# Research Methods 2

## *Lab session – Answer Key (Answers in green boxes)*

> Similar to how we conducted the lab session in Research Methods 1, please follow the steps below. Contrary to the previous lab session, though, **you are required to upload your work to Safe Assignment** this time in order to pass attendance for the lab session. This is because this time around, we do not have a separate graded test to make sure that you really understand the topics discussed here. Note that your work will not be graded, and any mistakes you may make are not a problem.

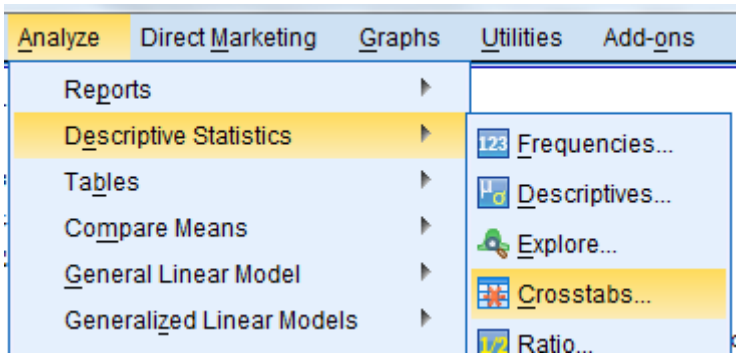
**As always, please feel free to ask for help on any of the steps in the exercises below.**

## 1. Measures of association & Crosstabs

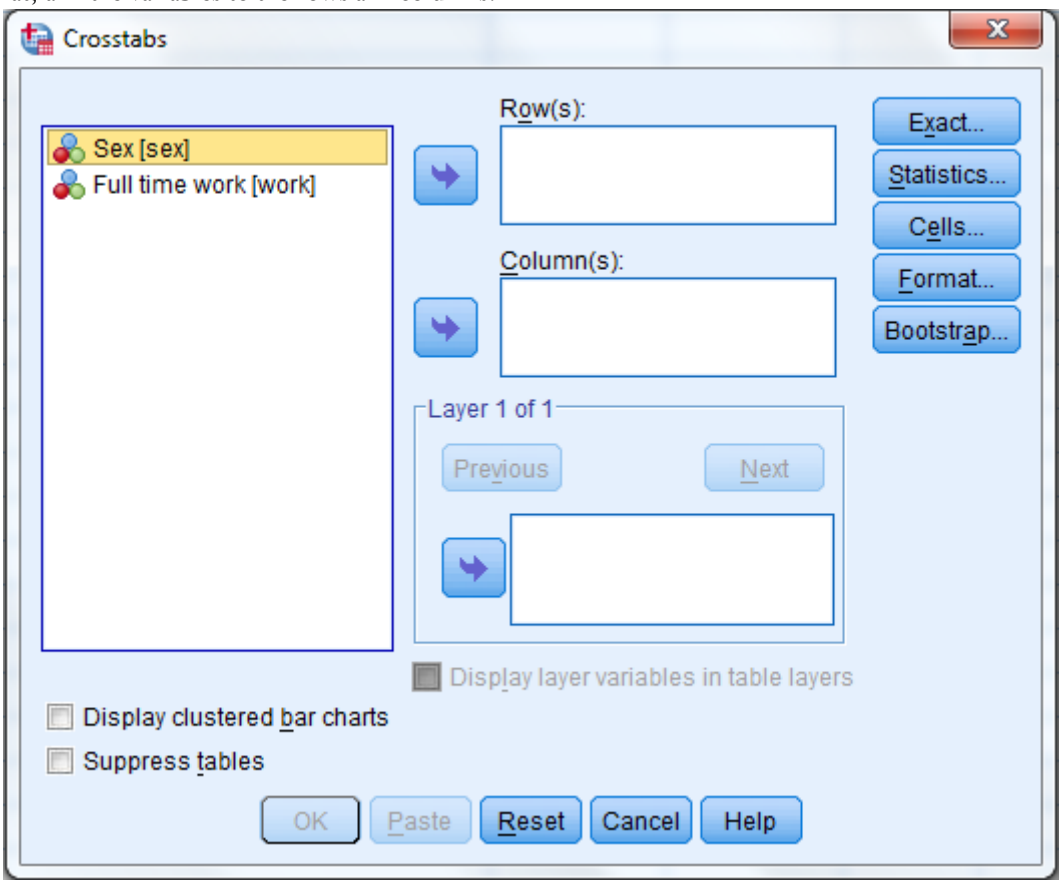
In the western world women have from time to time successfully fought social and economic inequalities. Especially in the Netherlands however, the percentage of women who work part time is still amongst the highest in Europe. In this exercise, we will first be making a simple table with percentages. After that we will test the association between full time work and gender.

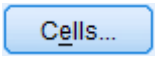
- Download the data set WORK.SAV from EleUM.

- Create a cross table with the variables WORK and SEX. The cells should contain both **absolute counts** and the **correct percentages**. The research question is about the inequalities between men and women (i.e., we like to compare the relative share of part time workers among men and among women). Follow the next couple of steps to create such a table:



After that, add the variables to the rows and columns:




Before proceeding, click this button: 


In the box that pops up, make sure that "**Observed**" is selected, as well as the correct **percentage** (either 'row or 'column', depending on which variable you decided to put where). **We want to have a table comparing the percentage of full-time employed men to the percentage of full-time employed women** (as a percentage of the respective gender). Experiment a bit until you get it right in the output window.

1. What can be concluded from this table? Does there appear to be a difference between men and women with regard to labour participation? Which percentages did you find for men and women in full time employment?

**Sex * Full time work Crosstabulation**

| | | | Full time work | | |
|---|---|---|---|---|---|
| | | | part time job | full time job | Total |
| Sex | male | Count | 45 | 842 | 887 |
| | | Expected Count | 186,8 | 700,2 | 887,0 |
| | | % within Sex | 5,1% | **94,9%** | 100,0% |
| | female | Count | 337 | 590 | 927 |
| | | Expected Count | 195,2 | 731,8 | 927,0 |
| | | % within Sex | 36,4% | **63,6%** | 100,0% |
| Total | | Count | 382 | 1432 | 1814 |

| | | | | |
|---|---|---|---|---|
| | Expected Count | 382,0 | 1432,0 | 1814,0 |
| | % within Sex | 21,1% | 78,9% | 100,0% |

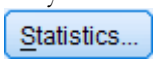
94,9% of men work full time, as compared to 63,6% of women. This suggests a gender divide indeed.

Note that in the table above I also included the 'expected' count in the cells. This is the 'expected' value that SPSS uses when calculating things like the Chi-Square and Cramer's V.

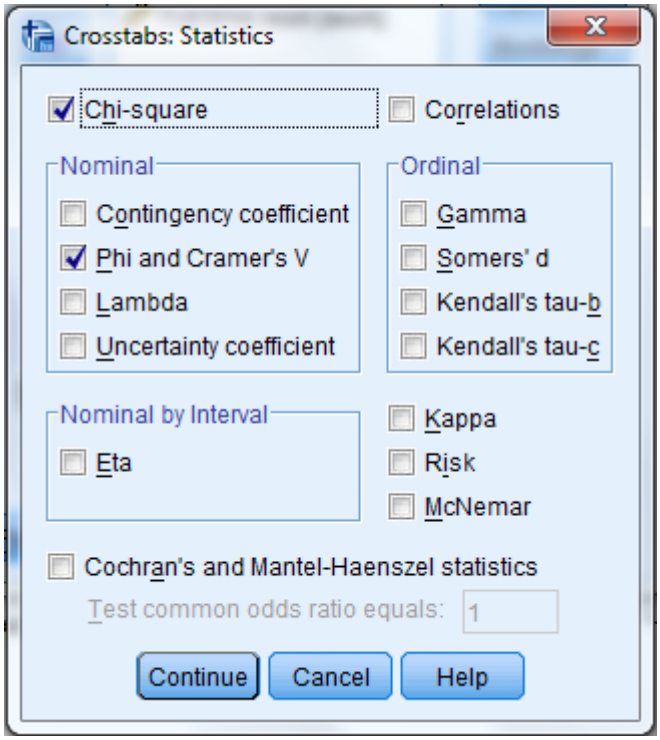
One may find in a table that indeed the relative share of full time workers among women is lower which indicates labour inequality in the Netherlands. However, we don't know for sure whether this relationship is also statistically significant. To get a more definite answer, we need a statistical test for the difference in percentages. In the lecture we argued that the Chi-square test and Cramer's V might be well suited.

**For future reference**: the height of the chi-square indicates how large the differences in percentages are between a (theoretical) cross table with no association (equal column percentages) and the observed (empirical) table. The larger the Chi-square, the larger the differences between the theoretical and empirical table.

- Compute the Chi-square and Cramer's V values in the cross table you created by going back to the Crosstabs analysis menu and clicking this button:

  In the window that pops up, make sure you select the Chi-square and Cramer's V:




Now click 'Continue', then 'OK', and have a look at your output window.

2. Have a look at the third table ("Chi-Square tests"). How large is chi-square, and is it significant ("Asymp. Sig.") at $\alpha = 0.05$ (5%)?

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | **266,798**[a] | 1 | **,000** | | |
| Continuity Correction[b] | 264,920 | 1 | ,000 | | |
| Likelihood Ratio | 296,295 | 1 | ,000 | | |
| Fisher's Exact Test | | | | ,000 | ,000 |
| Linear-by-Linear Association | 266,651 | 1 | ,000 | | |
| N of Valid Cases | 1814 | | | | |

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 186,79.

b. Computed only for a 2x2 table

The Chi-Square is 266,798, and it is indeed statistically significant. Note that this doesn't say anything about the strength of the relationship yet.

3. In the fourth and final table ("Symmetric measures") you will find the outcome for the Cramer's V test. What is the value of V? Is it statistically significant? How strong would you say that the relationship between gender and full time employment is (hint: think about the rule of thumb discussed in the lecture)?

**Symmetric Measures**

| | | Value | Approx. Sig. |
|---|---|---|---|
| Nominal by Nominal | Phi | -,384 | ,000 |
| | Cramer's V | **,384** | **,000** |
| N of Valid Cases | | 1814 | |

Cramer's V between these two variables is 0,384 (and statistically significant). According to our rule of thumb (discussed in the lecture), we can take this guideline:

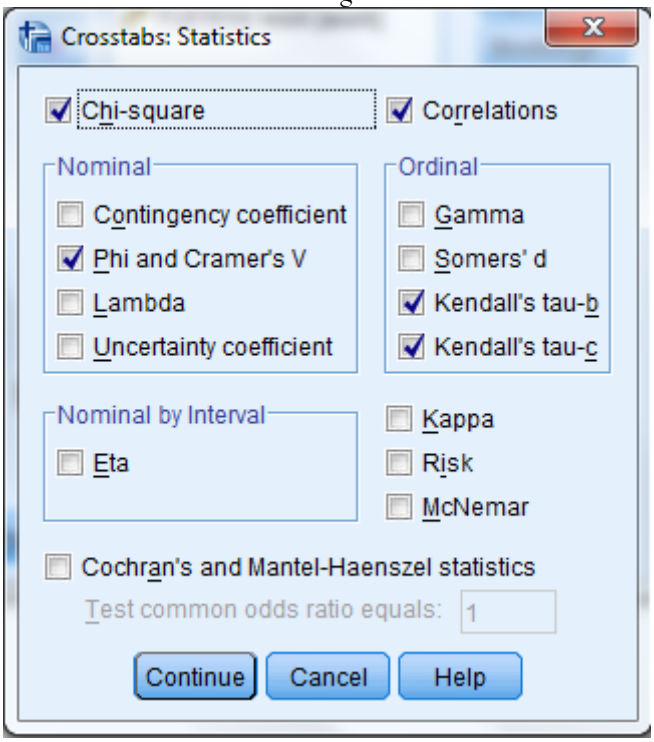
>0 - .10 = very weak
.10 - .25 = weak
.25 - .35 = moderate
.35 - .45 = strong
> .45 = very strong

0,384 would therefore be a 'strong' association.

4. Why would we use a chi-square test and/or Cramer's V in this case, and not – for example – a 'regular' Pearson's r, as we did in the previous lab session?

Because they are nominal variables. Pearson's r assumes interval or ratio measurements.

- Finally, go back one last time into the Crosstabs menu, enter the "Statistics" sub-menu again, and select all of the boxes shown in the image here:




This will produce many of the statistics for measures of association we discussed during the lecture.

5. In the lab session during Research Methods 1, we used Pearson's R to calculate correlations. Now, disregarding plus or minus signs, what do you notice about the values for the **Chi-Square, Cramer's V, Kendall's tau, Spearman Correlation, and Pearson's R**? In terms of ease of interpretation and comparison, would you rather use Chi-Square or Cramer's V? For those who really paid attention: why is the value for Tau-b different from Tau-c? Which one would you rather use in this case?

