervasive depression? Summarise the association below.
What are the joint effects of handicap and contact with friends on depression?
We have seen that never seeing one's friends is associated with depression. There is also a very strong association between handicap and depression. Poor health may result in loss of the ability to contact friends. It is possible therefore that the association between never seeing friends and depression may be confounded by handicap. The pathway would be….

Associations a and b have already been established. In order to be a candidate as a confounder, handicap should also be associated with lack of contact with friends (association c).

Q.10) Is there an association between handicap and lack of contact with friends?

The joint effect of these two variables can be assessed in a stratified analysis. Here was the output when we looked at the effect of never seeing friends on depression.
FRDCAT by DPDSCASE dpds case

Q.11) Is there evidence of confounding?

It is possible to use stratified analyses of this type to control for the effects of confounding. You will learn later in the course how to control for confounding using more complicated multivariate techniques such as logistic regression. In a stratified analysis it is possible to obtain an adjusted summary (Mantel-Haenszel) odds ratio, which is effectively a weighted average of the odds ratios across the different strata. Again it is not possible to obtain a Mantel-Haenszel summary odds ratio using SPSS, but you can calculate it using CIA. Try to obtain an estimate of the odds ratio for the association between never seeing friends and depression, adjusted for handicap.
Go into Chapter 6 of section 3 (Odds ratios in a series of unmatched case-control studies), type '2' for number of samples, and enter in data from the cross-tabs for each stratum.

Q.12) What is the Mantel Haenszel summary odds ratio?
How does it differ from the raw (unadjusted) OR?

Case Control Studies.

In this practical you will analyse a *nested* case control study assessing risk factors for developing dementia.
Nested means that cases and controls are selected from among a larger study cohort, often put together for another purpose.
In this case the cohort was a large population of hypertensive older people recruited for a randomised controlled trial of antihypertensive treatment.

The appendix includes information on the variables. The outcome of interest is CASE (diagnosis of probable or possible AD). Cases are coded as 1, controls as 0.

The exposures can be divided into several groups
a) risk factors for vascular disease (smoking, cholesterol levels, body mass index, systolic blood pressure)
b) evidence of vascular disease (ECG ischaemia)
c) urban residence
d) education, and pre-morbid intelligence (measured using the NART)
e) a family history of dementia
f) age and gender

The hypotheses to be tested are:
1. Family history of dementia is an independent risk factor for AD.
2. Cardiovascular risk factors are associated with AD.
3. Educational level and higher pre-morbid intelligence are inversely related.

Plan of action
1. Familiarise yourself with the coding and distribution of each variable. Are you happy about the way in which missing data is coded for each variable?

2. Characteristics of cases and controls.
Now we will compare the characteristics of cases and controls. Is increasing age associated with AD? What about female gender?

To look at age use the **compare means** command on **statistics**.
To look at gender and AD use **crosstabs**

How would you test for significance for these two variables?

Now continue this procedure comparing cases and controls for other exposures. Produce a summary table of results.

3. Now use crosstabs to get SPSS to give you an odds ratio for the association between.
To do this go to **crosstabs**, into **statistics** and click on **risk**. The output will look something like this:

Statistic	Value	95% Confidence Bounds	
Relative Risk Estimate (CASE 0 / CASE 1) :			
case control	.85973	.40304	1.83391
cohort (GENDER 1 Risk)	.91168	.57927	1.43485
cohort (GENDER 2 Risk)	1.06043	.78221	1.43760

Number of Missing Observations: 0

Which of these values is the odds ratio?
How many decimal places would you use when presenting the findings of this study?

4. Now do the same for the other risk factors you are interested in: which of them show clear associations with AD?

5. SPSS will not calculate an odds ratio for smoking: why? Can you do it by hand, or can you think of a way to overcome this?

Hint: one way around this is to recode it as a binary (0/1) variable. It is always best to create new variables if you manipulate the data so that you retain your original data. Select the fags column, go to **edit**, select **copy**, then move the cursor to an empty column, go to **edit**, select **paste** and you will have copied the fags category into the new column. Now go to **data** select **define** and rename the variable "smokebin" (for smoking binary). Now go to **transform**, select **recode**, select **into same variable** and then old and new values. Try to make it so that the old values 2/3 are recoded as 1.

