An SPSS companion book

to

Basic Practice of Statistics – 6th Edition.

Type text here

SPSS is owned by IBM.

Basic Practice of Statistics 6th Edition by David S. Moore, William I. Notz, Michael A. Flinger. Published by W.H. Freeman and Company.

Companion Book by **Michael "Jack" Davis** of Simon Fraser University, jackd@sfu.ca, factotumjack.blogspot.ca last updated 2015 January 3.

About SPSS

SPSS stands for **S**tatistical **P**ackage for **S**ocial **S**ciences. It was briefly called PASW, so you may also see that acronym tossed around.

It's a menu-based system for graphing and analyzing data. Having some experience which a spreadsheet program like Excel will be of some help.

Having experience with another menu-based statistical program like JMP or Statcrunch will help a lot.

SPSS versions are updated often. As of January 2015, the newest version was SPSS 23. This guide is based on SPSS 19.

However, basic usage changes very little from version to version. Many of instructions for SPSS 19-23 are the same as they were in SPSS 11.

SPSS is owned by IBM, and they offer tech support and a certification program which could be useful if you end up using SPSS often after this class.

http://www-03.ibm.com/certify/certs/47100101.shtml

Some datasets used in this guide are available at http://www.sfu.ca/~jackd/SPSS/Datasets/

The datasets for problems and examples in Basic Practice of Statistics are available at http://content.bfwpub.com/webroot_pubcontent/Content/BCS_5/BPS6e/Student/DataSets/SPSS/SPSS.zip

Knowing how to use SPSS is not the same as knowing statistics. It's becoming increasingly important to know what the most appropriate tools and analyses are for a given situation rather than rote memorization. Interpretation of terms, such as 'p-value' is also important, but is covered in your textbook instead of this guide.

Inputting Data

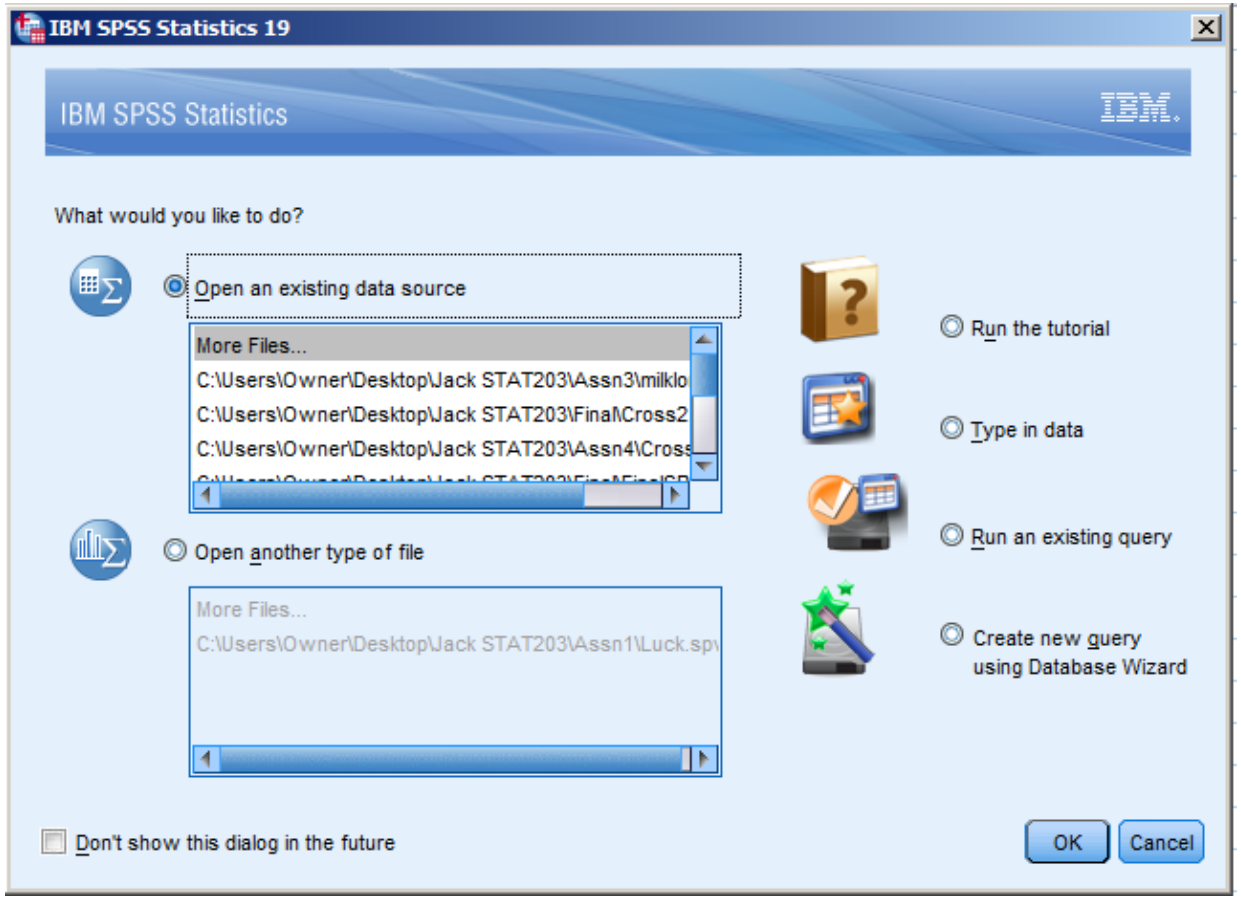
There are a few ways to input data in SPSS. The simplest way to input data is from its own format, the ***.sav file***.

Sometimes data doesn't come in the **.sav** format. Data can come from another program like Excel using the **.xls**, or **.xlsx** formats, It can come as a multi-program portable file in the **.por** format, or as text in the **.txt** or **.csv** formats.

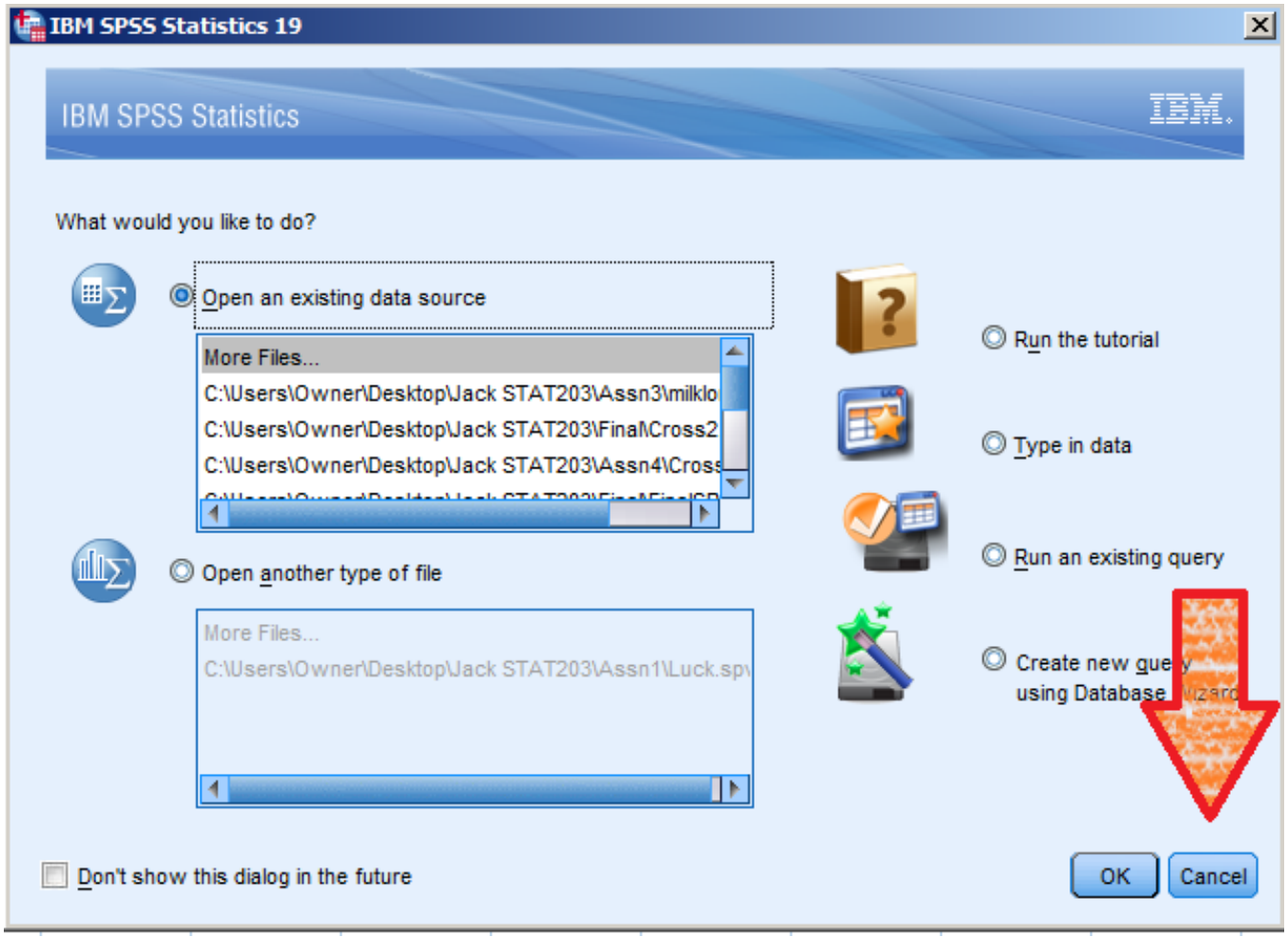
Basic Practice of Statistics datasets are in the .por format.

Inputting data in SPSS manually isn't ideal, but sometimes it needs to be done, so that is covered here as well.

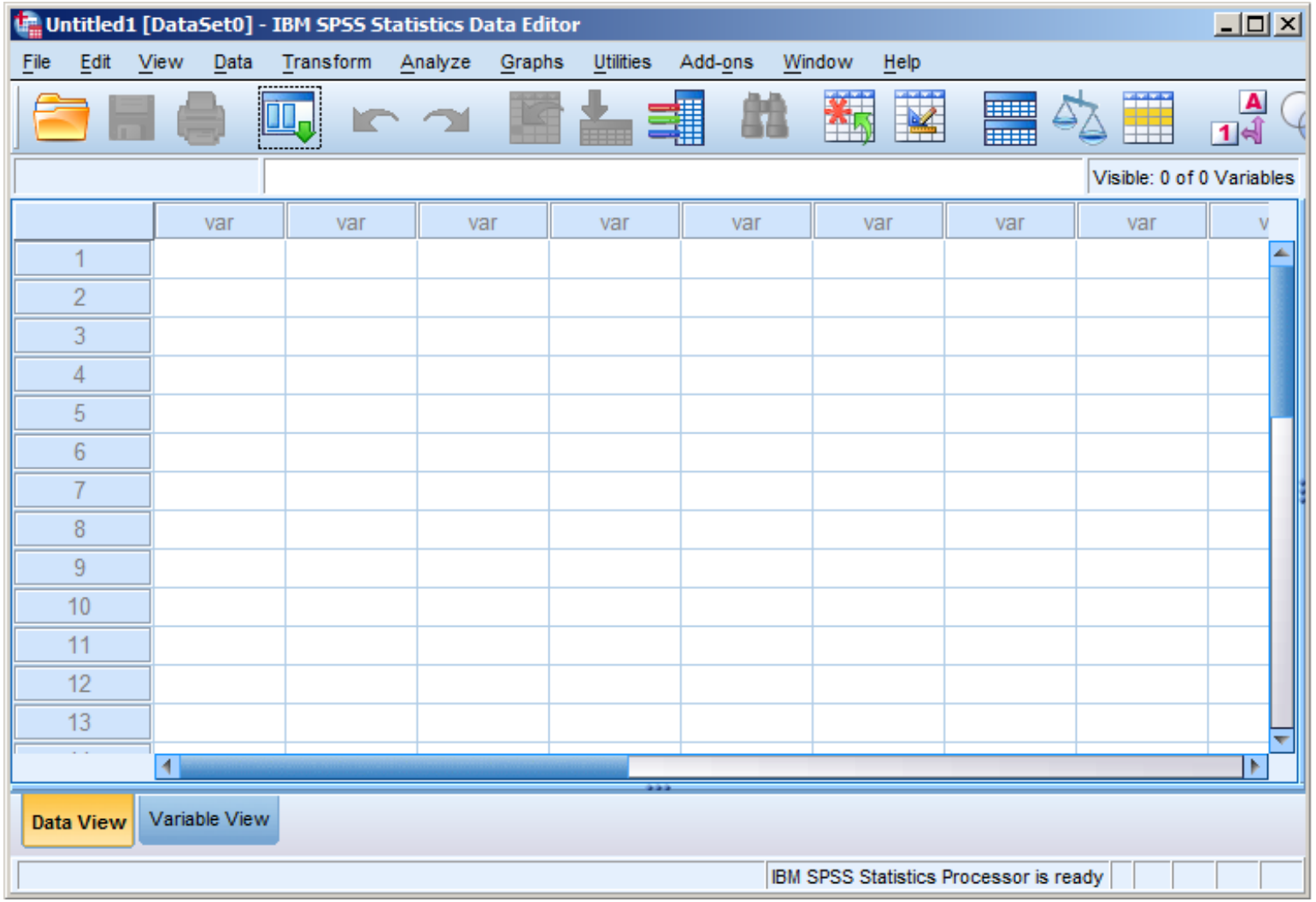
To input (import) data from a .sav file, first open SPSS. You'll first get a dialog like this:



This dialog can get your started quickly, but we're assuming SPSS is already running in the examples in this guide, so click ***Cancel*** in the lower right.

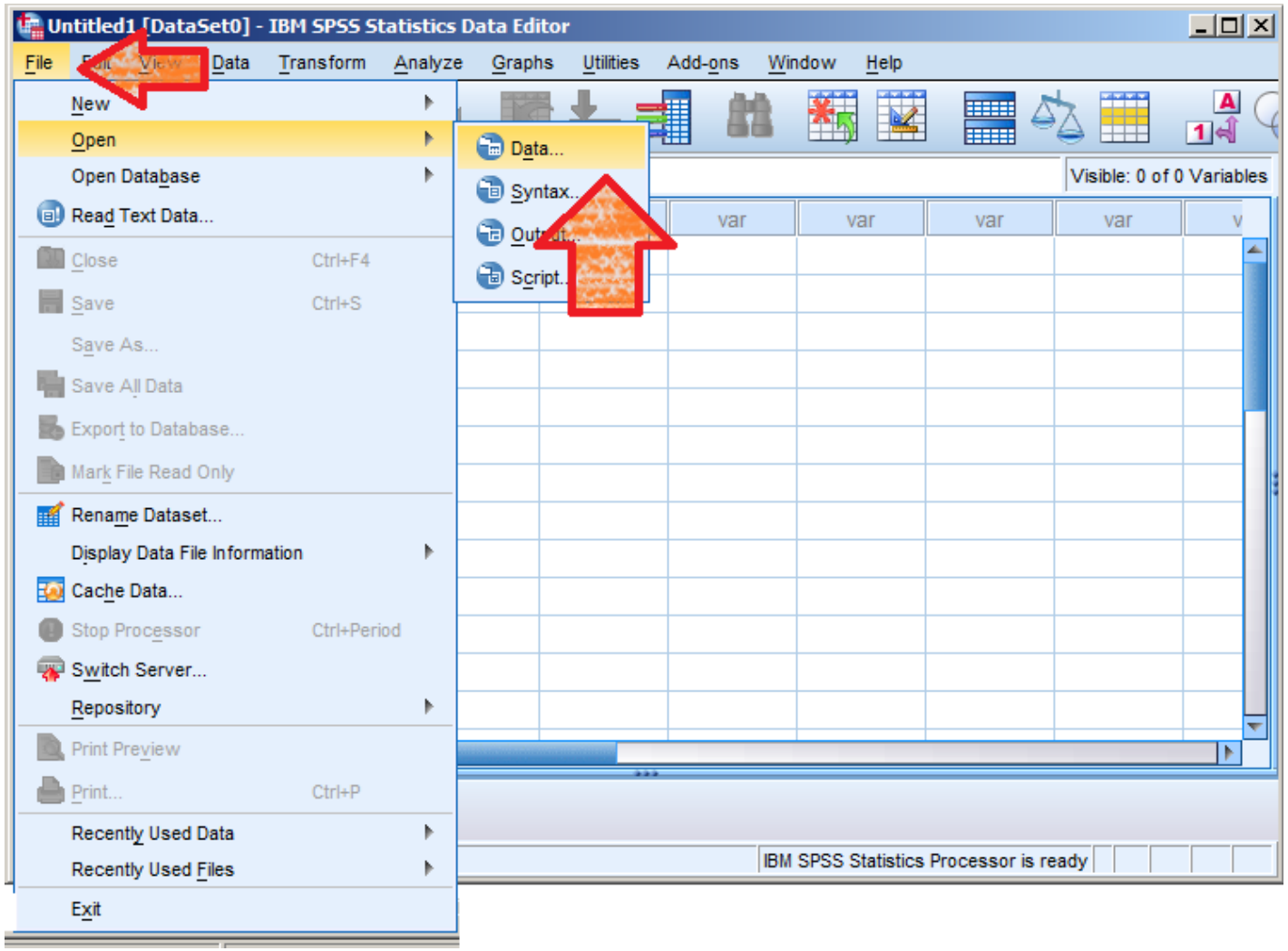


You'll get a window that looks like this. This is the ***data view*** window. Sometimes called the ***main window***

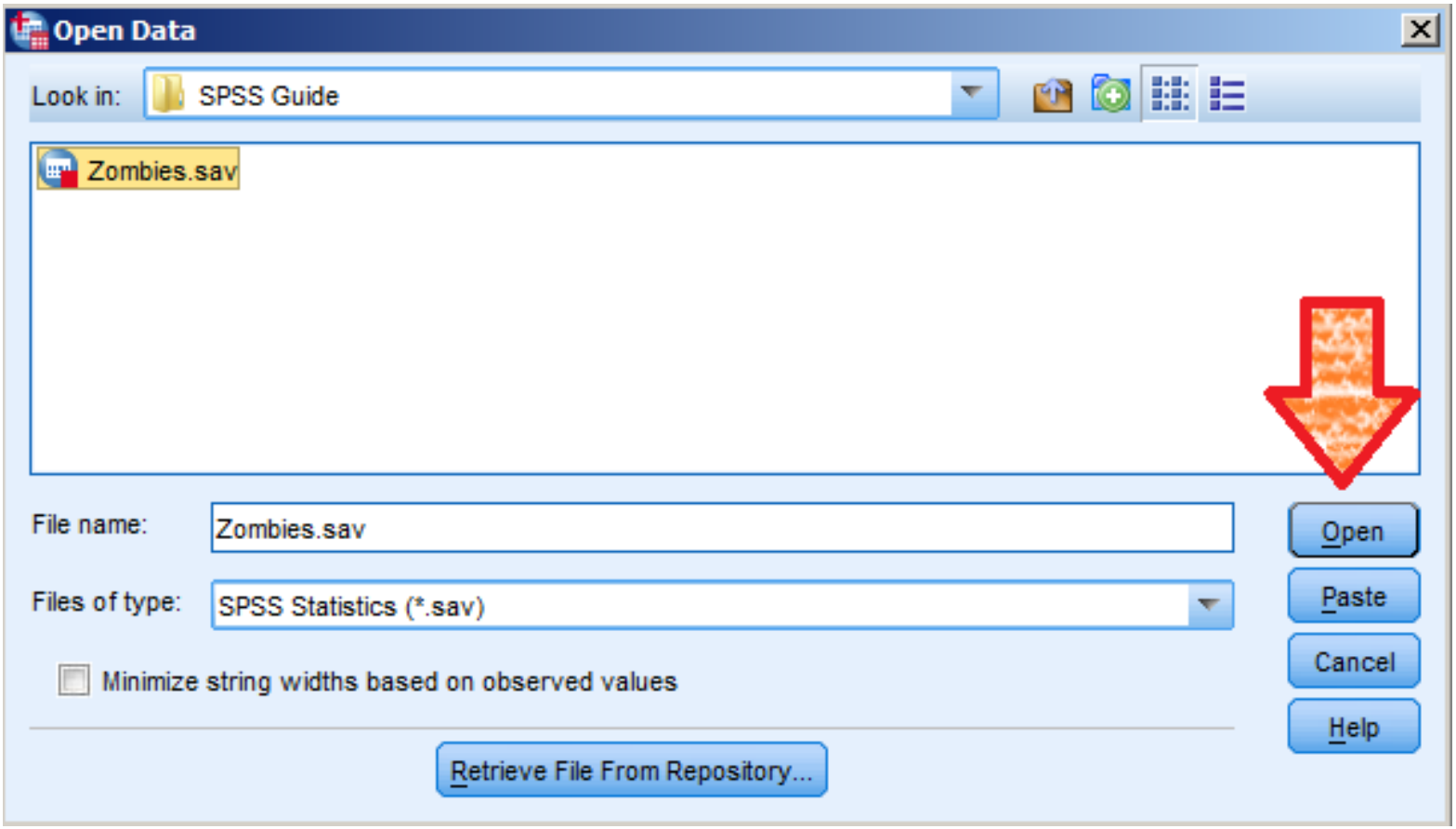


To open a file, click on the  icon in the upper left, or use ***file***

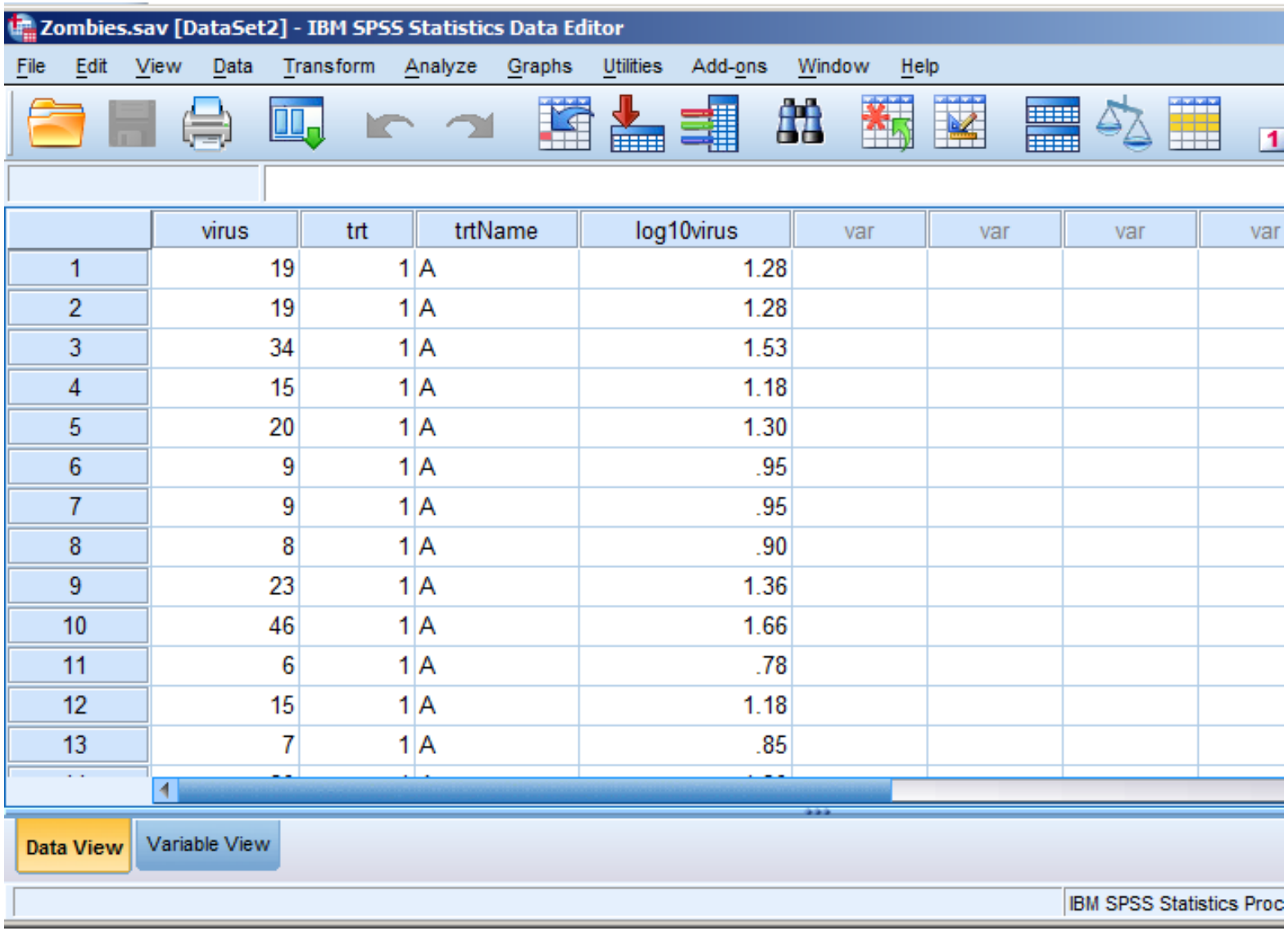
→open →data, also in the upper left.



Then, in the navigation dialog, navigate the folder containing the file you want and click ***Open.***



If you've done it correctly, the data should fill in the cells of the main window.

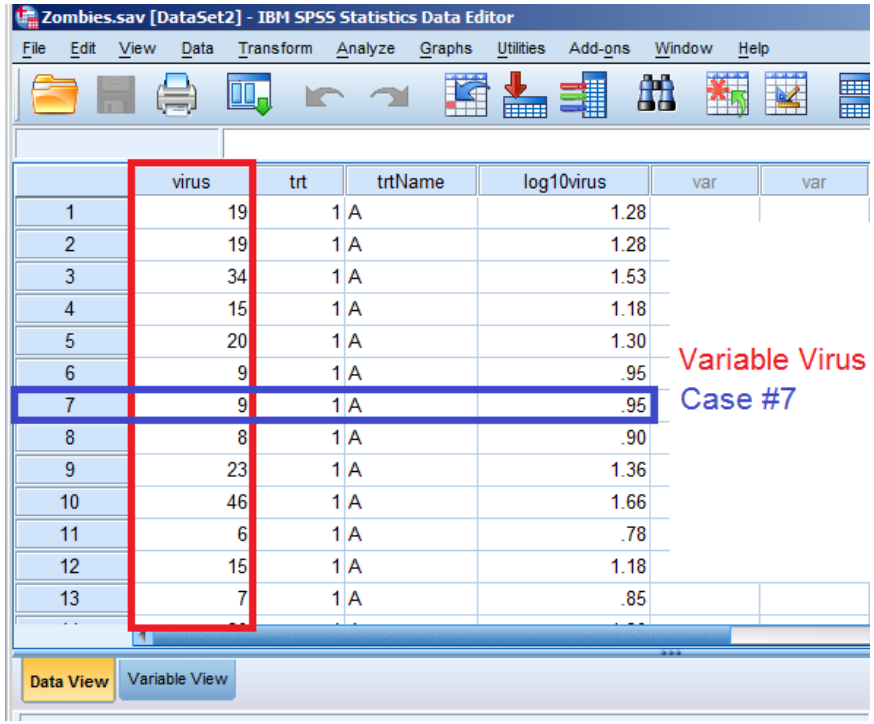


	virus	trt	trtName	log10virus	var	var	var	var
1	19	1	A	1.28				
2	19	1	A	1.28				
3	34	1	A	1.53				
4	15	1	A	1.18				
5	20	1	A	1.30				
6	9	1	A	.95				
7	9	1	A	.95				
8	8	1	A	.90				
9	23	1	A	1.36				
10	46	1	A	1.66				
11	6	1	A	.78				
12	15	1	A	1.18				
13	7	1	A	.85				

Additional notes

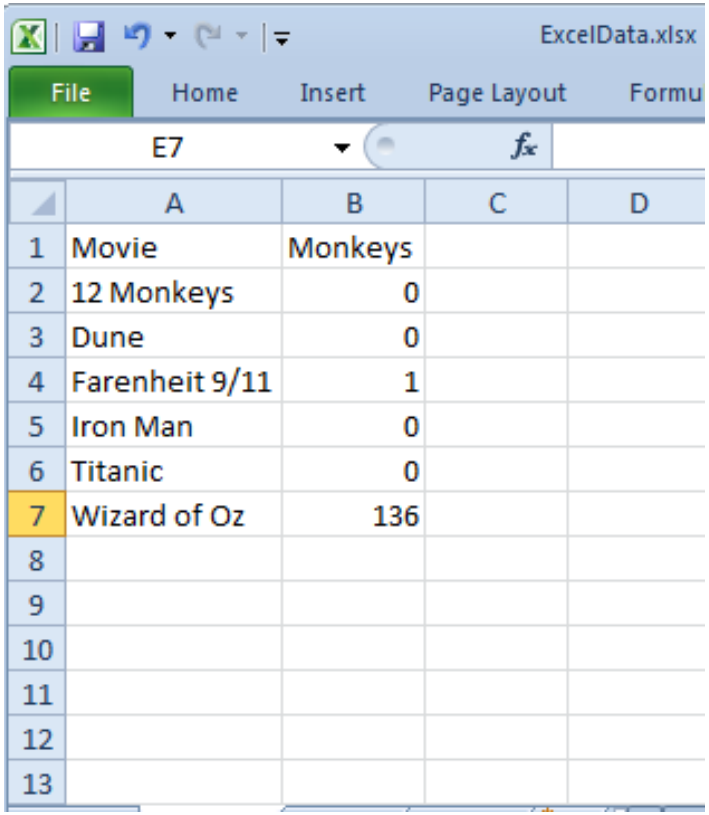
Each column is a ***variable,*** a feature of the data as a whole (e.g. how many viruses, what treatment has been given).

Each row is a ***case,*** which is all the information that was sampled from a single person/city/petri dish/unit.



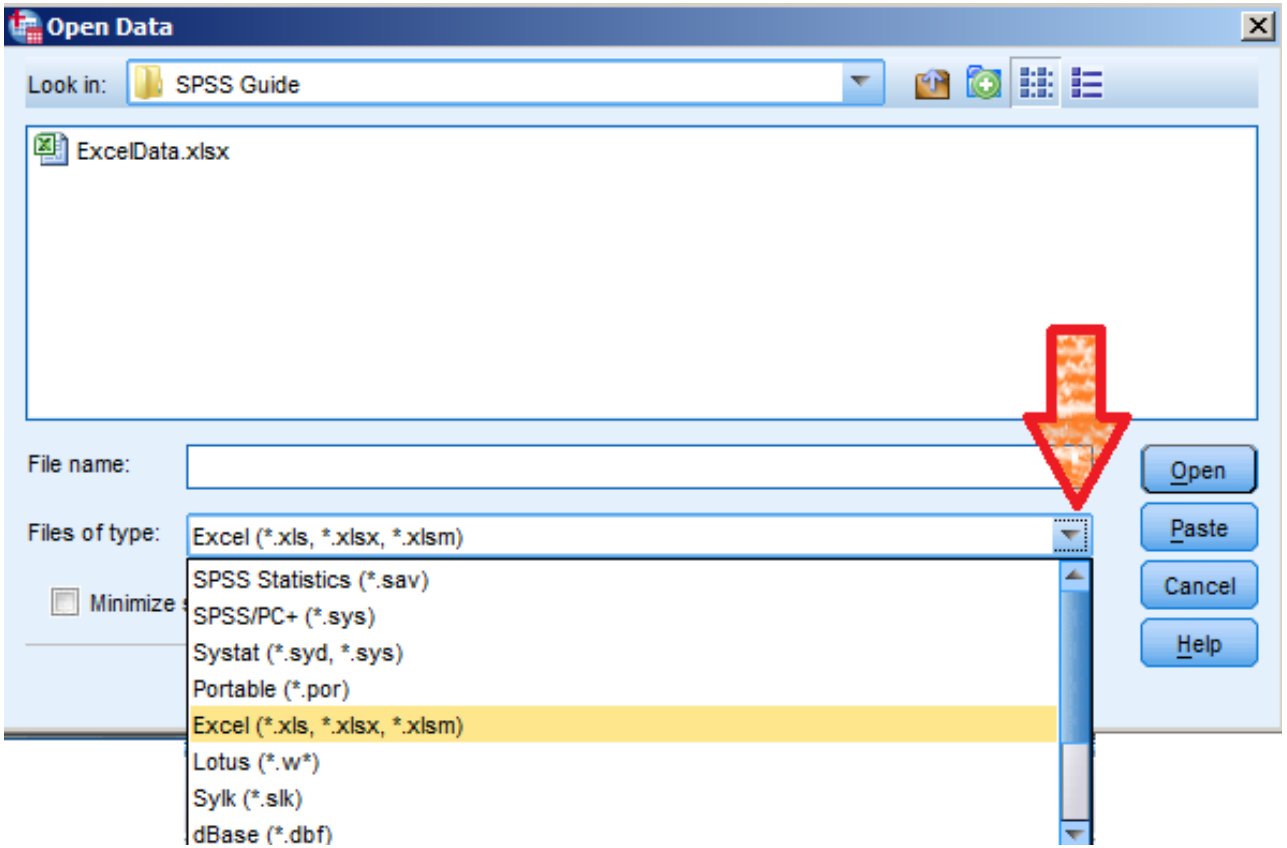
If you need to import something from Excel, first, open the data file in Microsoft Excel (screen below of MS Excel 2010) to see if the variable names are included in the first line.

It's recommended that data from Excel follow the same case/variable setup as the .sav data, just like this screen.

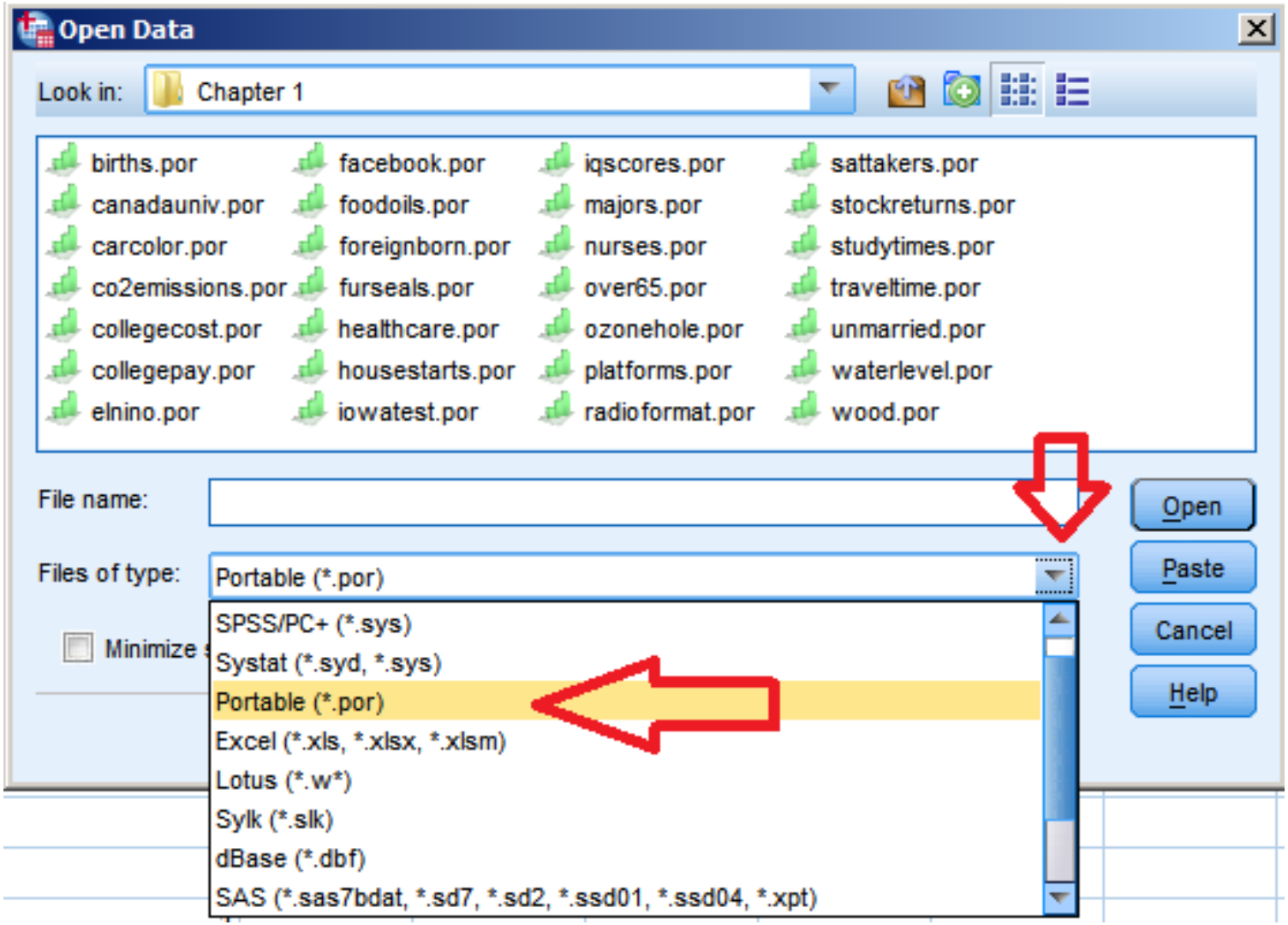


	A	B	C	D
1	Movie	Monkeys		
2	12 Monkeys	0		
3	Dune	0		
4	Farenheit 9/11	1		
5	Iron Man	0		
6	Titanic	0		
7	Wizard of Oz	136		
8				
9				
10				
11				
12				
13				

Then, in SPSS (screen from SPSS 19, use the  icon or File → Open → Data again, but this time in the dialog, change the ***Files of type*** pulldown to Excel before selecting the file you want and clicking Open. (Excel files won't appear otherwise)



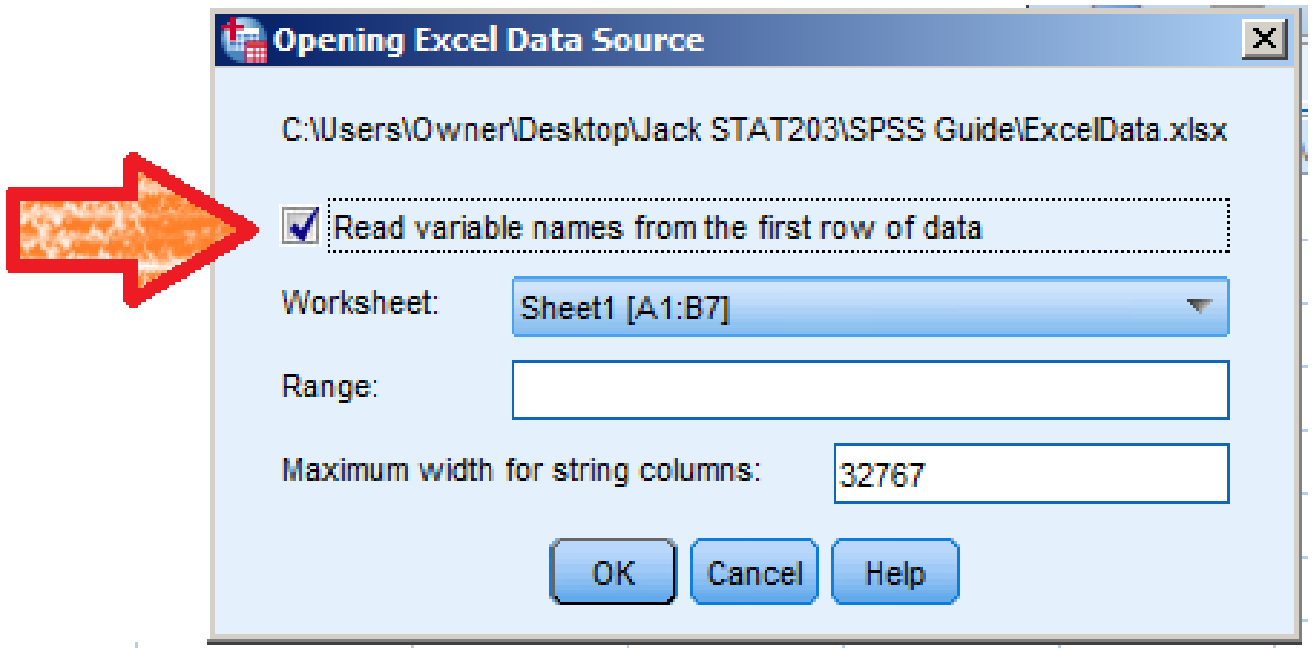
When loading **.por** datasets, make sure the file type ***Portable*** is set.



If the variable names were in the first row, make sure this box (see arrow) is checked. Otherwise, leave it unchecked.

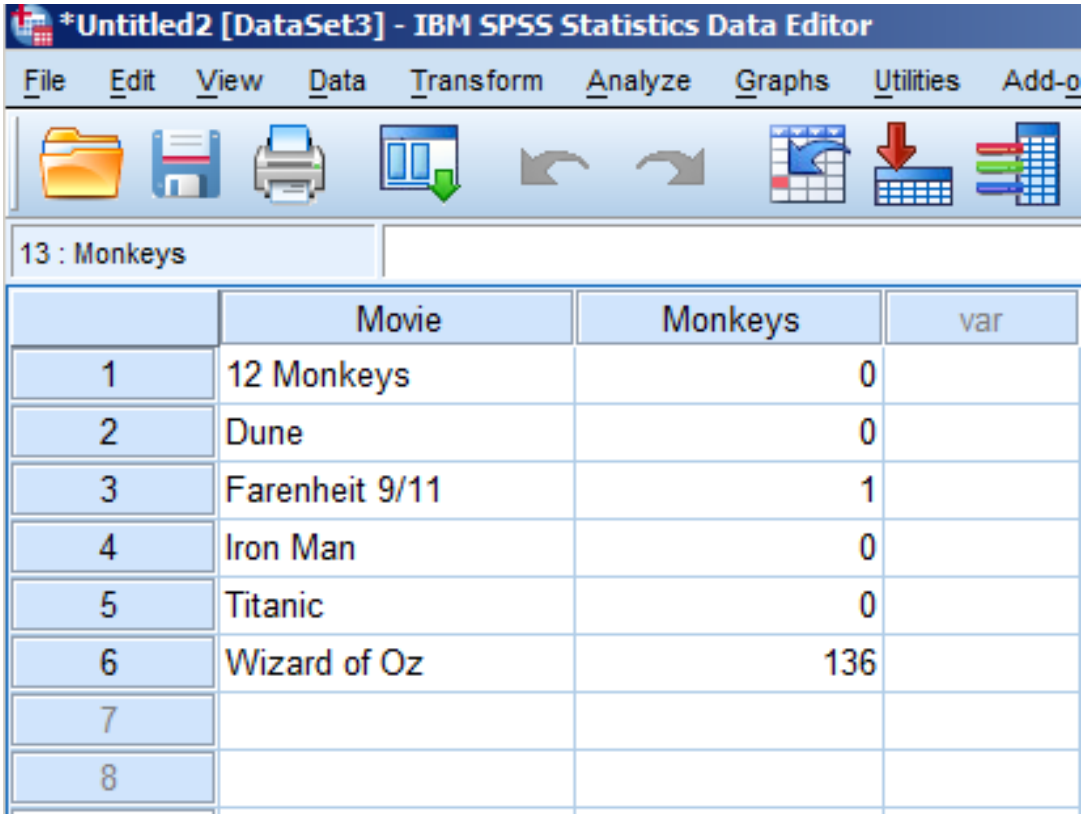
If your excel file has multiple sheets, use Worksheets to make sure you have the right one (by default it will usually be right)

Then Click OK



If everything was done right, your SPSS main window should look like the screen below. “Movie” and “Monkeys” have been interpreted as variables and not part of the data.