**Chi-Square:**

**Chi-Square Tests**

| | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
|---|---|---|---|---|---|
| Pearson Chi-Square | 266,798[a] | 1 | ,000 | | |
| Continuity Correction[b] | 264,920 | 1 | ,000 | | |
| Likelihood Ratio | 296,295 | 1 | ,000 | | |
| Fisher's Exact Test | | | | ,000 | ,000 |
| Linear-by-Linear Association | 266,651 | 1 | ,000 | | |
| N of Valid Cases | 1814 | | | | |

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 186,79.

b. Computed only for a 2x2 table

## Other measures:

**Symmetric Measures**

| | | Value | Asymp. Std. Error[a] | Approx. T[b] | Approx. Sig. |
|---|---|---|---|---|---|
| Nominal by Nominal | Phi | -,384 | | | ,000 |
| | Cramer's V | ,384 | | | ,000 |
| Ordinal by Ordinal | Kendall's tau-b | -,384 | ,018 | -17,941 | ,000 |
| | Kendall's tau-c | -,313 | ,017 | -17,941 | ,000 |
| | Spearman Correlation | -,384 | ,018 | -17,677 | ,000[c] |
| Interval by Interval | Pearson's R | -,384 | ,018 | -17,677 | ,000[c] |
| N of Valid Cases | | 1814 | | | |

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

The value for Cramer's V is the same (disregarding plus or minus signs) as it is for Pearson's R, Spearman Correlation, and Kendall's tau-b. The Chi-Square is a very different number. For ease of interpretation and comparison to other associations, we would therefore prefer Cramer's V in this case.

Kendall's tau-c assumes assumes 'rectangular' tables, i.e. tables where the number of categories in the columns and rows are not the same. Therefore, this value is slightly different here. The correct choice in this case would be tau-b, because we have two categories for sex (male, female) and two for work (part-time, full-time).

**For future reference**: also note that SPSS gives you a nice little hint on which measure to use in which case:

| | |
|---|---|
| Nominal by Nominal | Phi |
| | Cramer's V |
| Ordinal by Ordinal | Kendall's tau-b |
| | Kendall's tau-c |
| | Spearman Correlation |
| Interval by Interval | Pearson's R |

6. Now, considering the following fictional variables, which statistic(s) would you consider for the following associations?

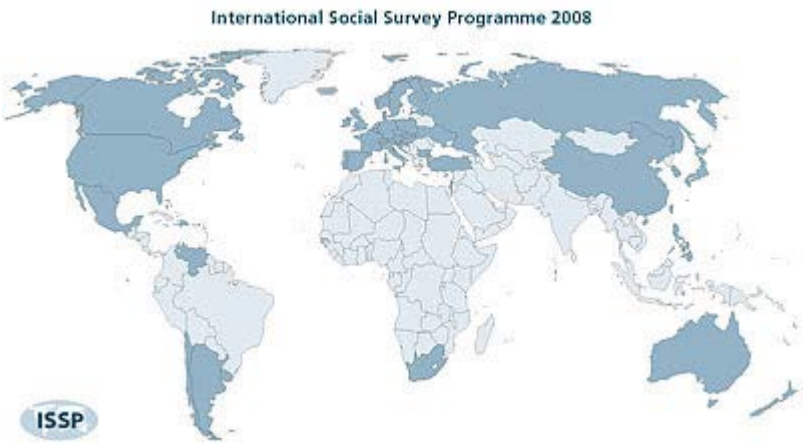
| Variable 1 | Variable 2 | Your choice(s) of measure |
|---|---|---|
| Age, in years | Number of children | **Pearson's R** (because both are interval/ratio/scale) |
| Level of education, in categories | Church attendance, in categories | **Tau-b, Tau-c, or Spearman Correlation** (because both are ordinal. Think about the number of answer categories to decide between Tau-b or Tau-c). |
| Height of interviewee | Weight of interviewee | **Pearson's R** again (scale vars) |
| Favourite beverage of interviewee | Country of origin of interviewee | **Cramer's V or Chi-Square** (because both are nominal) |

## 2. Downloading real data

For the remainder of the lab session, we will be using a 'real life' dataset instead of a practice file. A lot of quantitative data is available for free online. One of the largest data repositories for the social sciences is the GESIS Data Catalogue, maintained by the Leibniz Institute in Mannheim. In the next few steps you will create a free account there, and download a dataset.

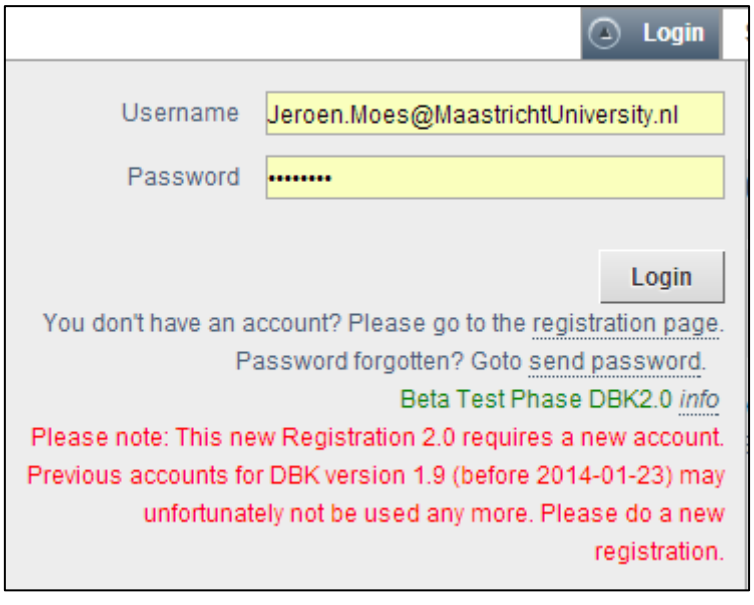
- Go to https://dbk.gesis.org/register/register.asp?db=E and fill in the form (preferably using your Maastricht University email address).
- Check your email inbox for the password they should have immediately sent you.

- Now we will download the data we will be using for the next few exercises. The International Social Survey Programme (ISSP) is a big survey that is repeated regularly with specific topic, and across the globe (see www.issp.org for more information). **We will be having a look at their 2003 survey on national identity**.



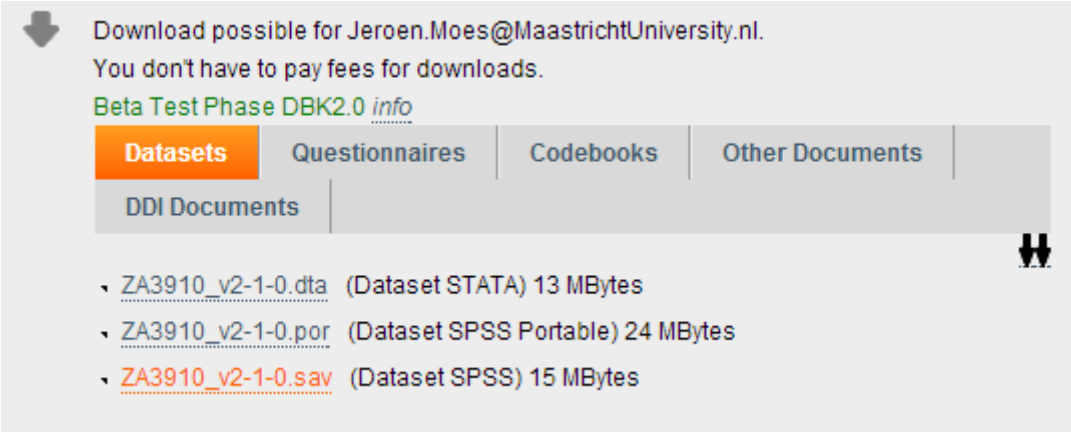

- Go directly to the following link: https://dbk.gesis.org/dbksearch/sdesc2.asp?no=3910

- Click the "Login" button at the top right corner, enter your email address and the password that was just sent to you, and click "Login":

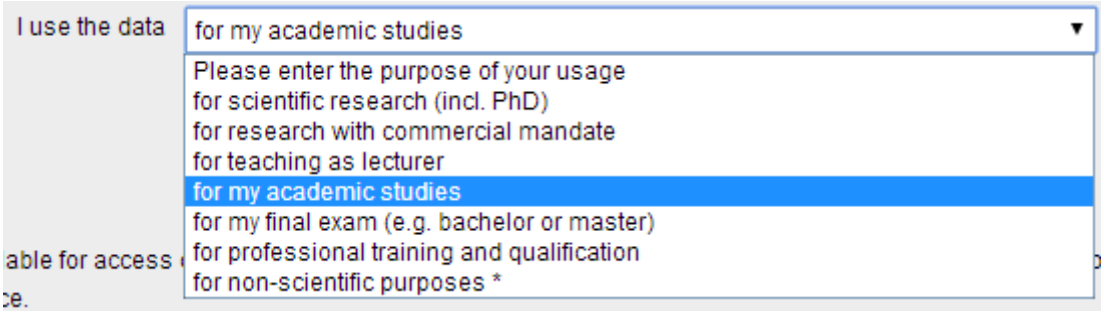



Once logged in, click "Data & Documents":

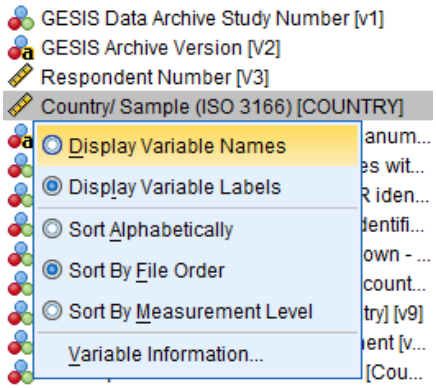
From there, download the 'ZA3910_V2-1-0.sav' dataset by clicking on it:




Indicate that you will use this data for your studies, tick the box, and click 'Download'.




- You can now directly open the data in SPSS. Have a look at the many variables included in this dataset. Don't worry – we won't be using all of them. Big datasets like these can be confusing at first glance, but once you get the hang of it you'll quickly be able to 'weed' through these datasets and find the things that interest you. Comparatively, the ISSP datasets are actually relatively 'small' in terms of the number of variables. For example, two other well-known global/European surveys, the Eurobarometer and World Values Studies are typically a lot more expansive. Those are also available for free online.
- **For future reference**: for the next steps, when looking at the long list of variables in the various menus, it may often be more helpful to write down some of the variable names on a piece of paper, and look at those names in SPSS instead of their labels. You can do that by right-clicking the list and selecting "Display Variable Names". Additionally, you can "Sort Alphabetically". This may make your life slightly easier at times.



7. Create a frequency table for the variable COUNTRY to have a look at the countries included in this dataset. How many people were interviewed in the Netherlands? And how many in Germany?

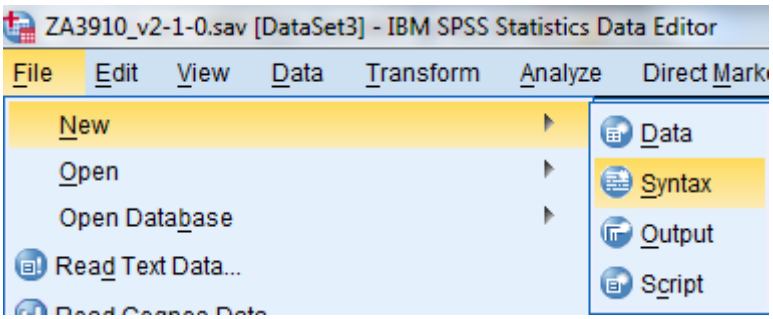
**Country/ Sample (ISO 3166)**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Australia (AU) | 2183 | 4,7 | 4,7 | 4,7 |
| | **Germany-West (DE-W)** | **850** | 1,8 | 1,8 | 6,6 |
| | **Germany-East (DE-E)** | **437** | 1,0 | 1,0 | 7,5 |
| | Great Britain (GB-GBN) | 873 | 1,9 | 1,9 | 9,4 |
| | United States (US) | 1216 | 2,6 | 2,6 | 12,1 |
| | Austria (AT) | 1006 | 2,2 | 2,2 | 14,3 |
| | Hungary (HU) | 1021 | 2,2 | 2,2 | 16,5 |
| | Ireland (IE) | 1065 | 2,3 | 2,3 | 18,8 |
| | **Netherlands (NL)** | **1823** | 4,0 | 4,0 | 22,8 |
| | Norway (NO) | 1469 | 3,2 | 3,2 | 26,0 |
| | Sweden (SE) | 1186 | 2,6 | 2,6 | 28,5 |
| | Czech Republic (CZ) | 1276 | 2,8 | 2,8 | 31,3 |
| | Slovenia (SI) | 1093 | 2,4 | 2,4 | 33,7 |
| | Poland (PL) | 1277 | 2,8 | 2,8 | 36,5 |
| | Bulgaria (BG) | 1069 | 2,3 | 2,3 | 38,8 |
| | Russia (RU) | 2383 | 5,2 | 5,2 | 44,0 |
| | New Zealand (NZ) | 1036 | 2,3 | 2,3 | 46,2 |
| | Canada (CA) | 1211 | 2,6 | 2,6 | 48,9 |
| | Philippines (PH) | 1200 | 2,6 | 2,6 | 51,5 |
| | Israel Jews (IL-J) | 1066 | 2,3 | 2,3 | 53,8 |
| | Israel Arabs (IL-A) | 152 | ,3 | ,3 | 54,1 |
| | Japan (JP) | 1102 | 2,4 | 2,4 | 56,5 |
| | Spain (ES) | 1212 | 2,6 | 2,6 | 59,2 |
| | Latvia (LV) | 1000 | 2,2 | 2,2 | 61,3 |
| | Slovakia (SK) | 1152 | 2,5 | 2,5 | 63,8 |
| | France (FR) | 1669 | 3,6 | 3,6 | 67,5 |
| | Portugal (PT) | 1602 | 3,5 | 3,5 | 70,9 |
| | Chile (CL) | 1505 | 3,3 | 3,3 | 74,2 |
| | Denmark (DK) | 1322 | 2,9 | 2,9 | 77,1 |
| | Switzerland (CH) | 1037 | 2,3 | 2,3 | 79,3 |
| | Venezuela (VE) | 1199 | 2,6 | 2,6 | 82,0 |
| | Finland (FI) | 1379 | 3,0 | 3,0 | 84,9 |

| | | | | | |
|---|---|---|---|---|---|
| | South Africa (ZA) | 2483 | 5,4 | 5,4 | 90,3 |
| | Taiwan (TW) | 2016 | 4,4 | 4,4 | 94,7 |
| | Korea (South) (KR) | 1315 | 2,9 | 2,9 | 97,6 |
| | Uruguay (UY) | 1108 | 2,4 | 2,4 | 100,0 |
| | Total | 45993 | 100,0 | 100,0 | |

Note that in many datasets, East and West Germany are still coded as separate countries. This obviously has historical reasons, but is also often done because socio-economic developments can be quite dissimilar in both parts of the country.