6. Summarise the results of these analyses in tabular form.

Now stop and think!

You will have found one or two clear associations which are statistically significant. You will also have found some possible, but non significant associations and you will have quite a few which look uninteresting. Family history is the strongest association in terms of effect size. Are there any possible confounders or biases which could explain this? How about some of the associations which don't reach statistical significance? Could *negative* confounding explain these? The aim of further analysis will be to try and control for confounding to explain better the associations between risk factors. To do this you will need to use logistic regression. You were introduced to this form of analysis in the last practical. Here is some revision.

8. Go to **statistics** and click **regression** then **logistic.** Now put in case as the **dependent variable** (which simply means outcome). Put in famhist into the independent variable box and press ok.

You should get an output which looks like this:

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)

FAMHIST	1.1878	.4998	5.6485	1	.0175	.1556	3.2800
Constant	-1.8165	.2411	56.7606	1	.0000		

What do the figures mean?

In logistic regression you are effectively modelling the odds of being a case according to exposure status. When doing it with only two variables (case and famhist) you are doing something very similar mathematically to the cross tab calculation in a 2x2 table.

For the moment ignore the row constant. Looking at famhist: "B" is *the log of the odds ratio*[1], SE is the standard error of the log odds ratio, Wald is a statistical test similar to the chi square value (check this: the Pearson chi square you did on the crosstab will be very similar to this value) df is degrees of freedom (one because this is effectively a 2x2 comparison) sig is the significance value (again compare this to what you got from crosstabs and you will have a very similar value). Ignore R. Exp(B) is the exponent (antilog) of B: in other words the odds ratio! Check this is the same as the odds ratio which you arrived at in crosstabs.

SPSS is not very epidemiological and does not give confidence intervals for odds ratios in logistic regression. Can you remember how to work out the confidence intervals using a calculator?

Now look at the constant value. This represents the value log (odds of being a case if you are not exposed). This agrees with the data from the 2x2 table. There were 20 cases who were unexposed and 123 controls who were unexposed.

Odds of being a case if unexposed= 20/123 = 0.163
Log(0.163) = - 1.82

The output from this logistic regression is represented by the regression equation:

log odds of being a case = 1.178x(exposure) - 1.816

In other word, when exposure = zero, the log odds of caseness = -1.816 because
1.178 x 0 = 0.

When the exposure is one the log odds of caseness increases to -0.638. Work it out for yourself using the logistic regression equation.

1.178x(1) - 1.816 = -0.638

Antilogging -0.638 we get 0.53. Check this against the 2x2 table. The odds of being a case if you were exposed was 8/15 = 0.53.

When logistic regression is used to assess the association between just one exposure and a dichotomous outcome, the process is very similar to that which we used when deriving an odds ratio from a crosstab (often referred to as a classical method of analysis). The advantage of logistic regression is that it can be used to handle more than one variable simultaneously.

In previous weeks we have used stratified analysis to adjust for the effect of just one potential confounding variable. Logistic regression, a form of multivariate analysis, can be used to address the effects of several exposures simultaneously, each adjusted for the effects of all of the others.

9. The next step will be to control the odds ratio estimate of family history for two crucial additional variables, age and gender. These are good candidates for confounders, which you will generallly want to account for in analyses. To control the effect of family history for these go back to the logistic window and add these variable to the independent variables box. You should get an output which looks like this:

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FAMHIST	1.2765	.5228	5.9610	1	.0146	.1621	3.5842
AGE	.1948	.0777	6.2933	1	.0121	.1688	1.2151
GENDER	-.3284	.4429	.5499	1	.4584	.0000	.7201
Constant	-16.5206	6.0935	7.3504	1	.0067		

This now is describing an equation:

log odds of caseness = 1.2765(famhist) + 0.1948(age) - 0.3284(gender) - 16.5206

Notice the following:
The variable age has an odds ratio of 1.2151. What does this mean (think carefully)?
The odds ratio for famhist has increased. What does this imply?