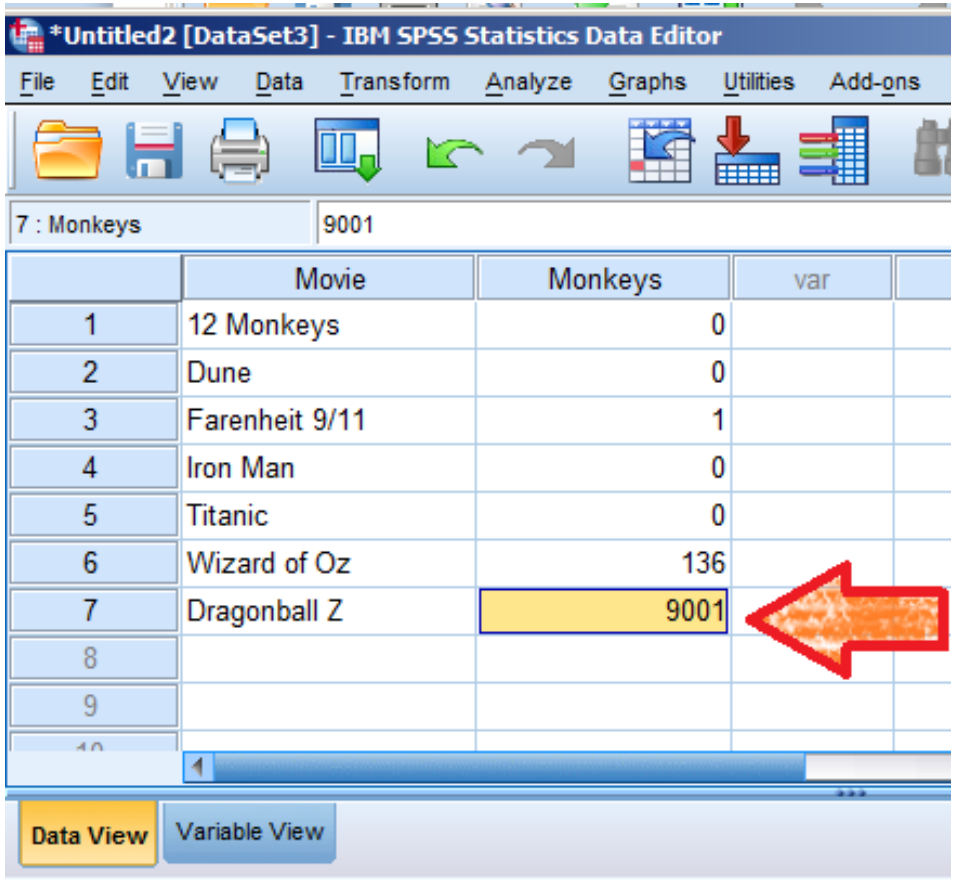
Note that the filename is Untitled; SPSS doesn’t open the Excel file, it makes a copy. Changes in SPSS won’t affect the original Excel file.



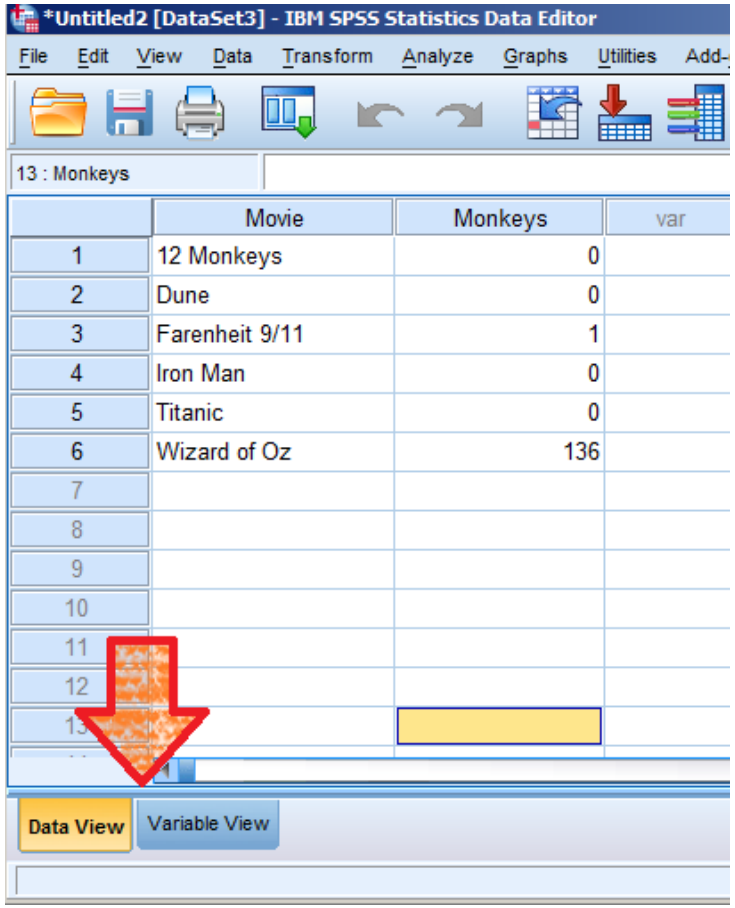
	Movie	Monkeys	var
1	12 Monkeys	0	
2	Dune	0	
3	Farenheit 9/11	1	
4	Iron Man	0	
5	Titanic	0	
6	Wizard of Oz	136	
7			
8			

You can also click on cells (where a column and row intersect) and type in new cases if you need to. “Dragonball Z” and “9001” have been typed in.

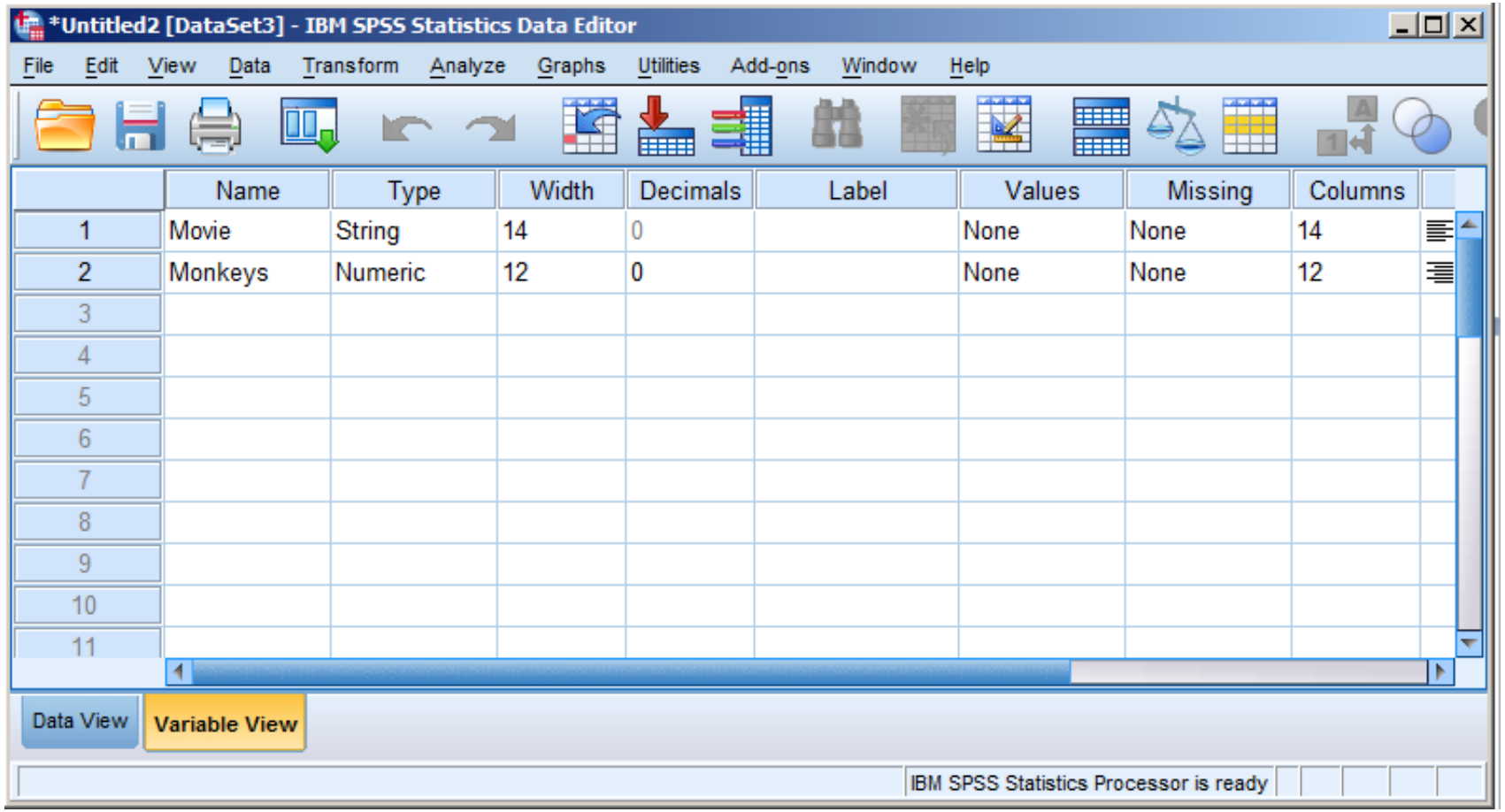
SPSS won’t let you type letters into a numeric variable.



Finally, note the two tabs in the bottom left of the main window.



We’re currently in Data View, but clicking the Variable View tab will bring up this:

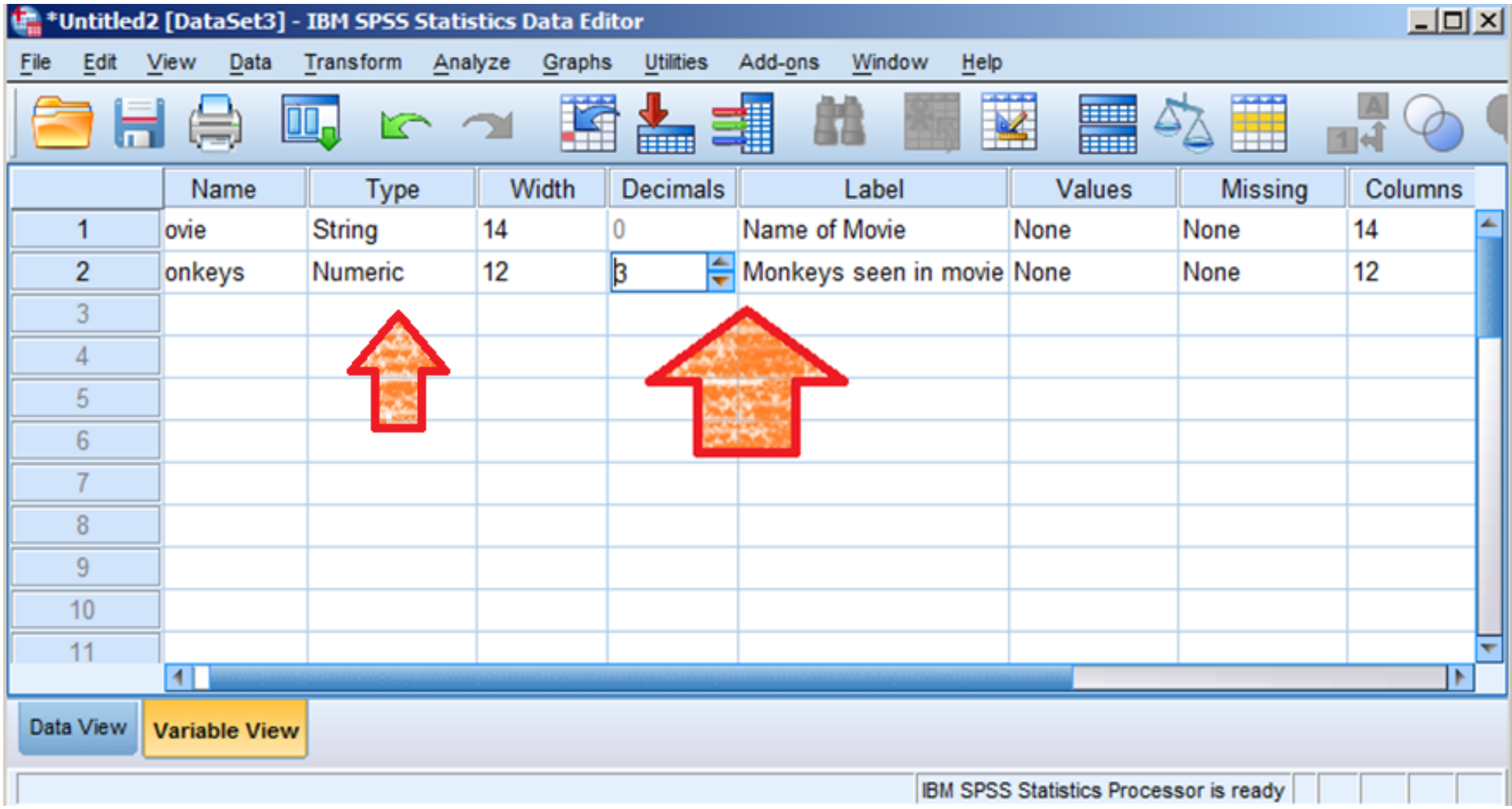


Instead of displaying the data, Variable View displays information about each of the variables.

If you want to see more/fewer decimals, you can click on the appropriate cell to change it.

You can change the type (***String*** is words, ***Numeric*** is numbers)

You can also give variables more descriptive names in the labels.



Transforming and Sorting Data

This part may or may not be part of your course. It's included here because it's useful knowledge for managing data in general, and because it helps solve an example in the crosstabs section.

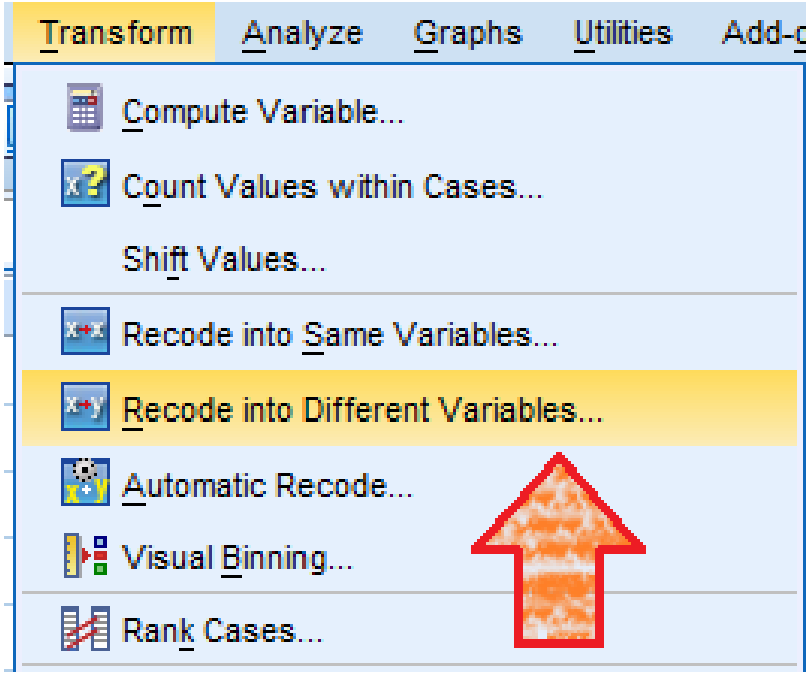
Here we take a nominal variable with three categories and transform it into a nominal variable with two categories, effectively merging two categories together.

(From the dataset Ch9_24.sav, based on Exercise 9.24)

To take the three category variable Young/Middle/Old

And make a two category variable Young/Not Young

Transform → Recode into Different Variables.

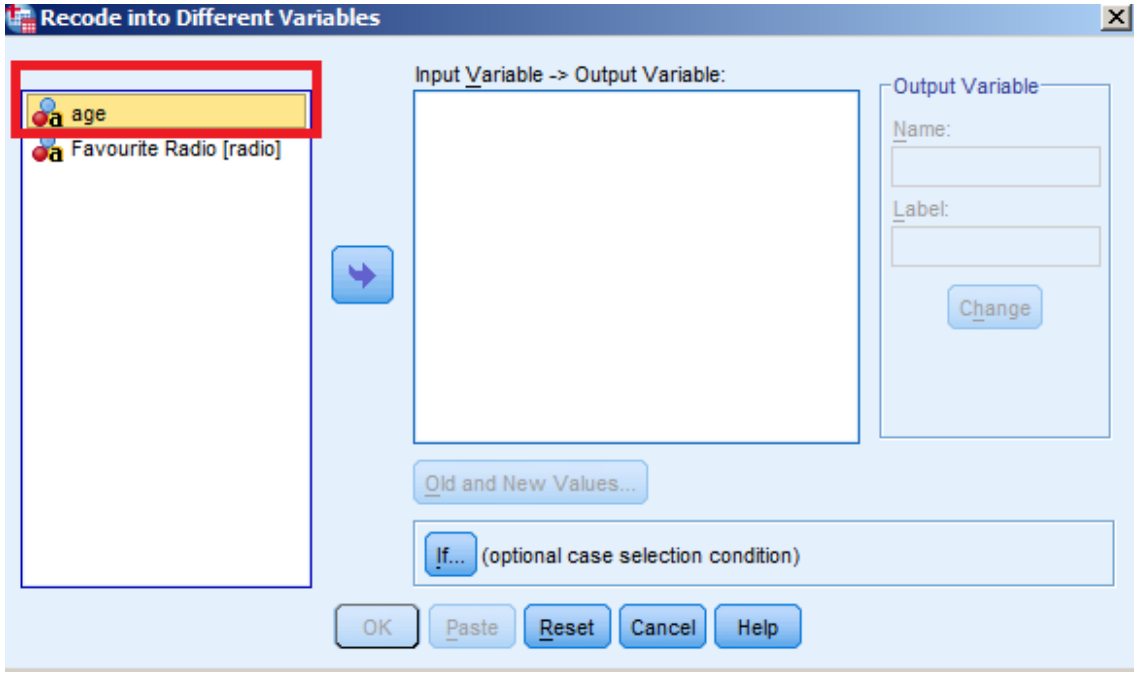


Select the variable you want to change (age) and drag it into ***"String Variable → Output Variable"***

Give the new variable a name in ***Output Variable: Name***,

Then click on ***Change.***

Before Dragging…

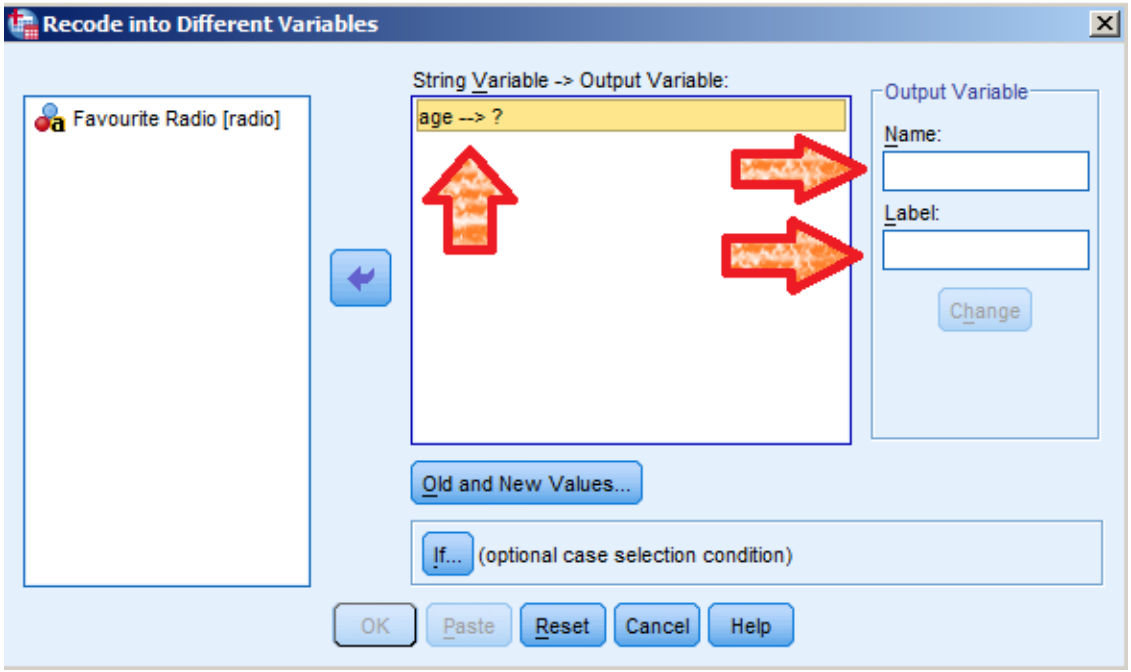


Select the variable you want to change (age) and drag it into ***"String Variable → Output Variable"***

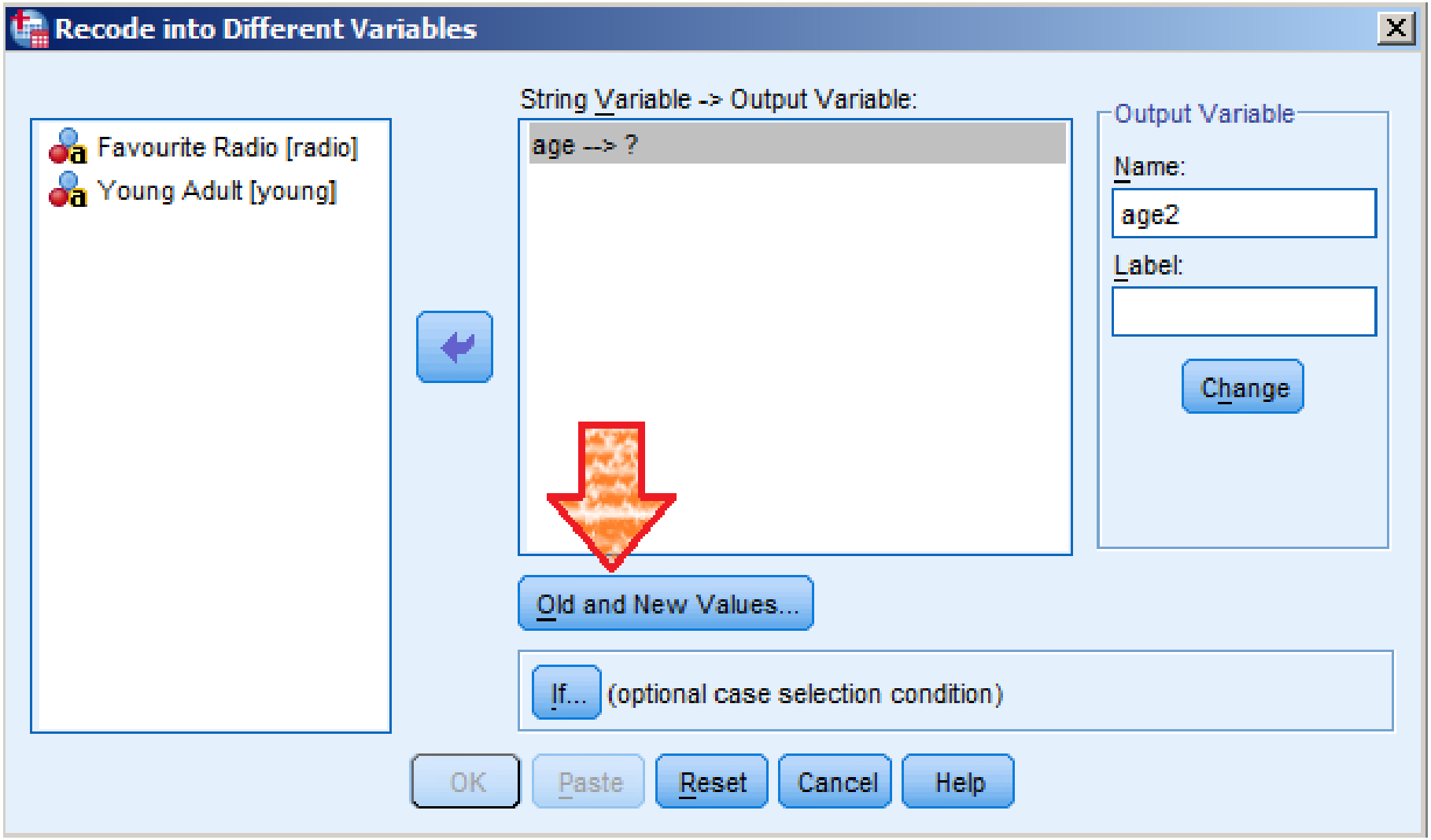
Give the new variable a name in ***Output Variable: Name***,

Then click on ***Change.***

After Dragging…

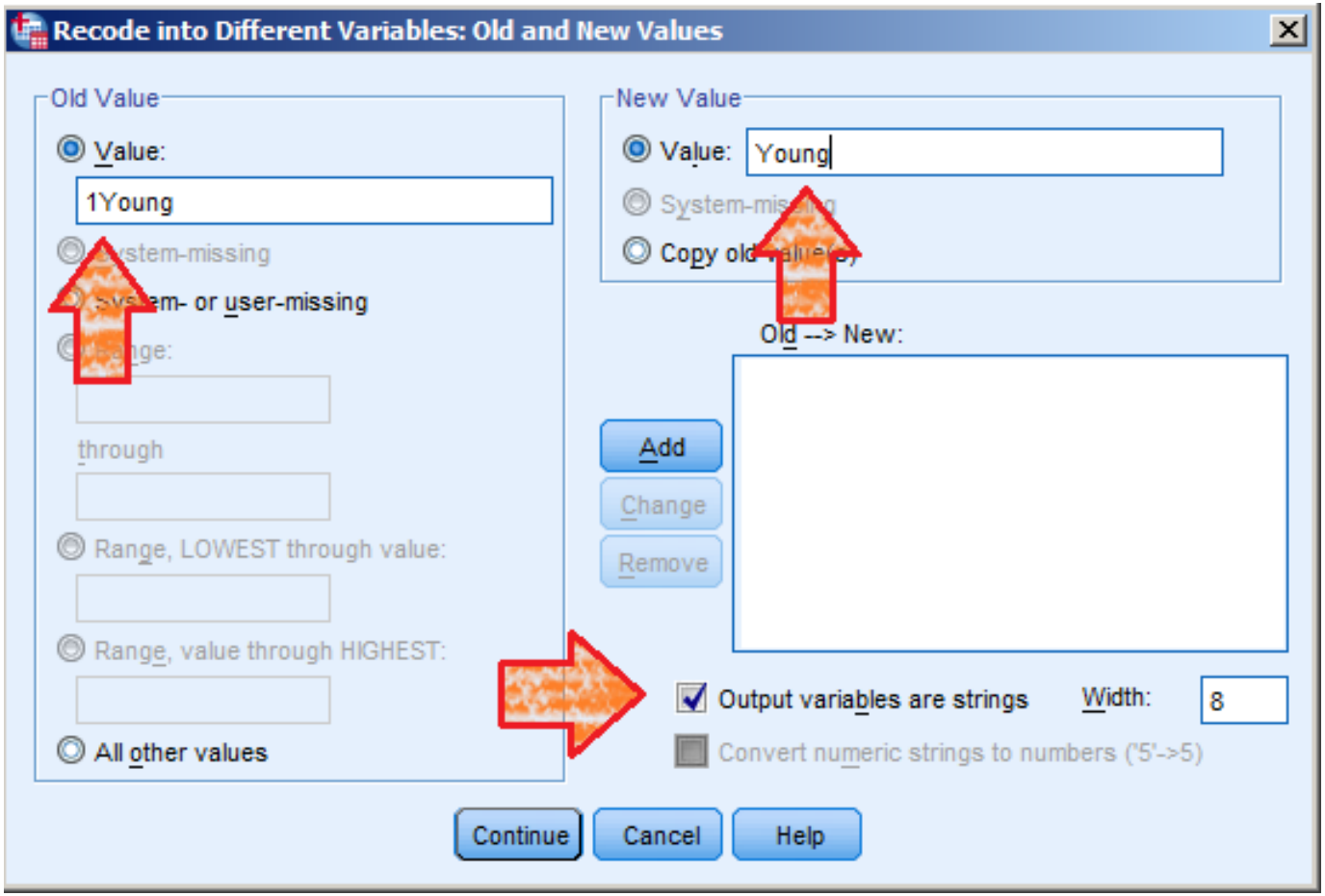


Then, click on ***Old and New Values.***



This brings up the menu to define the old categories you have and the new categories that you want.

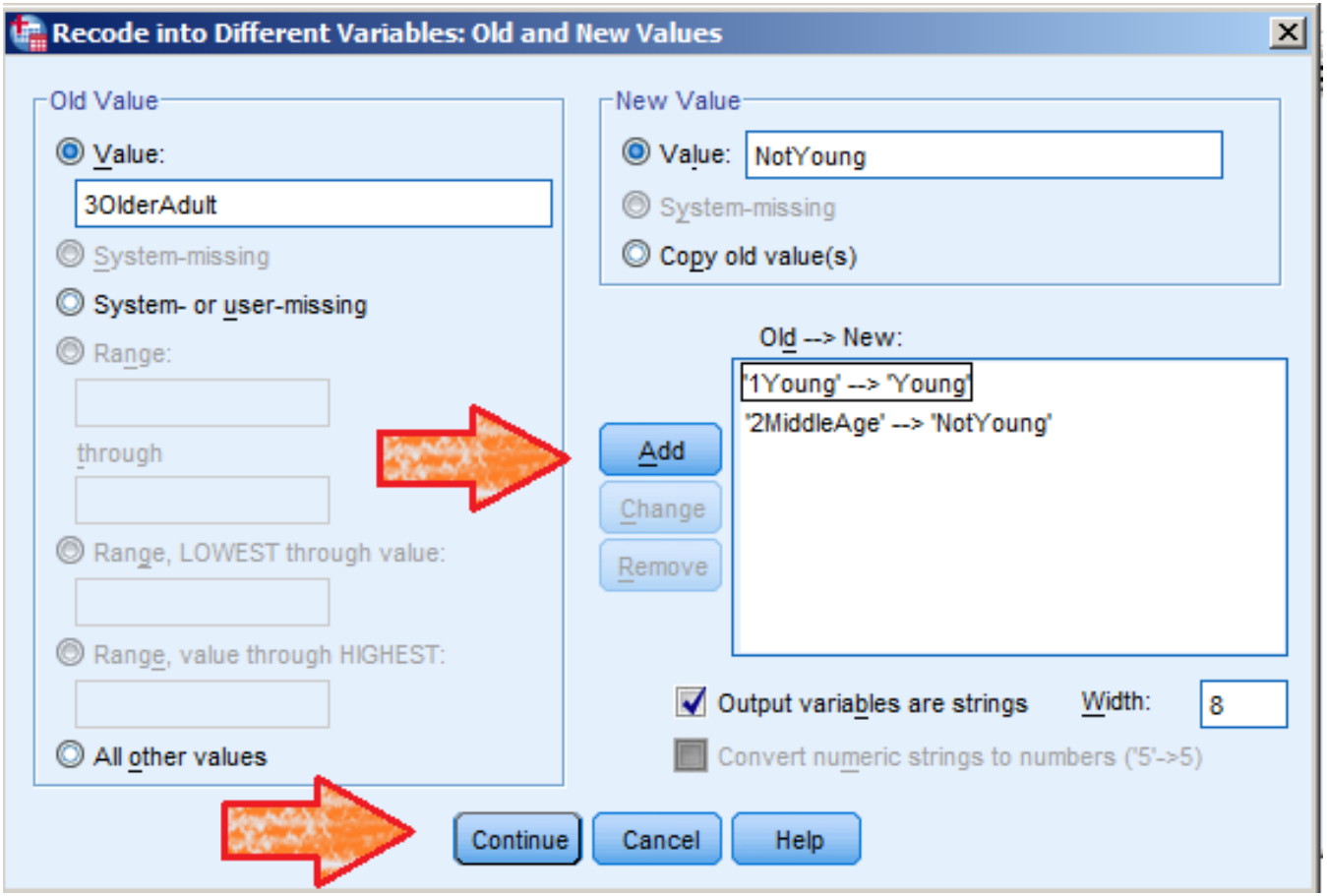
In the new dialog, check ***Output variables are strings*** first



Then enter the old category name in ***Old Value:*** ***Value***

And enter the new category name in ***New Value*****: Value**

Click ***Add*** and repeat the previous slide for each category.

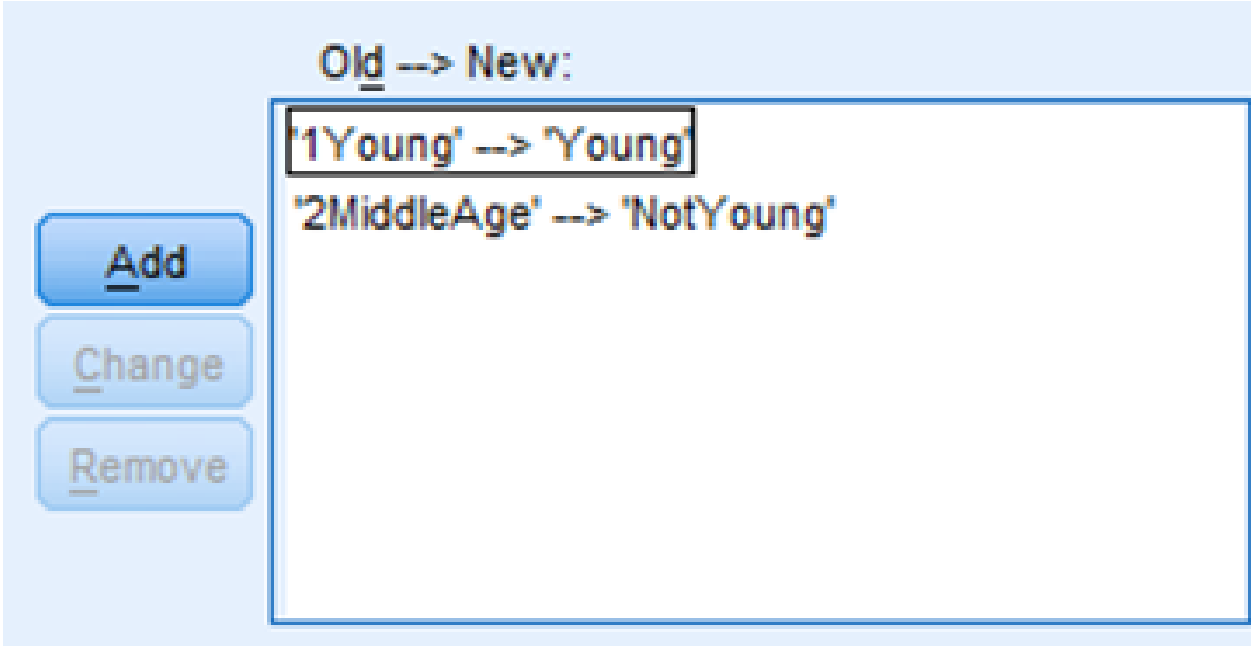


Repeat until the Old → New box reads:

‘1Young’ → Young’,

‘2MiddleAge’ → ‘NotYoung’,

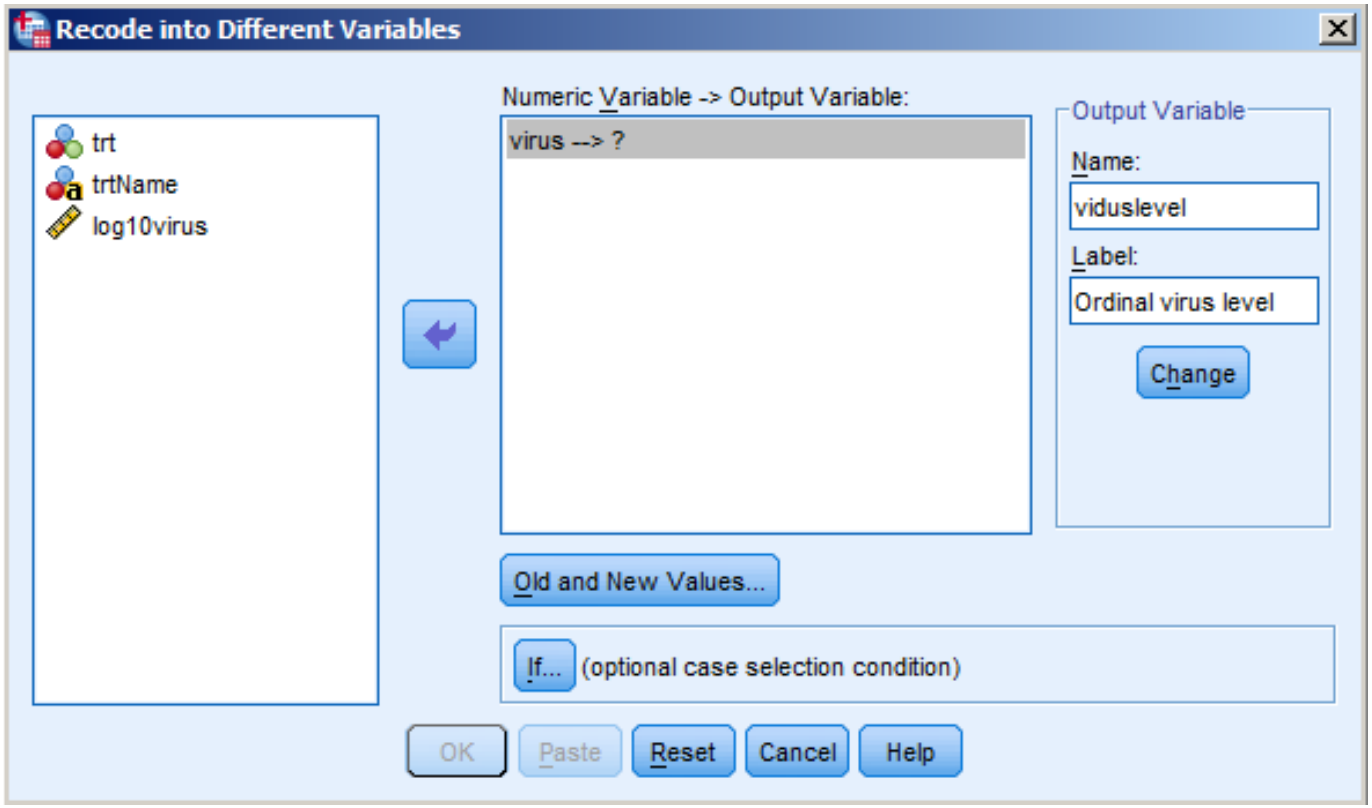
‘3OlderAdult’ → ‘NotYoung’.



Transforming – Example 2

You can also transform data from a continuous/interval format into a categorical, ordinal format. This is still done under

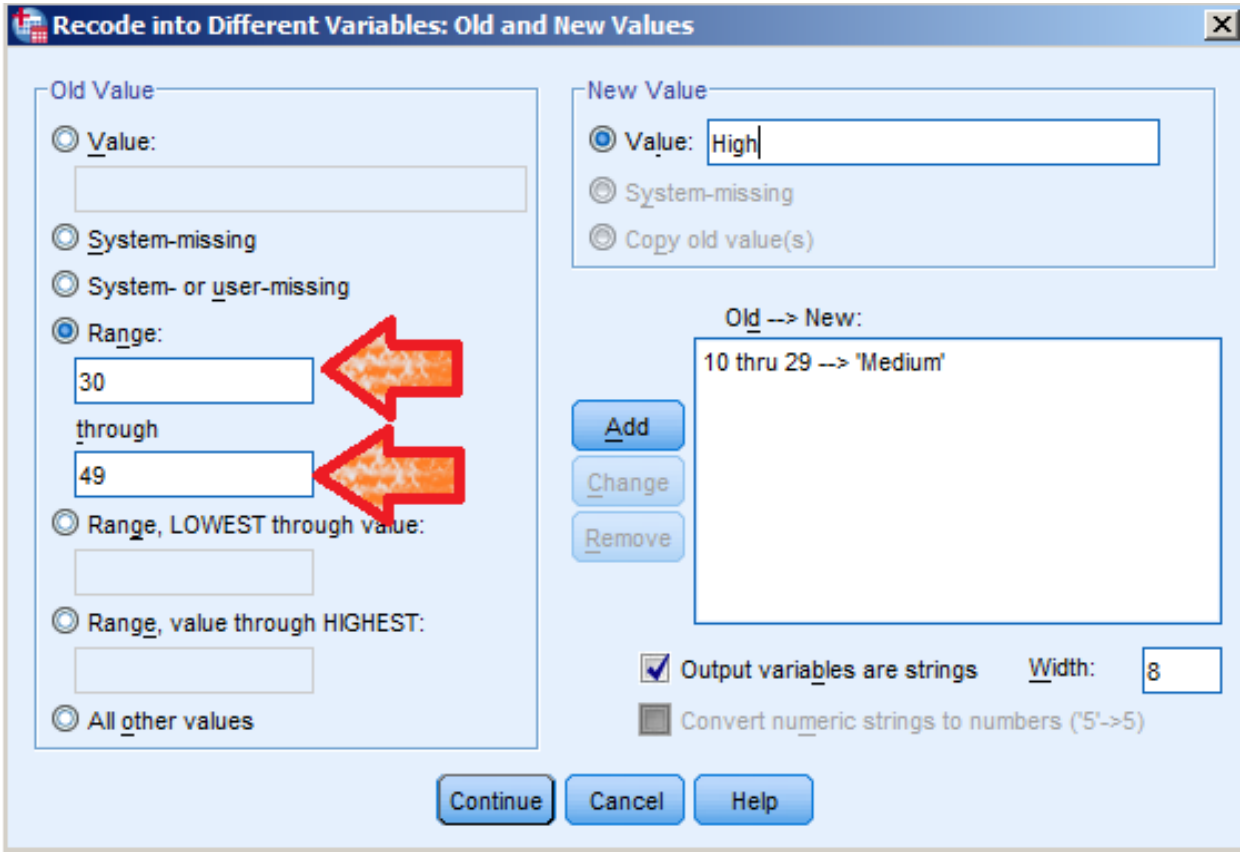
Transform → Recode into different variables



You can set a range of values to all be transformed into a single value using the ***Range*** and ***through*** options.