- Let's focus on just a couple of countries to simplify things a bit. Specifically, let's include Germany, the Netherlands, Poland, Russia, and the United States. An easy way to do procedures like these is by using the Syntax window we discussed during the previous lab session.
- Open a new Syntax window:



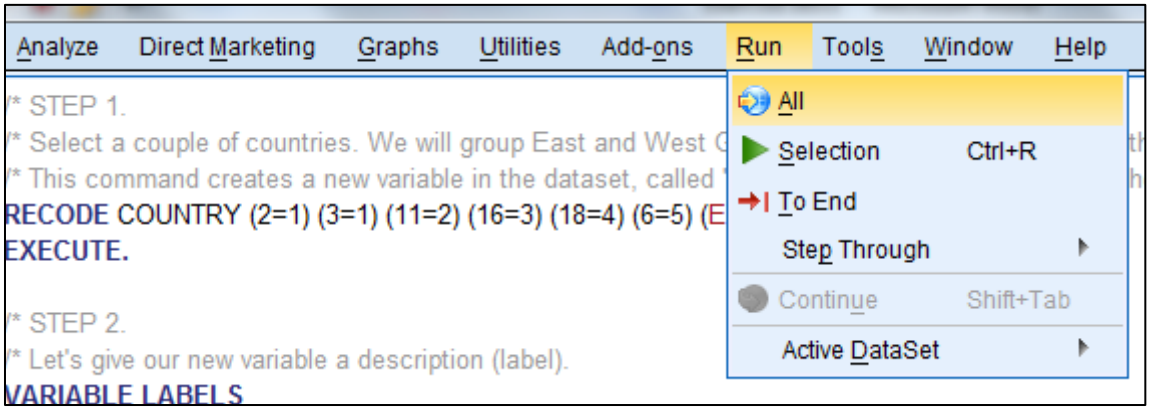

- I wrote the Syntax commands for you for this exercise. In your newly opened, empty Syntax window, copy and paste the following code:

```
/* STEP 1.
/* Select a couple of countries. We will group East and West Germany together
(this data was gathered in 2003).
/* This command creates a new variable in the dataset, called "country_new",
which is based on the existing variable "COUNTRY".
RECODE COUNTRY (2=1) (3=1) (11=2) (16=3) (18=4) (6=5) (ELSE=0) INTO
country_new.
EXECUTE.

/* STEP 2.
/* Let's give our new variable a description (label).
VARIABLE LABELS
country_new 'Selection of countries (recoded)'.
EXECUTE.

/* STEP 3.
```

```
/* Give names to those countries. We have numerical codes, but it will be
easier for us to know which number represents which country.
VALUE LABELS
country_new
0 'Other country'
1 'Germany'
2 'Netherlands'
3 'Poland'
4 'Russia'
5 'United States'.
EXECUTE.

/* STEP 4.
/* Let's also define that '0' is a missing value (these are all the other
countries).
MISSING VALUES country_new (0).
EXECUTE.
```

- Have a look at the Syntax. There are four steps we are taking. Each step is preceded by some comments (starting with "**/***") I wrote to explain what we are doing. In SPSS Syntax, the end of every command is indicated by a period.
    - **Step 1** creates a new variable for the countries we're interested in based on the existing COUNTRY variable.
    - **Step 2** gives a variable label to that new variable.
    - **Step 3** gives value labels to the codes in the new variable (i.e. what does "1" actually *mean*?).
    - **Step 4** tells SPSS that value '0' (all other countries) is a missing value for this variable.

- Let's go ahead and run this Syntax. Make sure that you only have one dataset open (the one we just downloaded. Click "Run"→ "All".



8. Create a frequency table for the new variable "country_new". How many respondents are there for Germany? How many respondents are there in "Other countries"?

**Selection of countries (recoded)**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | **Germany** | **1287** | 2,8 | 16,1 | 16,1 |
| | Netherlands | 1823 | 4,0 | 22,8 | 38,9 |
| | Poland | 1277 | 2,8 | 16,0 | 54,9 |
| | Russia | 2383 | 5,2 | 29,8 | 84,8 |
| | United States | 1216 | 2,6 | 15,2 | 100,0 |
| | Total | 7986 | 17,4 | 100,0 | |
| **Missing** | **Other country** | **38007** | 82,6 | | |
| Total | | 45993 | 100,0 | | |

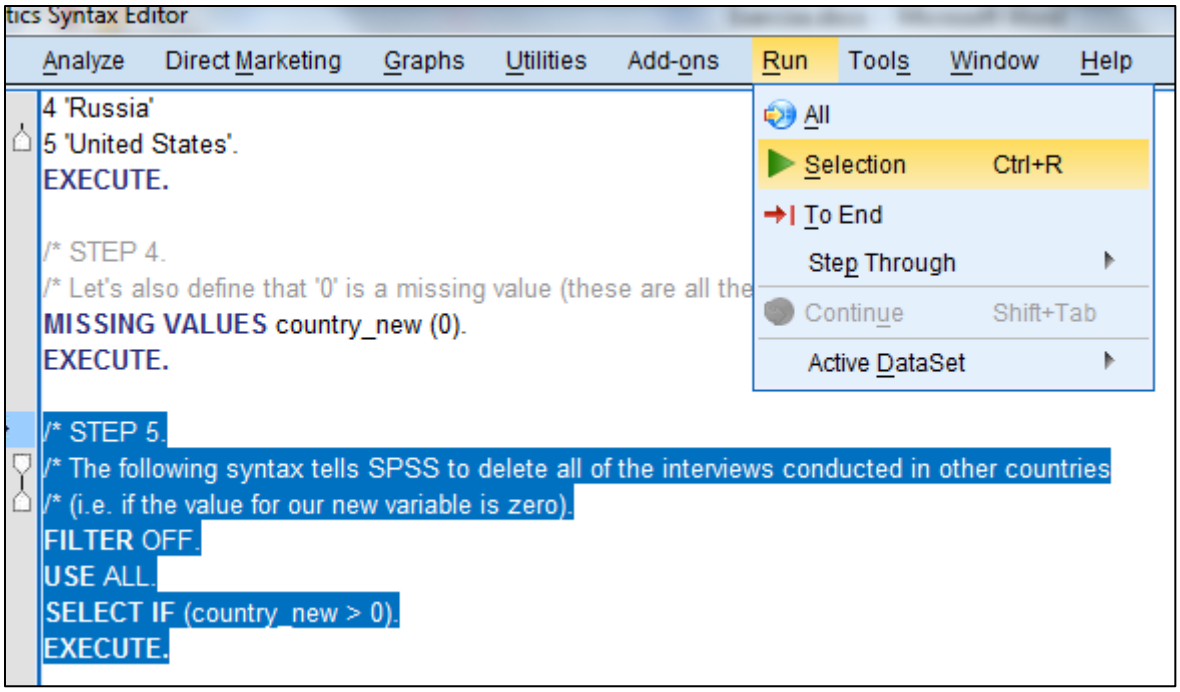
Note that I decided to group East and West Germany into one 'new' country called Germany.

The 'Other country' category is also coded as being missing values on this variable.

- Now let's remove all the interviews conducted in other countries than the ones we are interested in. Here is the Syntax for that:

```
/* STEP 5.
/* The following syntax tells SPSS to delete all of the interviews conducted
in other countries
/* (i.e. if the value for our new variable is zero).
FILTER OFF.
USE ALL.
SELECT IF (country_new > 0).
EXECUTE.
```

- To execute only that last bit of Syntax code, select it, and select "Run" → "Selection" (or press ctrl-R).

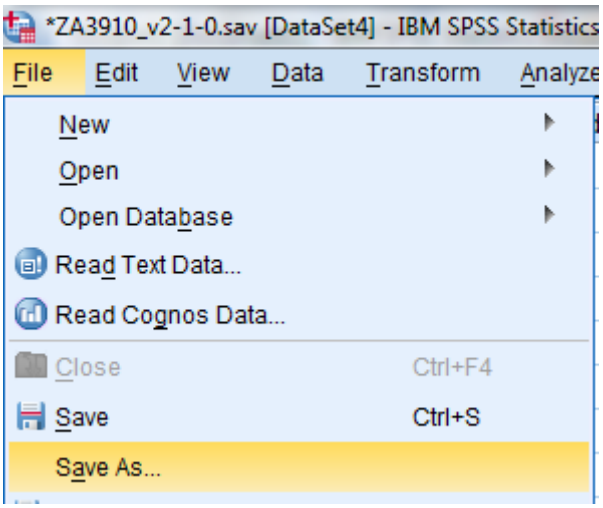



9. Run the same frequency table again. Notice that the category "Other countries" is no longer shown because there are now no respondents in that category anymore.

**Selection of countries (recoded)**

| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Germany | 1287 | 16,1 | 16,1 | 16,1 |
| | Netherlands | 1823 | 22,8 | 22,8 | 38,9 |
| | Poland | 1277 | 16,0 | 16,0 | 54,9 |
| | Russia | 2383 | 29,8 | 29,8 | 84,8 |
| | United States | 1216 | 15,2 | 15,2 | 100,0 |
| | Total | 7986 | 100,0 | 100,0 | |

- At this point, it may be wise to save your file under a different name, so as not to damage the original dataset if something goes wrong. From the data editor window:




## 3. Multiple regression analysis

Let's move on to conduct a simple multiple regression analysis. As you know, technically we can only use this technique with interval or ratio (scale) data. However, we will cheat a bit for this exercise and use the variable 'party_lr' ("R: Party affiliation: left-right (der.)") as the dependent variable. We're interested in looking at some of the determinants of right wing voting in the countries we selected above.

10. Which measurement type is this variable really? Set the proper "Measure" in SPSS, if necessary.

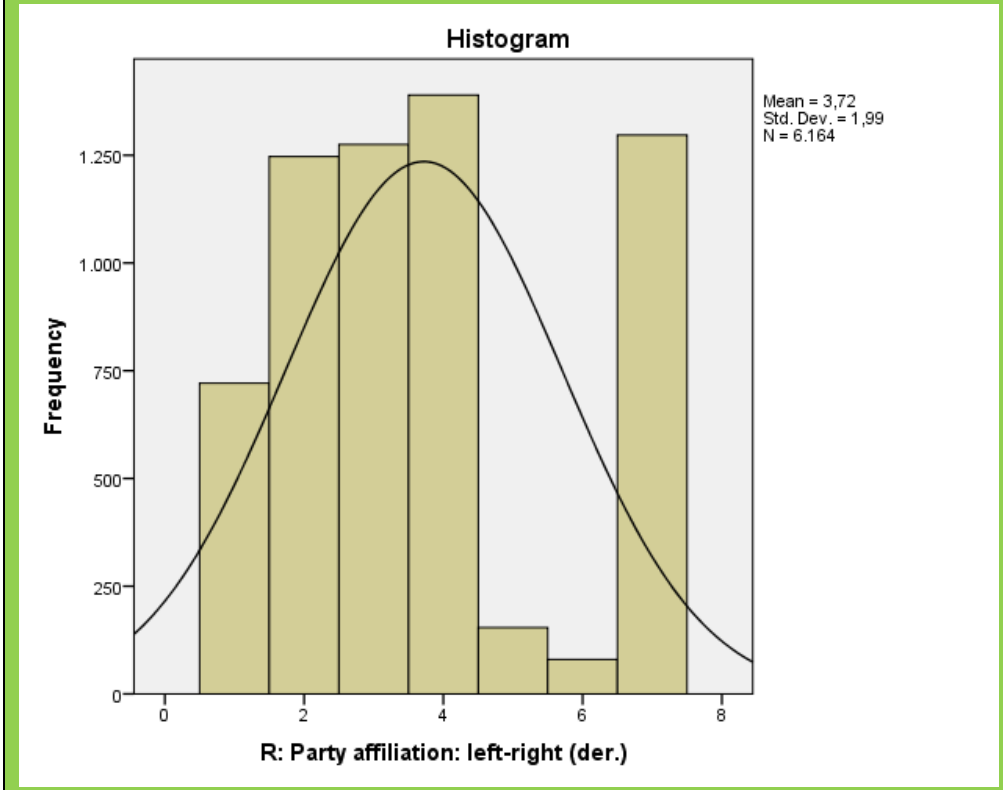
It's an ordinal variable. One is not necessarily 'higher' than the other, but they are definitely in a certain order.