10. What if we want to look at the effect of a variable which is coded on more than two levels. Take smoking for example. If you enter "fags" into the model as the only variable as it is you will get this:

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FAGS	-.2729	.3262	.7002	1	.4027	.0000	.7611
Constant	-1.8691	.2035	84.3705	1	.0000		

SPSS has estimated the change in log odds of being a case per unit change in the variable FAGS. However such an analysis will only be meaningful where a linear graded association can be anticipated (e.g change in log odds per one year increase in age). In the case of *ordered* categorical data, e.g

- age coded as five year age bands, or
- smoking coded as "none, ex, light, moderate, heavy" or
- alcohol consumption on five levels

One should create a *factored* variable, sometimes called (as in SPSS) an *indicator* variable. Do this by selecting "categorical" in the logistic regression window, selecting the variable "fags" moving it across to the middle box and selecting indicator from the contrast box. Click also on 'first' and then click on 'change'. Finally click on continue. If you now put it in the model you get this output:

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
FAGS			.8391	3	.8401	.0000	
FAGS(1)	.4140	.5913	.4901	1	.4839	.0000	1.5128
FAGS(2)	-.4798	1.0654	.2029	1	.6524	.0000	.6189
FAGS(3)	-6.2844	20.1448	.0973	1	.7551	.0000	.0019
Constant	-1.9181	.2100	83.4011	1	.0000		

What does this mean? Remember that FAGS was coded as below.

Valid Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
non-smoker	0	203	81.9	82.5	82.5
1-10 day	1	22	8.9	8.9	91.5
10+	2	12	4.8	4.9	96.3
pipe/cigar	3	9	3.6	3.7	100.0
	.	2	.8	Missing	
	Total	248	100.0	100.0	

The baseline category is the first i.e 'non-smoker'. FAGS(1) is the odds ratio for AD comparing those smoking 1-10 day with non-smokers. FAGS (2) compares those smoking 10+ with non-smokers. FAGS (3) compares pipe and cigar smokers with non-smokers.
The first line of the output gives a test of significance for the overall effect of smoking on odds of AD. In this case p=0.84, suggesting no effect. This is analogous to a chi-squared test for heterogeneity. The previous analysis in which FAGS was entered as a non-categorical variable was analagous to a chi-squared test for trend. Can you understand why?

Now spend the remainder of the practical experimenting with logistic regression, and seeing how robust any findings you have are when you correct for other variables.

Questionnaire design.
<u>Reliability and validity</u>

Teach1.sav contains data from a large population survey of 2,949 over 65 year old subjects. In this study a new brief depression scale has been derived from previously existing measures. The scale has 12 items which address the following areas
1. Depression
2. Pessimism
3. Suicide
4. Guilt
5. Sleep
6. Interest
7. Irritability
8. Appetite
9. Fatigue
10. Concentration
11 Enjoyment
12. Tearfulness

Responses to the individual items have been given weighted scores lying between 0 and 1, according to the level of depression typically associated with the response.

If this seems unclear look at the frequency distribution for one of the item variables EURO_D5 (sleep). Subjects have scored either 0.28, 0.58, 0.88 or 0.96.

The variable EURO_D contains the EURO_D total scale scores for each subject (the sum of the item scores EURO_D1 through EURO_D12).

Q.1) Describe the distribution of these scores, writing down the results below

Mean =

Standard deviation =

Minimum =

Maximum =

(Hint- Statistics/Summarize/ Descriptives)

Q.2) What is the shape of the distribution - normal? Skewed?

There are several ways to address this problem
1) Do a histogram plot of EURO_D
(Hint - Graphs/Histogram + tick display normal curve)
2) Identify the median value
(Hint - Statistics/ Summarize/ Frequencies and identify the 50% value)

Median =

Clearly the distribution is positively skewed - the median is lower than the mean.
The subjects in this study also completed a standard well validated self-report depression instrument, the CES-D (variable = CES_D). Check the distribution of this variable also using the procedures described above.

We will now look at the validity and reliability of the EURO_D scale.

<u>Criterion validity</u>

Q.3) What is the association between EURO_D and CES_D?
(Remember, you have already shown that each measure is non-normally distributed)

(Hint - Non-parametric statistics
Statistics/correlate/bivariate + tick both pearson's and spearman)

You can look at the same association visually by doing a scatter plot of EURO_D and CES_D
(Hint - Graphs/Scatter/Simple)

Q.4) What is the association between EURO_D and major depression?
Look at the frequency distribution of the variable majordep. What is the frequency of major depression?
There are two main ways of looking at the association between the continuous EURO_D measure and the dichotomous depression diagnosis
Firstly, does the mean EURO_D score differ significantly between depressed and non

depressed subjects?
(Hint - t-test. Statistics/ compare means/independent sample t-tests + EURO_D is test variable, majordep is grouping variable with values 0,1.)