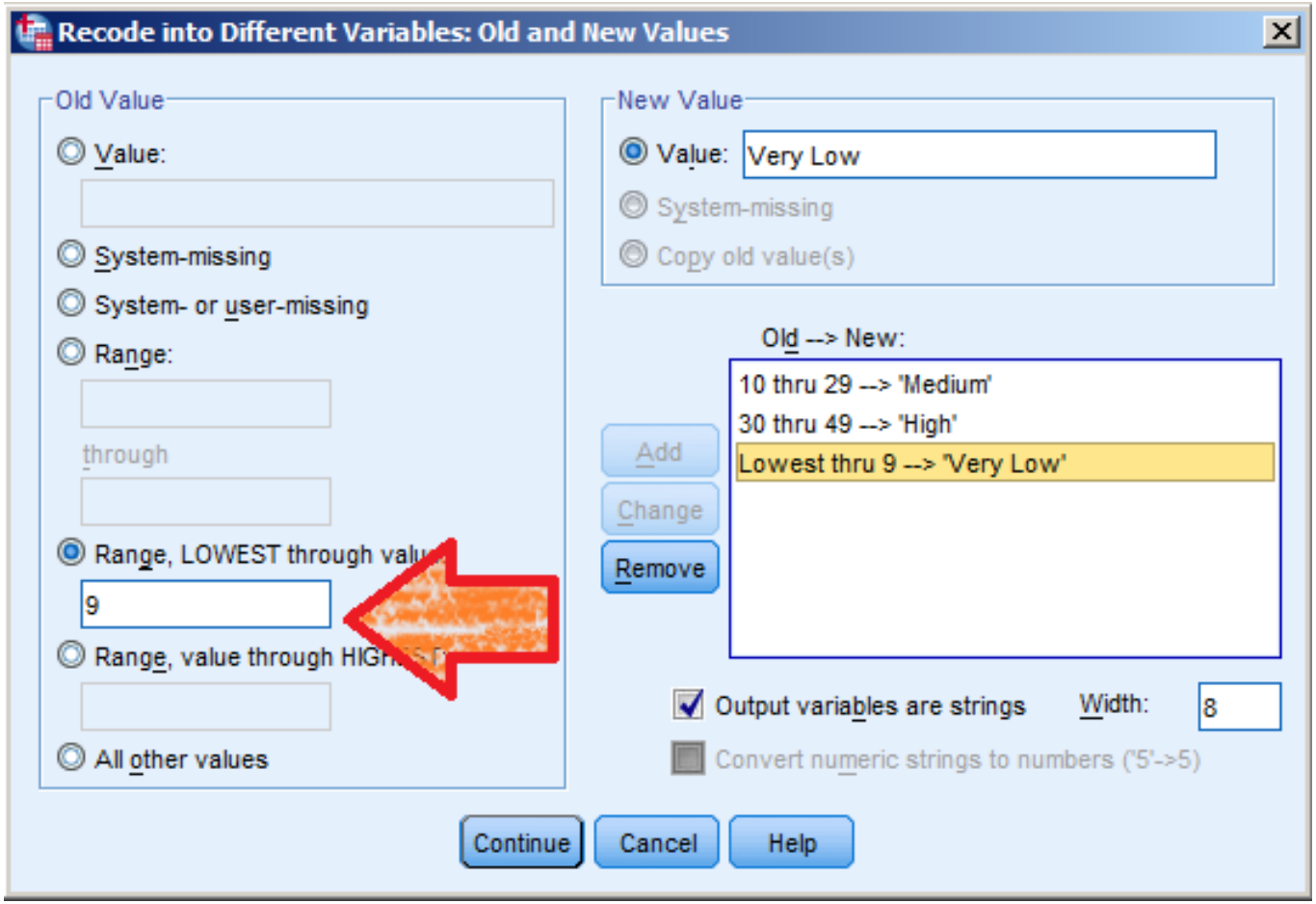
Here, all the values from 10-29 are being coded as “Medium”

And the values 30-49 are about to be coded as “High”.



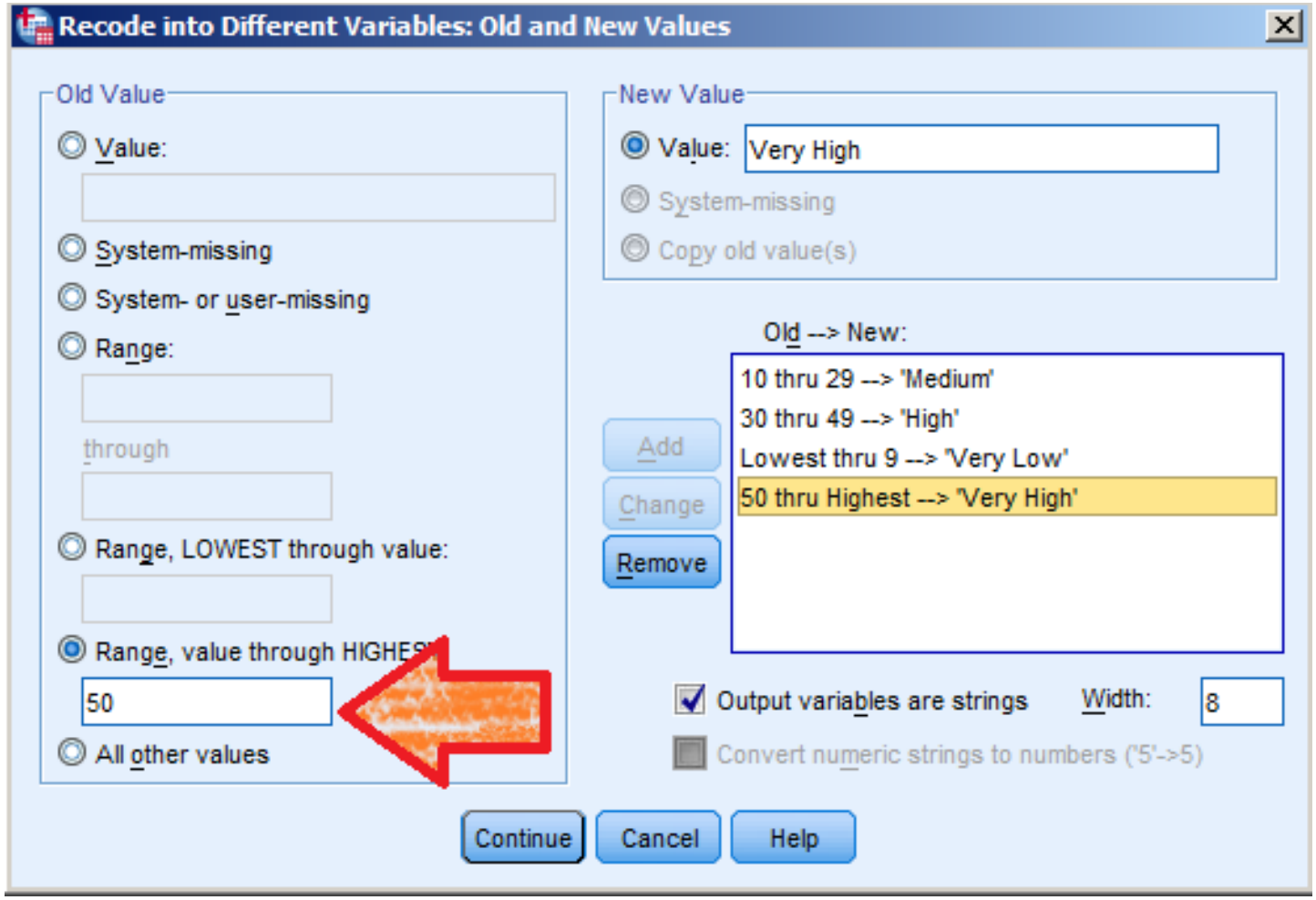
If you want “anything X or lower” to be coded together, use the ***LOWEST through value***

Here, anything of 9 or less is being coded as “Low”.



Likewise, you can have a code for “X or higher” with the ***value through HIGHEST*** option.

Here, anything 50 or higher is being coded as “very high”



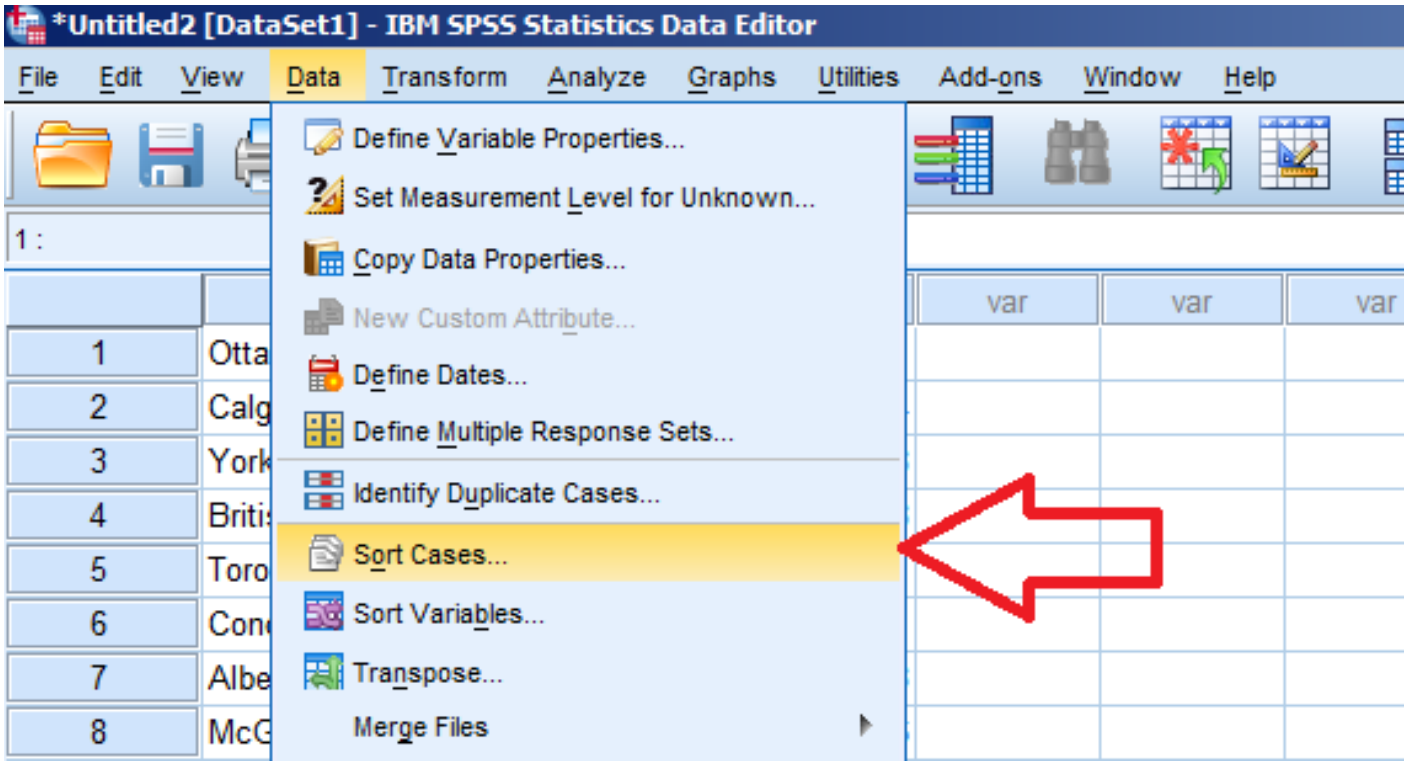
Now the numerical virus counts have been turned into categories.

virus	viruslevel
19	Medium
19	Medium
34	High
15	Medium
20	Medium
9	Very Low
9	Very Low
8	Very Low
23	Medium
46	High
6	Very Low
15	Medium

Sorting by a Variable

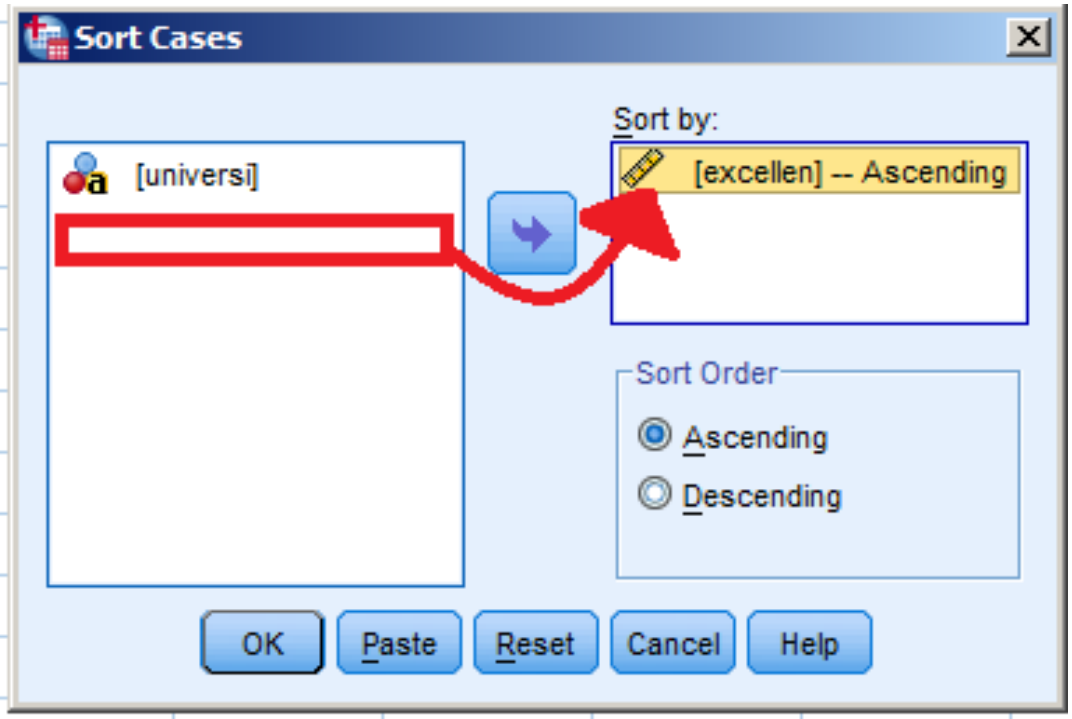
Problem 1.29 calls for a bar graph of universities ordered by excellence rating. To sort data, use the top menu and go to

Data → Sort Cases



In the dialog that pops up, move the variable you which to sort by into the “Sort by” box, by either dragging or using the → button. Then click OK.

In problem 1.29, you want to sort by [excellen].



Your data should now be sorted:

	universi	excellen
1	Ottawa	11
2	Calgary	14
3	York	18
4	BritishColumbia	18
5	Toronto	21
6	Concordia	21
7	Alberta	23
8	McGill	26
9	Waterloo	36
10	WesternOntario	38

Problem 1.29 has only 10 rows, so this could have been done manually, however, if there were 10,000 rows like there are in some professional datasets, that wouldn't have been possible.

One-Variable Graphs

Here we build a bar graph, histogram, a boxplot, and a side-by-side boxplot using the datasets radioformat.por, descriptives XYZ.sav and dragons.sav

Quick reference:

Analyze → Descriptive Statistics → Frequencies

Graphs → Legacy Dialogs → Bar

Graphs → Legacy Dialogs → Histogram

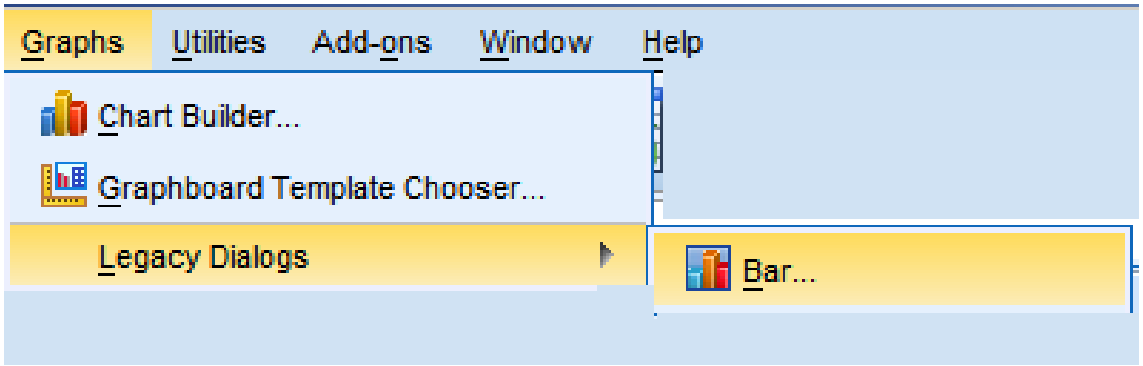
Graphs → Legacy Dialogs → Boxplot

Graphs → Legacy Dialogs → Pie

For the Bar Graphs in 1.3c,

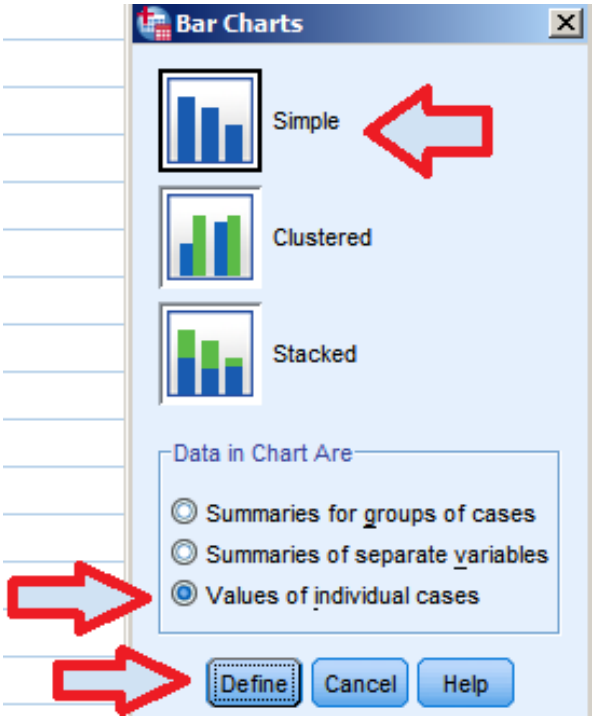
load the radioformat.por dataset and use the menu options:

Graphs → Legacy Dialogs → Bar



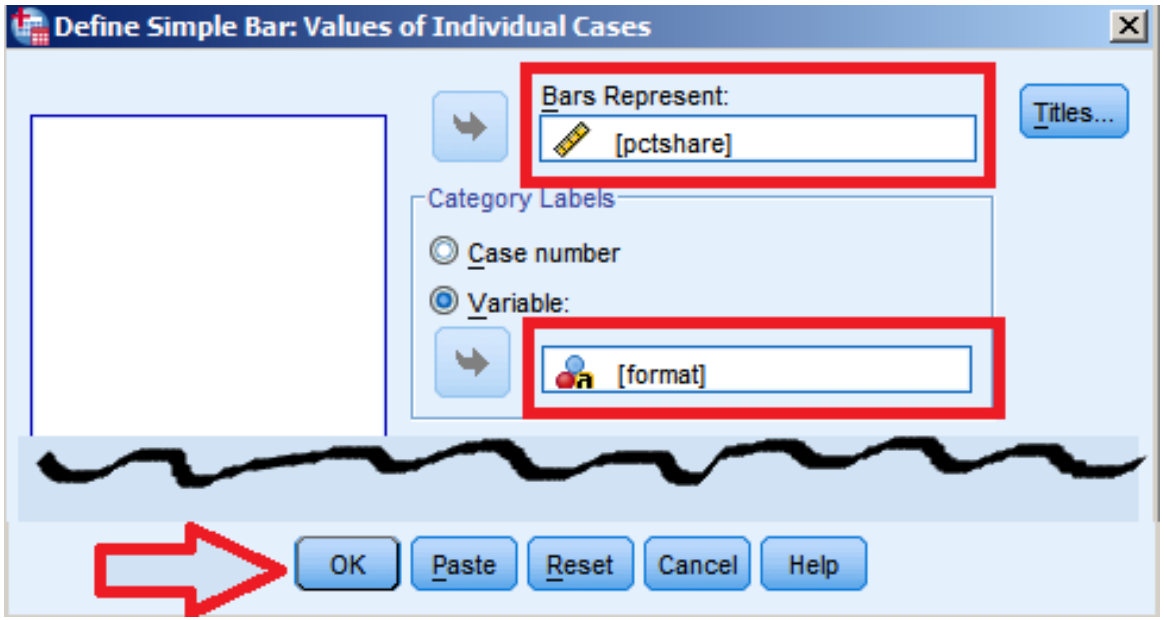
In the first dialog,

- use the **Simple** option *(only one variable to deal with)*
- Choose **Values of individual cases** *(each row is a category)*
- Click **Define**



The height of the bars should be the % of market share, so put ***pctshare*** in the ***Bars Represent*** box.

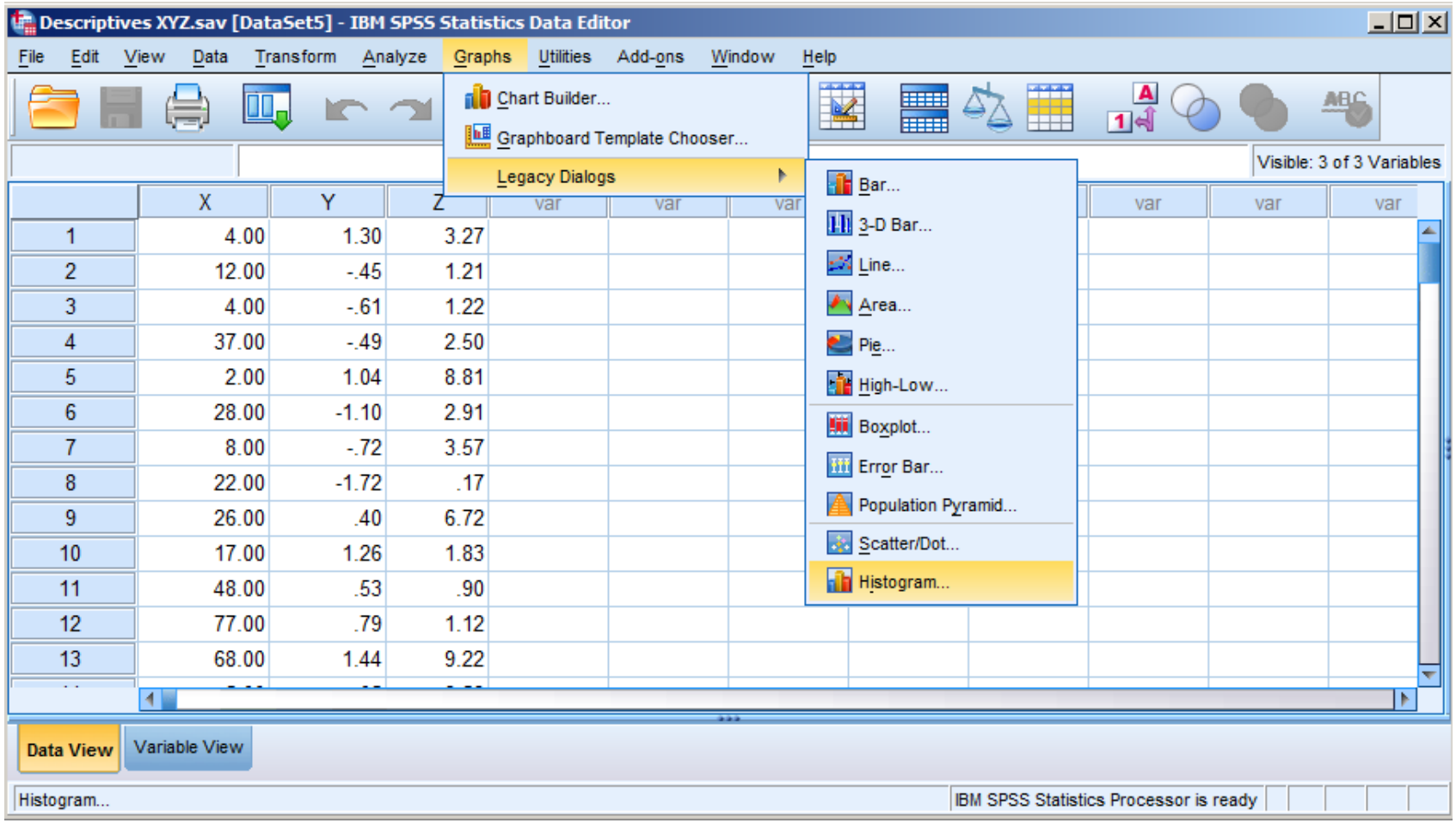
The labels should be of the music formats themselves, so under ***Category Labels***, choose ***Variable*** and put ***format*** in the box.



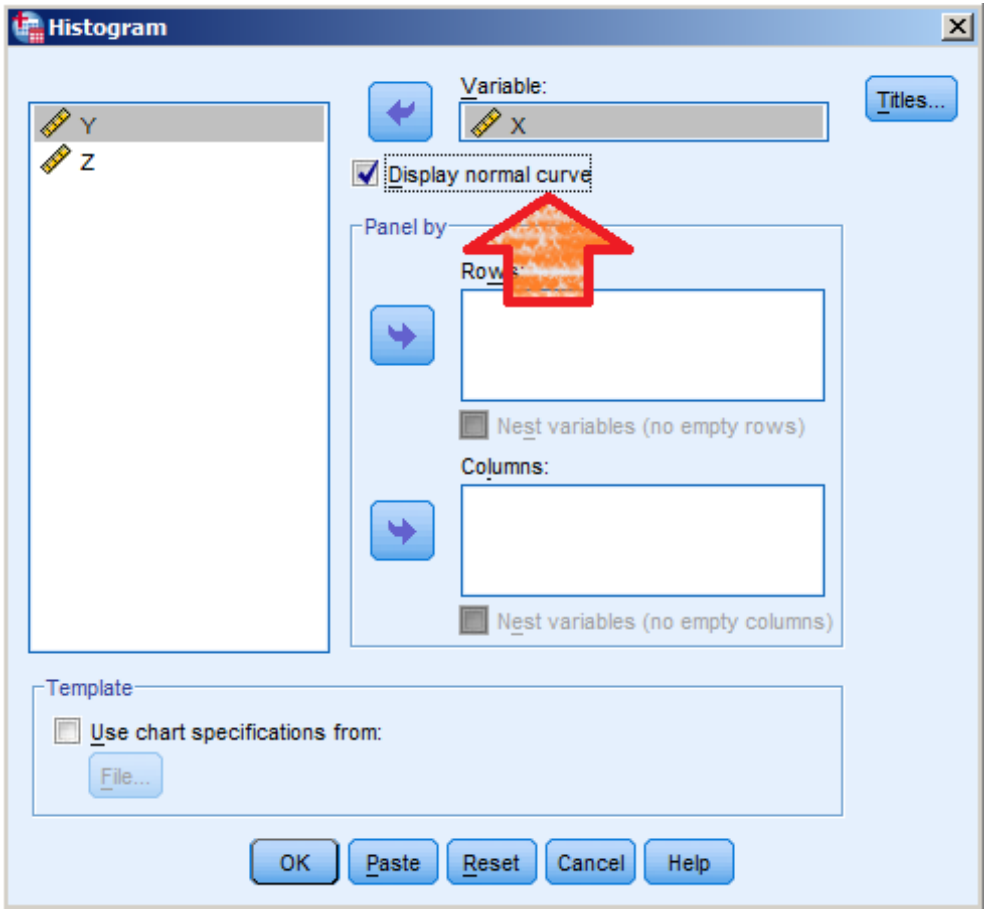
Click OK, and you're done!

To build a histogram

Option 1: In Graphs → Legacy Dialogs → Histogram…

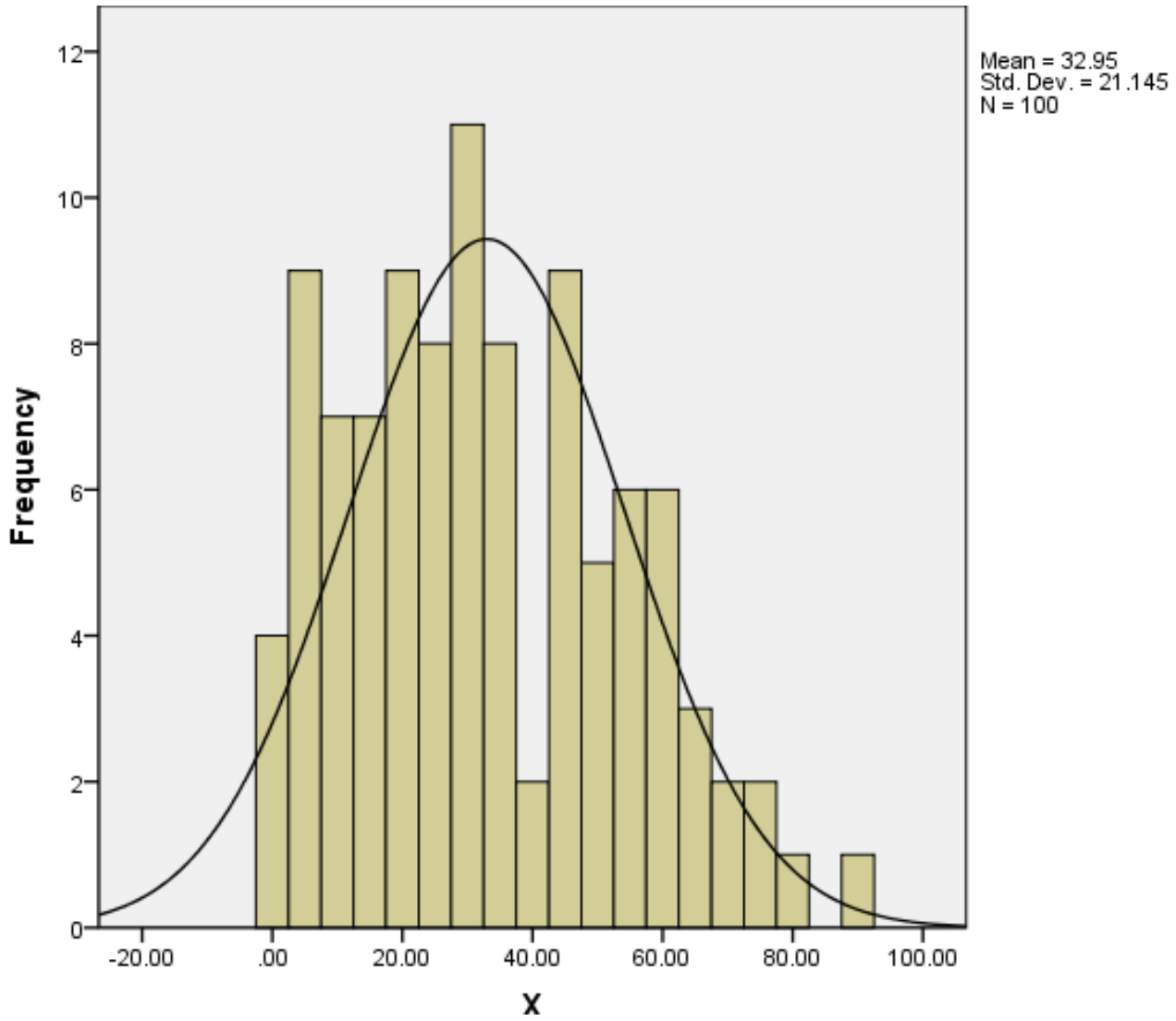


Select what you wish to make a histogram of, drag it to ***variable***

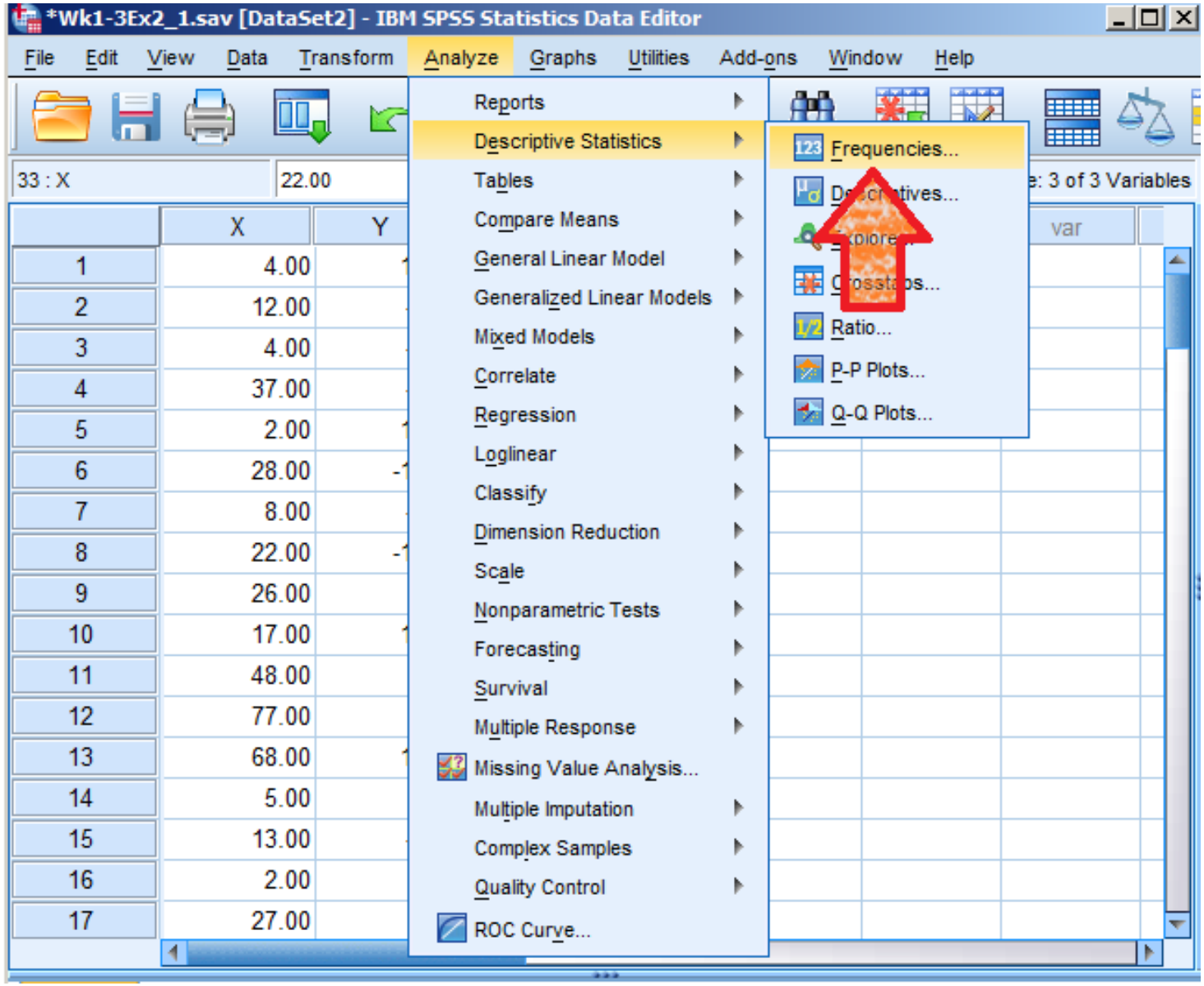


If you wish to compare the distribution in the histogram with the normal, check ***"display normal curve".*** Then click ***OK.***

In the output window, a histogram of your variable will appear.

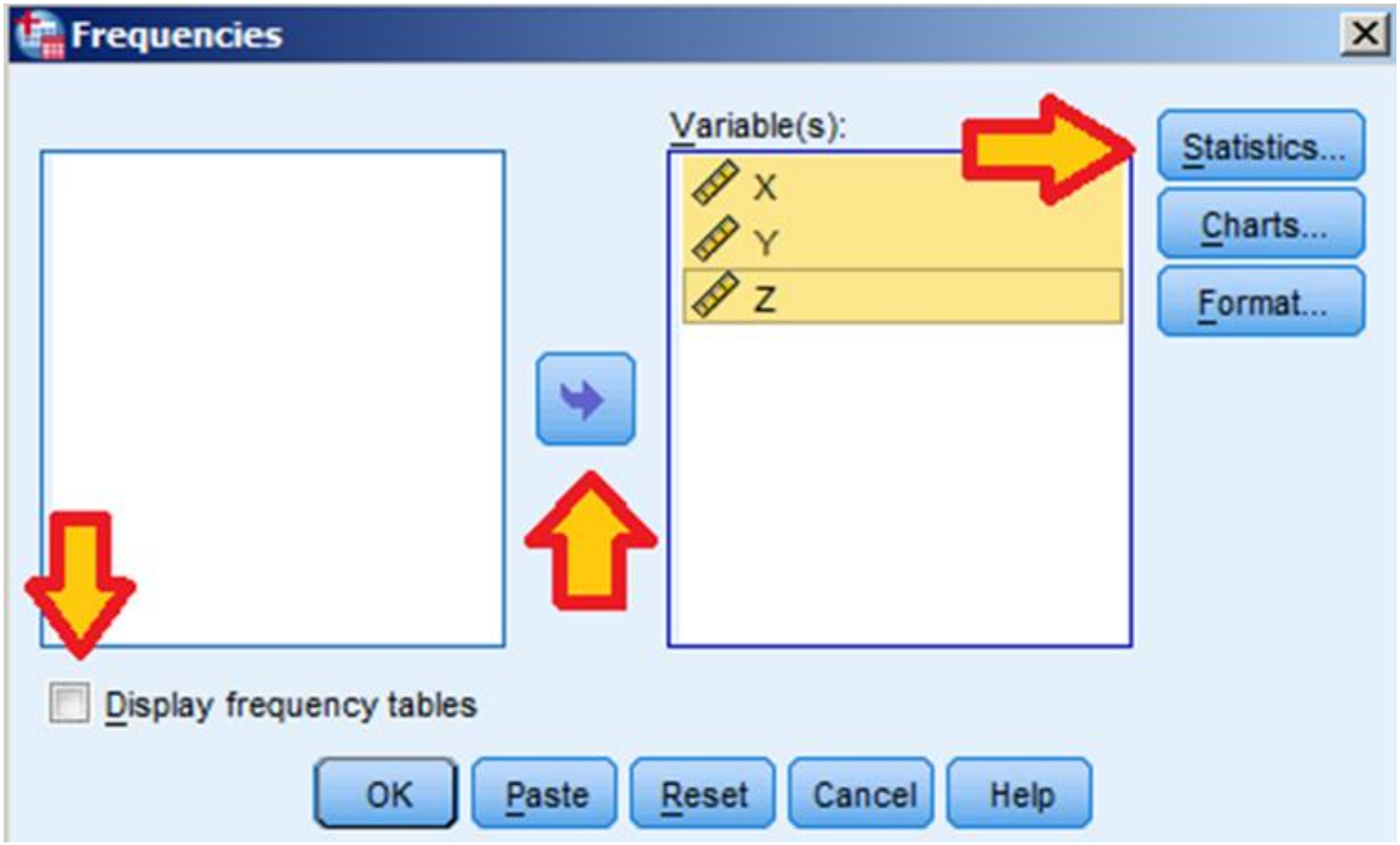


Option 2: In Analyze → Descriptive Statistics → Frequencies

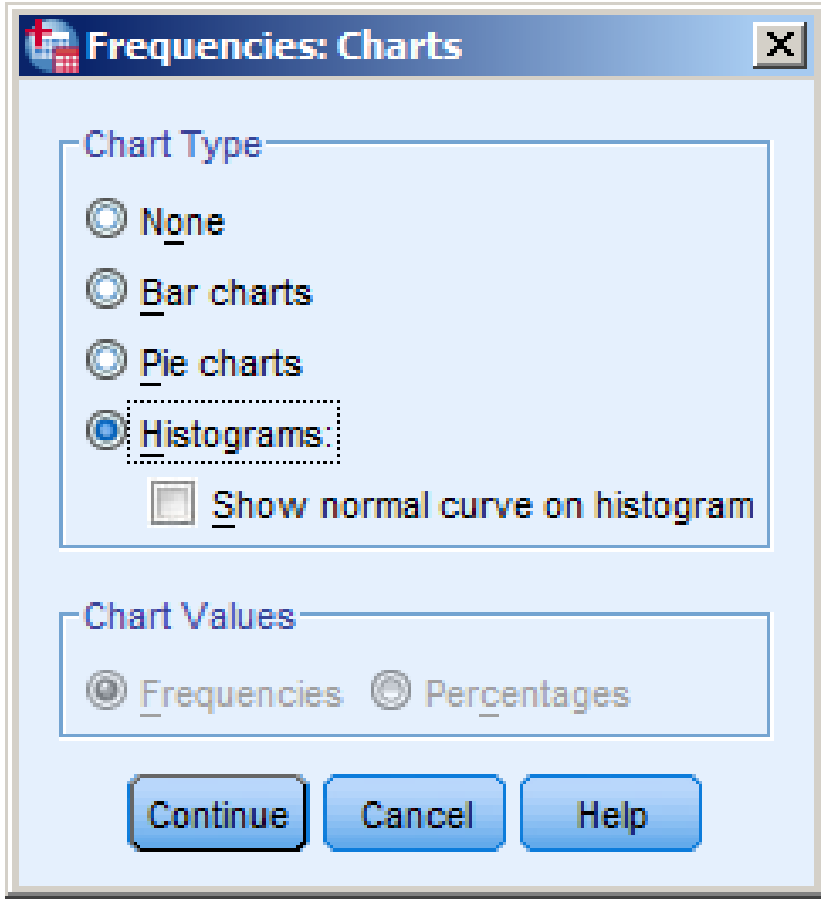


Drag any variables of which you wish to build histograms from the field on the left (empty in screen) into the ***variable(s)*** field.