11. Create a frequency table and a histogram for this variable and have a look at it. How many people say that they have no party preference? Why is this important to note?

**R: Party affiliation: left-right (der.)**

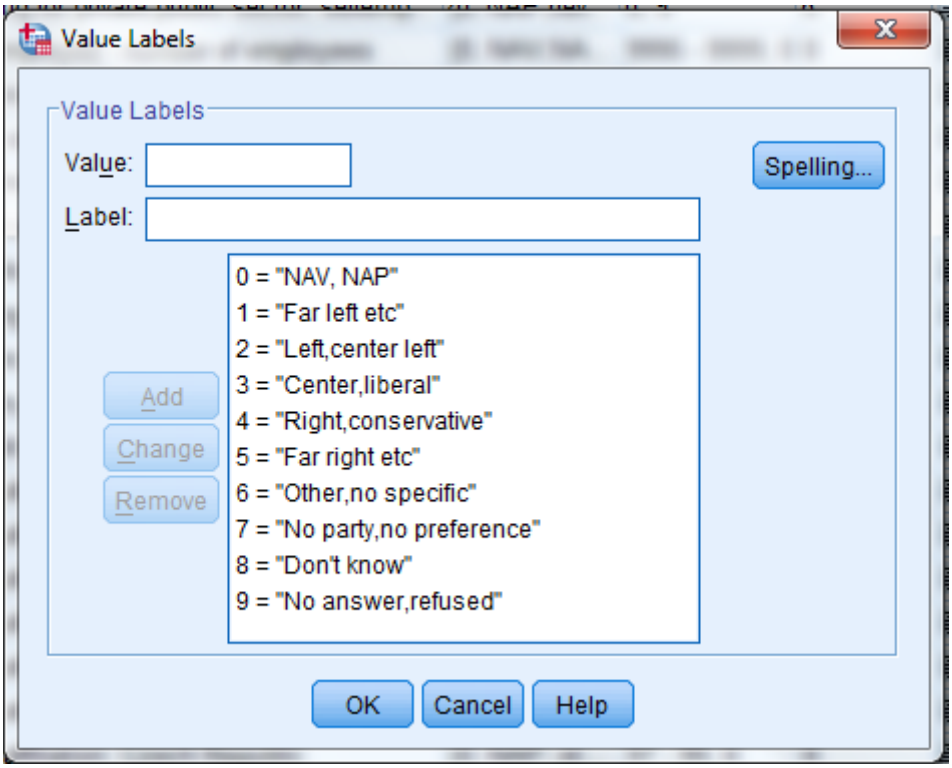
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Far left etc | 721 | 9,0 | 11,7 | 11,7 |
| | Left,center left | 1247 | 15,6 | 20,2 | 31,9 |
| | Center,liberal | 1275 | 16,0 | 20,7 | 52,6 |
| | Right,conservative | 1390 | 17,4 | 22,6 | 75,2 |
| | Far right etc | 154 | 1,9 | 2,5 | 77,7 |
| | Other,no specific | 80 | 1,0 | 1,3 | 79,0 |
| | **No party,no preference** | **1297** | 16,2 | 21,0 | 100,0 |
| | Total | 6164 | 77,2 | 100,0 | |
| Missing | NAV, NAP | 529 | 6,6 | | |
| | Don't know | 1198 | 15,0 | | |

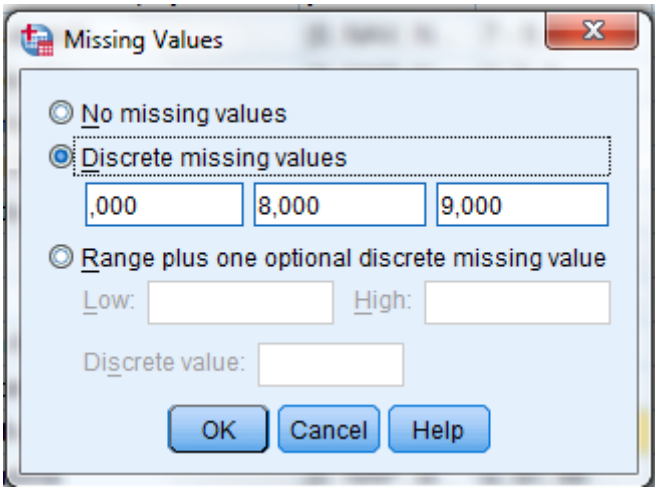
| | | | | | |
|---|---|---|---|---|---|
| | No answer,refused | 95 | 1,2 | | |
| | Total | 1822 | 22,8 | | |
| Total | | 7986 | 100,0 | | |




1297 people have no party preference. This is in principle fine, of course, but for our variable it disrupts the data. The code for this category in SPSS is '7' (see image below). This means that if we go ahead and cheat on regression analysis by including an ordinal variable as if it were a scale variable, SPSS things that 'no party preference' is actually a value above 'far right'. This would skew our data strongly towards the right and screw up our analyses.

- Before going into the analysis, we need to make sure that this variable has the correct codes set. The value labels and missing values are as follows:



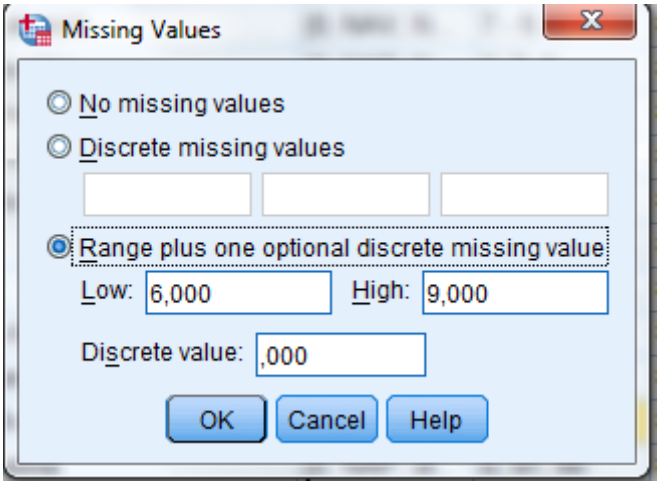




12. Have a look at the value labels for this variable. Thinking about what it means if we consider this variable an interval variable, what does a higher score on this variable mean exactly?

It means that the higher someone scores

As mentioned above, the code for this category in SPSS is '7' (see image below). This means that if we go ahead and cheat on regression analysis by including an ordinal variable as if it were a scale variable, SPSS things that 'no party preference' is actually a value above 'far right'. This would skew our data strongly towards the right and screw up our analyses.

13. Tell SPSS that 6 and 7 should also be considered missing values. You can do that like this:

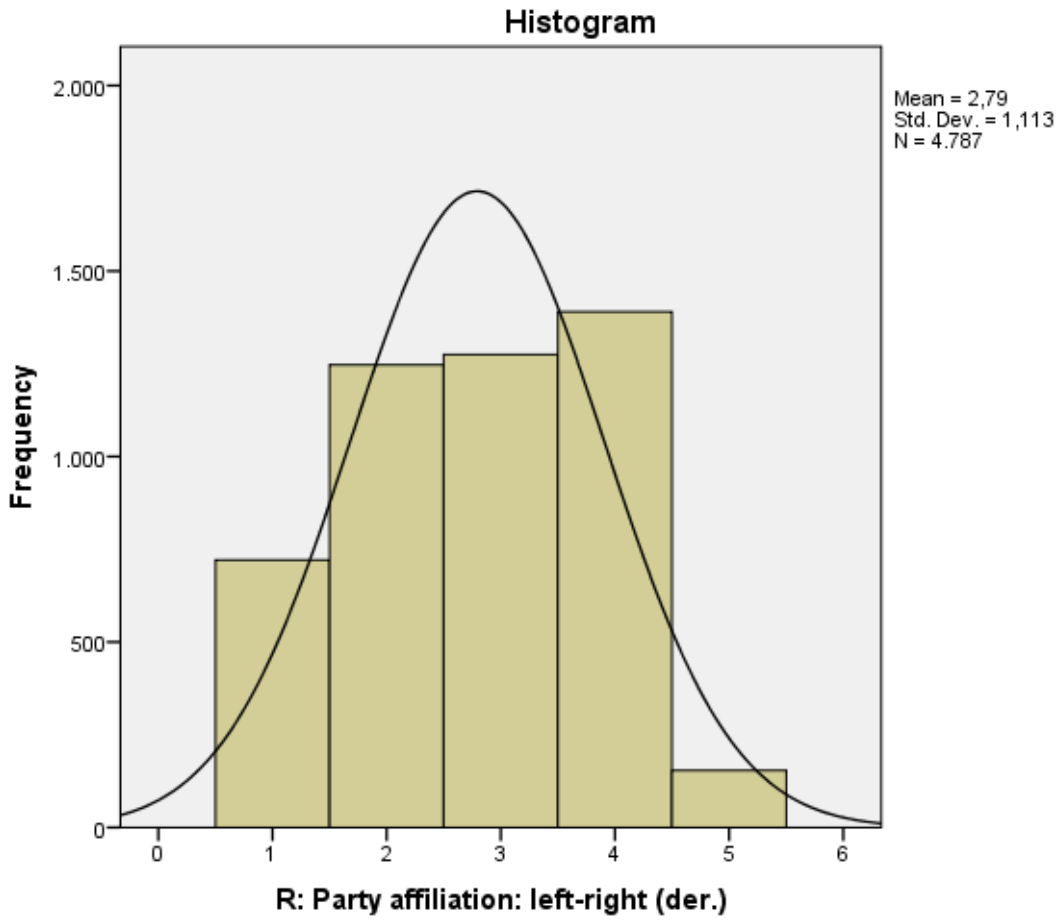



**Don't forget to include the discrete value zero above!**

14. Run the frequency table again, including histogram. Do you think this distribution will be more useful for our purposes?

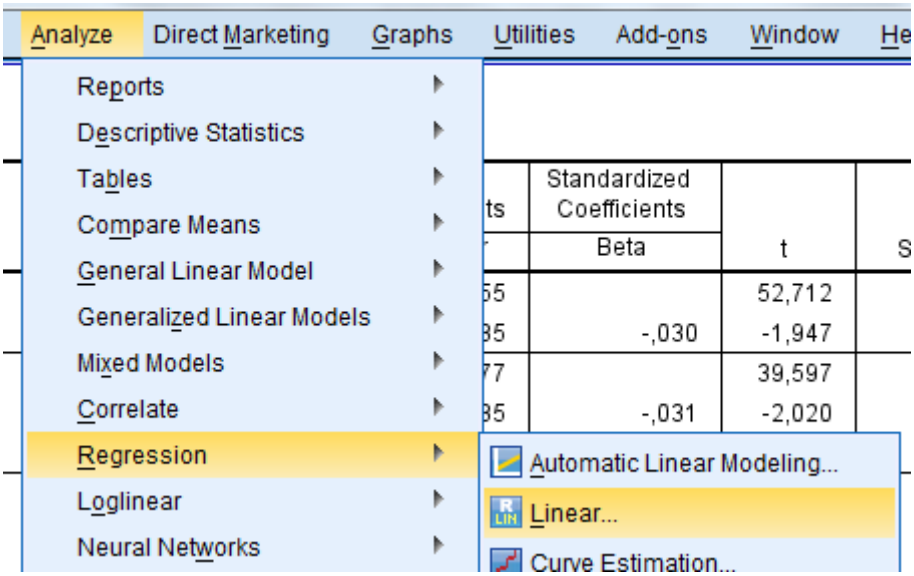
**R: Party affiliation: left-right (der.)**