Mean for non-depressed =

Mean for depressed =

Difference between means (with 95% confidence intervals) = (-)

T-value =

p-value =
Is there a significant difference ?

What was wrong with the analysis we have just carried out? What should we have done instead?
(Hint-statistics/non-parametric/2 independent samples + tick Mann-Whitney)

Has this made a difference?

Q.5) How does the probability of being a case of major depression vary with EURO_D score?

This is a tricky one as the EURO_D is not an integer scale (0,1,2,3 etc - all values are possible). One approach is to divide subjects into n equal groups according to their EURO_D score. The population can be divided into thirds, quarters or fifths, but in this case we are going to divide it into tenths. To do this click transform/rank cases + paste euro_d into variable box + click on rank types and tick Ntiles + type 10 . You have created a new variable called neuro_d. Look at its frequency distribution. 10% of subjects fall into each of 10 categories.

Now we can do a bar plot of the % of depression cases at each neuro_d level.
(Hint- Graphs/bar/simple/define, then paste neuro_d into category axis and tick 'other summary function' and paste majordep into 'bars represent' box. Then click on change summary and tick percentage above and type 0). This should give you a bar chart with the percentage of subjects scoring above 0 (i.e 1) on majordep at each of the 10 levels of euro_d.

What does this show? Could we use euro_d to screen for major depression. If so what should the cut point be?

<u>Reliability of EURO_D</u>
We also wanted to look at the test-retest reliability of euro_d and did a repeat assessment two weeks later. The results of this test are in the variable euro_drt.
Correlate euro_d and euro_drt using Pearson's and Spearman correlations. What is the correlation coefficient?

Q.6) Does this mean that we have perfect agreement?

We can check this out by doing a scatter plot of euro_d against euro_drt.
What emerges is that there is perfect association but not perfect agreement.

There is a systematic error in that the repeat measure is related to the first measure according to the equation euro_drt = 2x euro_d + 1.
Clearly this is fabricated data, to demonstrate that correlation is a measure of association, and not agreement.

The measure of agreement for continuous measures is the intra-class correlation coefficient. A useful trick for measuring this coefficient using spss requires some manipulation of the data set. A copy of the data set is attached to the bottom of the current data set, but the euro_d data is listed under euro_drt and vice versa. This manipulation has been performed in teach2.sav. Call up this data set. Note that there are twice as many cases as before. Now repeat the pearson and spearman correlations we carried out before and note the different results (you can ignore the minus signs).

Q.9) Is the scale internally consistent?
Next we can look at the internal consistency of the euro_d, assessing split half reliability and cronbach's alpha. Split -half reliability assesses the extent to which two halves of the scale are measuring the same thing, and Cronbach's alpha the extent to which individual items contribute similarly to a single dimension. A good scale should score at least 0.7 on each of these measures.

Split-half reliability
Statistics/scale/reliability analysis + paste euro_d1 through eurod_12 into the items box (hint hold down the left mouse button and drag the cursor through the relevant variables to highlight them). Then change model to split half and click on OK.

Split-half reliability =

Cronbach's alpha
(Statistics/scale/reliability analysis + change model to alpha + click on statistics and tick scale if item-deleted)
Cronbach's alpha =

Are all the items well correlated with the total scale score ?
Can the alpha be improved by deleting any of the scale items?

Factor analysis
Finally we will do a factor analysis to look at the constructs underlying the scale.
Statistics/data reduction/factor + paste euro_d1 through eurod_12 into the variables box (hint hold down the left mouse button and drag the cursor through the relevant variables to highlight them). Next click rotation and tick varimax (this makes it easier to identify items contributing [loading] on different factors).

Q.10) How many factors have been identified?

How much of the variance in the scale is accounted for by Factor 1, Factor 2, and Factor 3

Items with a factor loading above 0.6 are said to 'load on' (contribute to) a particular factor.

Which items load on

Factor 1	Factor 2	Factor 3
Item	Item	Item
Item	Item	Item
Item		

Item
Item

Do these seem to be unitary, intelligible, face valid underlying constructs for a depression scale?