Click on "Charts", on the right end of the dialog.



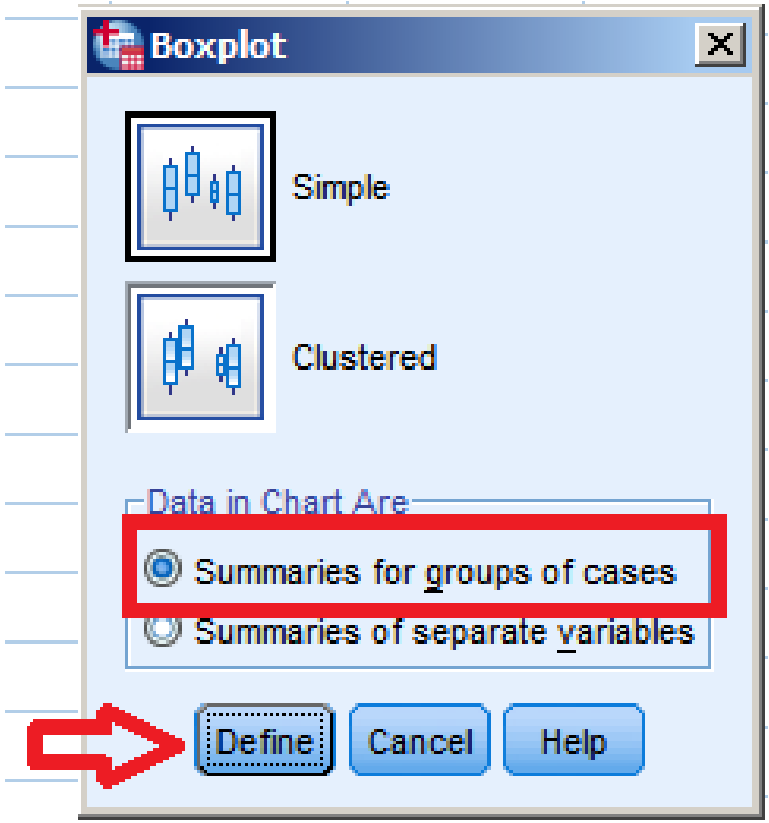
Choose the "Histograms:" radio button, click Continue, then OK.



The same histogram as before will appear in the output.

Boxplots (Grouped by case)

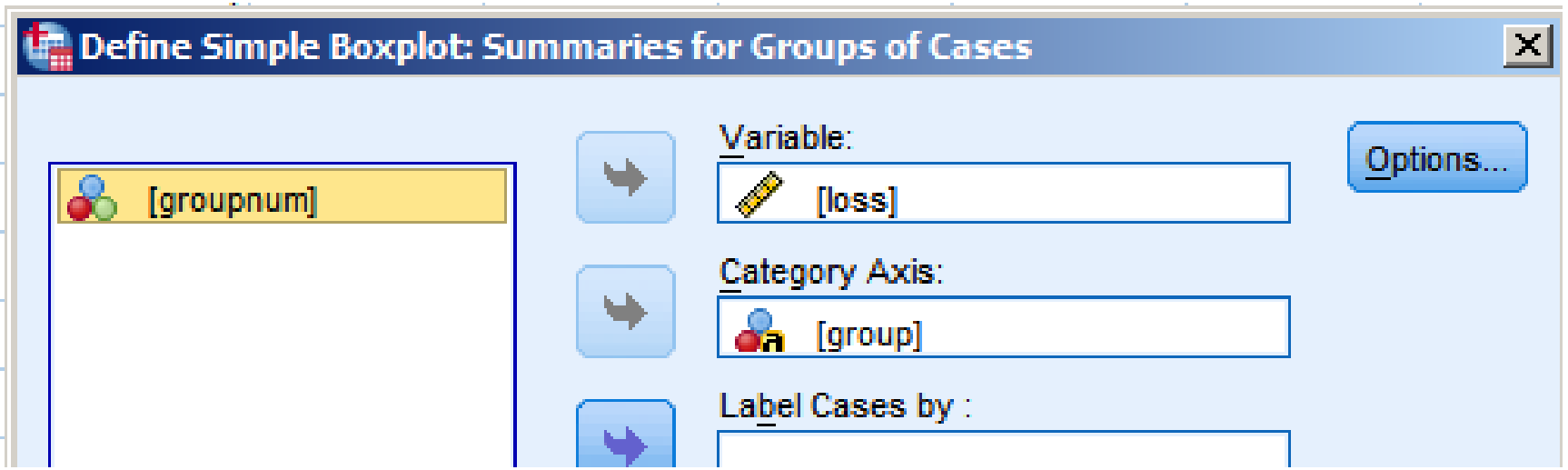
For the boxplots in 2.43, the **gastricbands.por** dataset has all the numeric data in a single column, so you can leave options as they are: ***Simple***, and ***Summaries for groups of cases***



Loss is the variable we're interested in, so it goes in ***Variable***.

The boxplots will be split by whatever we put in the ***Category Axis.***

In this case, either ***group*** or groupnum will do, but ***group*** is more informative.

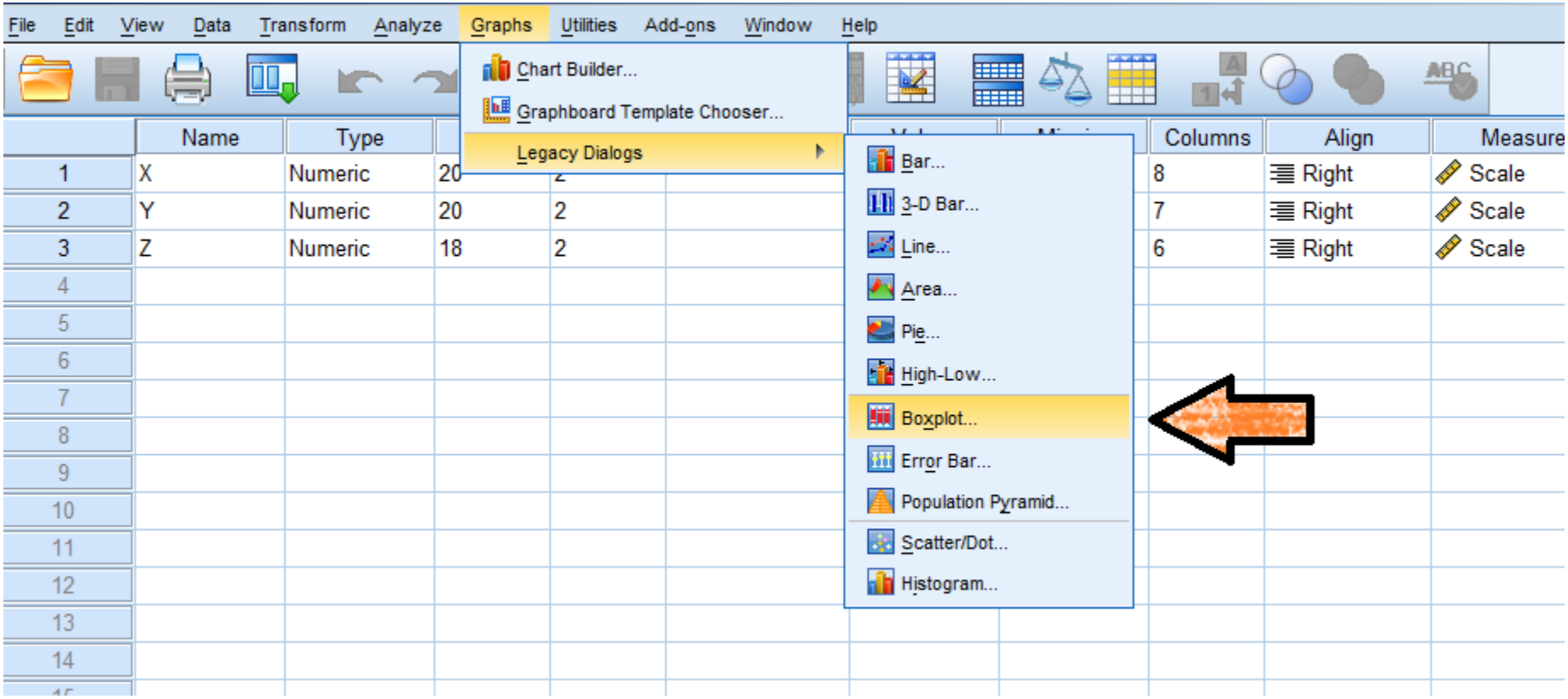


Click OK, and you're done!

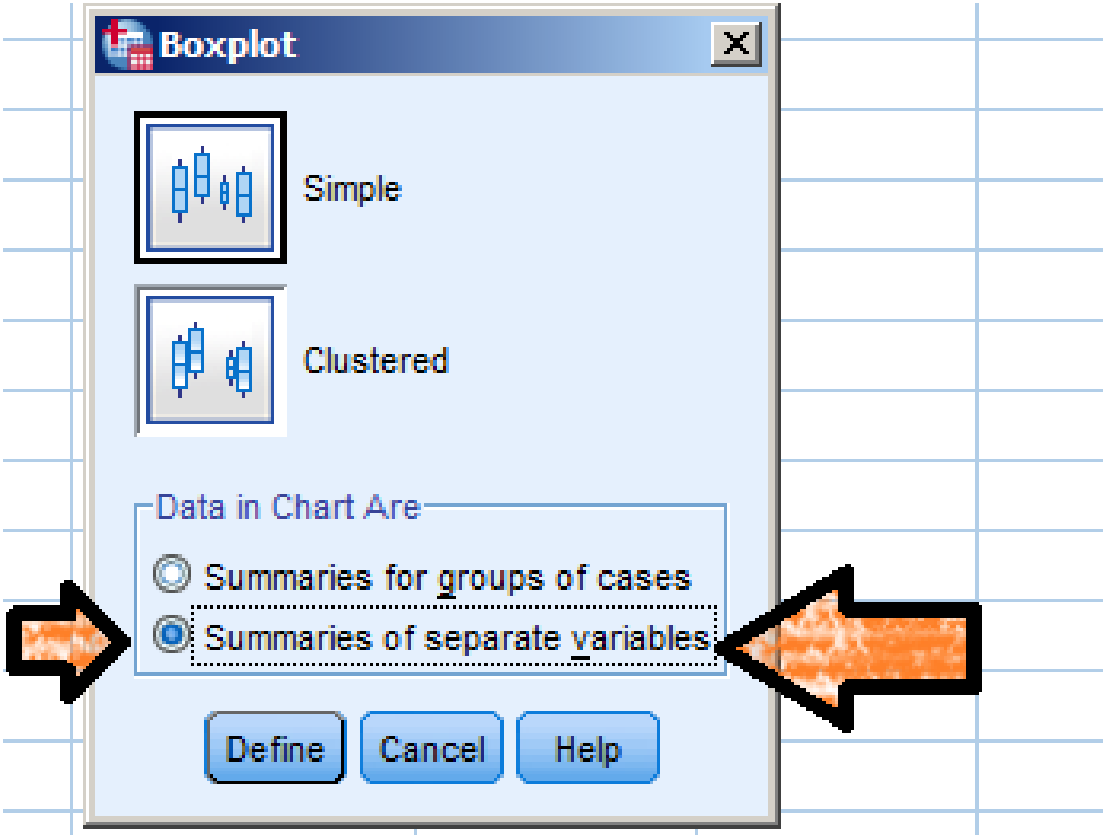
Boxplots (Separate Variables)

To build a boxplot in SPSS, go to

Graphs → Legacy Dialogs → Boxplot

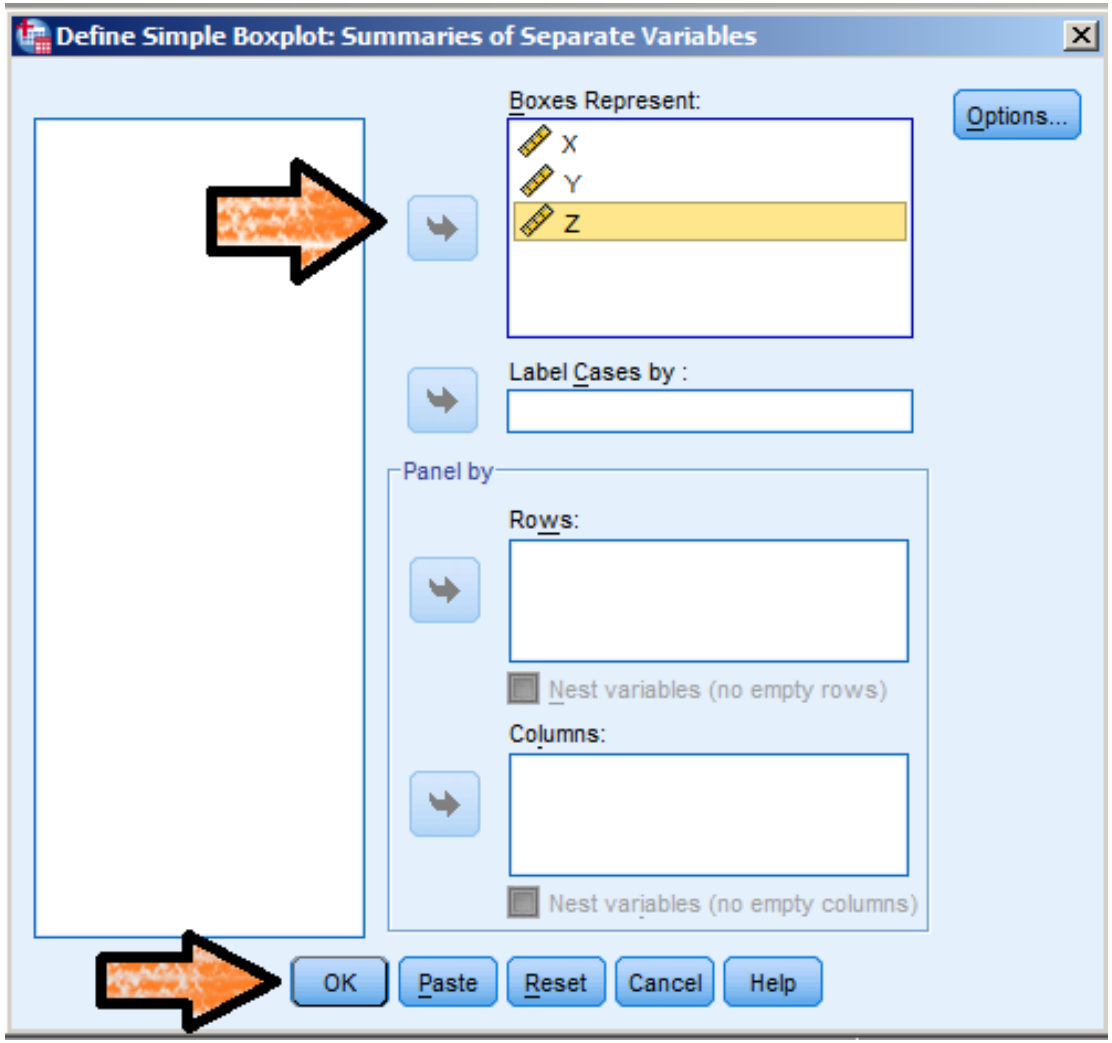


X, Y, and Z are separate variables. Therefore, in the boxplot dialog, we will switch the radio button to “Summaries of ***separate variables”*** before clicking ***“Define”.***

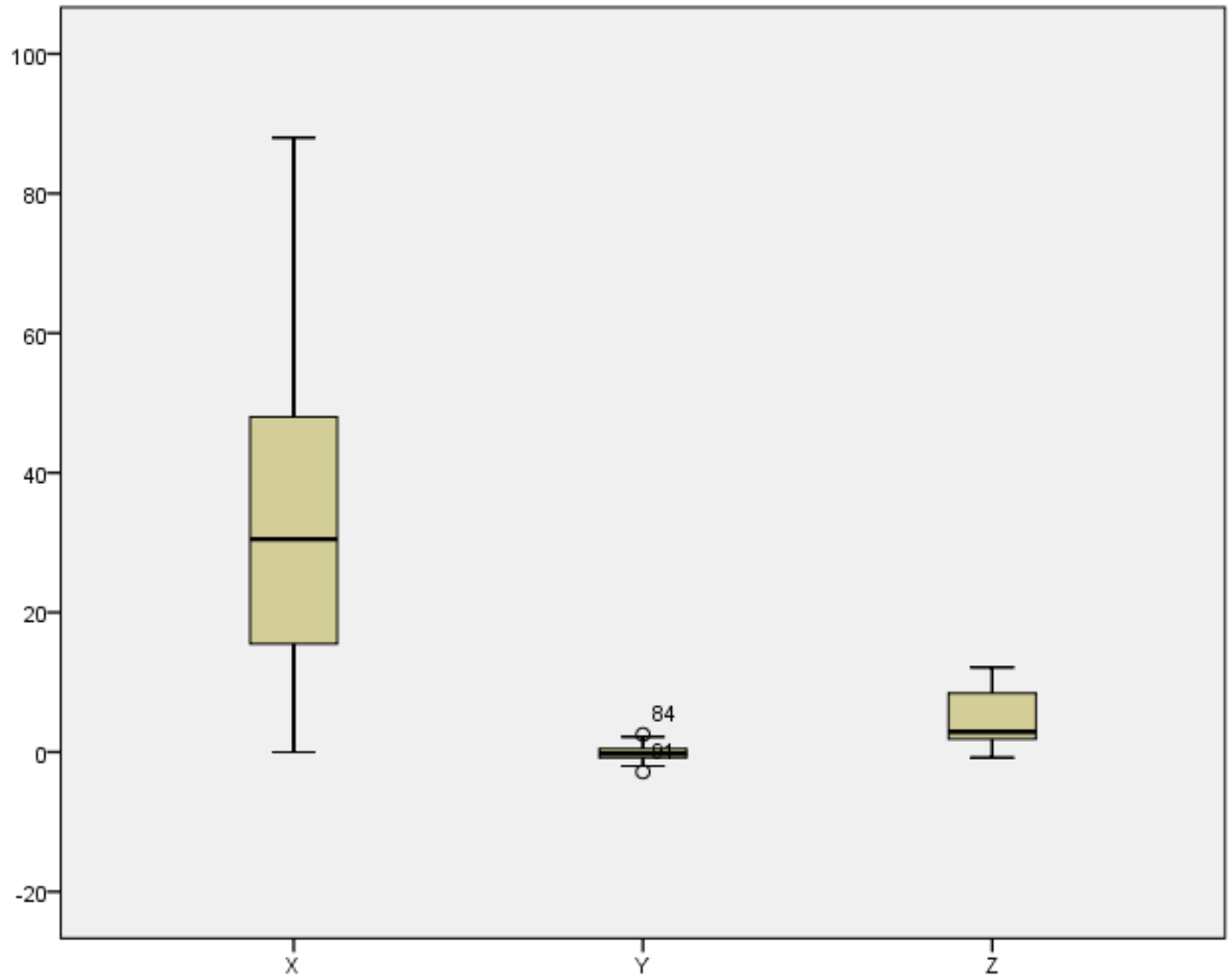


Move the variables you want into ***"Boxes Represent"***

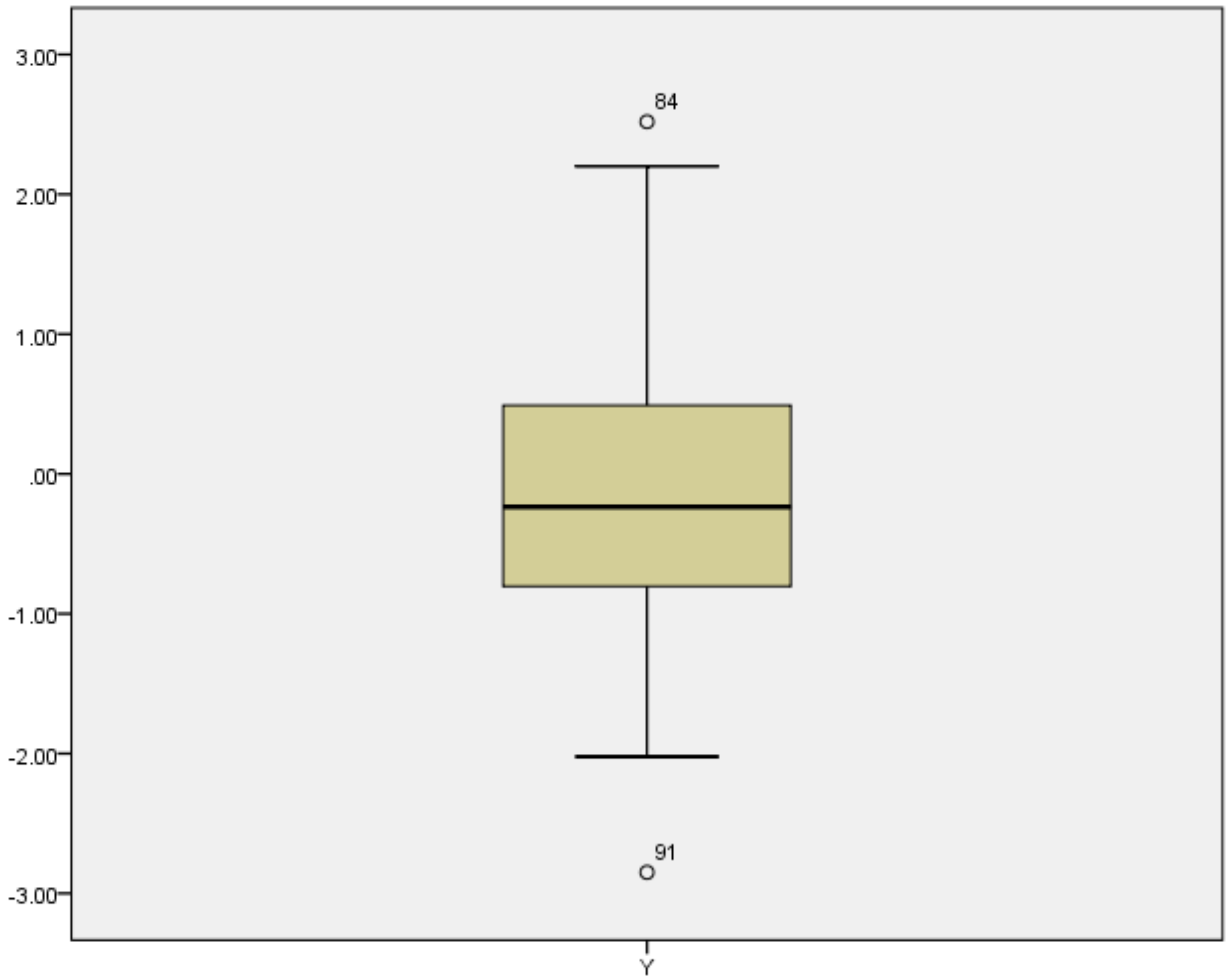
Then click OK



This is the result. Side-by-side boxplots can be used to compare more than 2 variables.

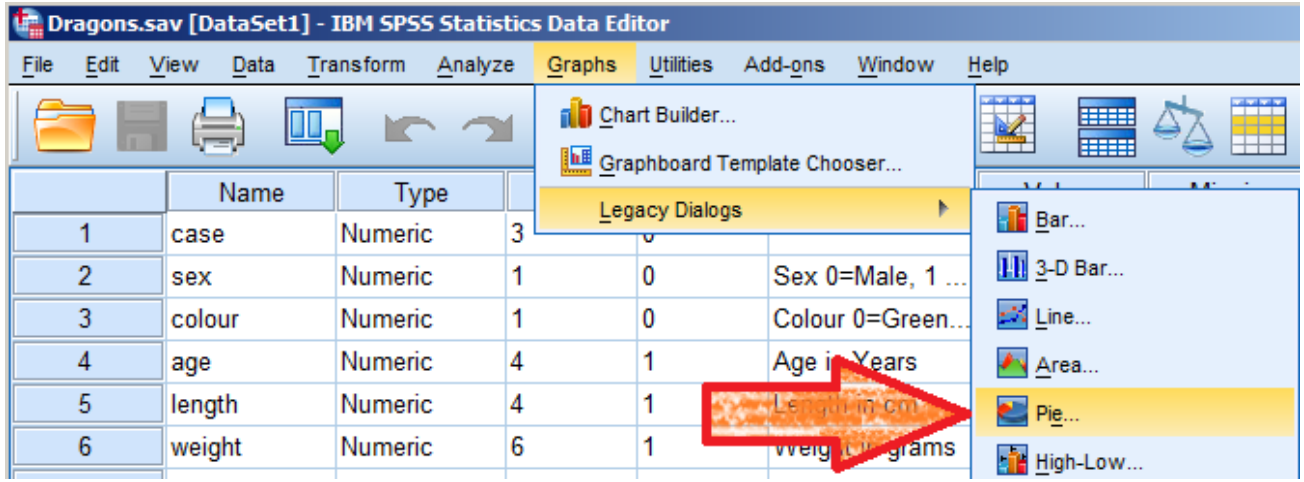


Including only one variable in boxplots will give you one boxplot, including multiple variables will give you side-by-side boxplots on the same scale.

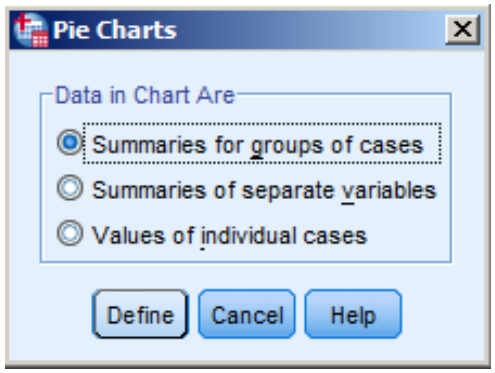


To build a **Pie Chart**,

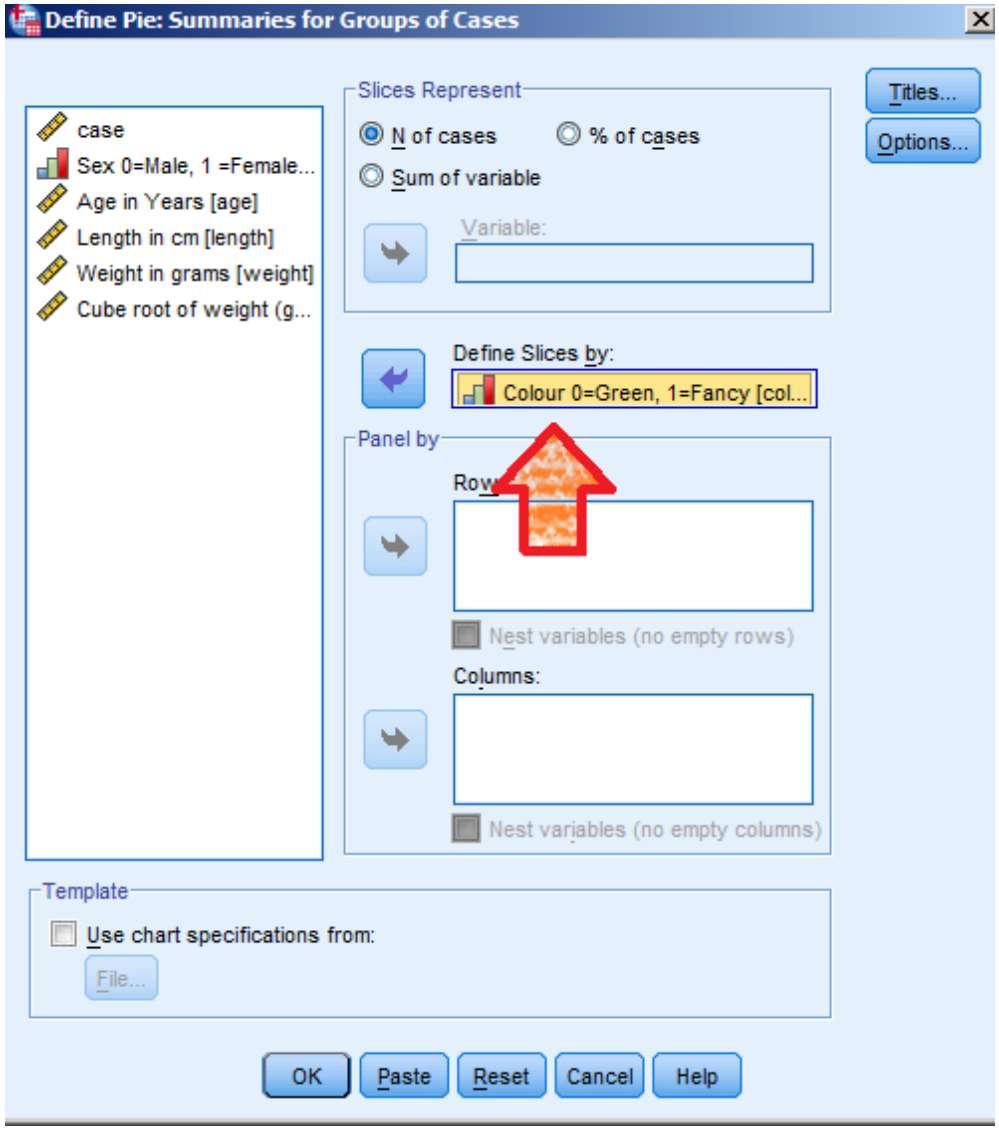
Graphs → Legacy Dialogs → Pie



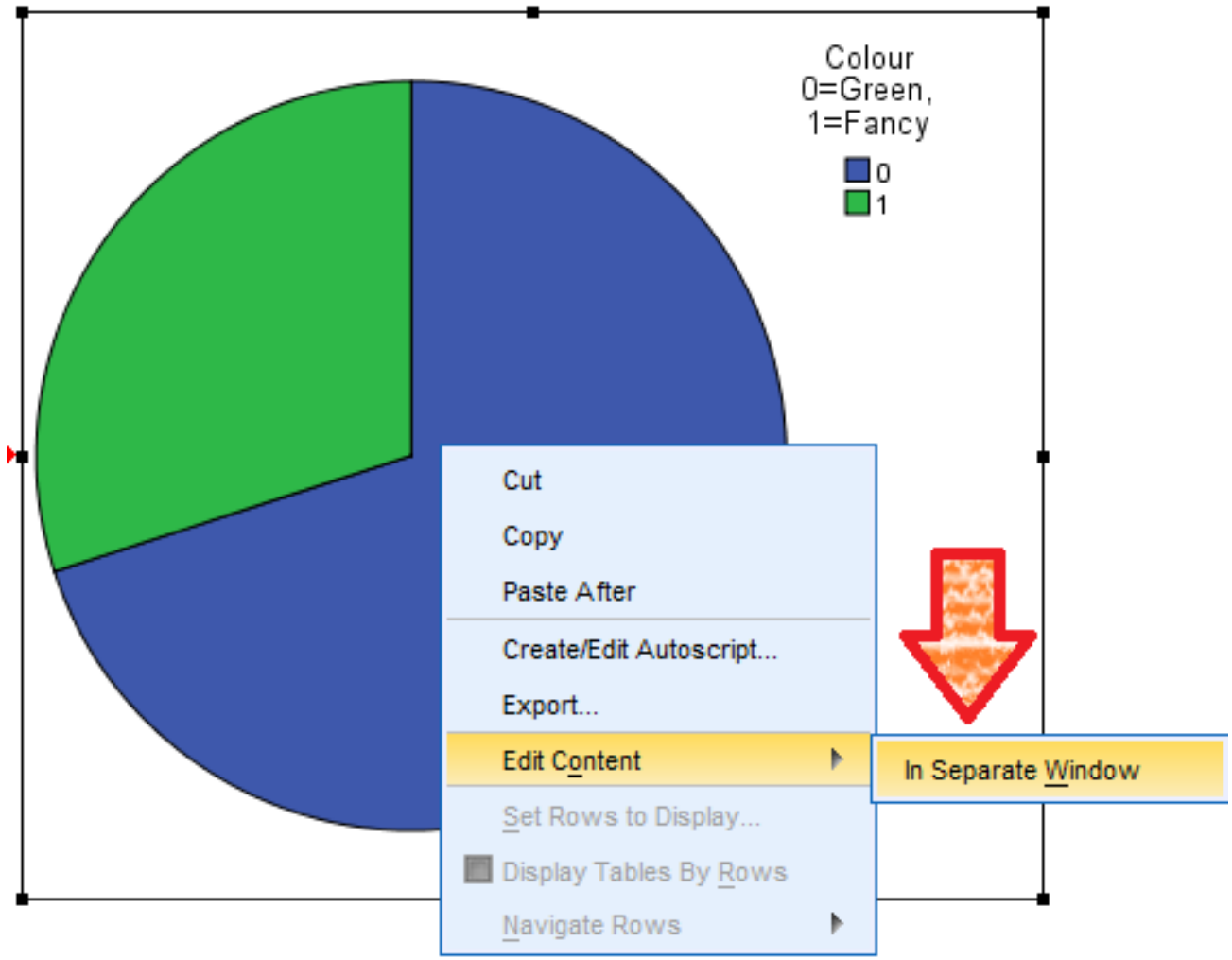
Then choose **summaries for groups of cases**, and **define.**



Choose a variable and drag it into **Define Slices by:** then click **OK**.



This will open the **output window**, which will include your pie chart. To add labels, percentages, and a title**, right-click the chart** and choose **Edit Content → In Separate Window**.



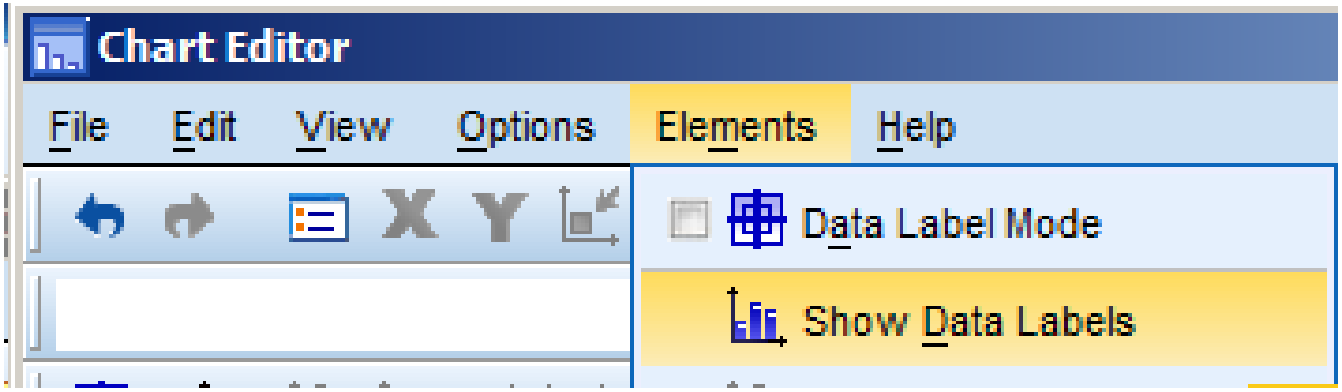
This will open the chart editor.

In the chart editor, you can add frequencies by double-clicking on the pie, and then choosing

Elements → Show Data Labels

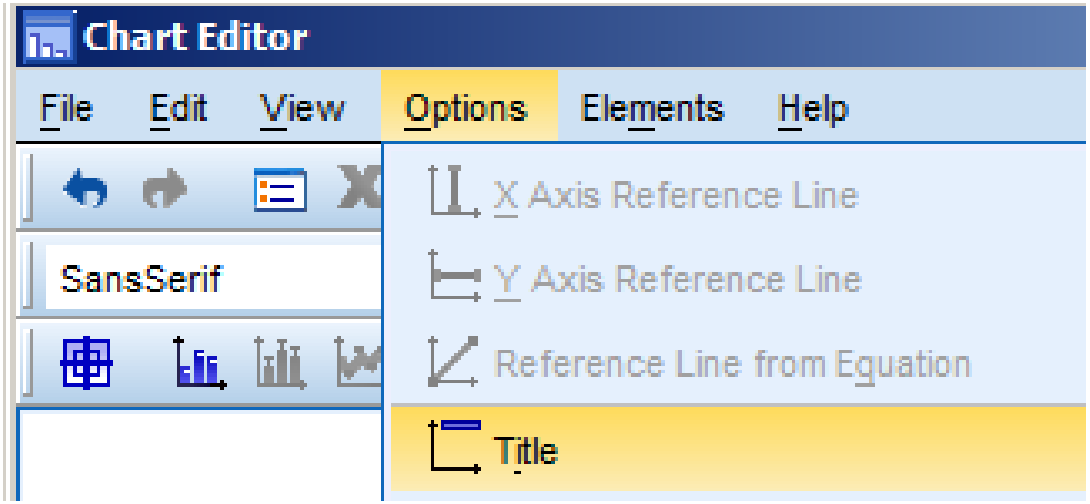
(or right-clicking the pie → Show Data Labels)

In the window that pops up, just click close. Percent is the default option, so nothing else is needed.

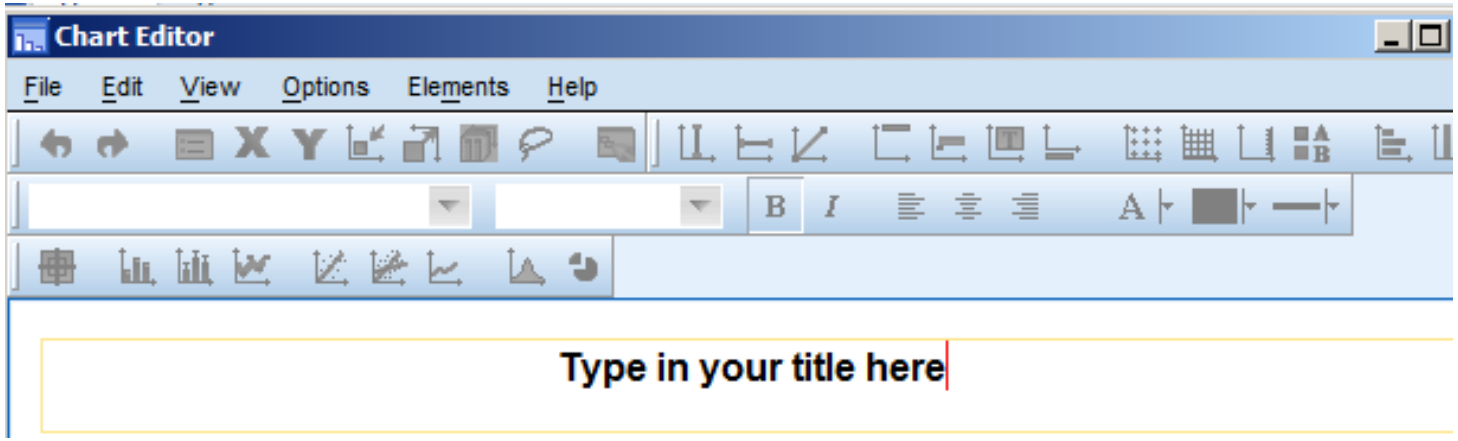


Also in the chart editor, you can choose to add a title in

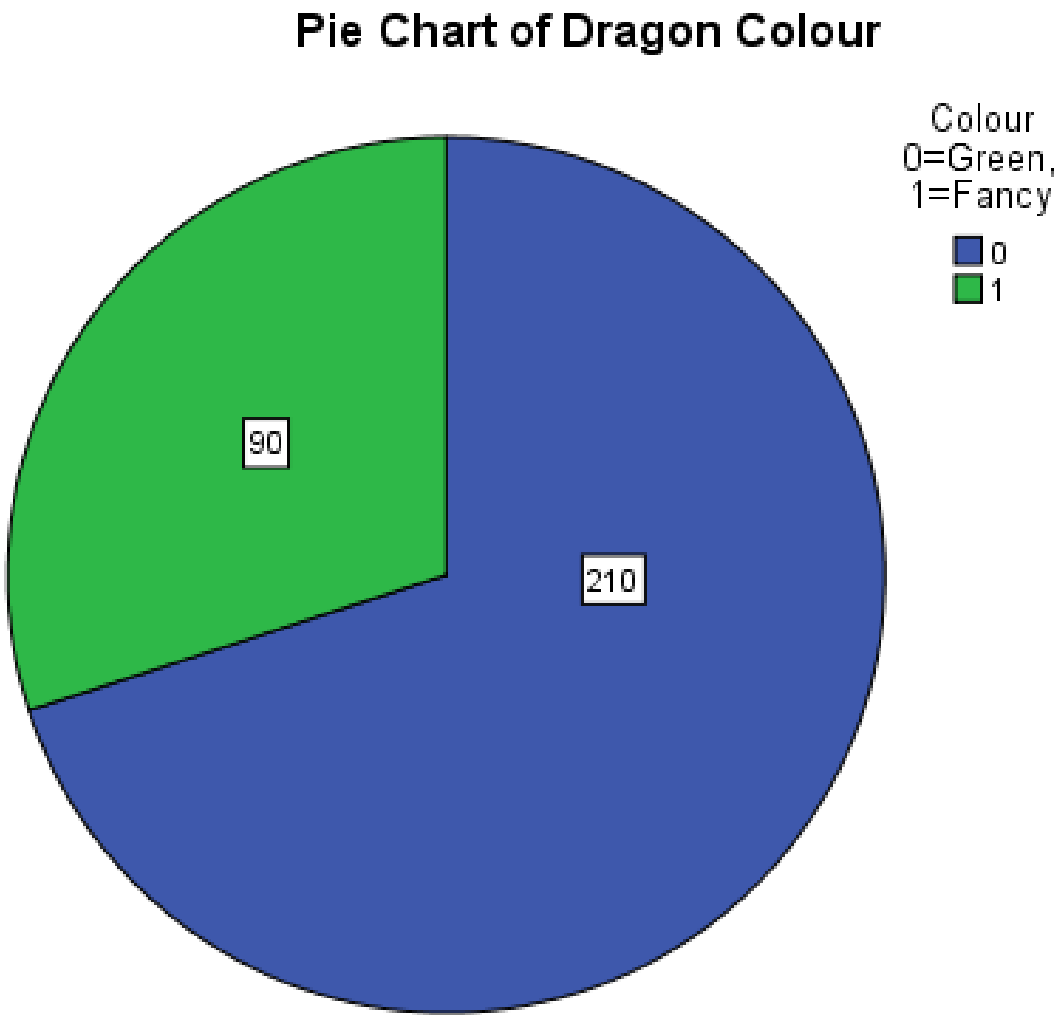
Options → Title



Then type in your title then closing the dialog that appears.