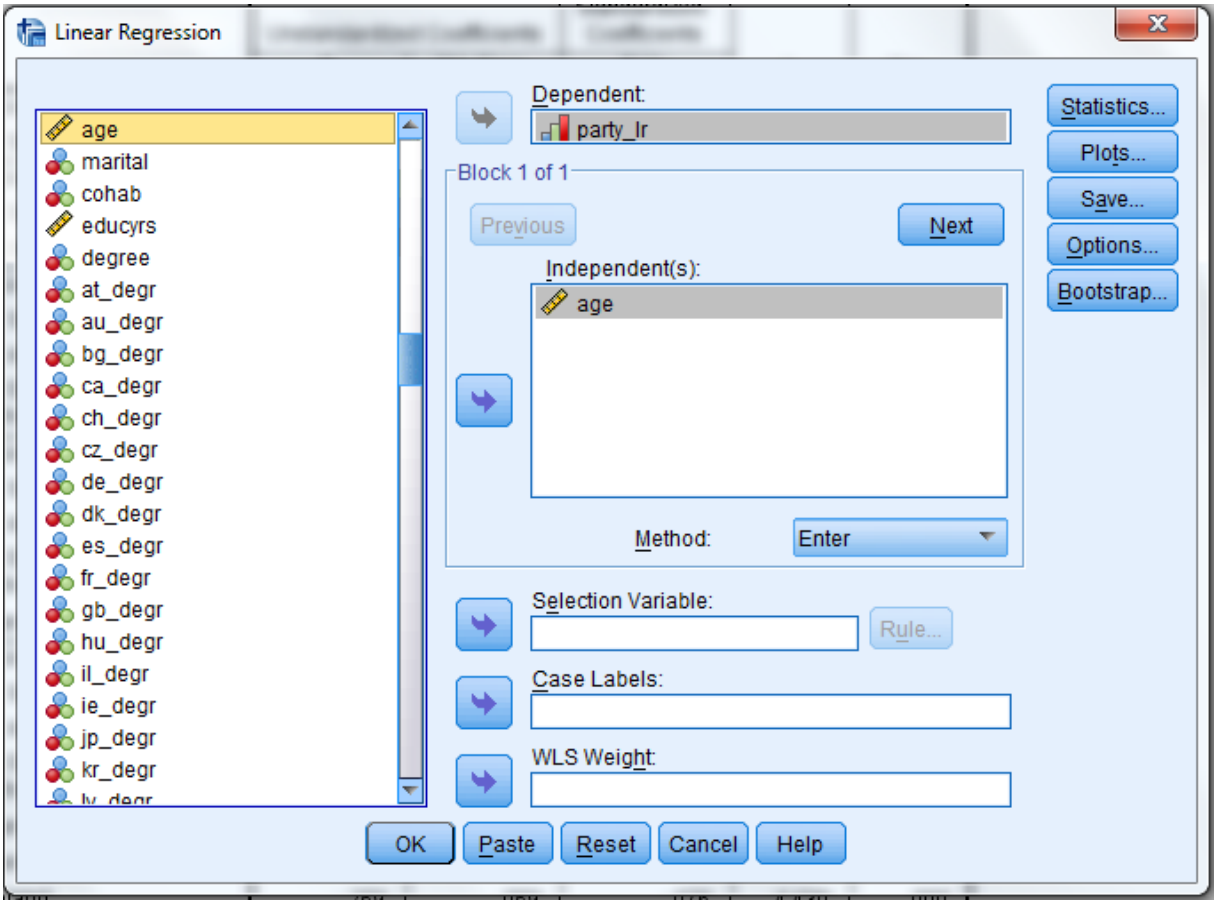
| | | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|---|
| Valid | Far left etc | 721 | 9,0 | 15,1 | 15,1 |
| | Left,center left | 1247 | 15,6 | 26,0 | 41,1 |
| | Center,liberal | 1275 | 16,0 | 26,6 | 67,7 |
| | Right,conservative | 1390 | 17,4 | 29,0 | 96,8 |
| | Far right etc | 154 | 1,9 | 3,2 | 100,0 |
| | Total | 4787 | 59,9 | 100,0 | |
| **Missing** | **NAV, NAP** | 529 | 6,6 | | |
| | **Other,no specific** | 80 | 1,0 | | |
| | **No party,no preference** | 1297 | 16,2 | | |
| | **Don't know** | 1198 | 15,0 | | |
| | **No answer,refused** | 95 | 1,2 | | |
| | Total | 3199 | 40,1 | | |
| Total | | 7986 | 100,0 | | |




Looks better, I think.

- Now that we have properly defined the dependent variable 'party_lr', let's start with a simple linear regression as we did during the first lab session. Our first hypothesis is that older people tend to be more right wing in these countries. Run a regression analysis with 'party_lr' as the dependent variable, and 'age' as the only independent variable. Some hints:


Analyze
Direct Marketing
Graphs
Utilities
Add-ons
Window
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Automatic Linear Modeling...
Linear...
Curve Estimation...
Standardized Coefficients
Beta
t
52,712
-,030
-1,947
39,597
-,031
-2,020


Linear Regression
age
marital
cohab
educyrs
degree
at_degr
au_degr
bg_degr
ca_degr
ch_degr
cz_degr
de_degr
dk_degr
es_degr
fr_degr
gb_degr
hu_degr
il_degr
ie_degr
jp_degr
kr_degr
Dependent:
party_lr
Block 1 of 1
Previous
Next
Independent(s):
age
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help

15. Looking at the Coefficients table, what is your tentative conclusion about the effect of age on voting right wing?

**Coefficients[a]**

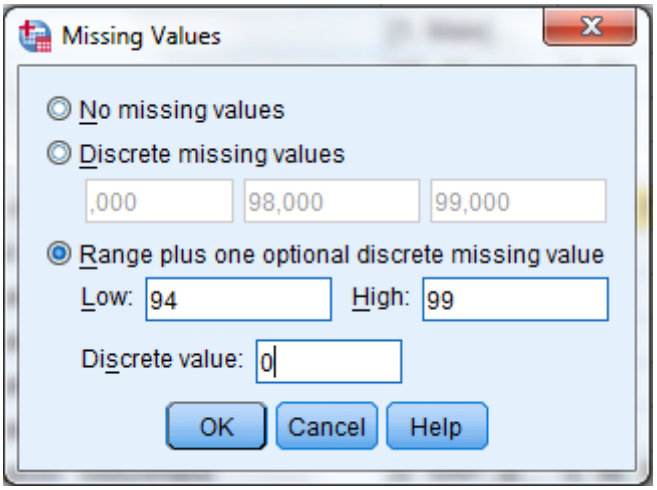
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 2,940 | ,050 | | 59,019 | ,000 |
| | R: Age | **-,003** | ,001 | -,045 | -3,087 | **,002** |

a. Dependent Variable: R: Party affiliation: left-right (der.)

There is a statistically significant effect of age on voting left/right wing. The effect is negative, which means that the older people get, the more likely it is to vote left wing (or the less likely it is they vote right wing).

- Now let's say that someone approaches you and says that it's rather silly to look at age. He claims that in fact men and women might be different, and more importantly, that the level of education will affect left-right wing voting. His hypothesis is that the higher someone's level of education, the less likely it is that they will vote right wing.

16. Let's test those ideas. We will run a linear regression analysis as before, but now we will add a variable for sex of the respondent (var: "sex"; coded 1=men, 2=women) and for years of education ("educyrs"). Unfortunately, we have to define some more missing values for educyrs first. Let's not waste time here. Make sure that the missing values for 'educyrs' are set as follows:




17. Now run a multiple regression analysis, the same way you did before, but this time include not only age, but also sex and educyrs. Have a look at the Coefficients table in your output. Based on the direction of the coefficients (positive or negative), and their level of significance, what are your conclusions? **Note: you can assume an α of 5%.**

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 2,916 | ,105 | | 27,814 | ,000 |
| | R: Age | **-,003** | **,001** | -,042 | -2,708 | **,007** |
| | R: Sex | **-,078** | **,033** | -,035 | -2,348 | **,019** |
| | R: Education I: years of schooling | **,010** | **,005** | ,033 | 2,167 | **,030** |

a. Dependent Variable: R: Party affiliation: left-right (der.)

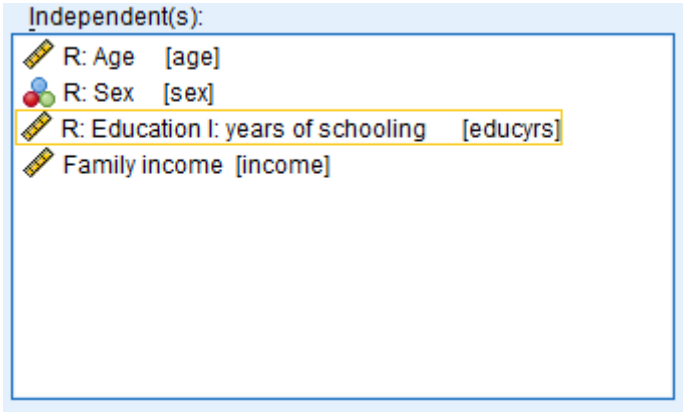
At α=5%, all three variables have a statistically significant effect on voting behaviour (they are all below 0,05). Age and sex have a negative effect, and years of education has a positive effect. Substantively, this implies that older people are likely to vote more left wing, women are more likely to vote left wing, and that higher education people tend to vote more right wing. This last conclusion is especially puzzling. Earlier research consistently suggests that higher educated people generally vote more left wing. Why is that the case then? Let's look at the next steps to find out.

18. Based on this information alone, would you confirm or reject your critic's hypothesis that education negatively affects right wing voting? Is this in line with your own expectations?

You would reject this hypothesis. In fact, the reverse seems to be true based on this analysis.

- Considering neo-liberal right wing parties as opposed to xenophobic (extreme) right wing parties, the effects may be quite different for both types of "right". Taking the neo-liberal right in mind, another control variable that we may have mistakenly omitted may be the household income of the respondents. Let's also add that as a control variable and run the regression again.

This is what we have in the linear regression menu now:



19. Have a look at the new Coefficients table. What is your conclusion about the effect for Family income?

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 2,923 | ,110 | | 26,502 | ,000 |
| | R: Age | -,003 | ,001 | -,039 | -2,456 | ,014 |
| | R: Sex | -,074 | ,035 | -,033 | -2,145 | ,032 |
| | R: Education I: years of schooling | ,004 | ,005 | ,011 | ,693 | **,488** |
| | Family income | **2,838E-006** | ,000 | ,088 | 5,488 | **,000** |

a. Dependent Variable: R: Party affiliation: left-right (der.)

Family income has a positive, statistically significant effect on voting behaviour. In other words, the higher someone's family income, the more likely it is they vote right wing.

20. Have another look at what the effect is for the level of education. Especially consider its significance level. What do you notice? In this regression model, what is the effect of controlling for income, according to you?

The effect of education is no longer statistically significant after we **control** for the level of income. You can look at this effect this way; people with a higher education also tend to have better jobs, and as a result a higher income (the same applies to older people and, sadly, men, which explains why those values change as well – though not significantly). Adding income as a control variable 'disentangles' the influence of education and income. After disentangling it, the conclusion here is that it is not the level of education causing people to vote right wing (no significant effect), but rather their level of income.

# 4. Dummy variables

As a final addition to our multiple regression model, we want to see whether there are statistically significant differences in left/right wing voting between the countries in our dataset (remember that we only still have Germany, the Netherlands, Poland, Russia, and the United States left in our data). For that purpose, we will be creating dummy variables for each of those countries.

21. Why can't we simply include the variable 'country_new' in the multiple regression analysis?

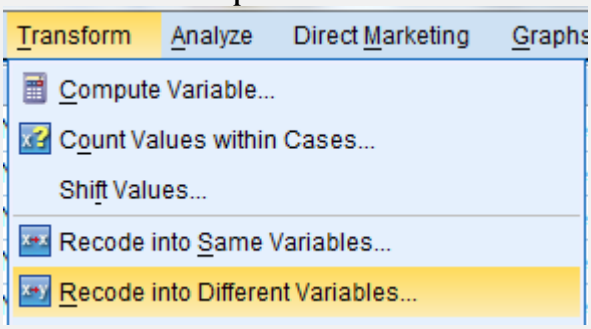
Because it is a **nominal** variable. If we were to include it anyway, interpreting it would be impossible. If you do try it, SPSS will even give you a statistically significant result for that variable, which illustrates that these analyses are always valid, but not always reliable (i.e. can be nonsensical). Look at it this way; the value for United States

is 6, while Germany is 1, and the Netherlands is 2. Those numbers don't represent anything other than separate categories. However, the coefficient (b) in the regression analysis for this nominal variable would imply the change in value per value of y (voting behaviour). Since these are just categories, that doesn't make any sense.