The result is a pie chart with a title and frequencies.



Descriptives

In this chapter, we calculate central measures like the ***mean and median,*** and measures of spread like the ***standard deviation*** and ***interquartile range (IQR)*** from the dataset Descriptives XYZ.sav

Quick reference:

Analyze →Descriptive Statistics → Frequencies

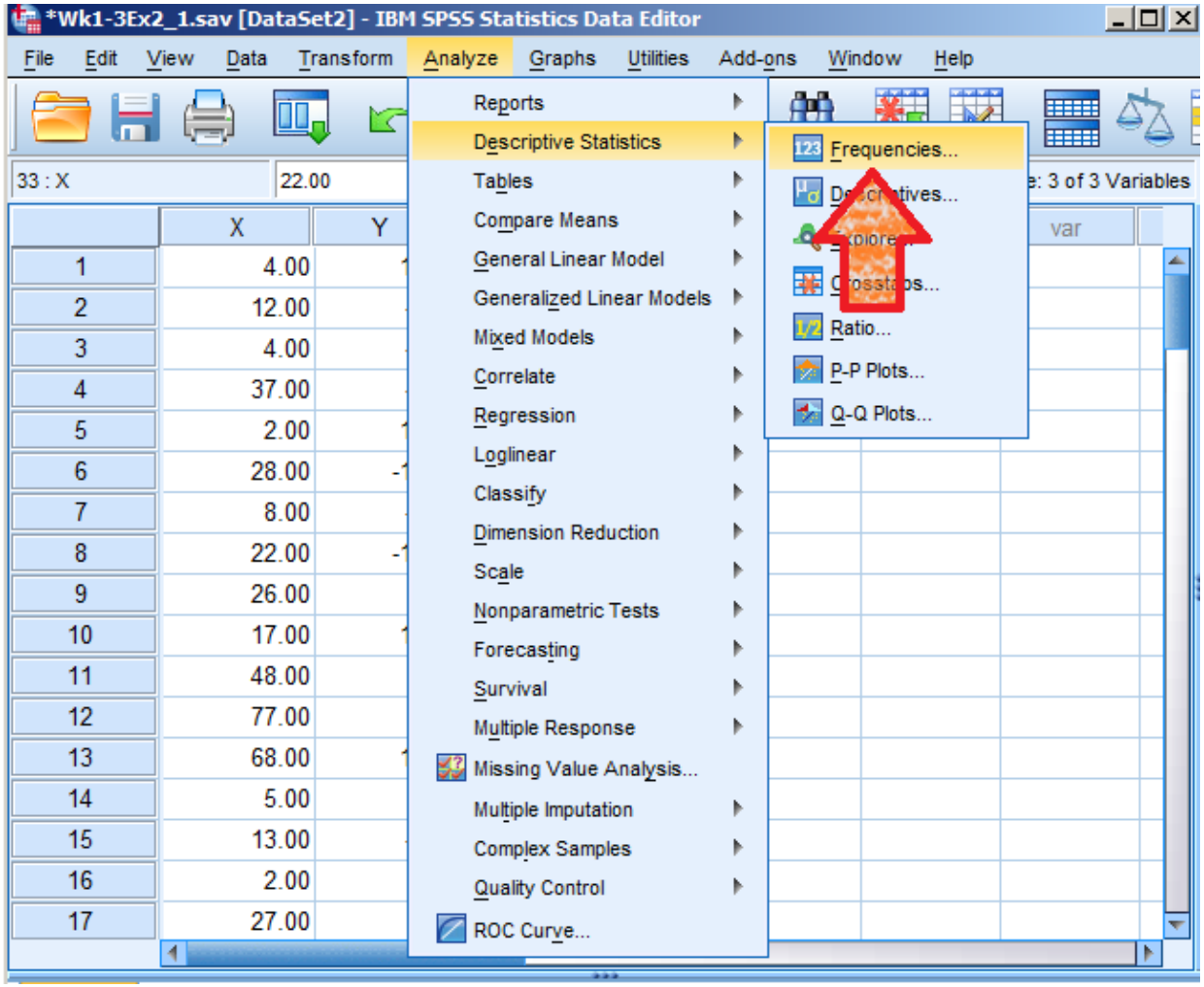
Analyze →Descriptive Statistics → Descriptives

Analyze →Descriptive Statistics → Explore

- First, open a file as shown in Inputting Data.

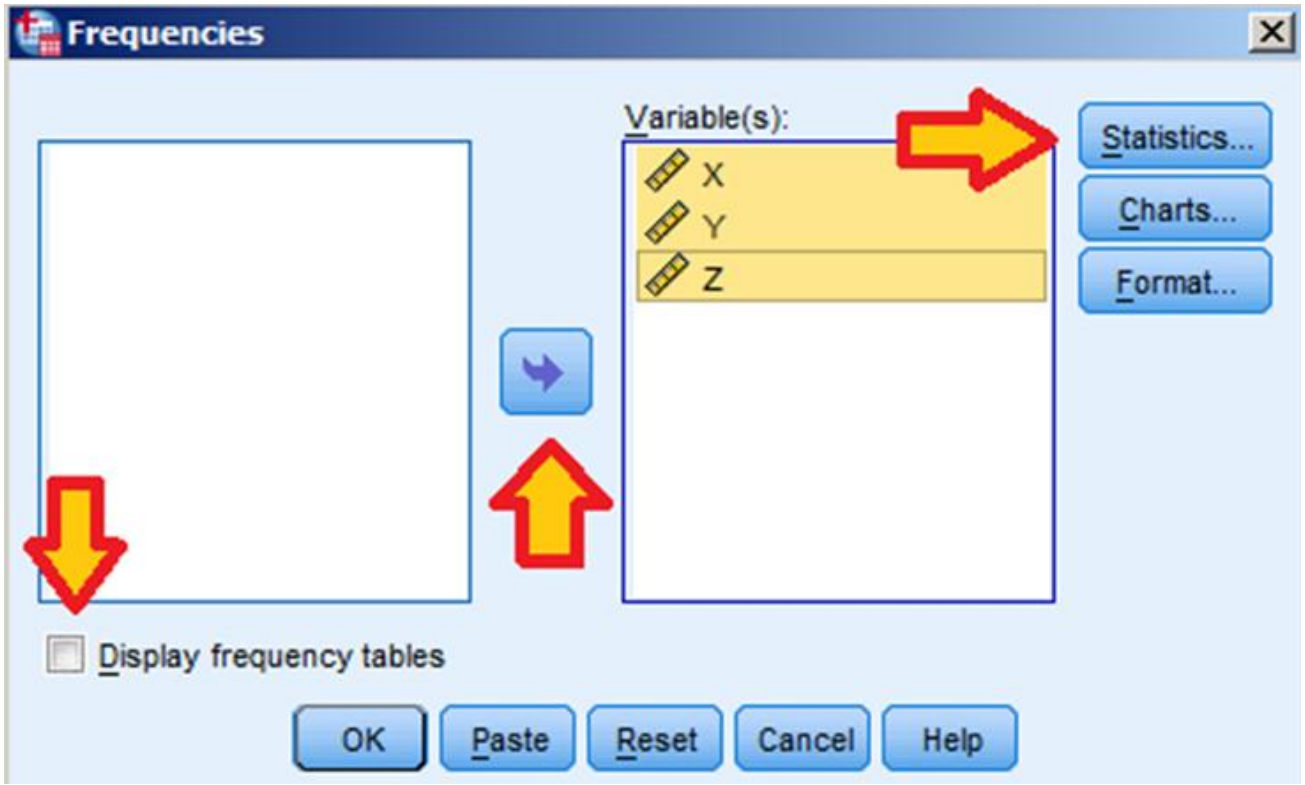
To calculate the mean, median, IQR, or standard deviation, go to

Analyze → Descriptive statistics → Frequencies



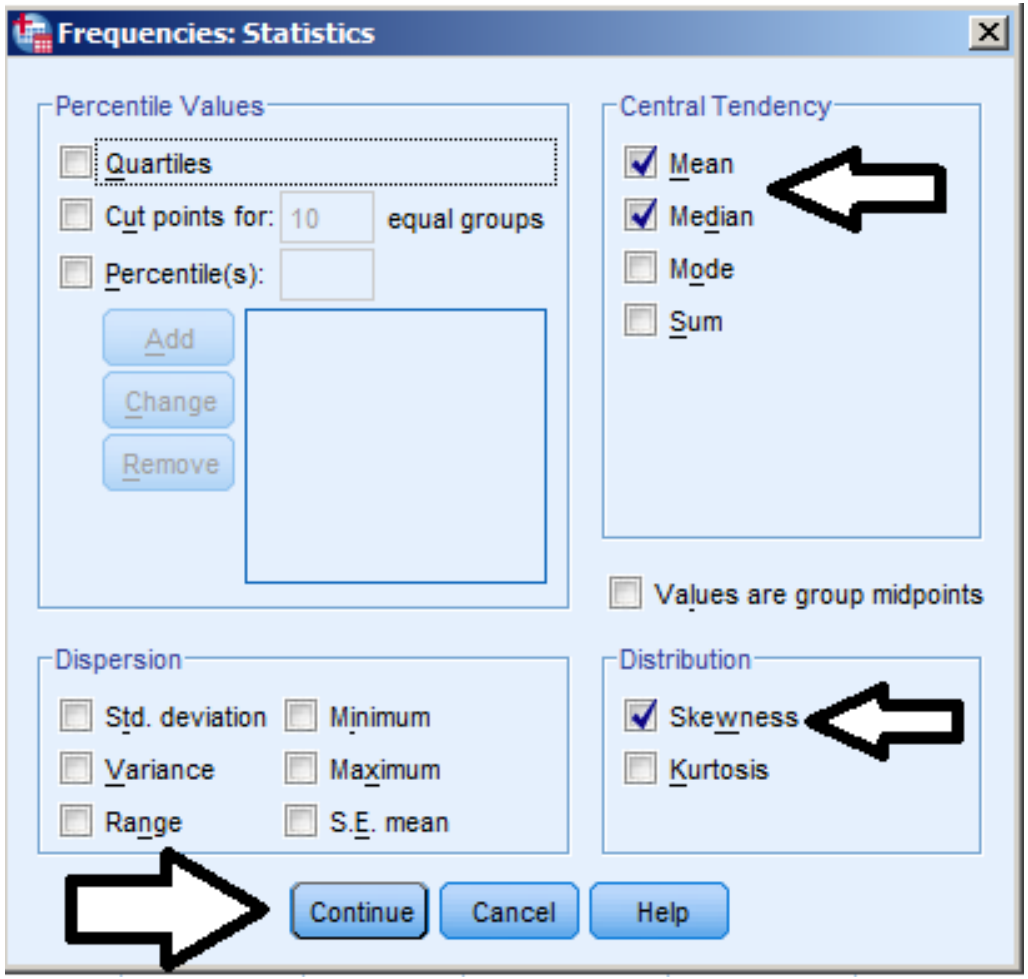
In the dialog that appears, uncheck ‘Display Frequency Tables’.

Select all the variables you’re interested in and move them to the right by dragging or using the → button in the middle.

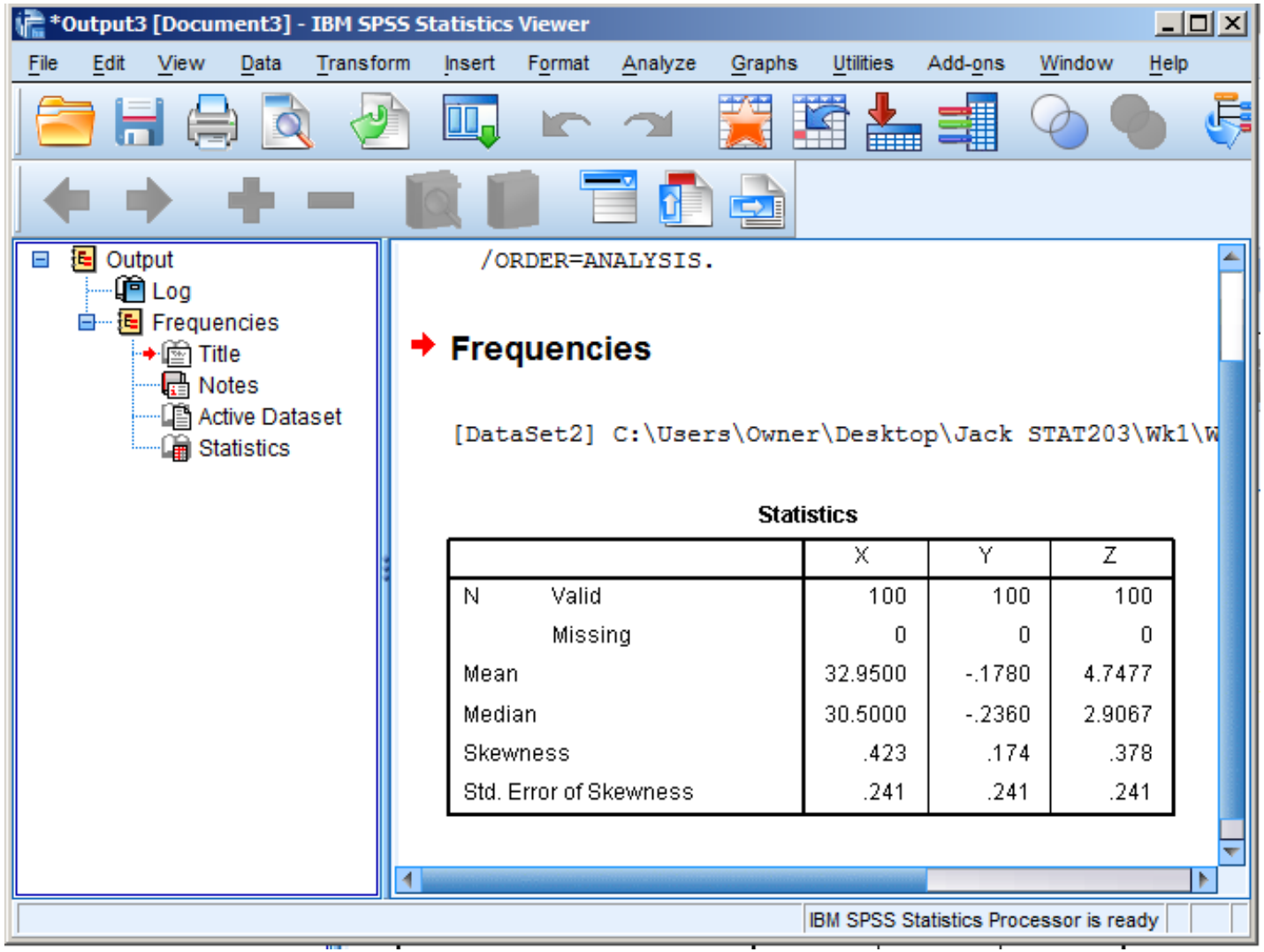


Click on “Statistics” in the upper right of this dialog window, and a second dialog window will open.

Check “Mean”, “Median” (upper right), and “Skewness” (lower right), then click “Continue” in the lower left. to close this dialog. Click “OK” in the dialog with the variables listed.



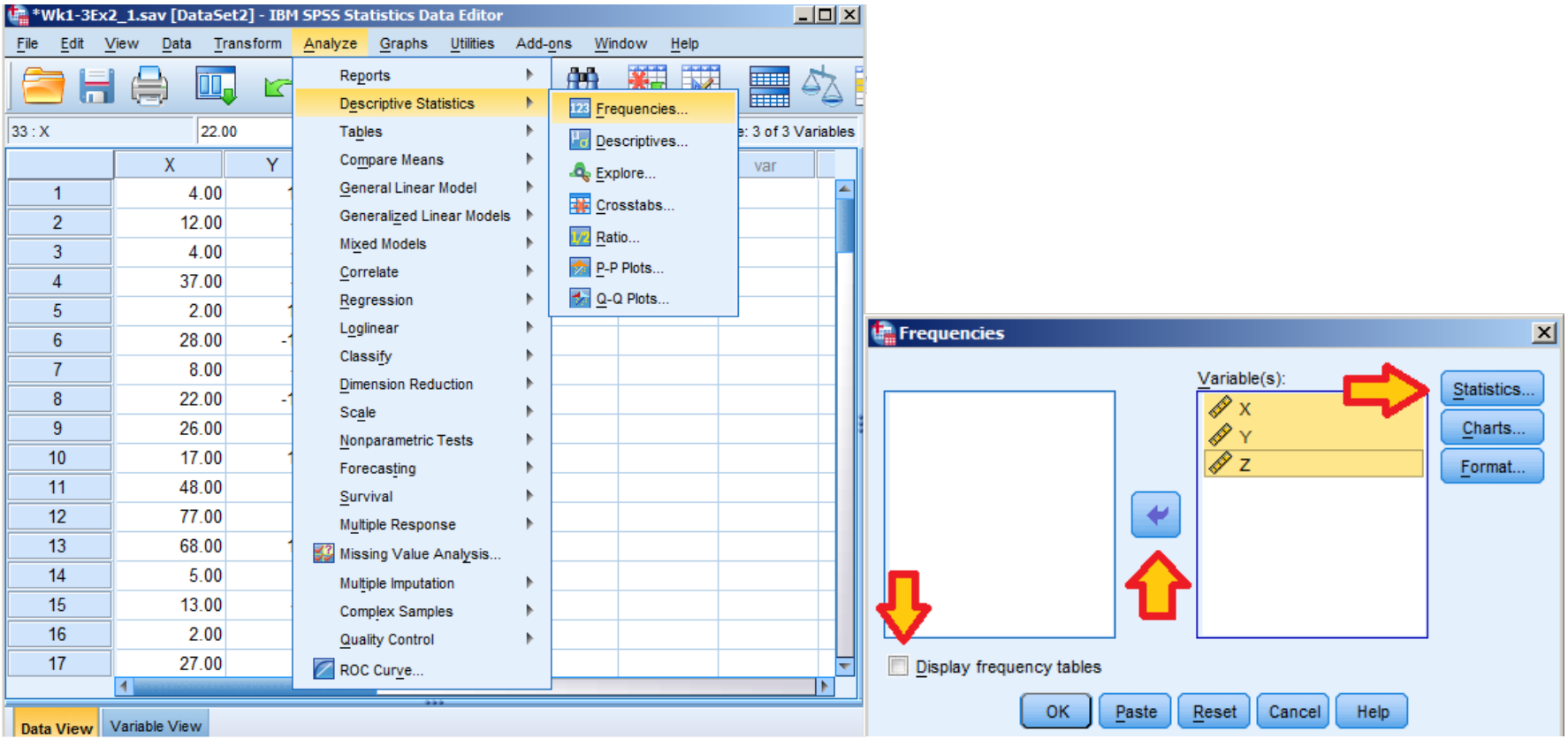
A results window should open, giving you the mean, median, and skew of our three variables.



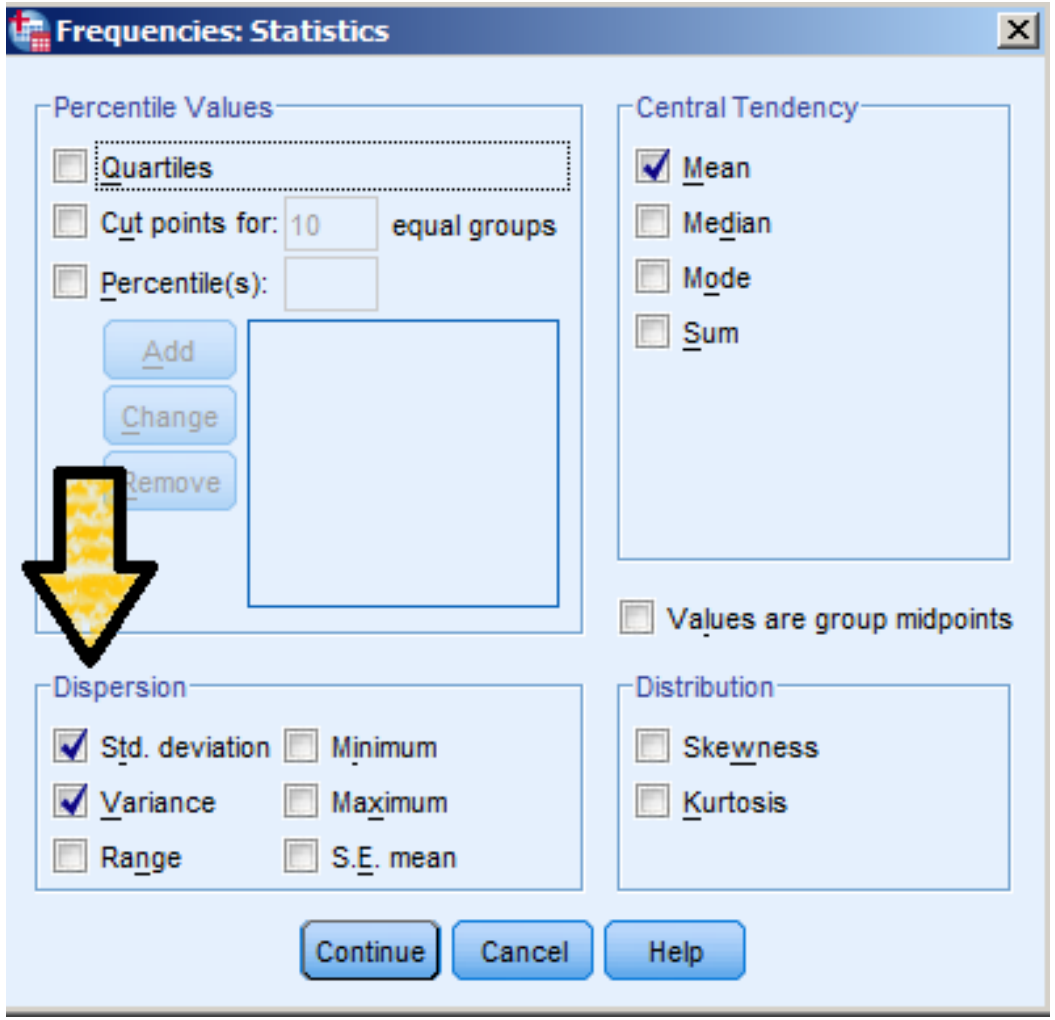
Calculating the standard deviation from SPSS is the same as calculating the mean, median, and quartiles:

Analyze → Descriptive Statistics → Frequencies

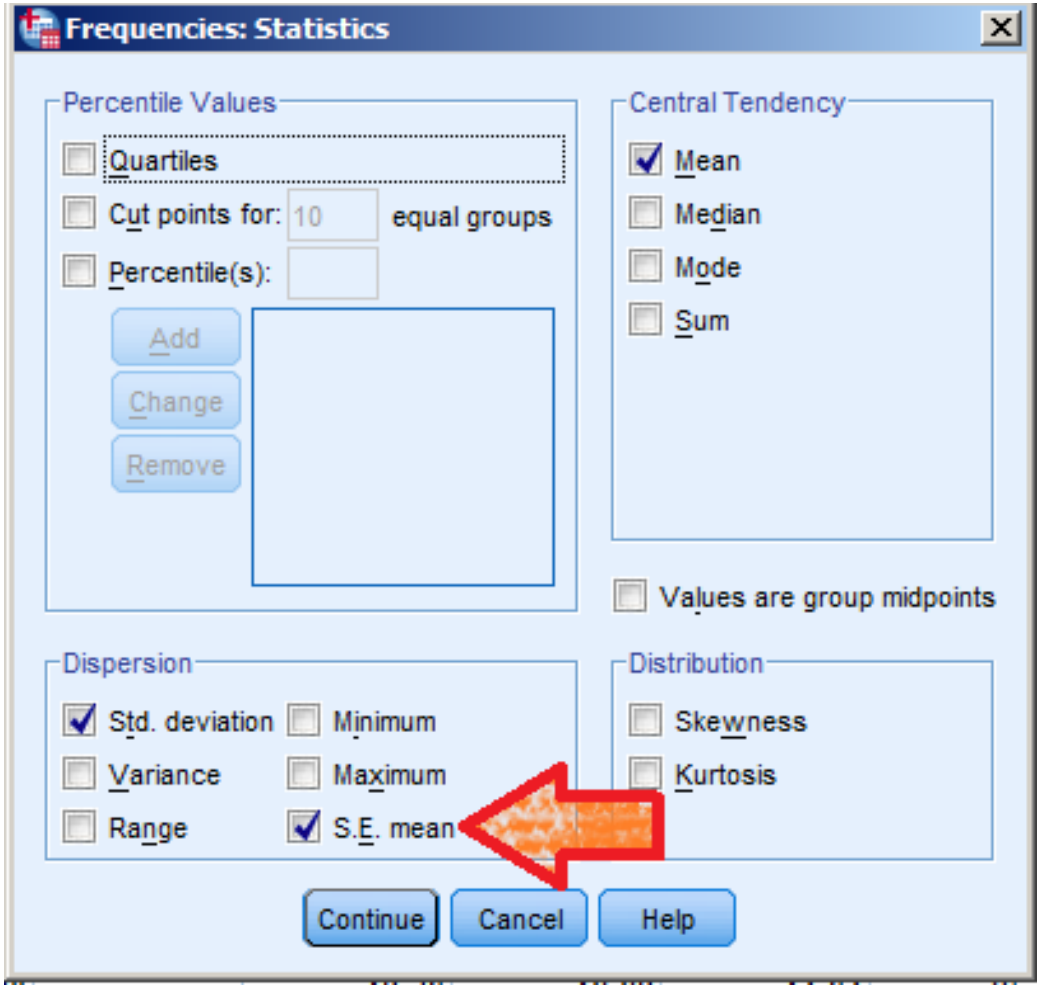
Choose your variables, click on “Statistics”



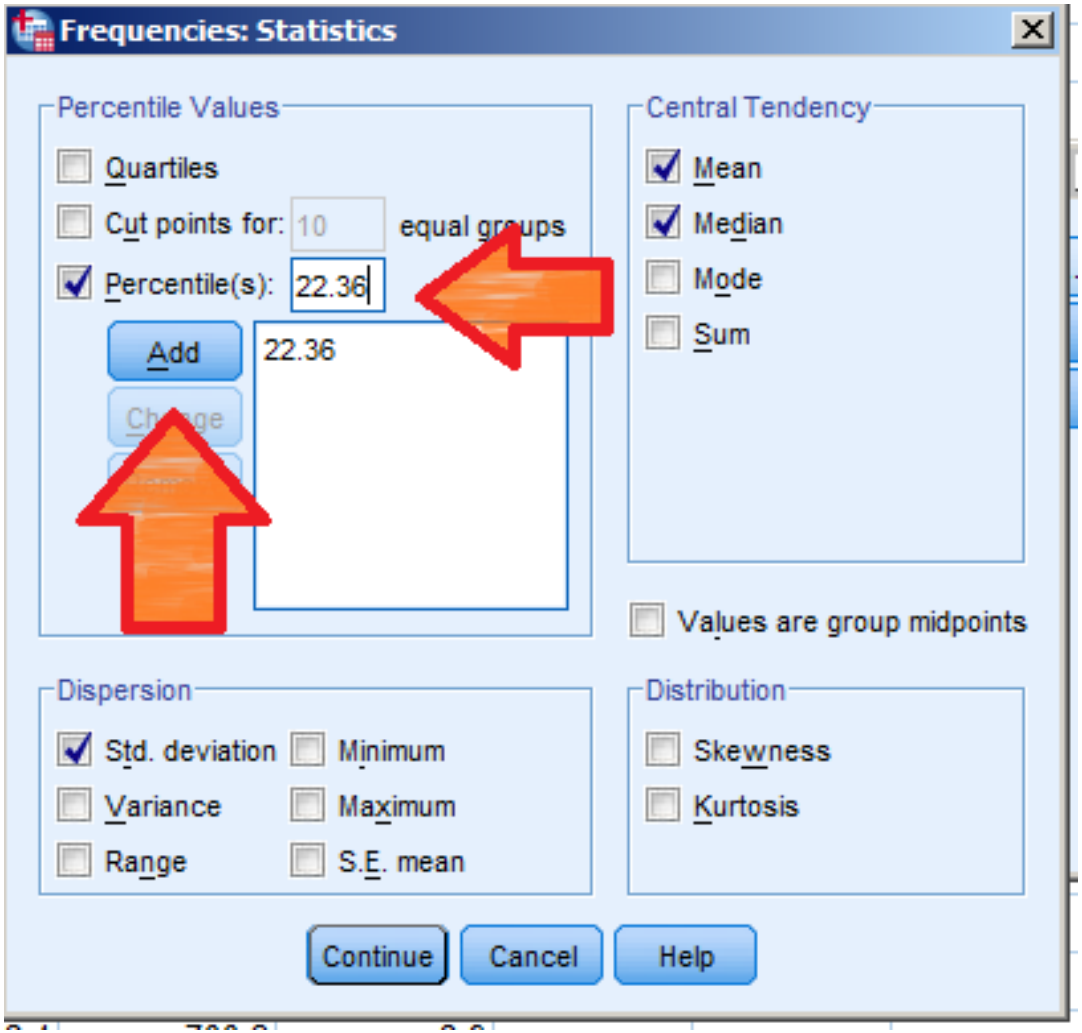
Check off “Std. deviation” and “Variance”



Standard Error of the mean (Sometimes called Standard Error) is also listed here as ***S.E. Mean.***



You can also find percentiles by checking Percentile(s), typing in a value from 0 to 100, and clicking Add.



Final note: You can right-click on, and copy-paste tables and graphs from SPSS into a word document or MS paint program.

Here are the results from directly copying the table from the exercise above and increasing the font size.

Statistics

		X	Y	Z
N	Valid	100	100	100
	Missing	0	0	0
Mean		32.9500	-.1780	4.7477
Median		30.5000	-.2360	2.9067
Std. Deviation		21.14518	.98133	3.48949
Variance		447.119	.963	12.177
Percentiles	25	15.2500	-.8246	1.8547
	50	30.5000	-.2360	2.9067
	75	48.0000	.4913	8.4302

Correlation and Scatterplots

Here we will find correlations and construct scatterplots using the Dragons.sav dataset.

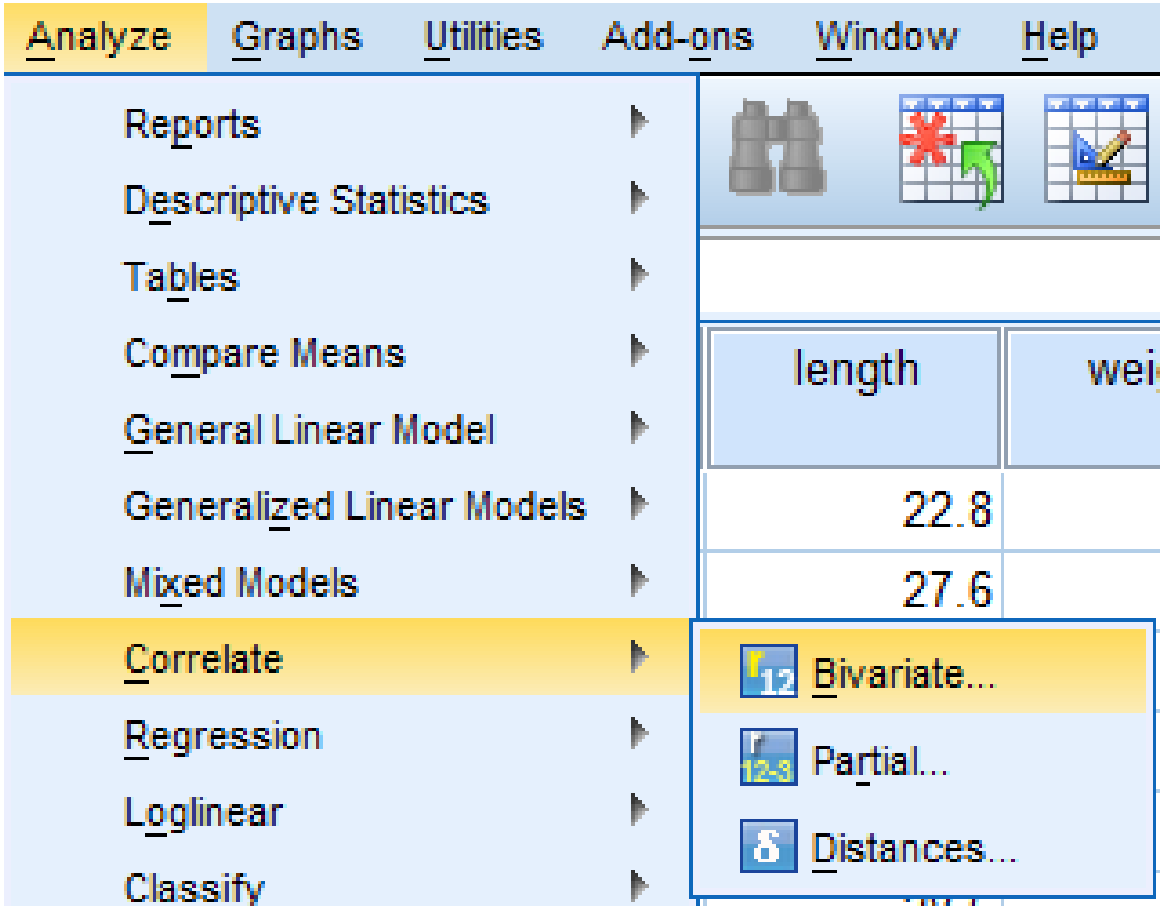
Quick Reference:

Analyze → Correlate → Bivariate

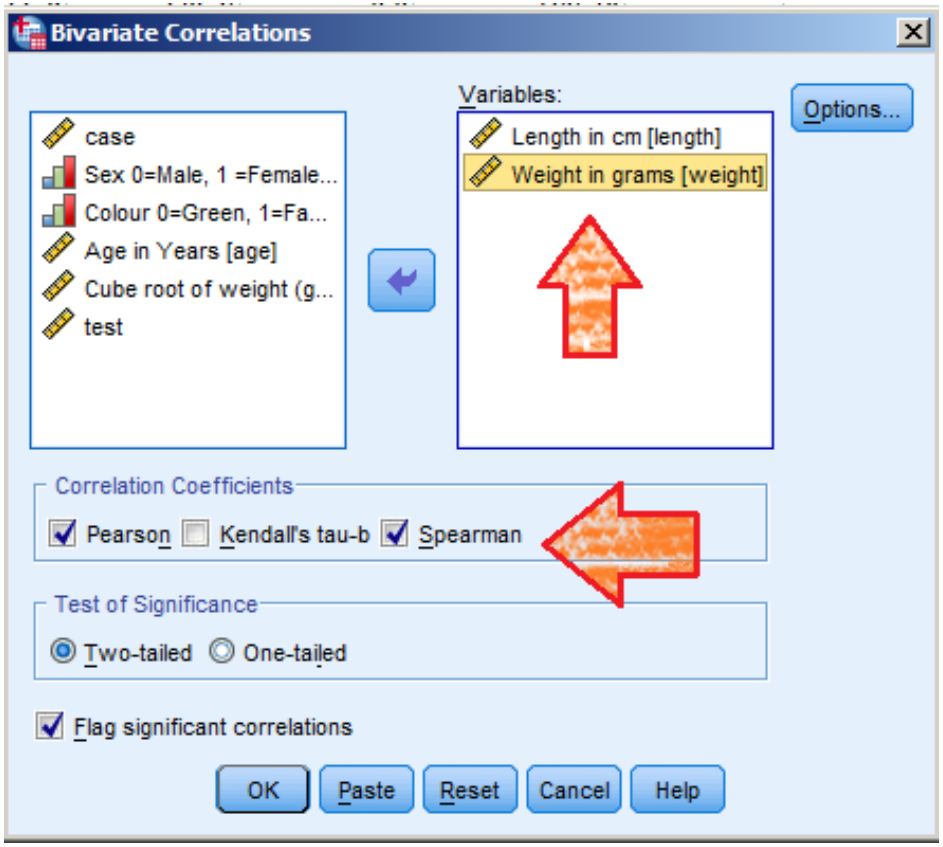
Graphs → Legacy Dialogs → Scatter/Dot

To find a correlation (using Dragons.sav) in SPSS, go to

Analyze → Correlate → Bivariate



Pick the variables you want to correlate, drag them to ***variables.*** Then click OK.



The default coefficient is ***Pearson.***

If you also want the Spearman coefficient check its box.

There is a correlation of ***r = .940*** between weight and height.

It's a significant correlation, with a p-value of less than .001

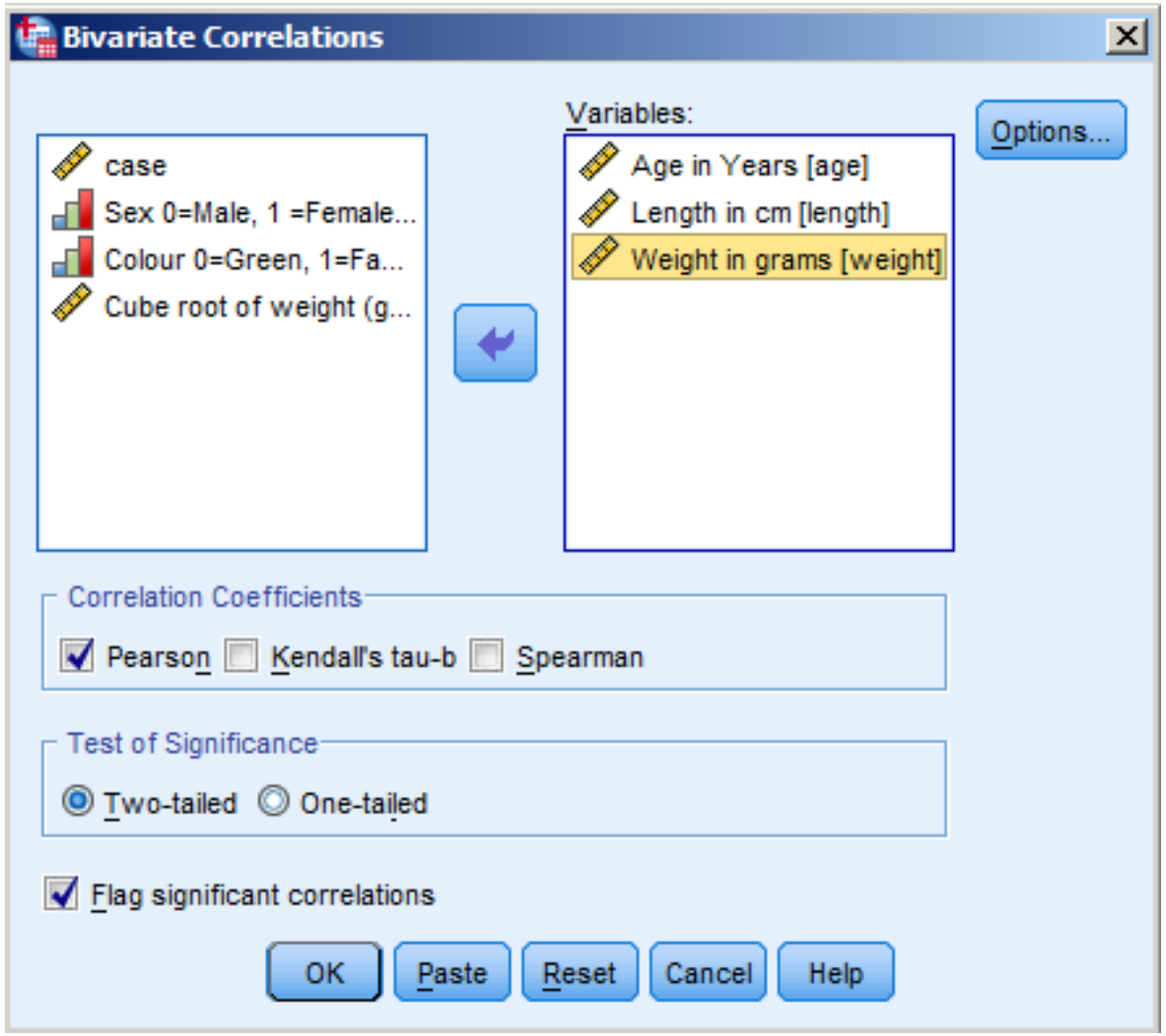
(appears as Sig. (2-tailed) = .000)

Correlations

		Length in cm	Weight in grams
Length in cm	Pearson Correlation	1	.940**
	Sig. (2-tailed)		.000
	N	300	300
Weight in grams	Pearson Correlation	.940	1
	Sig. (2-tailed)	.000	

Also, anything correlates with itself perfectly, so the correlation between length and length is ***r= 1***

You can calculate the bivariate correlation of more than two variables at once by dragging them into ***variables*** .



The table given in the output will be of every pair of variables.

Correlations

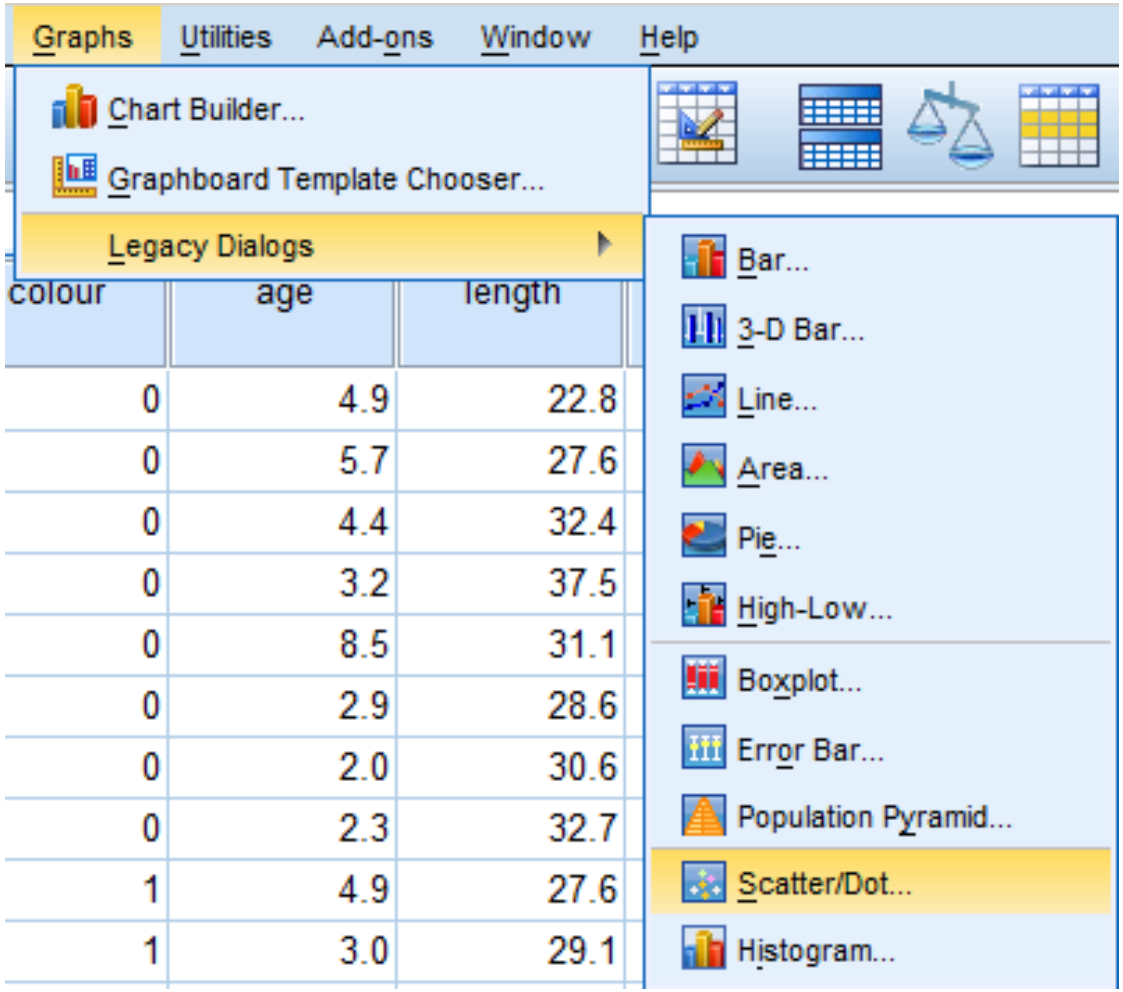
		Age in Years	Length in cm	Weight in grams
Age in Years	Pearson Correlation	1	-.023	.143*
	Sig. (2-tailed)		.690	.013
	N	300	300	300
Length in cm	Pearson Correlation	-.023	1	.940**
	Sig. (2-tailed)	.690		.000
	N	300	300	300
Weight in grams	Pearson Correlation	.143*	.940**	1
	Sig. (2-tailed)	.013	.000	
	N	300	300	300

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

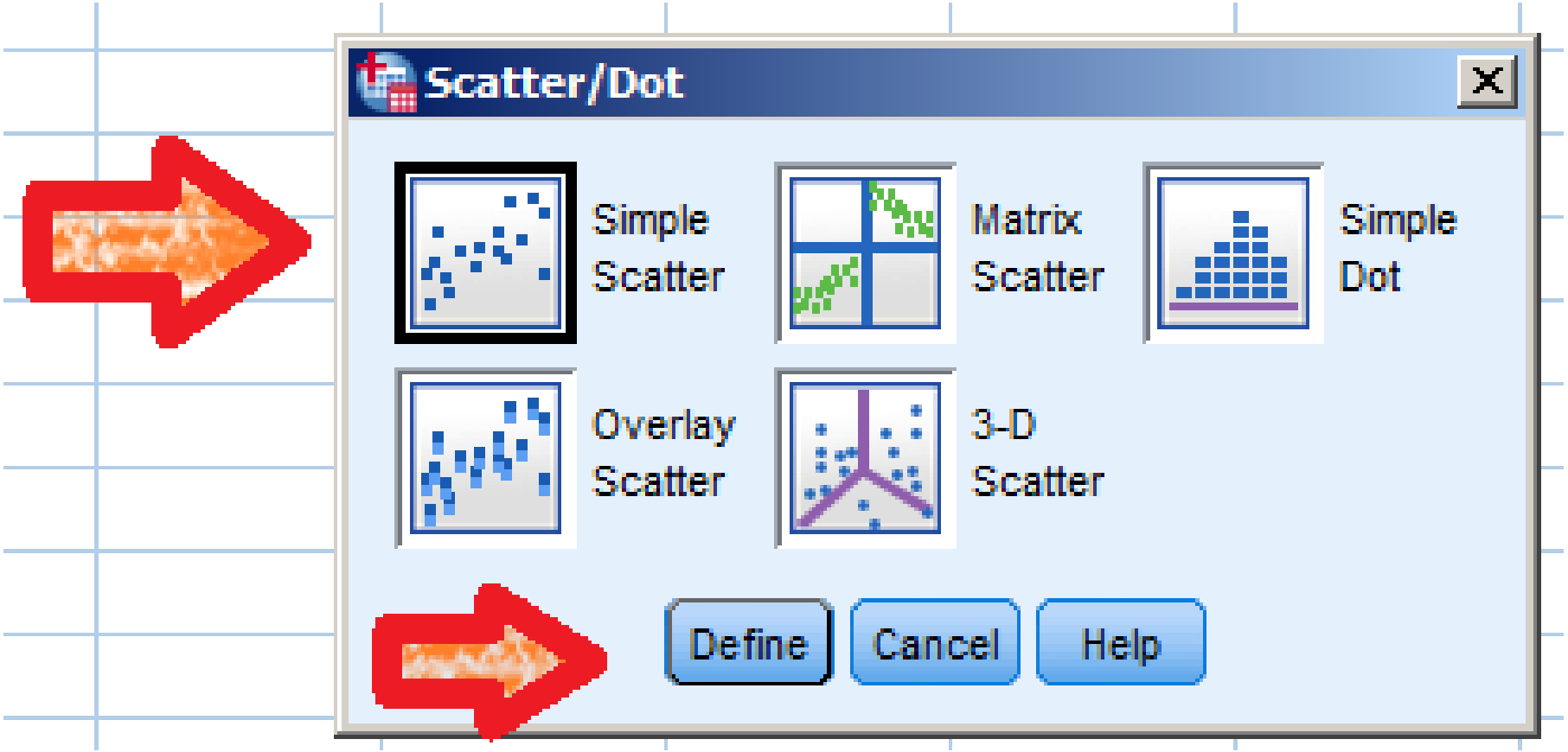
N is the number of cases with BOTH variables.

To build a scatterplot, go to

graphs → legacy dialogs → Scatter/Dot

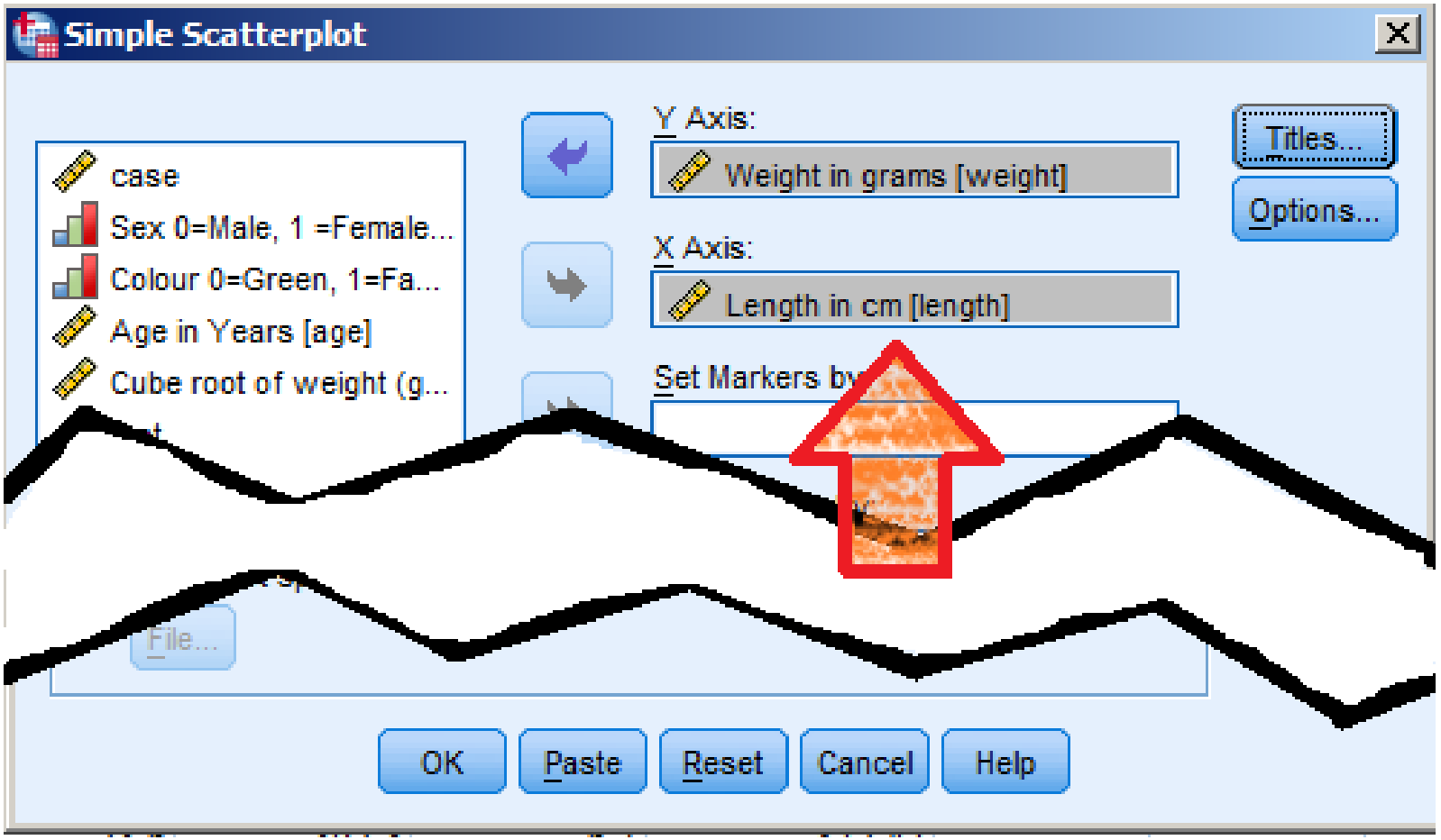


In the dialog, choose Simple Scatter if it's not already picked, and click ***Define.***



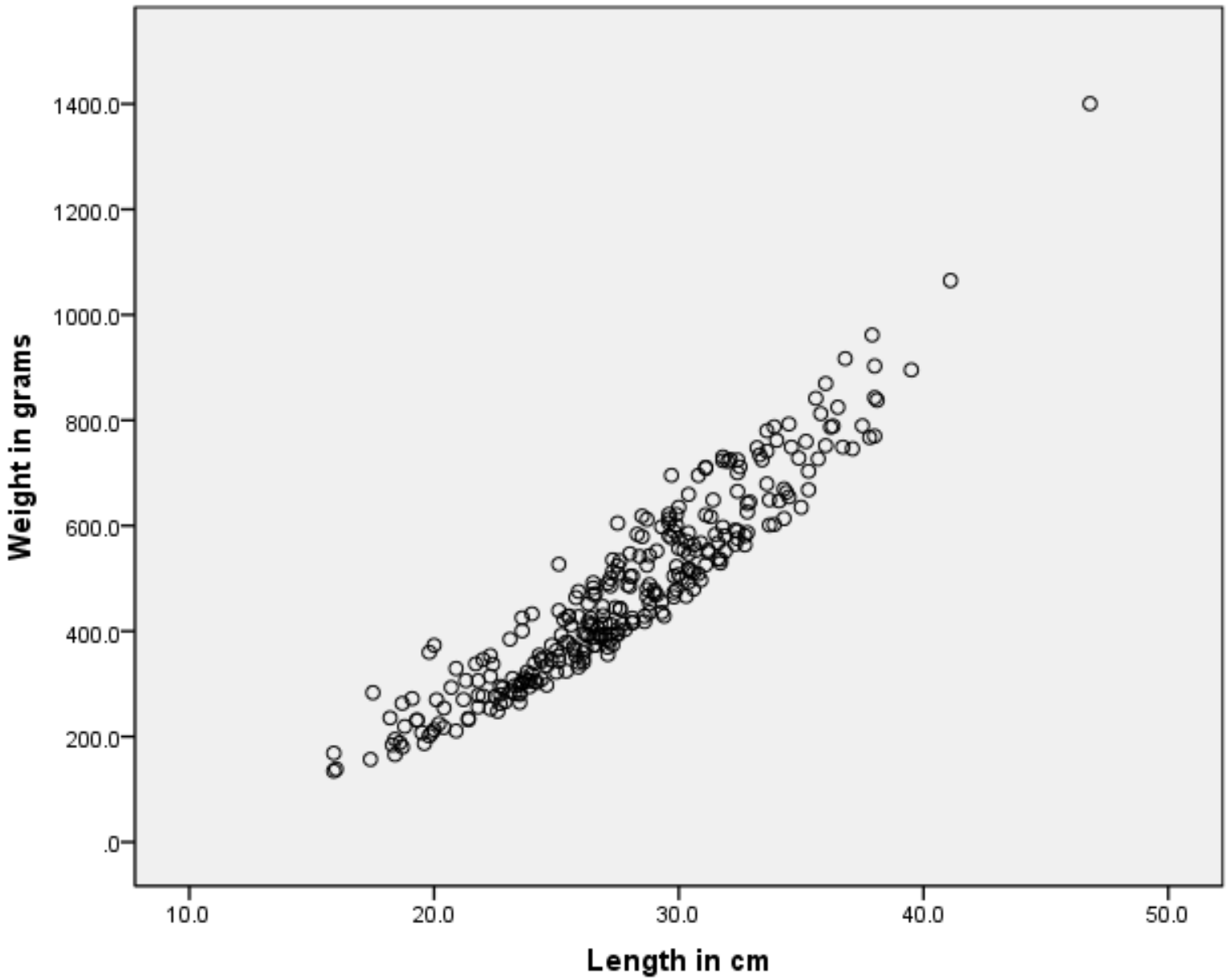
Move the independent variable into the x-axis,

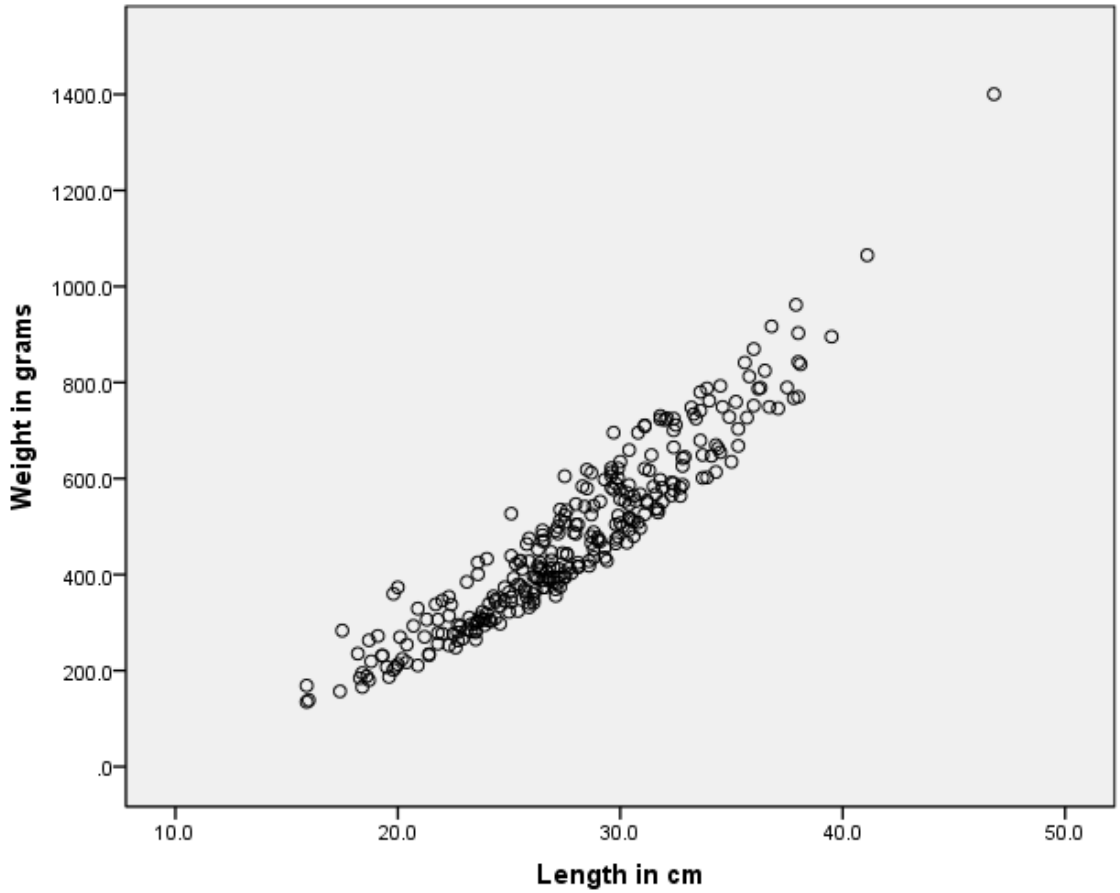
And the dependent variable into the y-axis,



then click OK (way at the bottom)

The result:


Weight in grams
1400.0
1200.0
1000.0
800.0
600.0
400.0
200.0
.0
10.0
20.0
30.0
40.0
50.0
Length in cm



Each dot represents a case, the farther right it is the longer that dragon. The farther up it, the heavier that dragon. Compare the scatterplot to the correlation between length and weight shown earlier in this section.

Regression

In this section we investigate further the relationship between the variables in the Dragons.sav dataset.

We look at simple regression,
regression on a dummy variable,
multiple regression,
drawing the regression line and
building the residual plot.

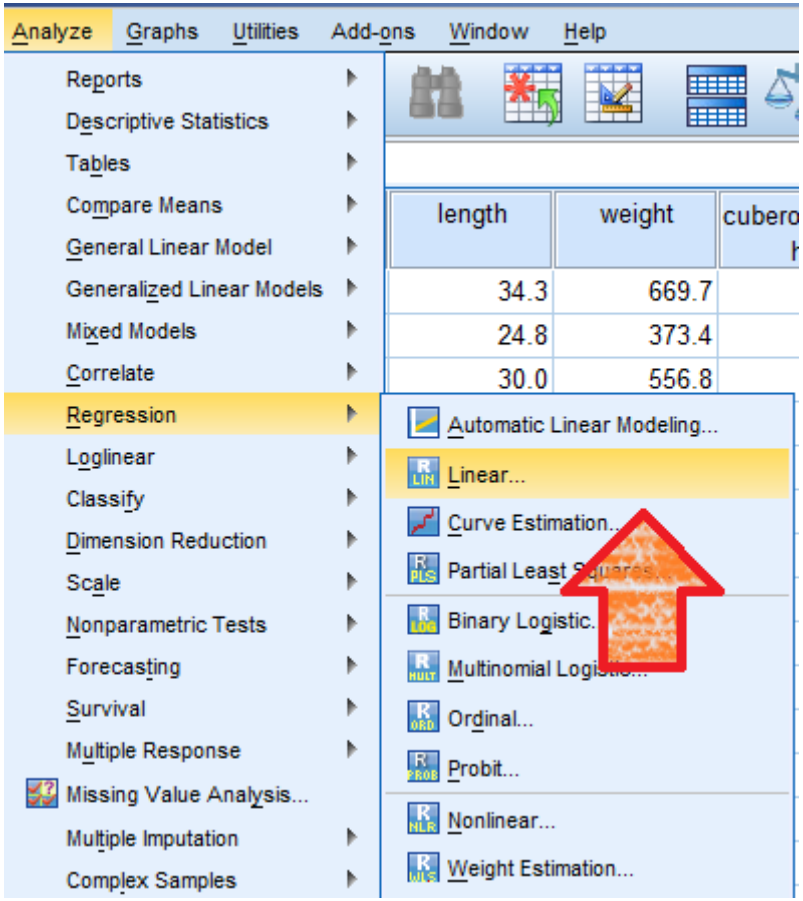
Quick Reference:

Analyze → Regression → Linear

Graphs → Legacy Dialogs → Scatter/Dot

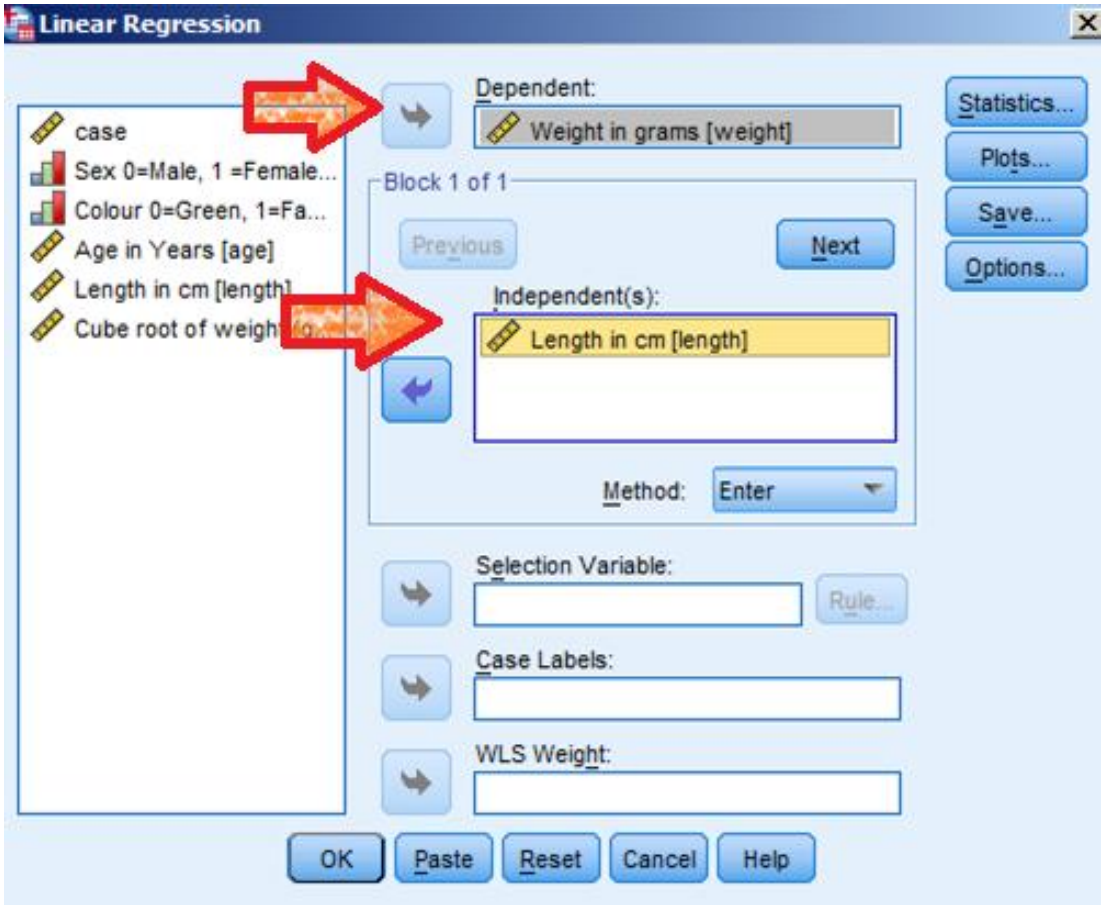
The regressions will be done through

Analyze → Regression → Linear



For a simple regression, put your response variable (Weight) in the ***dependent*** slot.

Put your explanatory variable in the ***Independent*** box.



After clicking OK, these are the results.

The model summary tells you what proportion of the variance in the response was explained by your explanatory variable as ***R-squared***

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.940[a]	.885	.884	62.6228

a. Predictors: (Constant), Length in cm

In a simple regression, this should be the Pearson correlation squared.

The coefficients table is the important one.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-470.788	20.220		-23.283	.000
	Length in cm	34.154	.715	.940	47.771	.000

a. Dependent Variable: Weight in grams

In the ***Unstandardized Coefficients B*** column,

(Constant) is the intercept.

Length in cm is the slope.

The Dependent variable is mentioned at the bottom.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-470.788	20.220		-23.283	.000
	Length in cm	34.154	.715	.940	47.771	.000

a. Dependent Variable: Weight in grams

In this case, a bearded dragon with zero length weights -470 grams on average.

For every increase of 1 cm of length, the average weight increases by 34.154 grams.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-470.788	20.220		-23.283	.000
	Length in cm	34.154	.715	.940	47.771	.000

a. Dependent Variable: Weight in grams

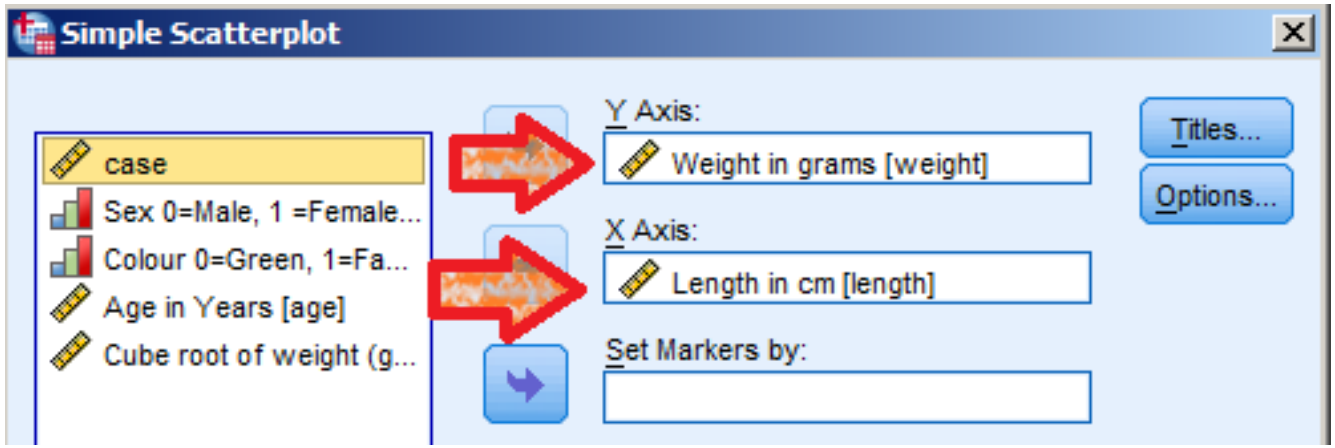
We can also see that p-value is less than .001 (Sig. = .000), this is strong evidence against the null hypothesis that the true slope is zero.

The large t-score of 47.771 of the slope also indicates a slope. t = 0 would indicate absolutely no evidence

We can draw a scatterplot with the regression equation.

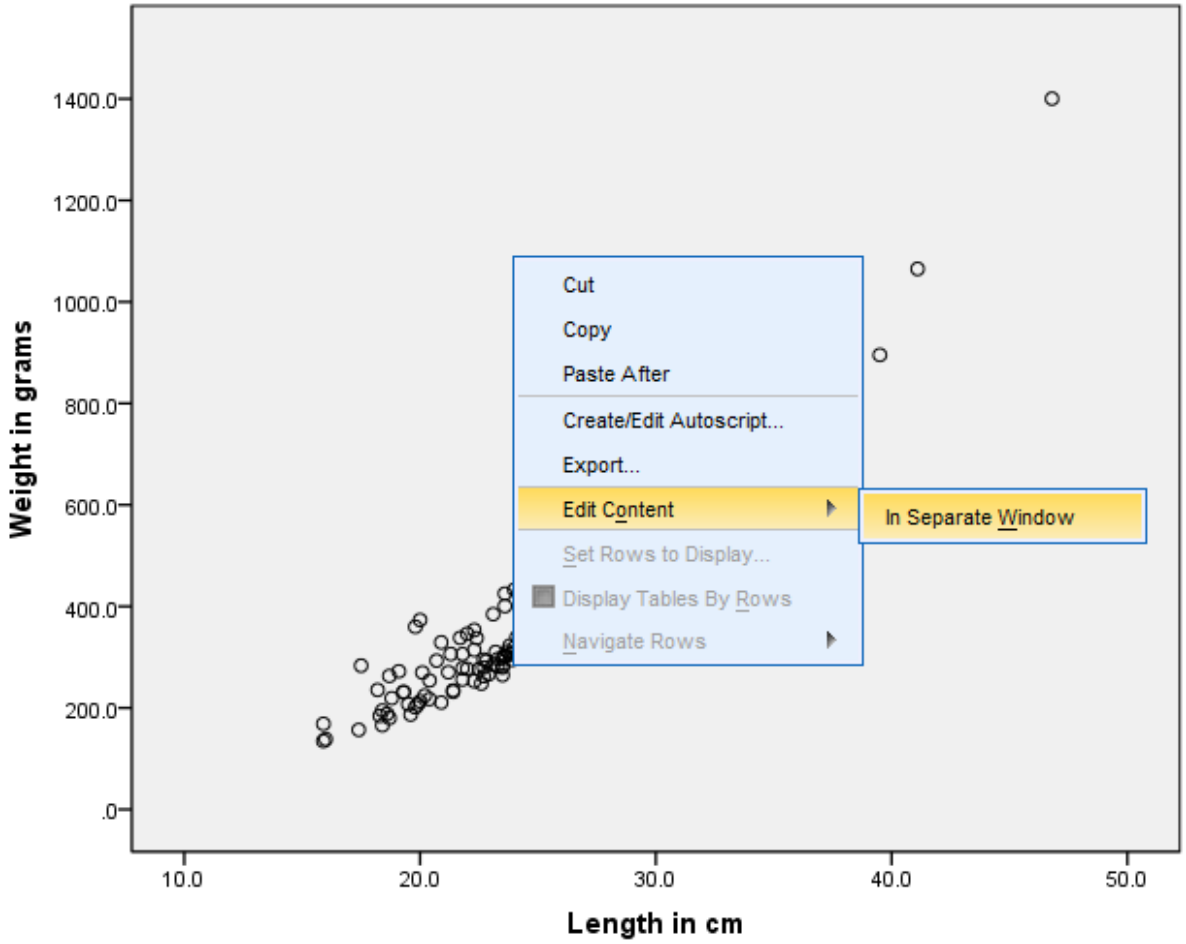
First, build a scatterplot with Graphs → Legacy Dialogs → Scatter/Dot.

Use the same response/dependent for Y and explanatory/independent for X as you did in the regression.



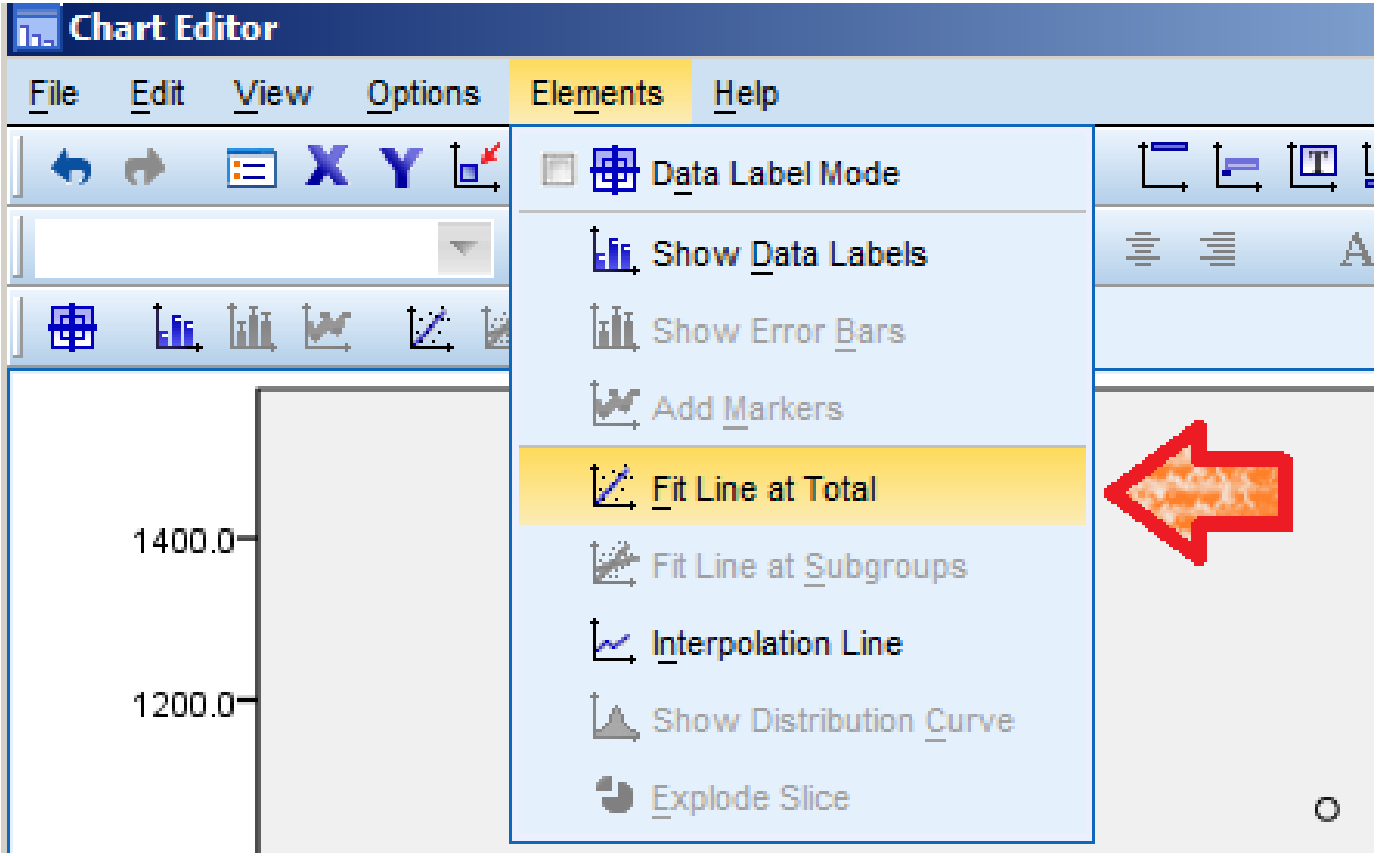
In the output, Right-Click on the Scatterplot and choose

Edit Content → In Separate Window.

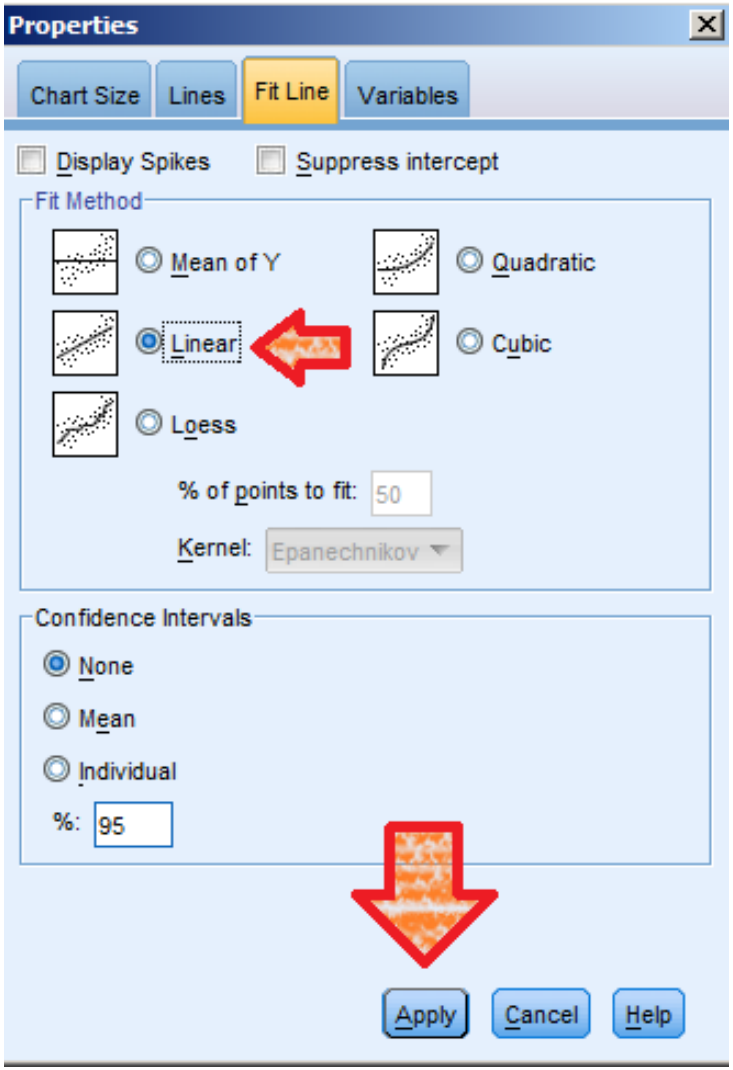


In the chart window that pops up, choose

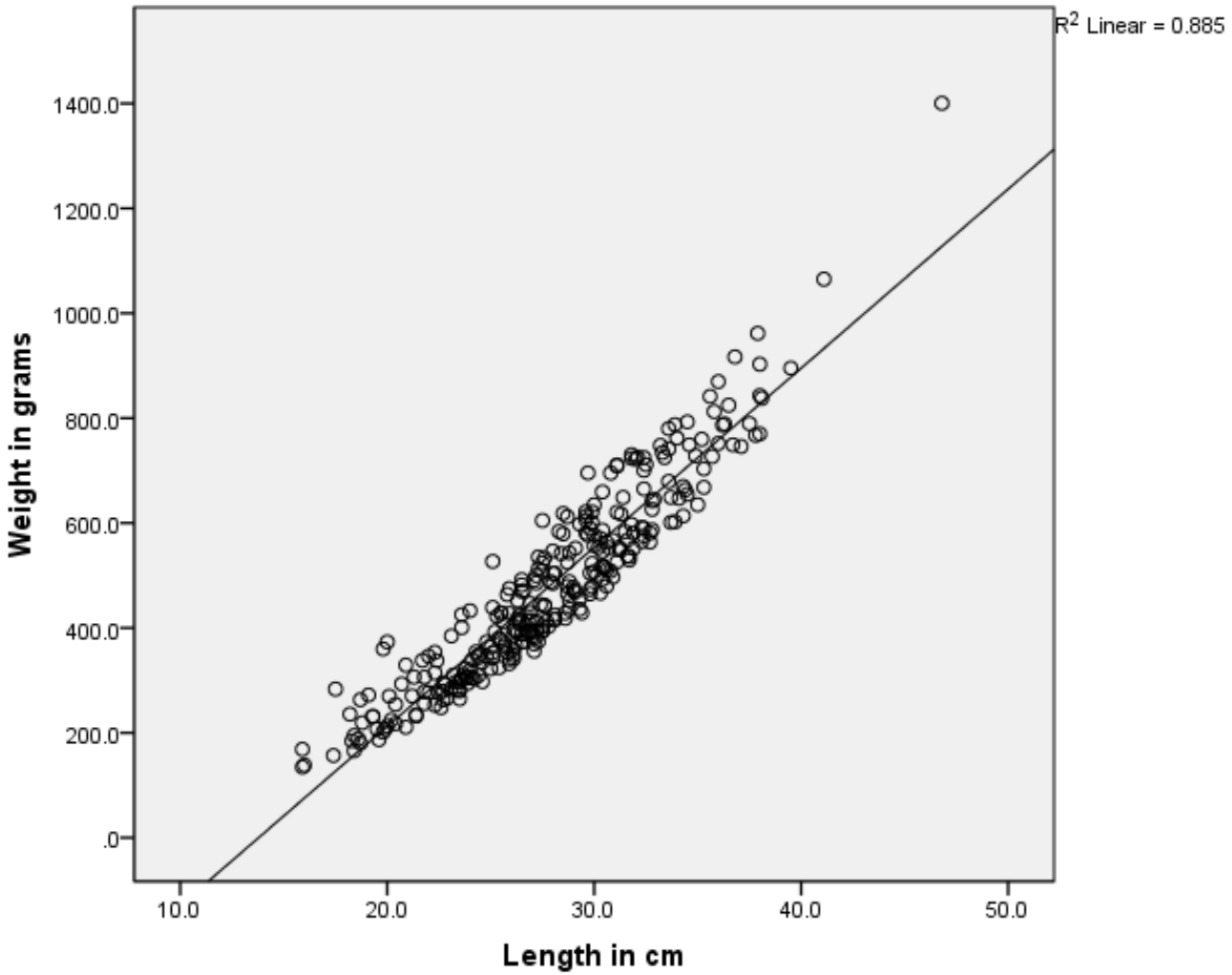
Elements → Fit Line at Total



We're doing a linear regression, so the fit line should be set to ***Linear*** (this is the default), then click ***Apply***.