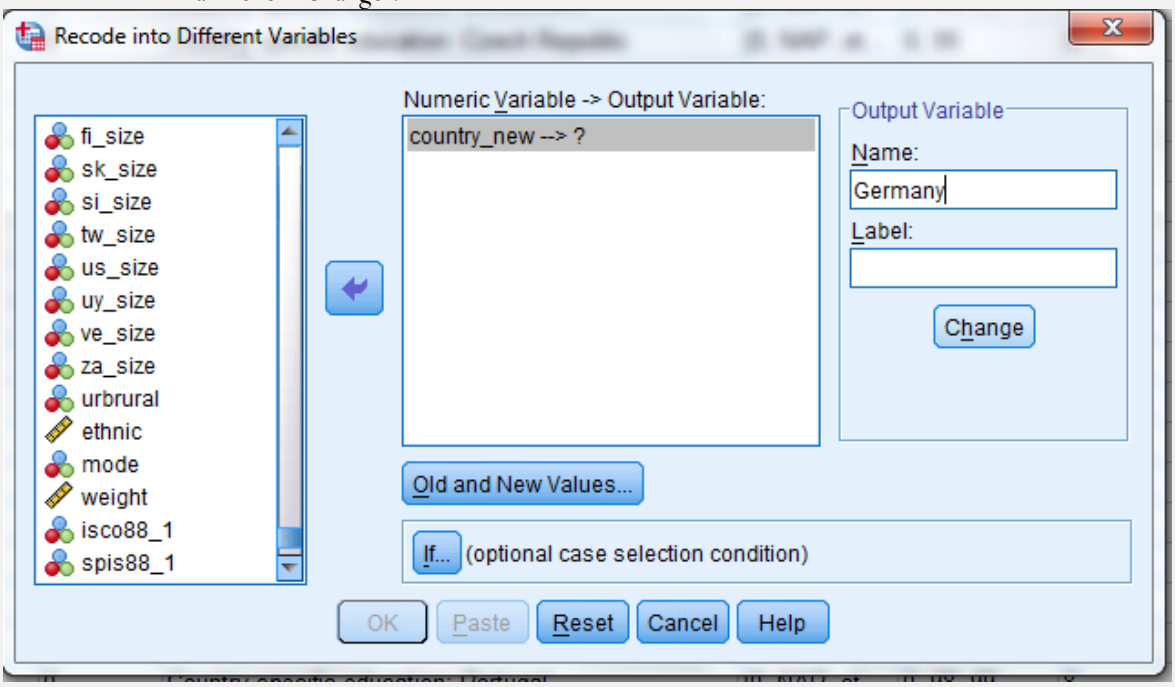
To create dummy variables, we have to follow the next couple of steps once for each country. **However, I have created a short Syntax code to save some time and do it for you below these instructions.** The below is reference material for you to use during your project if you need to.

## *Creating dummies 'by hand'*

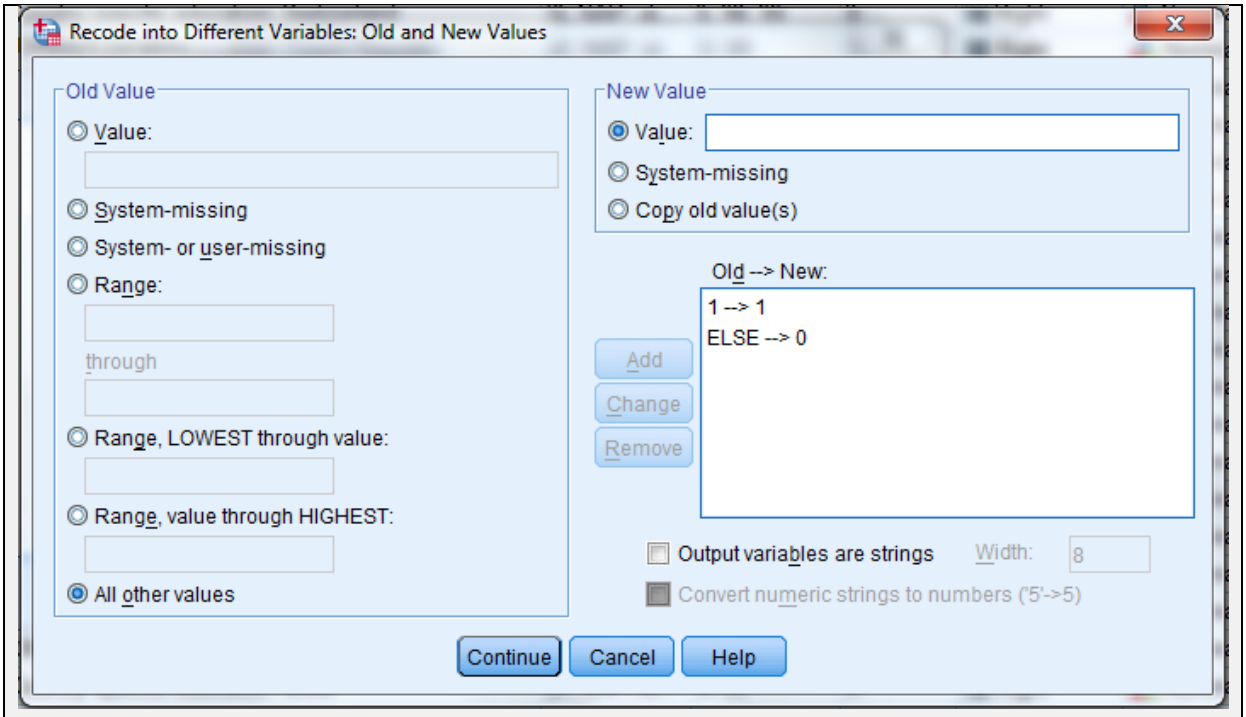
- **Step 1** – Click Transform → Recode into Different variables




- **Step 2** – Add 'country_new' to the box in the middle, type a name for the new dummy variable, and click 'Change'.




- **Step 3** – Click 'Old and New Values'. In this menu, you tell SPSS which value in the original variable should become value 1 (i.e. 'yes'), and tell it to code all other values as 0. Click 'Add' for both, and click Continue.

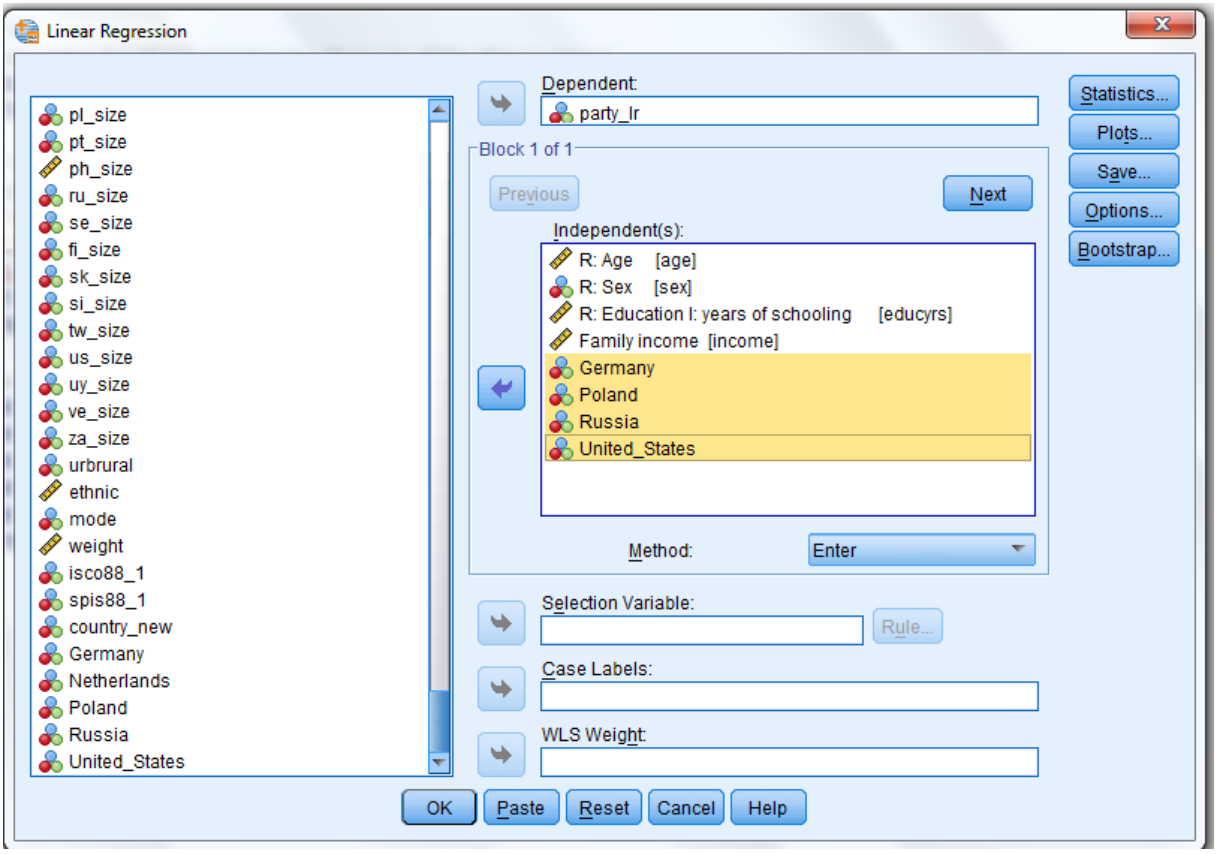



- **Step 4** – Click 'OK' and repeat the process for each country. Make sure that you indicate the correct value (i.e. 1 for Germany, 2 for Netherlands, etc.) in Step 3.

**Instead of doing the steps above, copy/paste and run the syntax below to create dummy variables for each country automatically (run selection):**

```
/* Create dummies per selected country.
RECODE country_new (1=1) (ELSE=0) INTO Germany.
RECODE country_new (2=1) (ELSE=0) INTO Netherlands.
RECODE country_new (3=1) (ELSE=0) INTO Poland.
RECODE country_new (4=1) (ELSE=0) INTO Russia.
RECODE country_new (5=1) (ELSE=0) INTO United_States.
EXECUTE.
```

- With the dummy variables ready, let's run a full multiple regression model again, including all of the variables from before, and all dummy variables **EXCEPT** the one for the Netherlands.




**For future reference**: remember that we **need** to exclude one of the categorical dummy variables in order to have a **reference category**. In your interpretation, this means that the effects shown per country are that particular country as compared to the reference category (i.e. the Netherlands).

22. Have a look at the significance levels of the country dummies, and at whether their coefficients are positive or negative. What do these mean? How does Germany compare to the Netherlands according to this? How about Russia? And what about the United States?

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 2,962 | ,119 | | 24,987 | ,000 |
| | R: Age | -,002 | ,001 | -,032 | -2,009 | ,045 |
| | R: Sex | -,059 | ,034 | -,026 | -1,706 | ,088 |
| | R: Education I: years of schooling | ,002 | ,005 | ,007 | ,384 | ,701 |
| | Family income | 2,225E-006 | ,000 | ,069 | 3,140 | ,002 |
| | **Germany** | **,181** | ,055 | ,059 | 3,284 | **,001** |
| | **Poland** | **-,259** | ,059 | -,076 | -4,430 | **,000** |

| | | | | | | |
|---|---|---|---|---|---|---|
| | **Russia** | **-,409** | ,056 | -,126 | -7,258 | **,000** |
| | **United_States** | **,012** | ,060 | ,005 | ,206 | **,837** |

a. Dependent Variable: R: Party affiliation: left-right (der.)

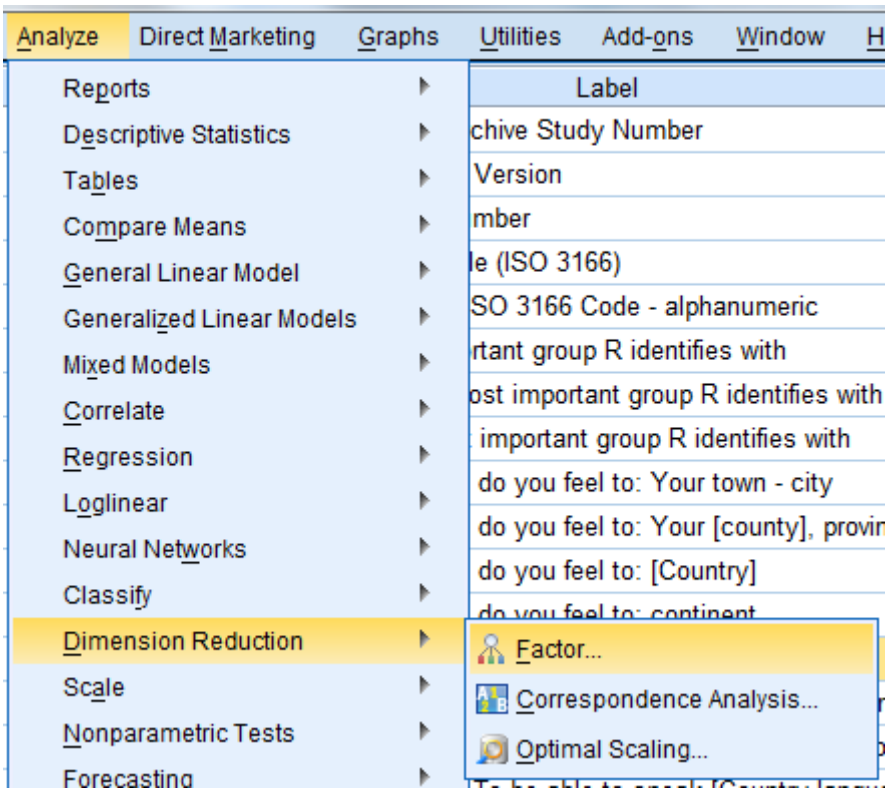
Note that when interpreting these dummy variables, you are comparing them to the reference category, which in this case is the Netherlands. Germany, Poland, and Russia are statistically significantly different from the Netherlands, but the United States is not. This means that the US are not statistically different from the Netherlands in this particular model. Also, the effects (b coefficients) of Poland and Russia are negative (more left-wing), while Germany is positive (more right wing/conservative). Rest assured that this effect changes further if we would include specific interaction effects, and don't forget that we are not only measuring the extreme right but also conservative parties like the Christian Democrats etc.