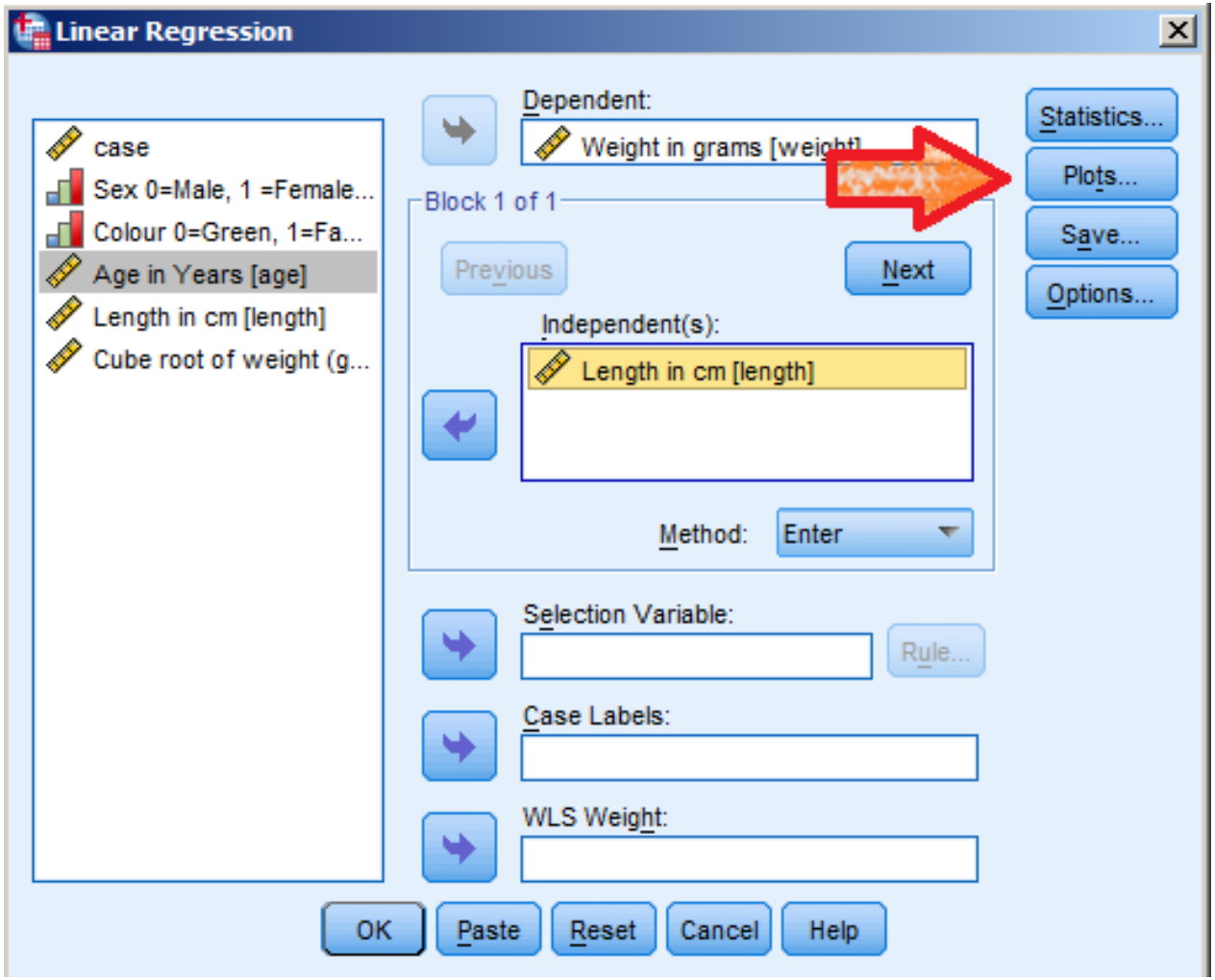


Close the chart editor and the scatterplot with the regression line will remain.



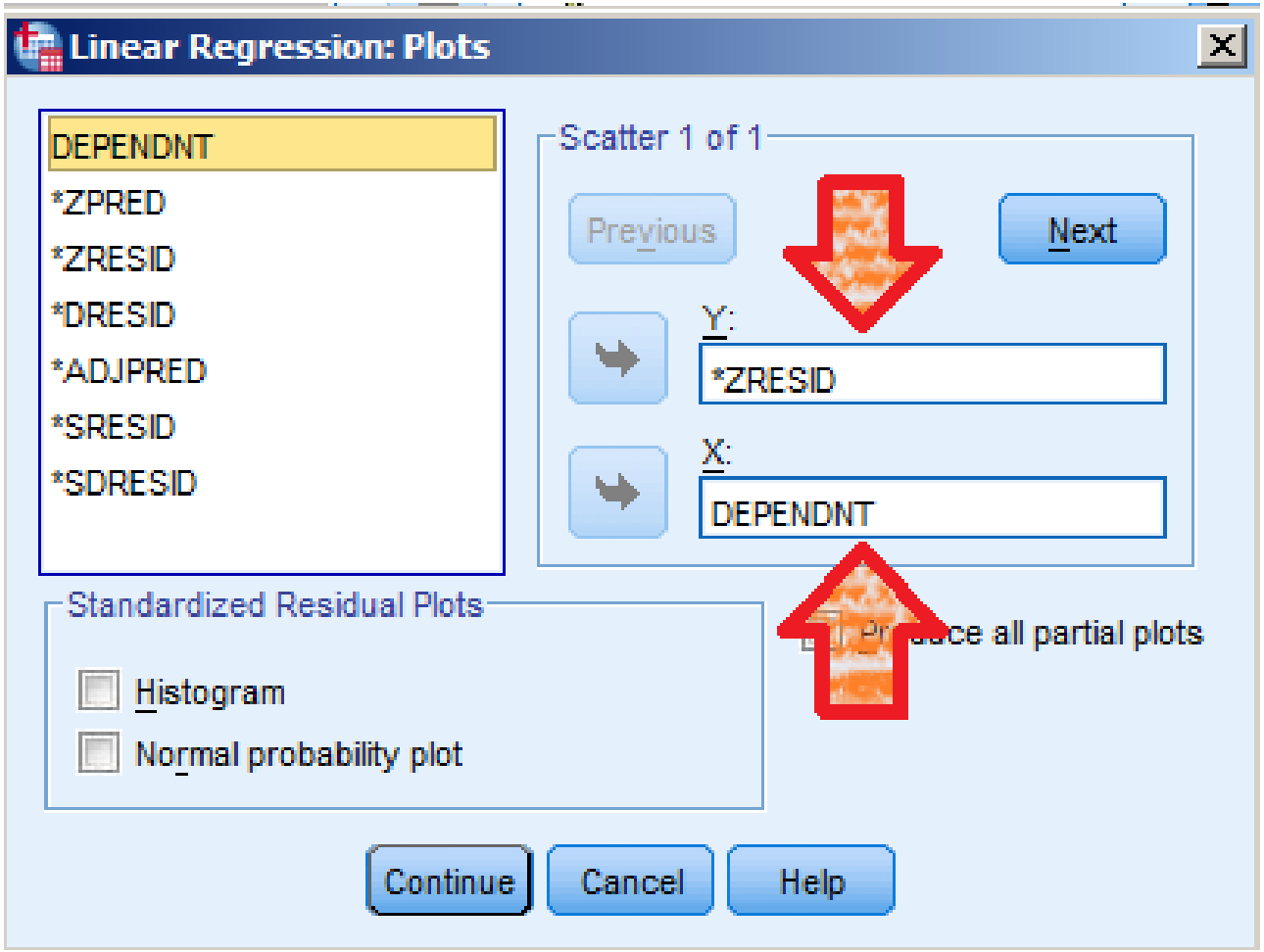
Another graph we may be interested in is the residual graph.

When setting the variables for regression, click on ***Plots***

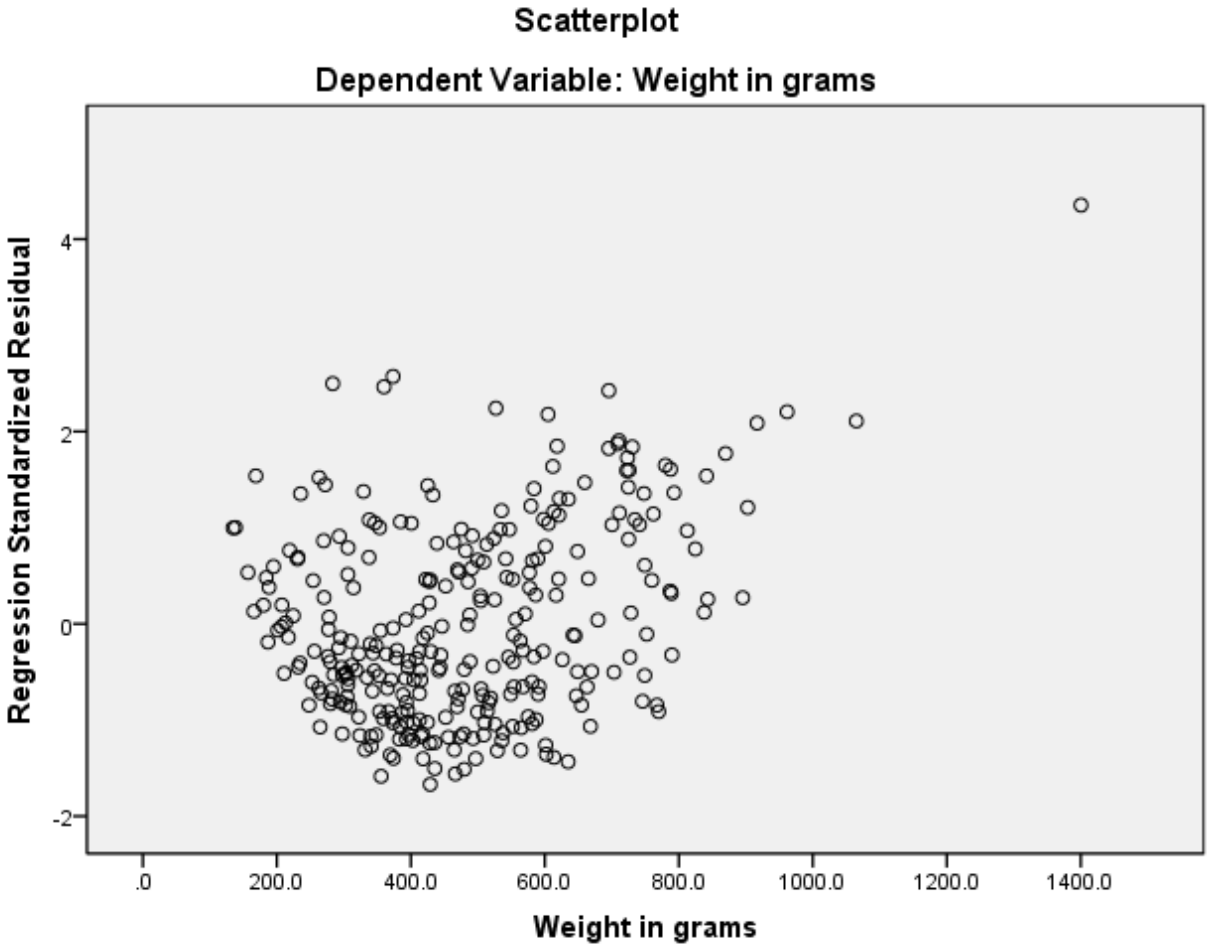


Then put ***ZRESID*** in the Y slot, and

DEPENDNT in the X slot. Then click Continue and OK.



Along with the rest of the regression output, the residual graph will appear. There should be no obvious patterns if the regression works. In weight vs. length, there could be issues.



When editing a plot in a separate window, some other options you have include…

Changing the bounds of the graph.

Edit → Set Y-Axis, and ***Edit → Set X-Axis***

Put a reference line (especially useful for residual plots)

Options → Y-Axis reference line, then click ***Apply***

Other Regression Examples

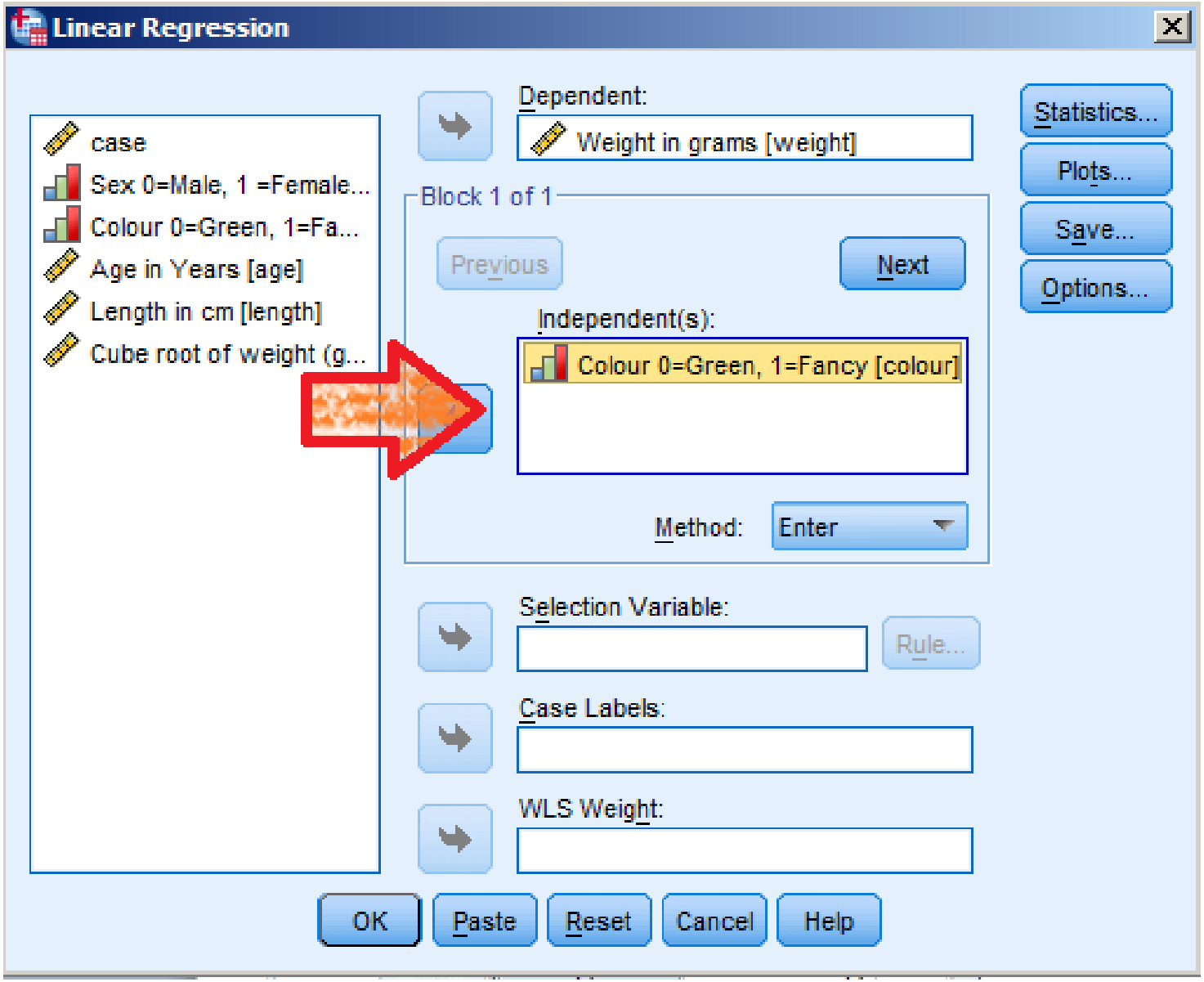
We can also do regression on dummy variables. These are variables coded so that 0 means one category and 1 means another. Example: Weight as a function of colour.

Weight is interval, and colour is a dummy variable.

0 = Green

1 = “Fancy”, anything but green.

Here it is in Analyze → Regression → Linear. Then Click OK.



Here is the output for Weight vs. Colour.

Coefficients[a]

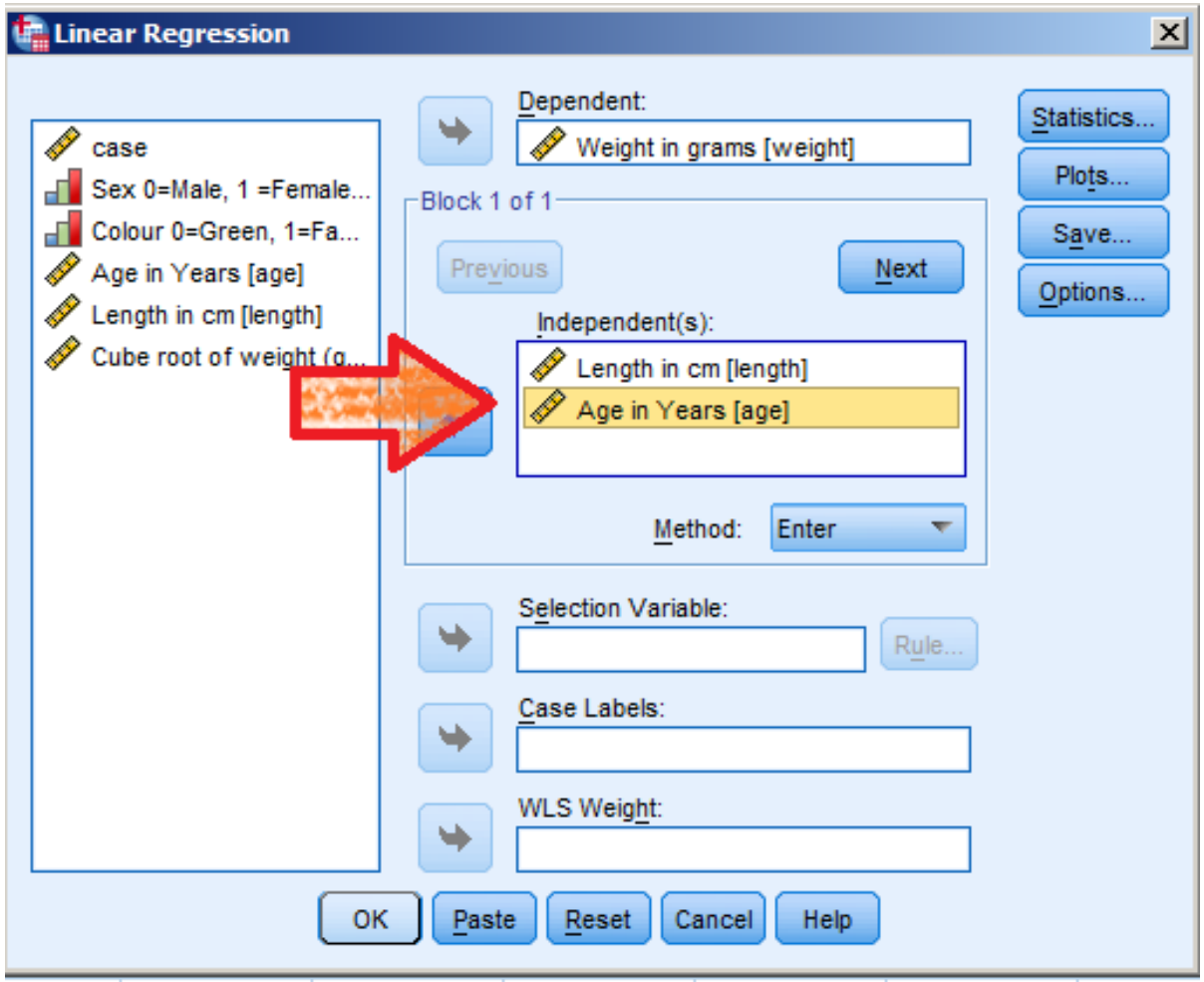
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	448.061	12.271		36.514	.000
	Colour 0=Green, 1=Fancy	105.059	22.403	.262	4.689	.000

a. Dependent Variable: Weight in grams

The intercept (448.061) is the average weight when all the explanatory variables are 0. In this case that just means the average green dragon weight.

The slope (105.059) is the average amount a fancy dragon is heavier than a green dragon.

Multiple explanatory variables can be included by putting then both in the ***independent(s)*** box.



This is weight as explained by Length AND Age together.

The output gives an intercept and two slopes, one for each explanatory variable.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-550.339	19.594		-28.088	.000
	Length in cm	34.292	.627	.944	54.699	.000
	Age in Years	17.336	1.820	.164	9.527	.000

a. Dependent Variable: Weight in grams

The slope for length here is the amount weight increases on average as length increases AND while age stays the same.

Crosstabs, Odds Ratio, Chi-Squared

Crosstabs are tables of combinations of two categorical variables. Odds ratio and Chi-Squared are tool we can use to investigate the relationship between these variables.

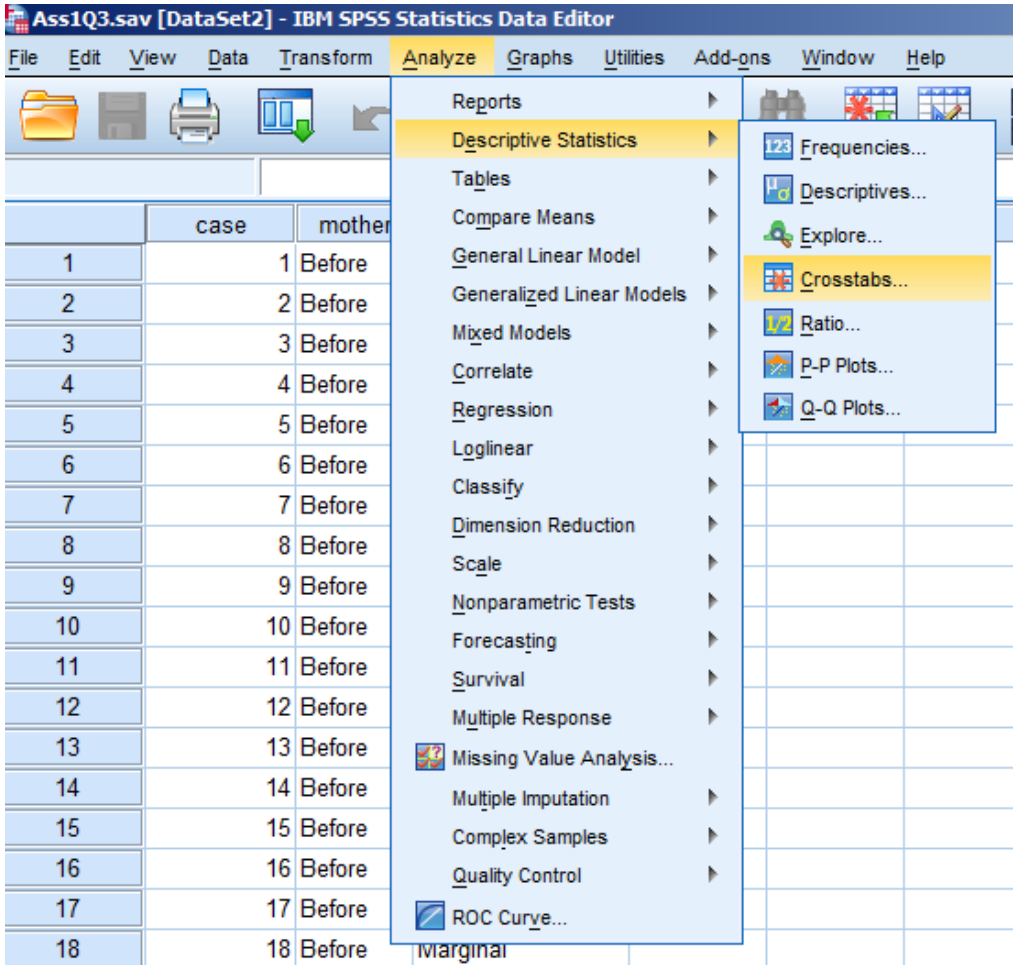
Here we do a 2x2 crosstab and compute the odds ratio, Do a 3x3 crosstab and calculate the expected frequencies and chi-squared, and merge two categories together to fix an potential problem.

Quick Reference:

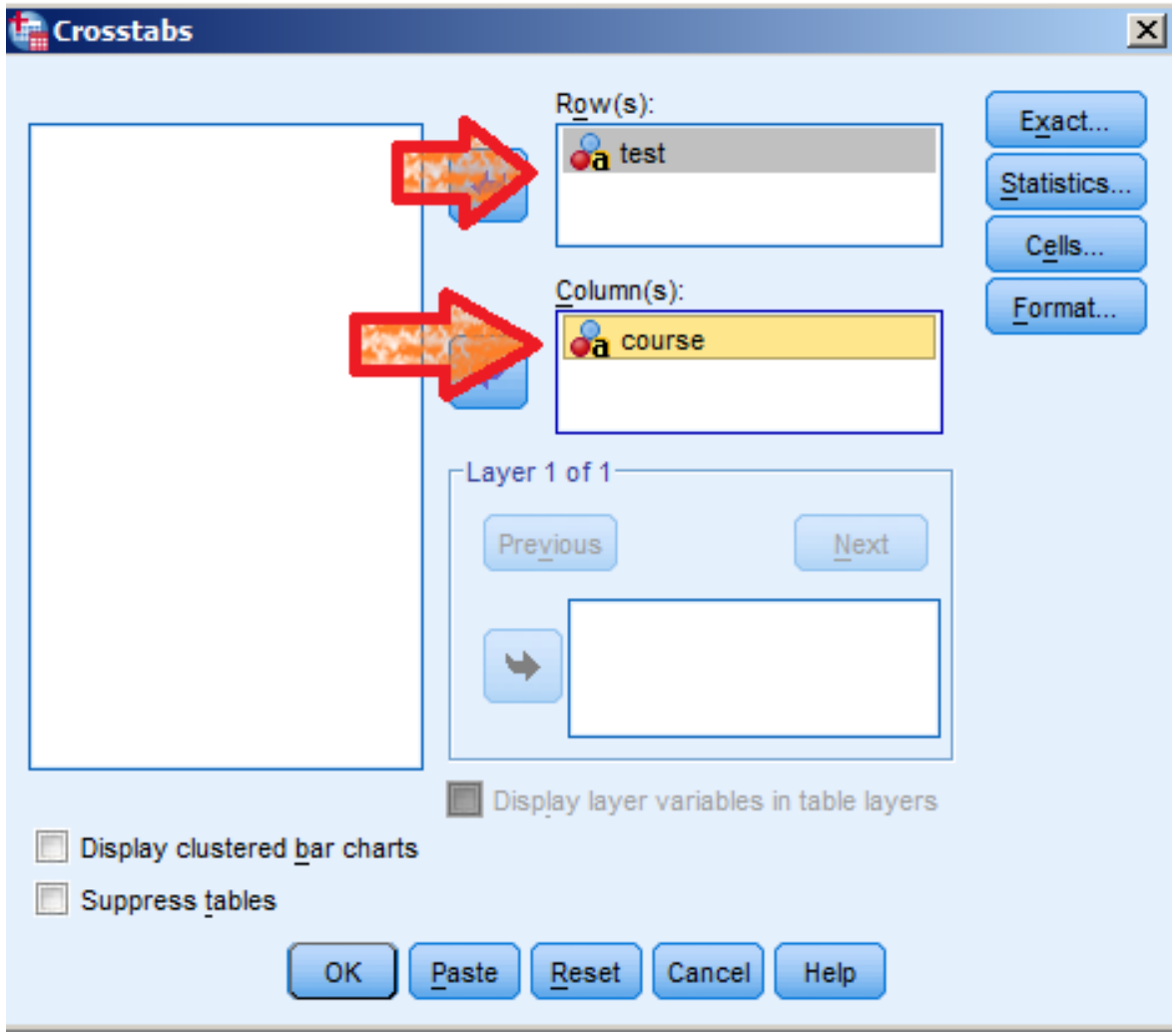
- Analyze → Descriptive Statistics → Crosstabs

First, a 2x2 crosstab from Ch9_15.sav, based on the textbook exercise 9.15. (Taken Driver course vs. Passing Test)

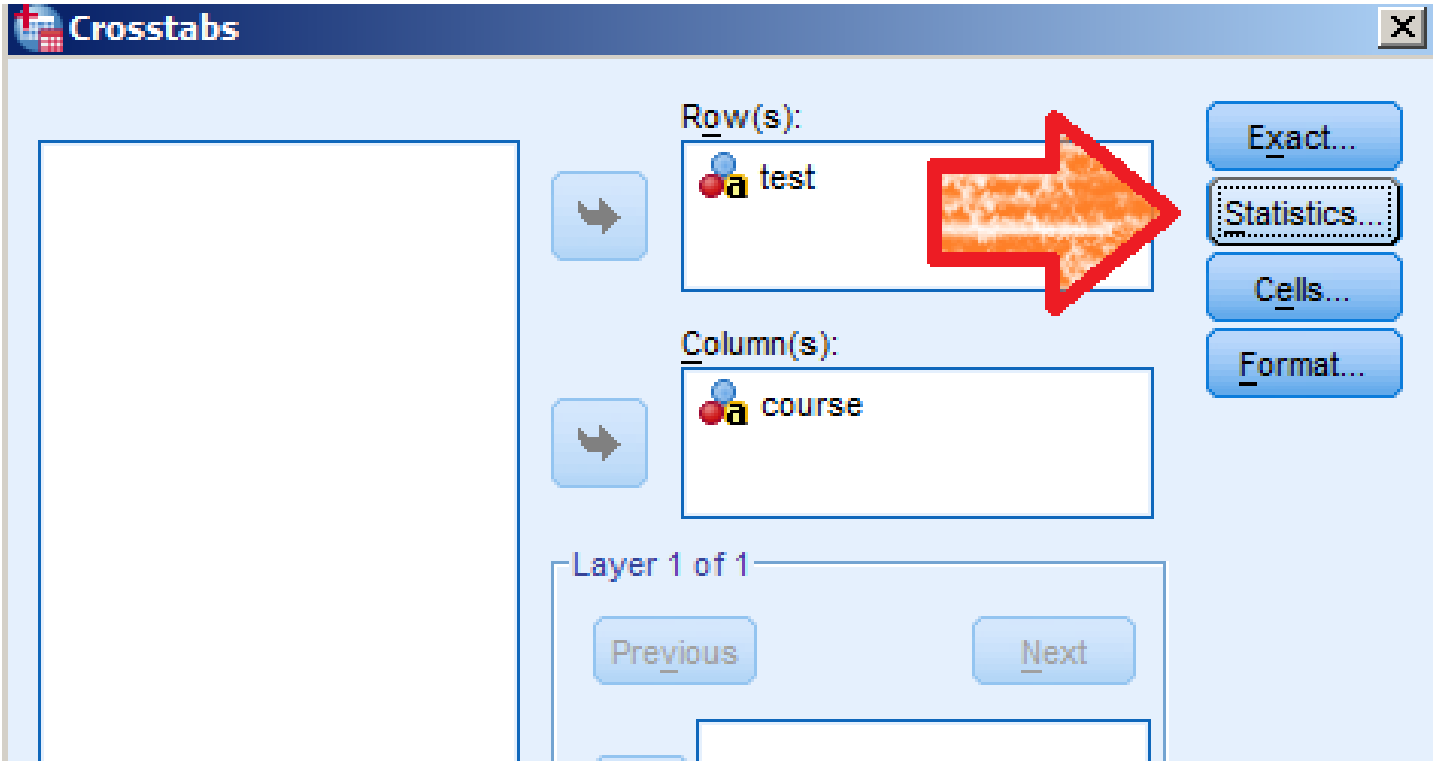
Analyze → Descriptive Statistics → Crosstabs



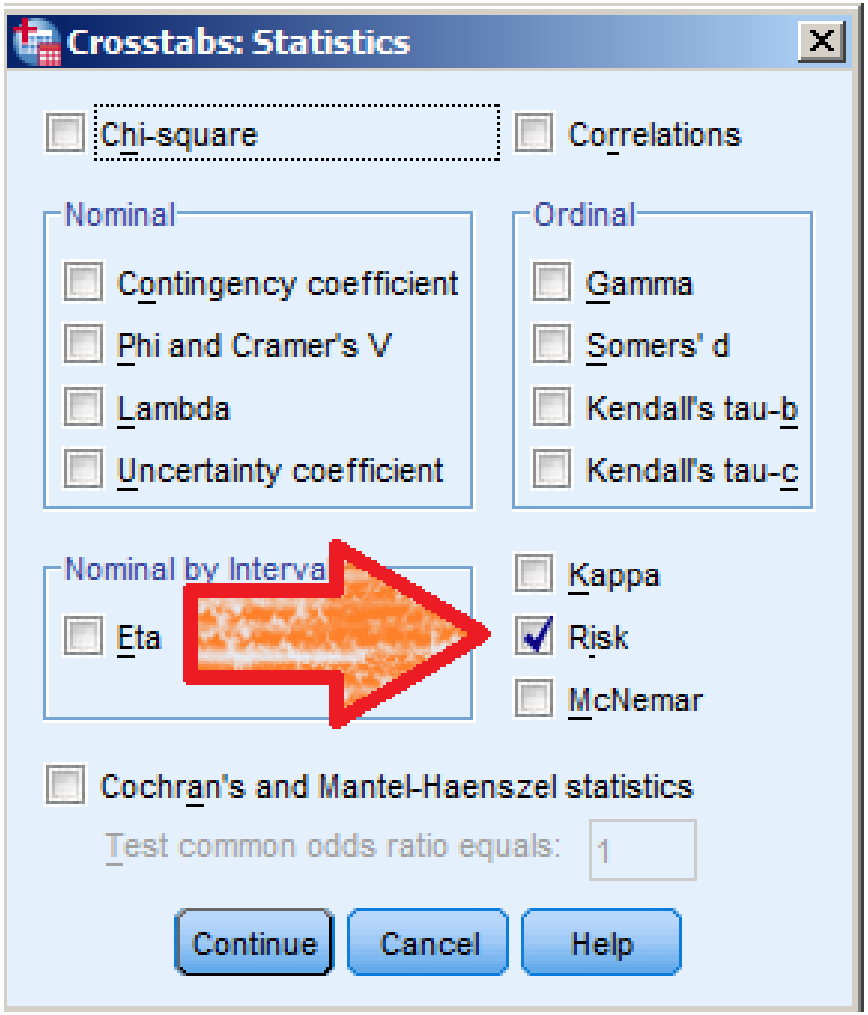
- In the crosstab dialog, move one variable to ***rows*** and one to ***columns***.



- For the odds ratio, click on the ***statistics*** button in the upper right.



- To calculate the odds ratio, check off ***risk.***
- It's under "risk" because odds ratio is related to relative risk.



-

- In the output window, the crosstab will appear with the labels instead of the variable names if you set them.

test * course Crosstabulation

Count

		course		
		No	Yes	Total
test	Fail	11	8	19
	Pass	7	16	23
Total		18	24	42

Another name for a crosstab is a ***contingency plot.***

Here, there are 24 people that took the driver's course.

Of those 24, 16 of them passed.

The output will also include the odds ratio since we checked the Risk box. Here is the odds ratio table.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for test (Fail / Pass)	3.143	.881	11.215
For cohort course = No	1.902	.919	3.936
For cohort course = Yes	.605	.335	1.095
N of Valid Cases	42		

In the sample, the odds are 3.143 times as high that someone will pass their driver exam if they've taken the course.

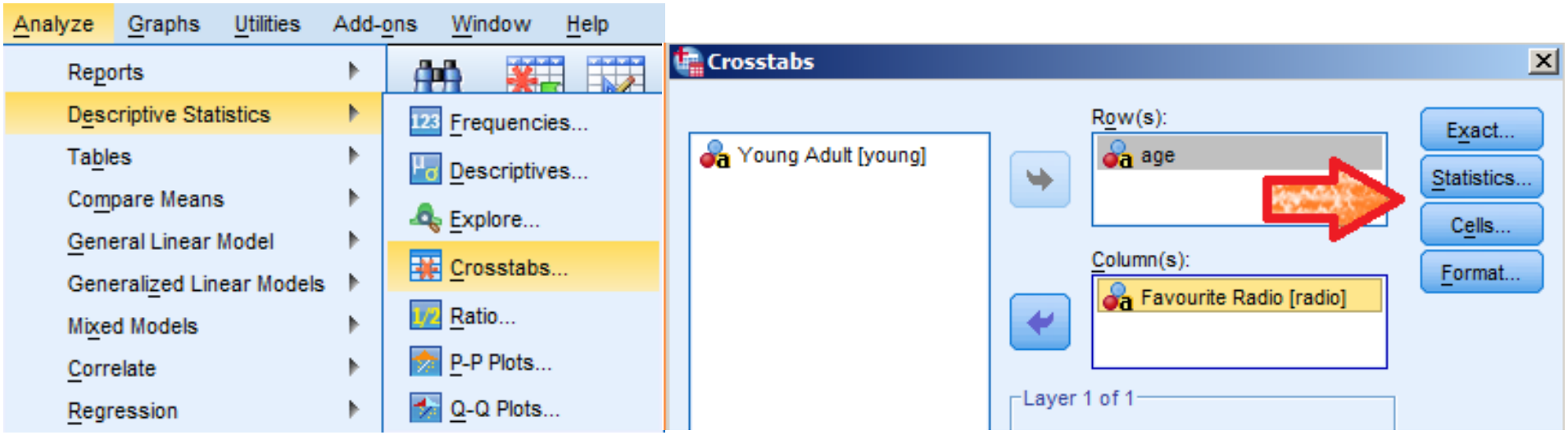
We also have the confidence interval of the population odds ratio.

Now let's try with a 3x3 cross tab, Ch9_21.sav, based on music choices and age groups.

We start the same way:

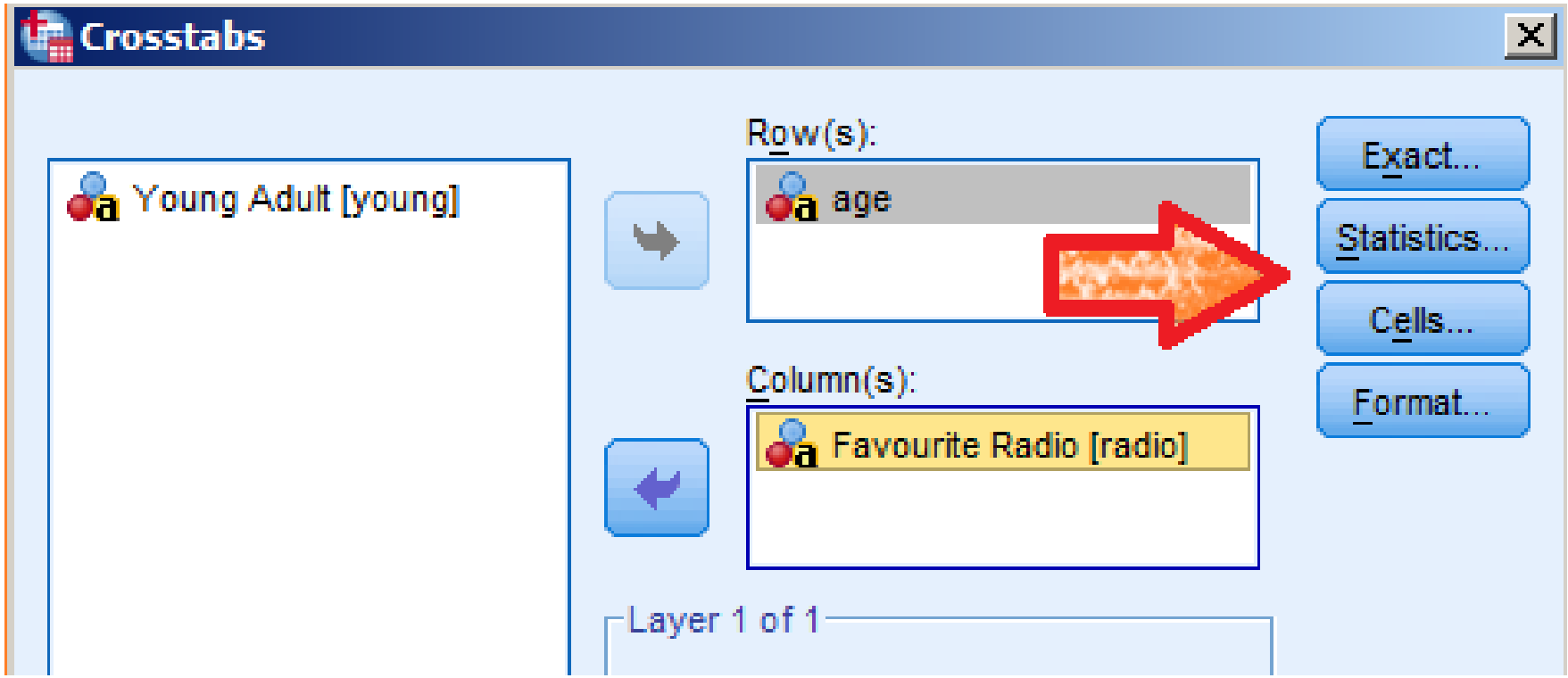
Analyze → Descriptive Stats → Crosstabs

Then put one variable in row, and the other in column.

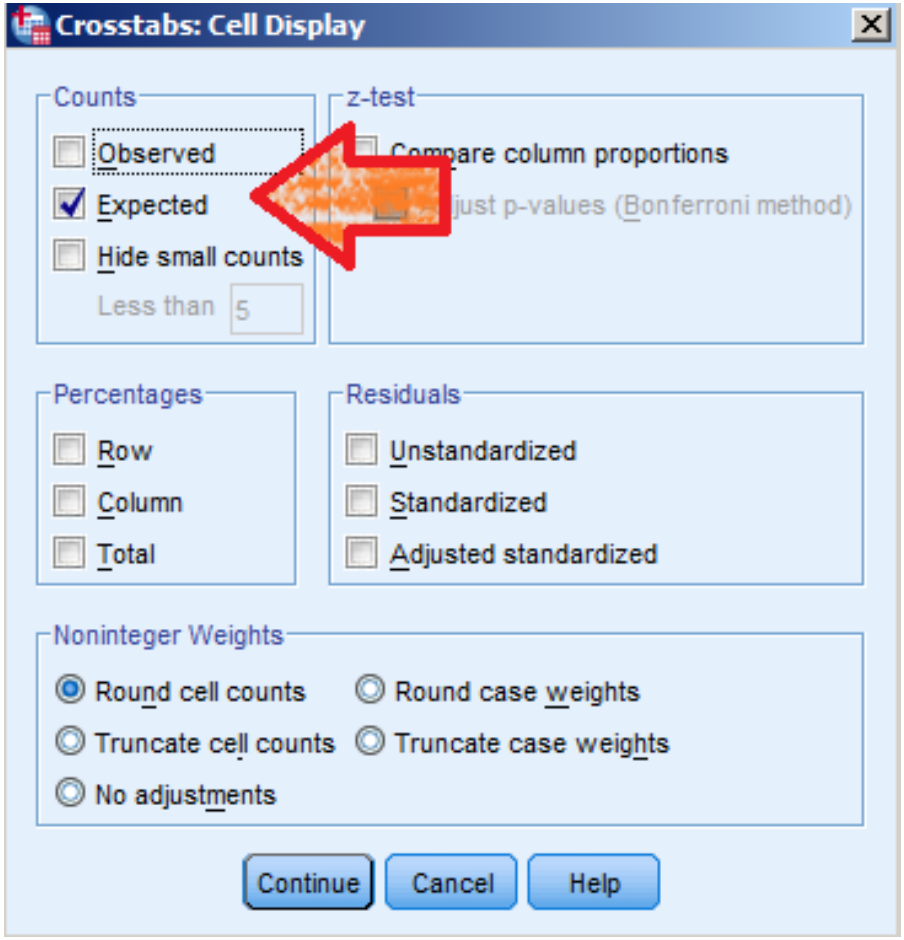


The odds ratio doesn't make sense for a 3x3, but we can calculate the expected values and the chi-squared statistic.