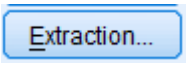
Also, don't forget that when you're reporting results like these in a paper you should always mention which category the reference category is. Often, this variable is simply added to the table, with "(ref.)" instead of a value for the b coefficient.

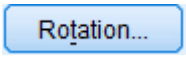
# 5. Factor analysis

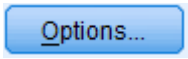
You might be getting fed up with doing statistics for today. Don't forget that you managed to pull off quite an impressive analysis today! This last section will be very quick, and acts mostly as a demonstration and manual for future reference. We will be conducting a factor analysis to see if we can distinguish different factors in respondents' attitudes about what it means to be from their respective country.

- Open the factor analysis menu:

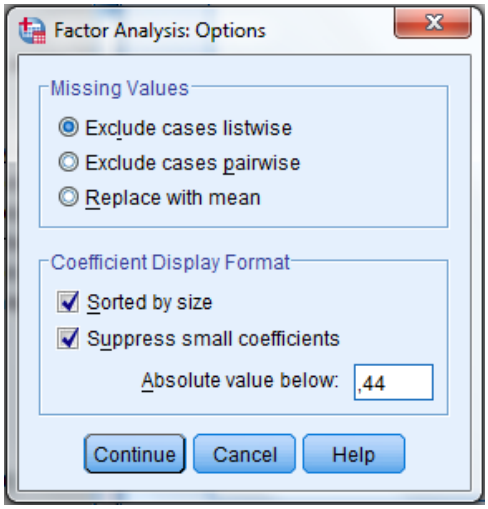


- Add V11 to v18 (i.e. Q3a to Q3h) to the variables list. Remember that you can switch between variable names and variable labels by right clicking the list of variables on the left.
- Click the 

button on the right, and don't change anything there except checking the 'scree plot' tick box.
- Click 'OK', and then click the 

button on the right. Don't change anything, except selecting 'Direct Oblimin'.
- Click 'OK' and then click the 

button on the right. Make sure that box looks like this (remember also our **0.44 rule of thumb** in the lecture):




- Click 'Continue', and then 'OK' to run the Factor Analysis.

23. Have a look at your Output window, and remember what we discussed in the lecture:
    a. Look at the second table ('Total Variance Explained'). Looking at the Eigenvalues, how many factors would you distinguish based on this?
    b. Look at the Scree plot. How many factors would you distinguish based on this?

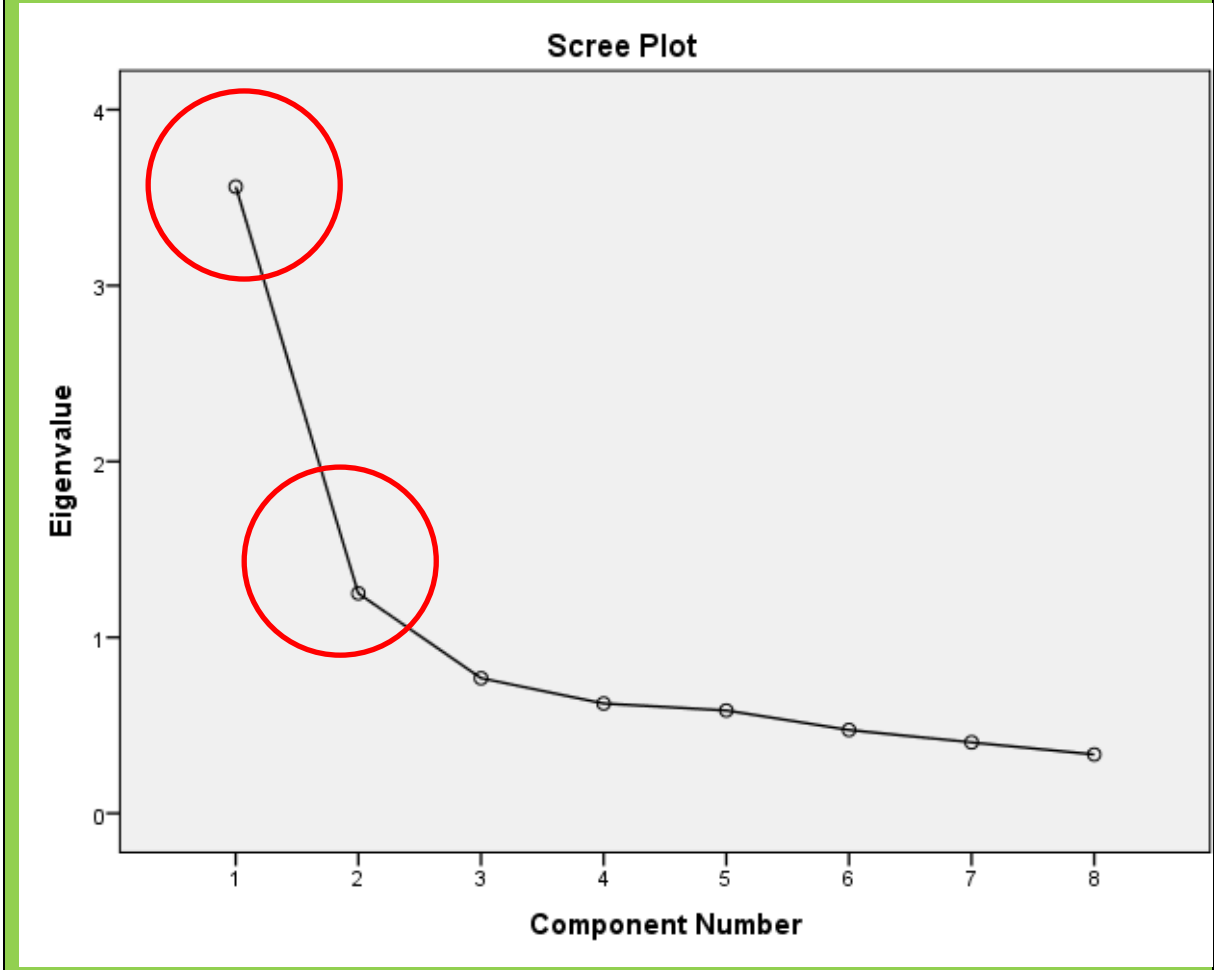
**A.**

**Total Variance Explained**

| Component | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | | Rotation Sums of Squared Loadings[a] |
|---|---|---|---|---|---|---|---|
| | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % | Total |
| **1** | **3,562** | 44,522 | 44,522 | 3,562 | 44,522 | 44,522 | 3,321 |
| **2** | **1,250** | 15,627 | 60,149 | 1,250 | 15,627 | 60,149 | 2,065 |
| 3 | ,768 | 9,601 | 69,750 | | | | |
| 4 | ,624 | 7,800 | 77,550 | | | | |
| 5 | ,585 | 7,309 | 84,858 | | | | |
| 6 | ,474 | 5,921 | 90,779 | | | | |
| 7 | ,403 | 5,041 | 95,820 | | | | |
| 8 | ,334 | 4,180 | 100,000 | | | | |

There are two factors ('components') with Eigenvalues over 1, which is our rule of thumb (see lecture).

**B.**




There are two factors/components identified before the 'bend'. We'd reach a similar conclusion here.

24. Finally, have a look at the Pattern Matrix (the third table from the bottom). We told SPSS that we only want to see values of over 0.44, which explains the blank spots. Do the factors ('components') that SPSS distinguished make sense to you? How would you interpret them?

**Pattern Matrix[a]**

| | Component | |
|---|---|---|
| | 1 | 2 |
| Q3h Important: To have [Country Nationality] ancestry | ,878 | |
| Q3a Important: to have been born in [Country] | ,835 | |
| Q3e Important: To be a [religion] | ,747 | |
| Q3c Important: To have lived in [Country] for most of one's life | ,728 | |
| Q3b Important: To have [Country Nationality] citizenship | ,483 | ,443 |
| Q3g Important: To feel [Country Nationality] | | |
| Q3f Important: To respect [Country Nationality] political institutions and laws | | ,834 |
| Q3d Important: To be able to speak [Country language] | | ,719 |

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.
a. Rotation converged in 6 iterations.

I outlined the two identified factors in red and green above. Let's call them 'Factor Red' and 'Factor Green' for now. Let's also disregard Q3b (citizenship) for now because it seems to be a part of both factors.

What the above means is that SPSS calculated that the group of variables in Factor Red 'go together'. One way of looking at this, is that the correlation between all of the variables *within* Factor Red will correlate highly with *each other*, but comparatively not as much with the variables in Factor Green. Conversely, the variables in Factor Green (so respect for institutions and being able to speak the language) will correlate strongly with each other, but not that much with the variables in Factor Red.

Whether this grouping makes sense substantively is a decision for the researcher to make. It requires thinking and analysis on your part, and SPSS will not tell you (which is why it simply calls them 'Component 1' and 'Component 2'). Personally, I'm tempted to consider Factor Red a measure for symbolic nationalism, and Factor Green a more 'instrumental' nationalism. Or perhaps a cultural versus a civic type of nationalism/identity.

# 6. ANOVA and t-test (optional, but recommended)

You might be getting fed up with doing statistics for today. Don't forget that you managed to pull off quite an impressive analysis today! This final section is also optional, but short and highly recommended. T-tests and ANOVA's are considered two fairly basic techniques in statistics, and are in many ways actually less versatile and complex than things like regression analysis and factor analysis. Here, we will only look at the bare essentials of these techniques, and only at ANOVA.[1] If and when you need to apply them during your project period (or later), you have a starting point for your work and Jeroen will be happy to help you further.

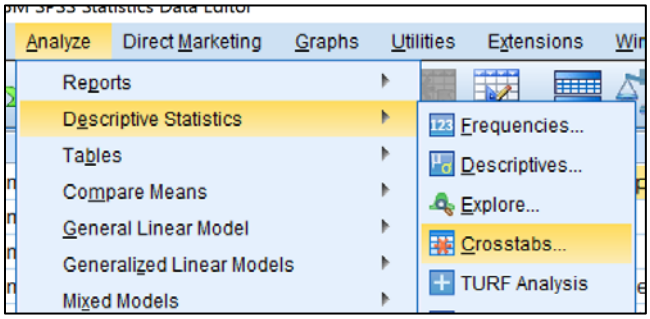
As we discussed in the lecture, these two techniques are used to assess whether there is a statistically significant difference in the average scores on a variable between two (or more) groups in the data. For this reason, these techniques are often used for experimental research, where a control group is compared to an experimental group.

This is exactly what two PEERS students were trying to achieve when they collected data amongst the Research Methods 2 students who volunteered to participate. Their research was fairly straightforward: they wanted to find out w**hether students would be more likely to judge a crime suspect as guilty after seeing a video of their confession rather than reading the transcript**. To this end, they showed the video confession to one half of the students, and had the other half read the written transcript of that confession.

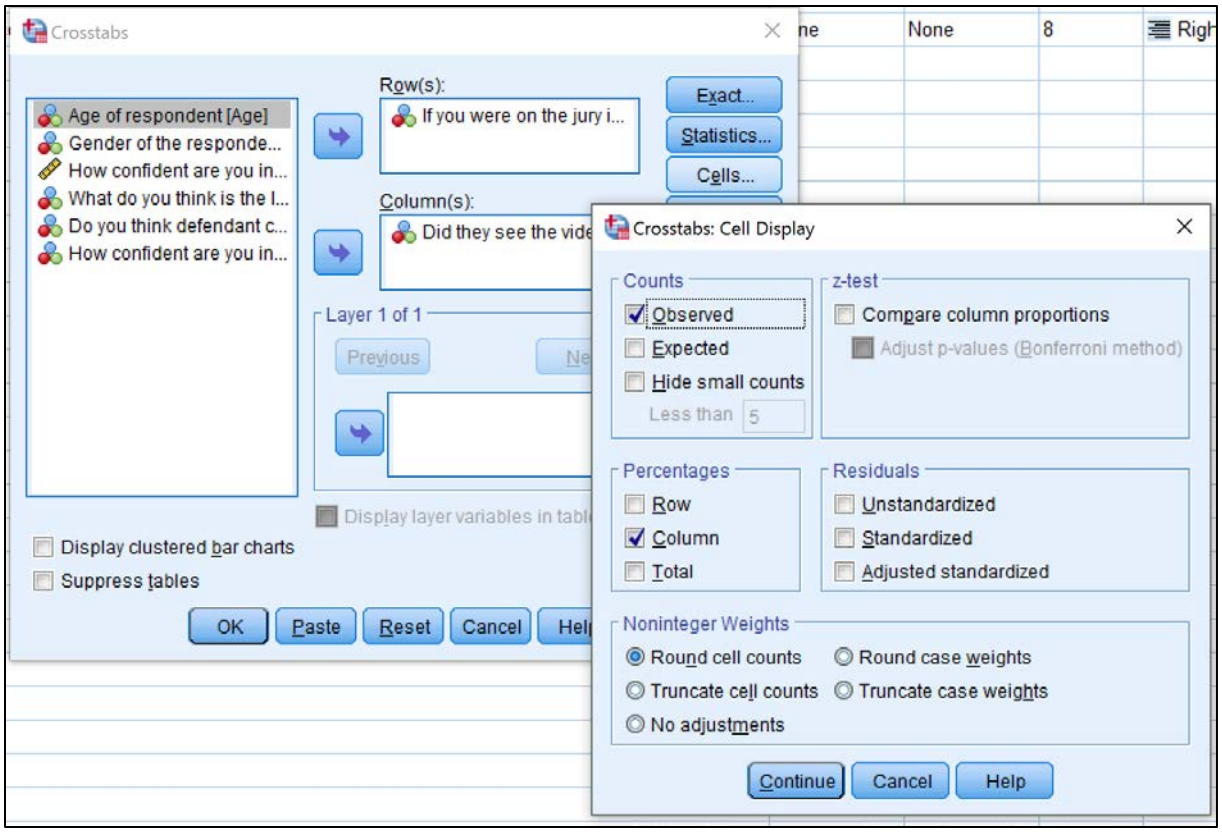
Their anonymized (I removed some variables that might identify individual students) data is stored in the file 'PEERS data (anonymous).sav' (on Student Portal).

Once you have downloaded and opened their data, let's first look at their results:

21. Make a Crosstabs between the variables 'Version' and 'Verdict', and show the percentages of guilty verdicts per version (i.e. video vs. transcript). Here are the steps (you've done this before):




[1] The application and interpretation of t-test is largely similar, and for most purposes an ANOVA is more versatile for you to use.




**If you were on the jury in this case, would you vote that the defendant, is guilty or not guilty? * Did they see the video or transcript? Crosstabulation**

| | | | Did they see the video or transcript? | | |
|---|---|---|---|---|---|
| | | | Transcript | Video | Total |
| If you were on the jury in this case, would you vote that the defendant, is guilty or not guilty? | Not Guilty | Count | 23 | 32 | 55 |
| | | % within Did they see the video or transcript? | 36.5% | 60.4% | 47.4% |
| | Guilty | Count | 40 | 21 | 61 |
| | | % within Did they see the video or transcript? | 63.5% | 39.6% | 52.6% |
| Total | | Count | 63 | 53 | 116 |
| | | % within Did they see the video or transcript? | 100.0% | 100.0% | 100.0% |

Not in the assignment here, but if you had calculated a Cramer's V for this relationship, you would have gotten the following:

**Symmetric Measures**

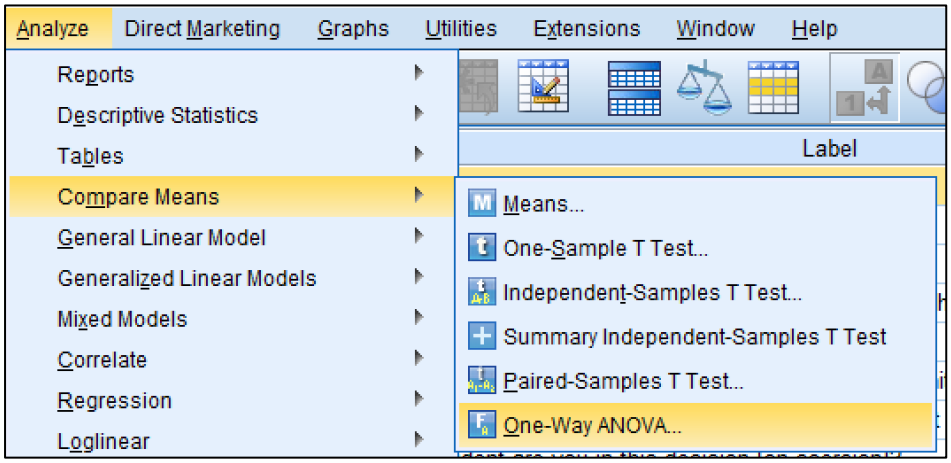
| | | Value | Approximate Significance |
|---|---|---|---|
| Nominal by Nominal | Phi | -.238 | .010 |
| | Cramer's V | .238 | .010 |
| N of Valid Cases | | 116 | |

This indicates a statistically significant relationship between the two variables at $p<0.05$ (5%).

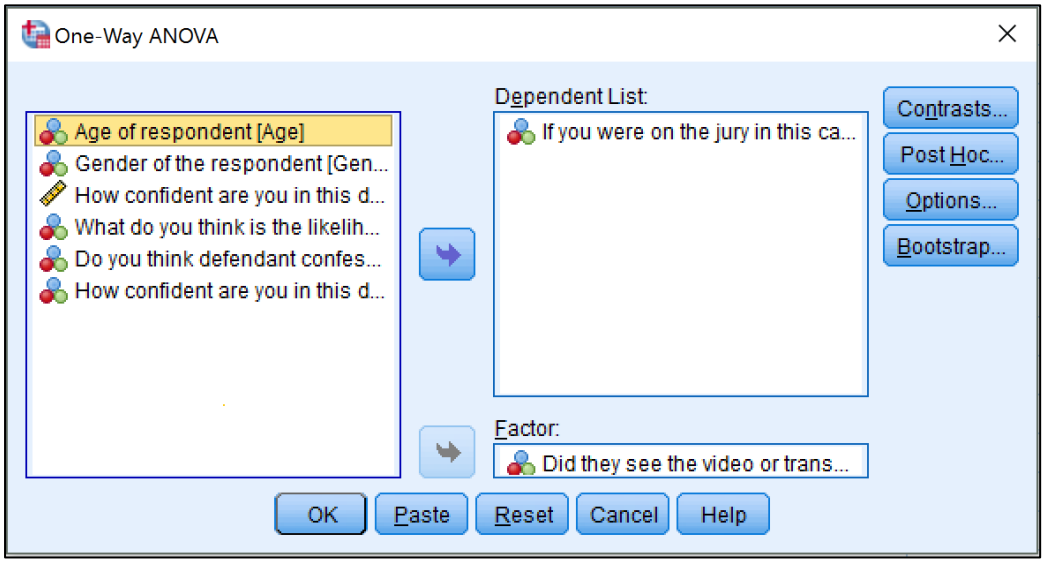
22. What is the percentage of people who decided that the defendant was guilty for those who saw the video? And for those who read the transcript?

39.6% of those who saw the video instead of the transcript gave the verdict 'guilty'. For those who only read the transcript, this was 63.5%.

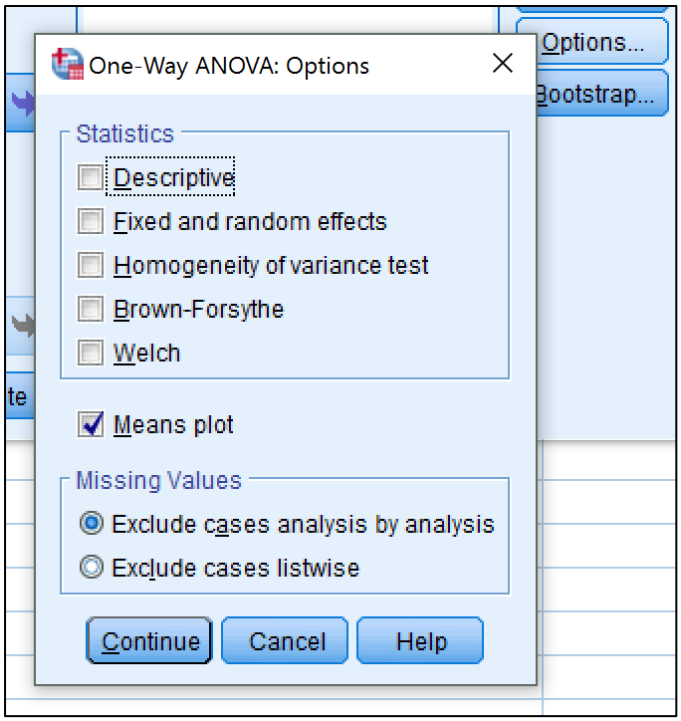
Next, let's execute our ANOVA analysis. You can find it under 'Analyze' > 'Compare Means' > 'One-Way ANOVA':




In the screen that pops up, you can set the grouping variable as the 'Factor', and the variable you want to compare the means *of* as the 'Dependent List':




Before you click 'OK', click 'Options', and select 'Means plot':

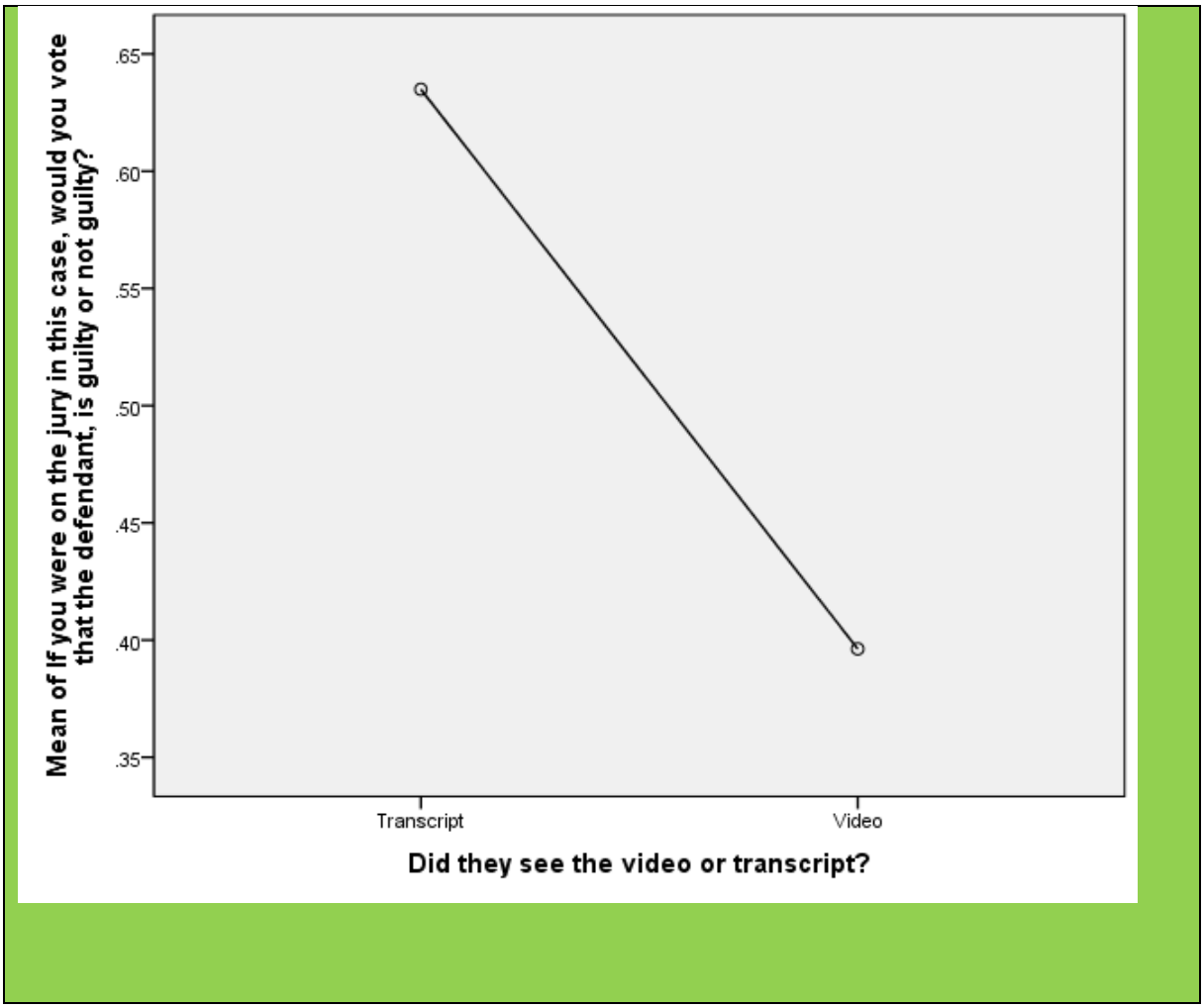



Click 'Continue' and then 'OK'. You now get a table and a plot.

**ANOVA**

If you were on the jury in this case, would you vote that the defendant, is guilty or not guilty?

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Between Groups | 1.640 | 1 | 1.640 | 6.853 | .010 |
| Within Groups | 27.282 | 114 | .239 | | |
| Total | 28.922 | 115 | | | |




23. What is the calculated p-value ('Sig.')?

p = 0.010

24. The p-value here represents the test whether the average scores between the groups are significantly different from one another. Would you conclude that the two groups (i.e. seeing the video vs. reading the transcript) are significantly different in terms of whether they thought the defendant was guilty?

Yes, because p = 0.01 < 0.05, so significant at a=5%. Not at 1%, though (that would be smaller than this value, and the actual p value if you double click the table is 0.010052, so slightly over p = 0.01).

25. The plot shows the average scores for the two groups again. Which group (video vs. transcript) has the higher number of people who thought the defendant was guilty (0=innocent, 1=guilty)? Think of this plot as a bar chart, if that helps.

Those who only read the transcript. They have an average of roughly 0.63 (i.e. 63% answered guilty), while the video group has an average of roughly 0.40 (i.e. 40% answered guilty).

26. What is your conclusion overall? What is the effect of seeing the video confession versus just reading the transcript?

Seeing the video confession as compared to reading the confession transcript makes students statistically significantly less inclined to pass a guilty verdict on the defendant.