First, go to ***Cells.***



From the cells menu, you can decide you want to see the observed frequencies (on by default), the expected frequencies (off by default), or both.



Here are the observed frequencies from the output window.

Favourite Radio * age Crosstabulation

Count

		age			Total
		1Young	2MiddleAge	3OlderAdult	
Favourite Radio	Music	14	10	2	26
	News	4	15	8	27
	Sports	7	9	3	19
Total		25	34	13	72

Here are the expected frequencies. Note that the totals are the same in both tables.

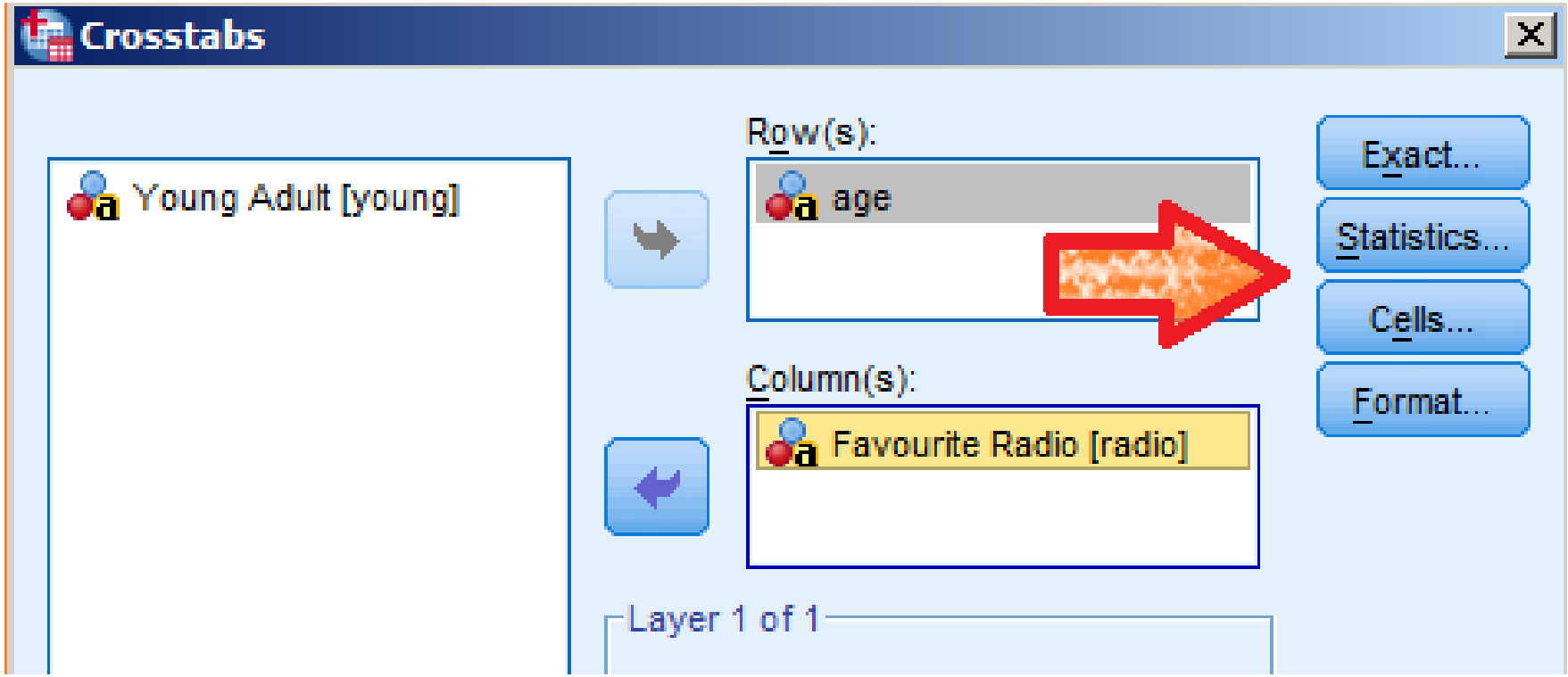
age * Favourite Radio Crosstabulation

Expected Count

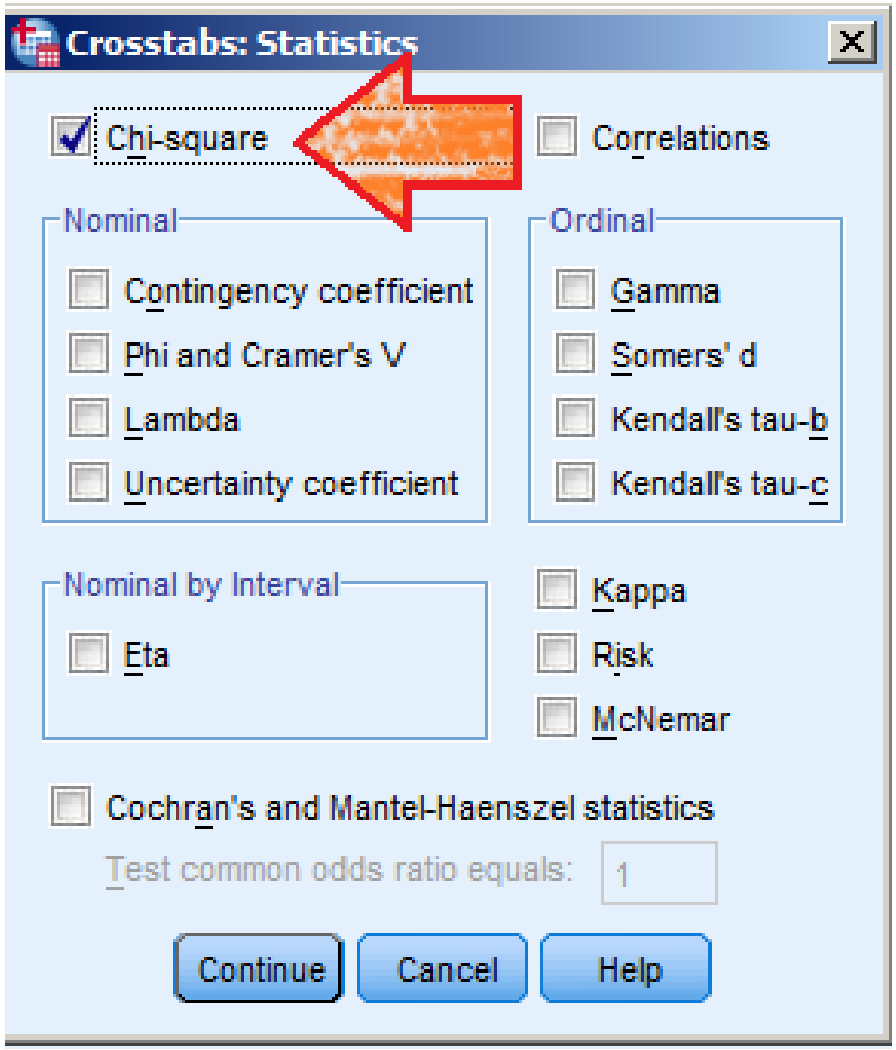
		Favourite Radio			Total
		Music	News	Sports	
age	MiddleAge	12.3	12.8	9.0	34.0
	OlderAdult	4.7	4.9	3.4	13.0
	Young	9.0	9.4	6.6	25.0
Total		26.0	27.0	19.0	72.0

Click ***Continue*** to get out of the Cells menu.

From the main crosstabs dialog, you can also calculate the chi-squared statistic by clicking on the ***Statistics*** button.



Then, put a check next to ***Chi-Square*** in the upper left.



Then click Continue, then OK.

Checking Chi-Squared produces the following table.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.268[a]	4	.036
Likelihood Ratio	10.835	4	.028
N of Valid Cases	72		

a. 3 cells (33.3%) have expected count less than 5. The minimum expected count is 3.43.

We want the ***Pearson Chi-Square***

χ^2 = ***10.268*** and df = ***4***.

The p-value against independence is ***.036.***

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	10.268[a]	4	.036
Likelihood Ratio	10.835	4	.028
N of Valid Cases	72		

a. 3 cells (33.3%) have expected count less than 5. The minimum expected count is 3.43.

The Chi-Square test also tells us of potential problems.

The test assumes there is a large number of respondents in each cell. The standard rule is that every cell should have a frequency of ***at least 5.***

There are ways to deal with cells with small n. The easiest one is to find a logical way to ***group categories together.***

Favourite Radio * age Crosstabulation

Count

		age			Total
		1Young	2MiddleAge	3OlderAdult	
Favourite Radio	Music	14	10	2	26
	News	4	15	8	27
	Sports	7	9	3	19
Total		25	34	13	72

We could ***merge*** the middle age and older adult categories into a "not young" category. This is done by coding a new variable, as outlined in the transformation section.

Favourite Radio * age2 Crosstabulation

Count

		age2		Total
		NotYoung	Young	
Favourite Radio	Music	12	14	26
	News	23	4	27
	Sports	12	7	19
Total		47	25	72

We can look at the expected frequencies.

(Crosstabs menu, Statistics button, Check “Expected”)

Favourite Radio * age2 Crosstabulation

			age2		Total
			NotYoung	Young	
Favourite Radio	Music	Count	12	14	26
		Expected Count	17.0	9.0	26.0
	News	Count	23	4	27
		Expected Count	17.6	9.4	27.0
	Sports	Count	12	7	19
		Expected Count	12.4	6.6	19.0
Total		Count	47	25	72
		Expected Count	47.0	25.0	72.0

Even though one cell has observed frequency less than 5, its expected frequency is more than 5, so the potential problem is lessened.

We can also do the chi-squared test again and see if there's a problem or a change in the p-value.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.954[a]	2	.011
Likelihood Ratio	9.432	2	.009
N of Valid Cases	72		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.60.

0/6 cells are too small instead of 3/9.

We went from 4 df to 2 because we now have a 2x3 crosstab.

(2 – 1) x (3 -1) = 2.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.954[a]	2	.011
Likelihood Ratio	9.432	2	.009
N of Valid Cases	72		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.60.

Also, the most important part, the p-value, hasn't changed dramatically. (In the 3x3 table it was .036)

This implies that merging middle age and older didn't change anything major.

We ***reject the null*** ; radio choice depends on age.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.954[a]	2	.011
Likelihood Ratio	9.432	2	.009
N of Valid Cases	72		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 6.60.

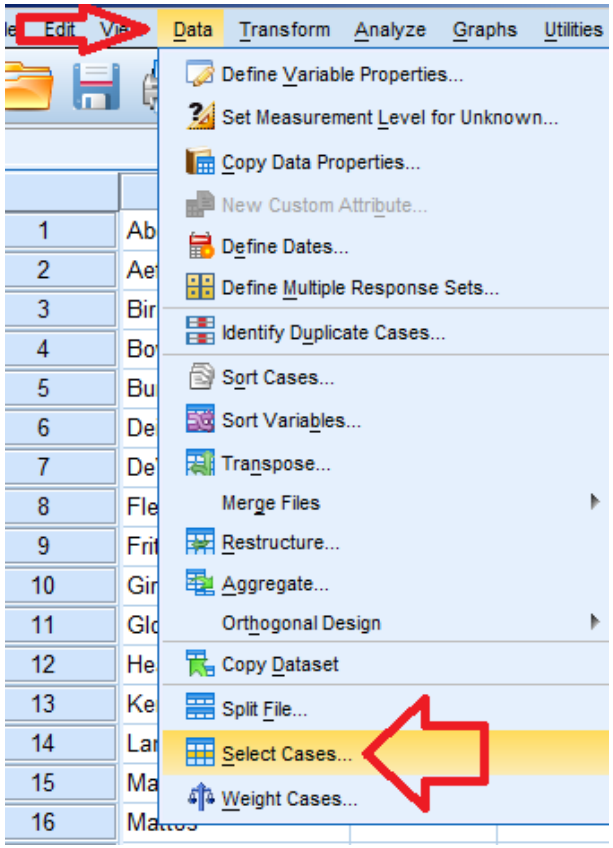
It’s easier to detect differences in larger groups, so we would expect the p-value to go down a little, but not to something dramatically lower like .001 or .000.

If the p-value had increased much we would have lost the ability to reject the null. (A bad merge can do this).

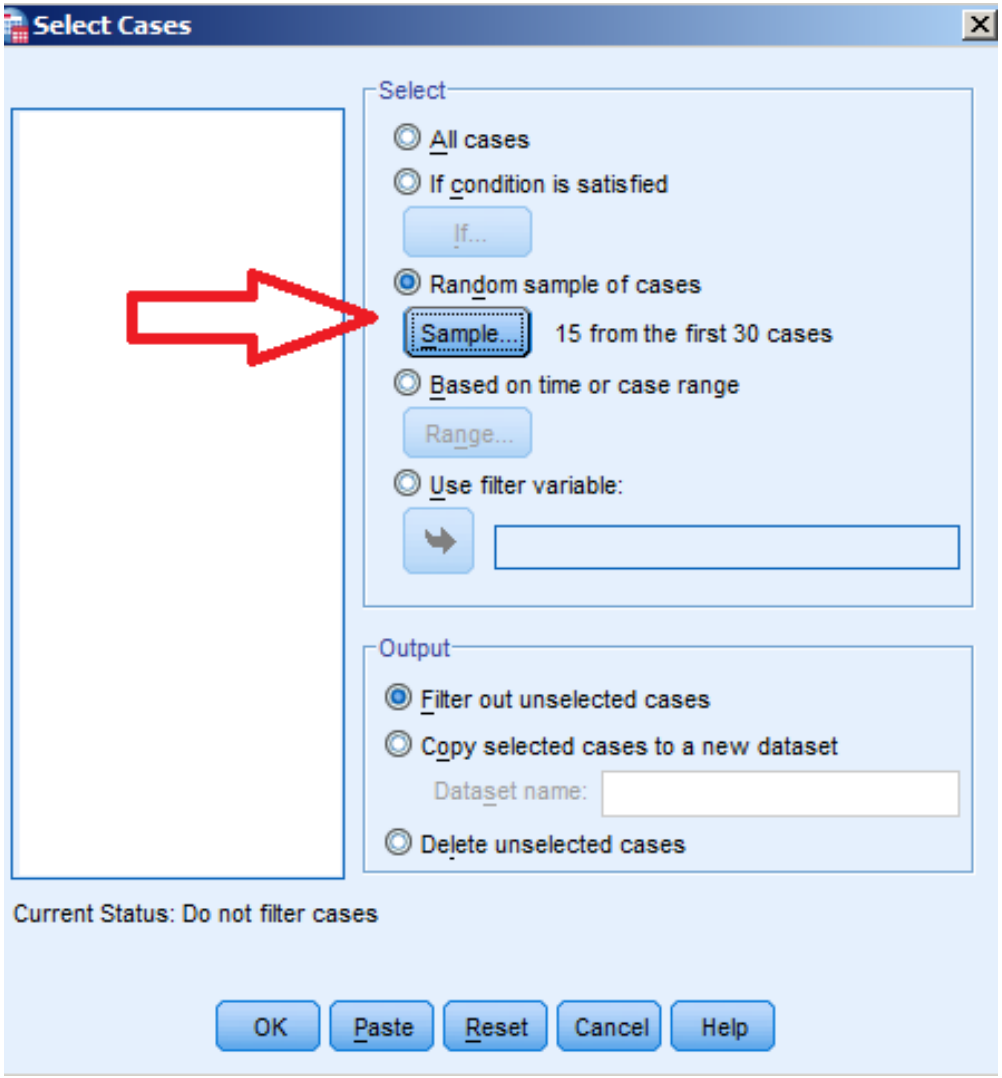
Random Selection

In the data window, choose

Data → Select Cases

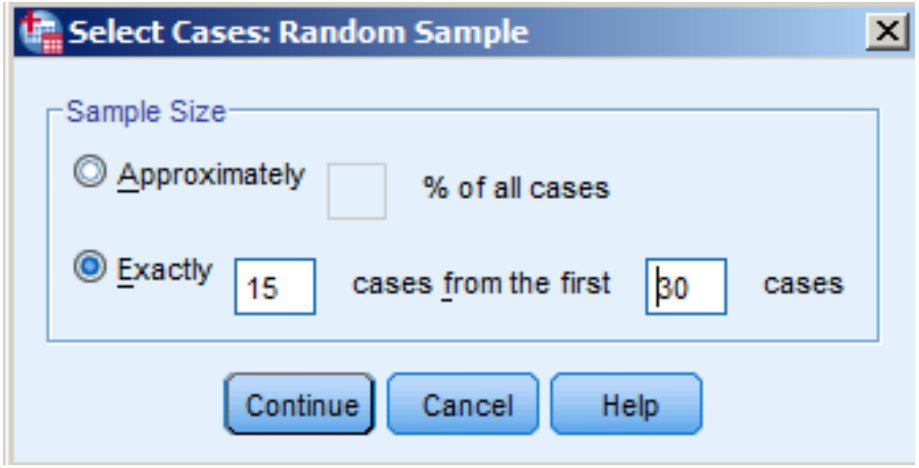


In the dialog box that comes up, choose the “Random sample of cases”, and click “Sample”



In this secondary dialog, choose whatever % or sample size you need.

For example, in problem 9.34, choose “Exactly 15 cases from the first 30” as shown and click “Continue”.



The cases that are in your sample are the ones that aren't crossed out.

In 9.34, the selected ones can be group 1, the rest group 2.

	Name	filter_$
~~1~~	Abel	0
~~2~~	Aeffner	0
3	Birkel	1
~~4~~	Bower	0
5	Burke	1
6	Deis	1
~~7~~	DeVorce	0

If you do any analysis on this data without resetting the cases you've selected, then the analyses will ignore cases that are crossed out.

One-Sample T-Tests

Here we run a one-sample t-test on the variable X from Descriptives XYZ.sav to test against the null hypothesis that the population mean is 30.

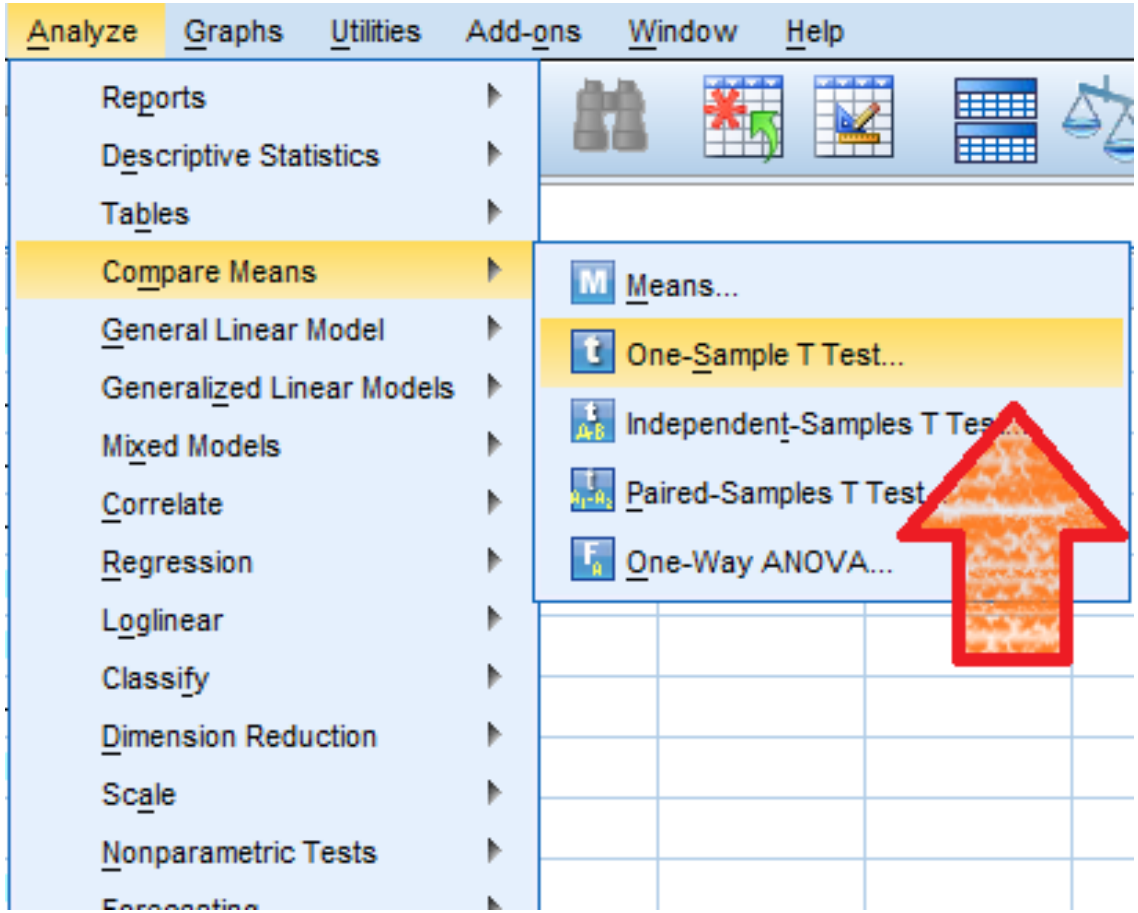
We're also going to produce a confidence interval of the mean.

Quick Reference:
Analyze → Compare Means → One Sample T-Test

Most “get a specific answer” functions are done in ***Analyze.***

T-tests give specific answers, they are in ***Compare Means.***

We’ll start with a ***One-Sample T-Test***

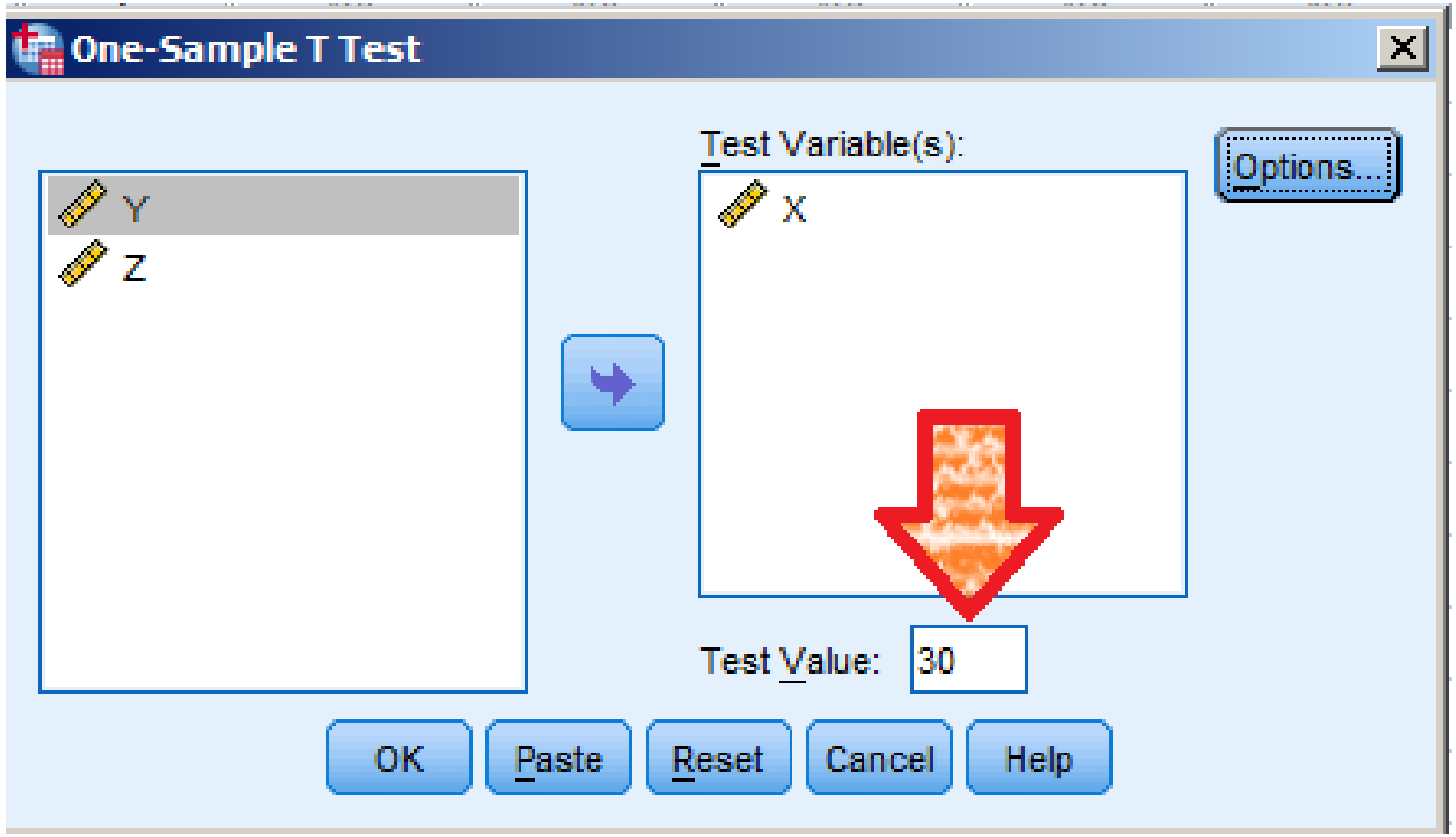


In the one-sample T Test dialog,

move X over to ***Test Variable(s):***

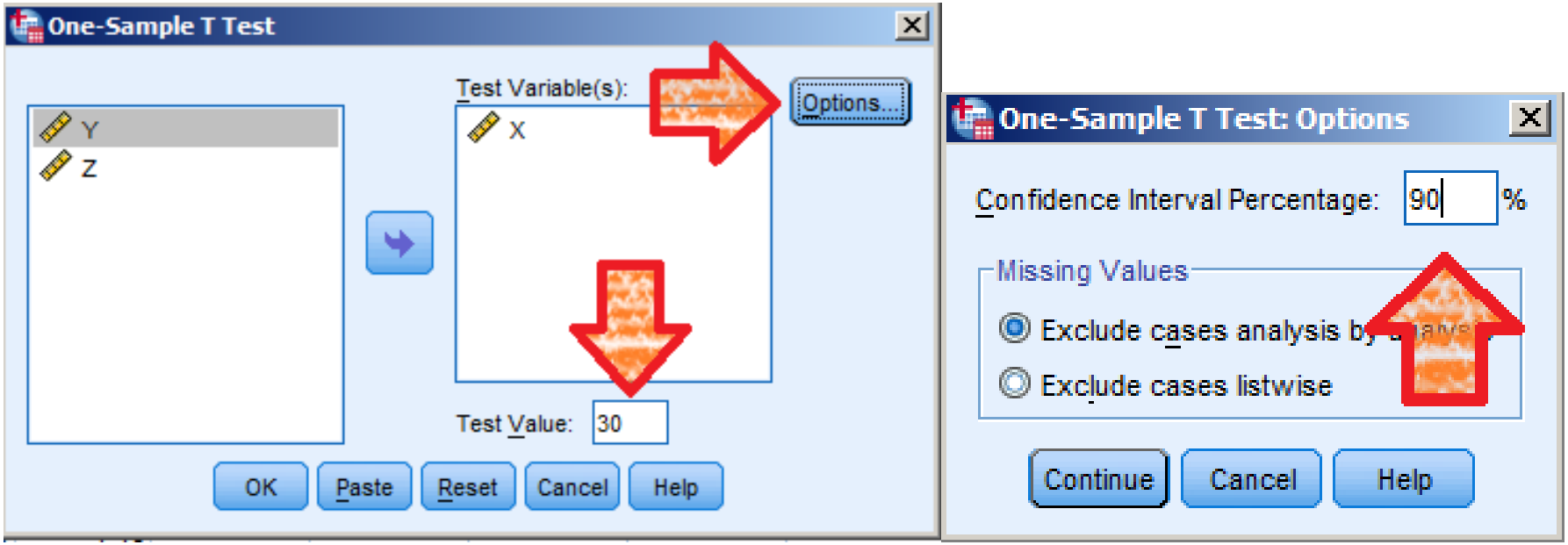
Then, since we're testing against a null hypothesis of mean=30,

put 30 in for the ***Test Value***



This test will also produce a 95% confidence interval by default.

If you wish to change the confidence level, click on ***Options.***



In the options dialog, change the confidence level to whatever you need and click **Continue, then OK**.

This is the output table.

Everything here is in relation to the null hypothesis and a two-tailed or two-sided alternative hypothesis.

One-Sample Test

	Test Value = 30					
					90% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
X	1.395	99	.166	2.95000	-.5609	6.4609

Sig. (2-tailed) is the p-value of the t-test. "Sig. " stands for "Significance"

Mean Difference is the difference between the sample mean and the test value. (The sample mean is 32.95)

One-Sample Test

	Test Value = 30					
					90% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
X	1.395	99	.166	2.95000	-.5609	6.4609

The confidence interval is also in relation to the null hypothesis, that the mean is 30.

The confidence interval if the mean is

30 – 0.5609 to 30 + 6.4609

Or

29.4391 to 36.4609.

One-Sample Test

	Test Value = 30					
					90% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
X	1.395	99	.166	2.95000	-.5609	6.4609

Finally, t is the t-score, and df is the degrees of freedom.

You can use these to do a t-test on a table from this information as well.

Two-Sample T-Tests

There are two kinds of two sample t-tests we'll cover in this section. Paired samples t-tests, and independent t-tests.

An additional check is done in independent t-tests for equal variance, or pooled variance.

Quick Reference:

Analyze → Compare Means → Independent T Test

Analyze → Compare Means → Paired T Test

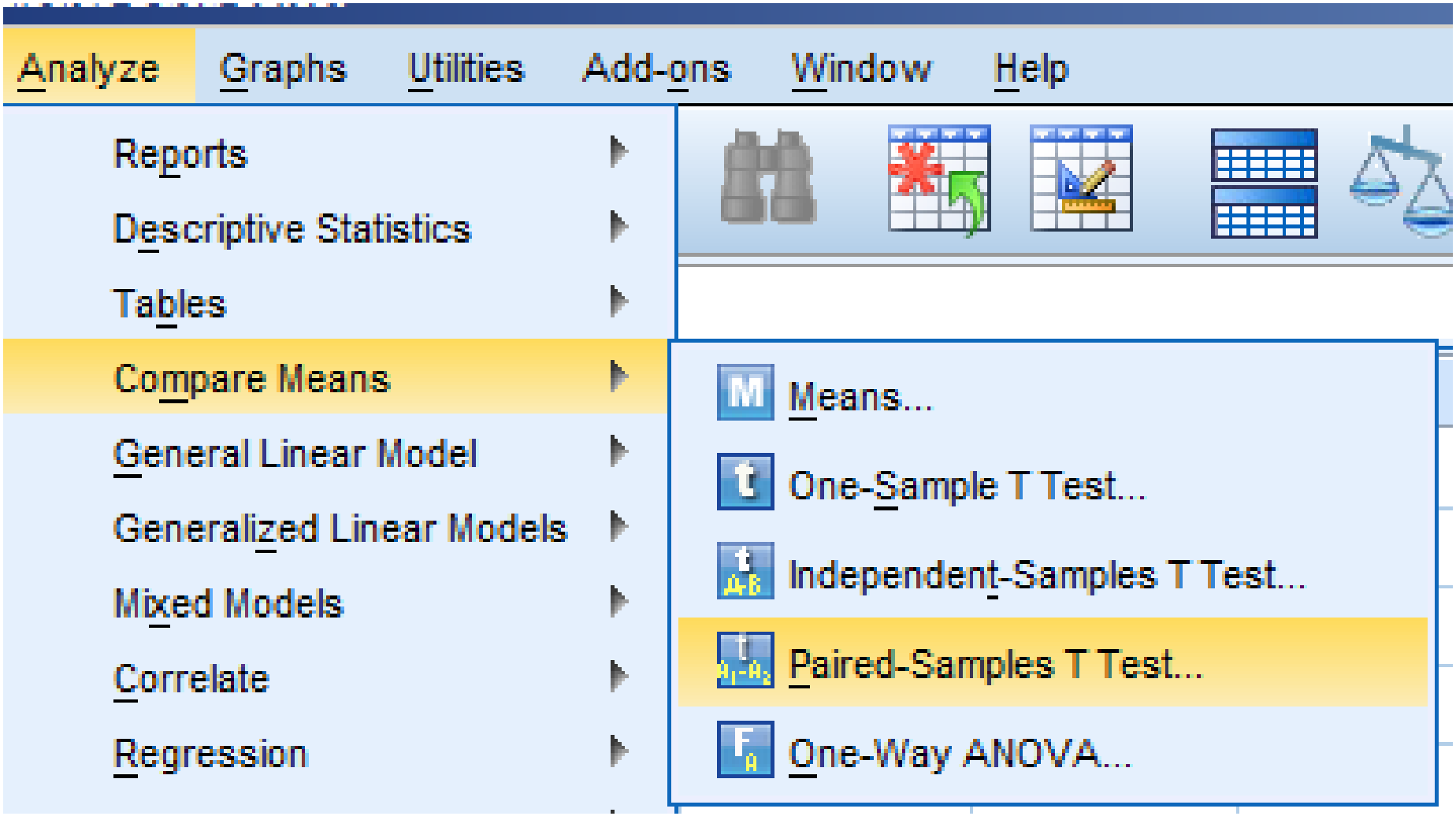
For samples that have a pairing structure between them, use the paired samples t-test. Paired t-tests can only be done on data that's in two side-by-side columns.

This is using the dataset Gas.sav

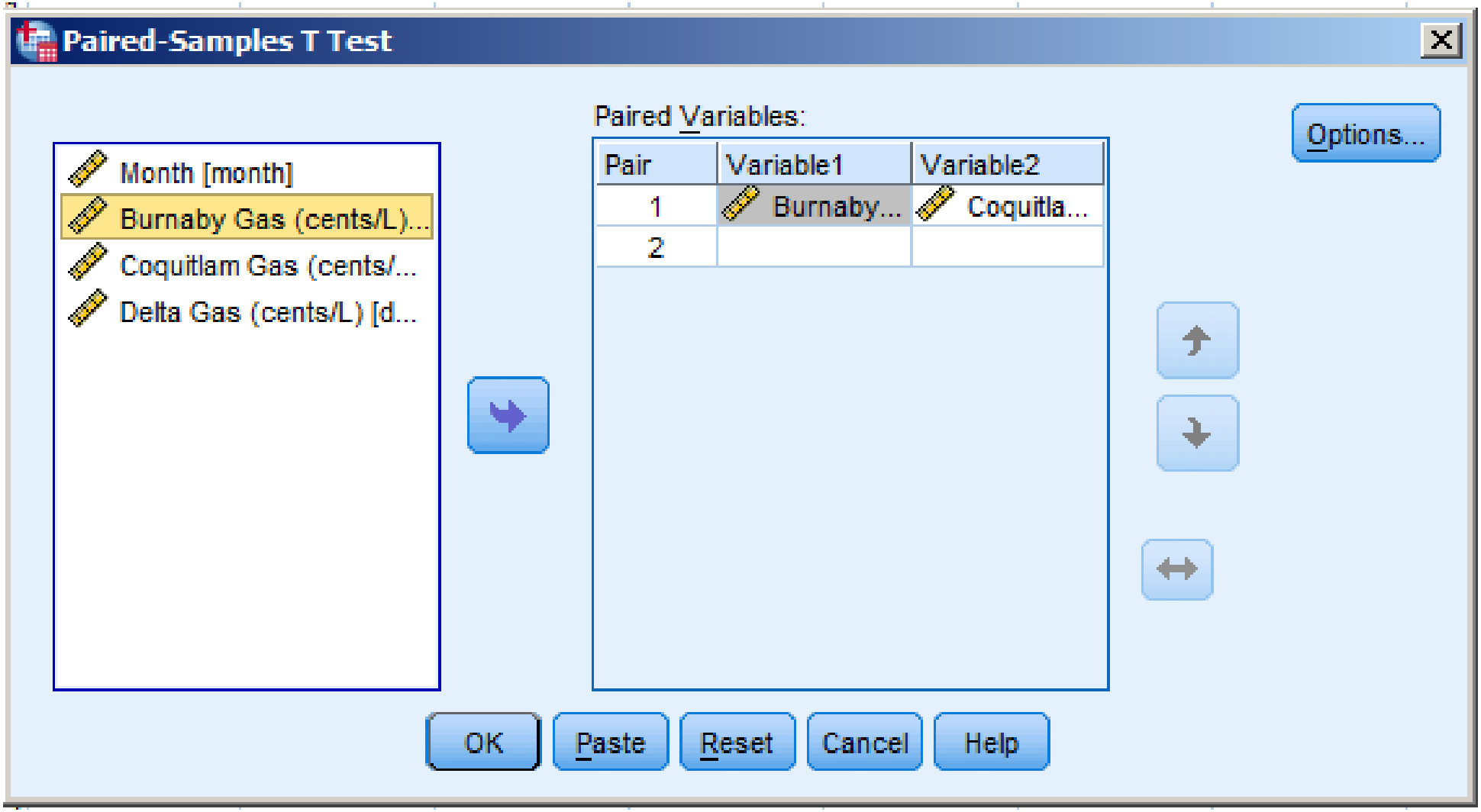
	month	burn	coq	delta
1	1	125.3	125.4	125.3
2	2	139.2	133.4	129.5
3	3	133.6	152.2	134.9
4	4	106.7	110.6	83.1
5	5	131.1	143.1	143.3
6	6	111.3	122.2	98.1
7	7	141.7	121.6	142.0
8	8	128.8	136.8	138.6

To perform a paired t-test, go to

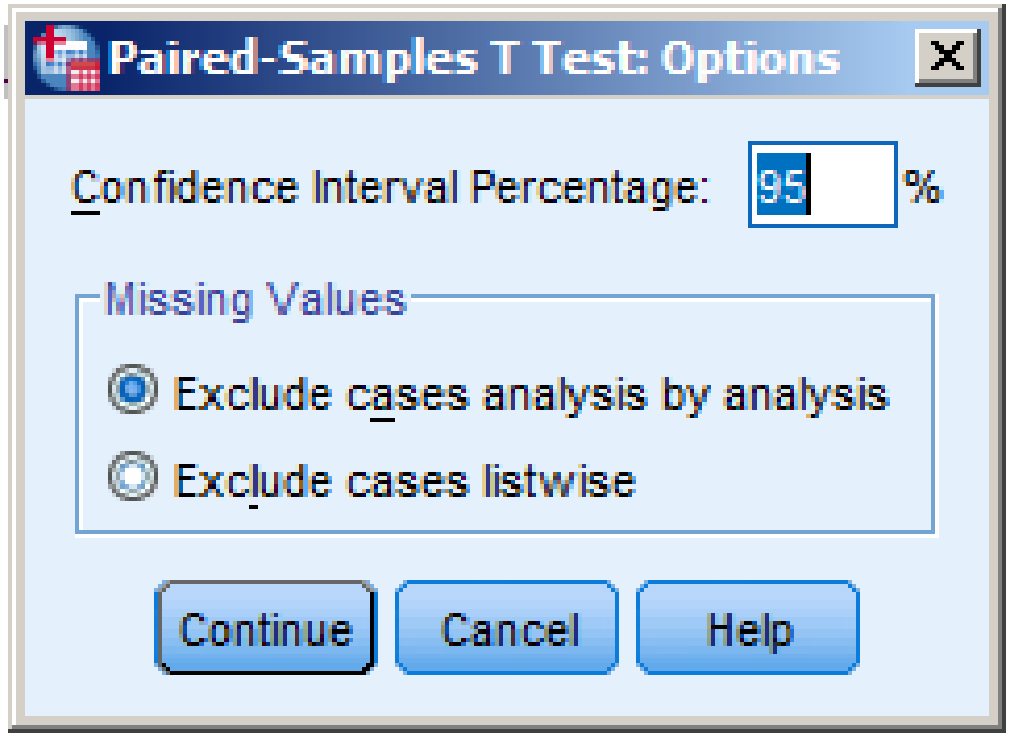
Analyze → Compare Means → Paired-Samples T Test…



Then drag the paired variables into the same pair. (Order doesn't matter for two-tailed tests) Then click OK.



If you want to change the confidence interval, press the options button, change it, then click Continue.



When you're ready, click OK on the main dialog.

The table we want is the ***Paired Samples Test***

Paired Samples Test

		Paired Differences			
		Mean	Std. Deviation	Std. Error Mean	95%
Pair 1	Burnaby Gas (cents/L) - Coquitlam Gas (cents/L)	-4.5323	13.6584	1.7346	

: **Test**

es				
95% Confidence Interval of the Difference				
Lower	Upper	t	df	Sig. (2-tailed)
-8.0008	-1.0637	-2.613	61	.011

The paired test only looks at the differences between values, so the mean is the mean difference. A negative mean implies that the second group is larger on average.

Paired Samples Test

		Paired Differences			95%
		Mean	Std. Deviation	Std. Error Mean	
Pair 1	Burnaby Gas (cents/L) - Coquitlam Gas (cents/L)	-4.5323	13.6584	1.7346	

Likewise, Std. Deviation and Std. Error Mean are the standard deviation and the standard error of the mean difference between the values.

The confidence interval is of the differences, t is the t-score against the mean difference being zero, Sig. (2-tailed) is the p-value for a two-tailed alternative.

: **Test**

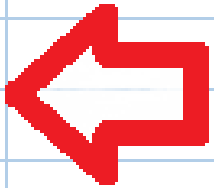
es				
95% Confidence Interval of the Difference				
Lower	Upper	t	df	Sig. (2-tailed)
-8.0008	-1.0637	-2.613	61	.011

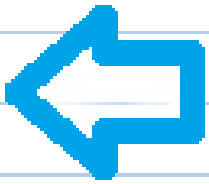
The degrees of freedom is of the variance of the differences, it's the number of pairs minus 1.

For unpaired data, we use the independent t-test. As found in RedCars.sav

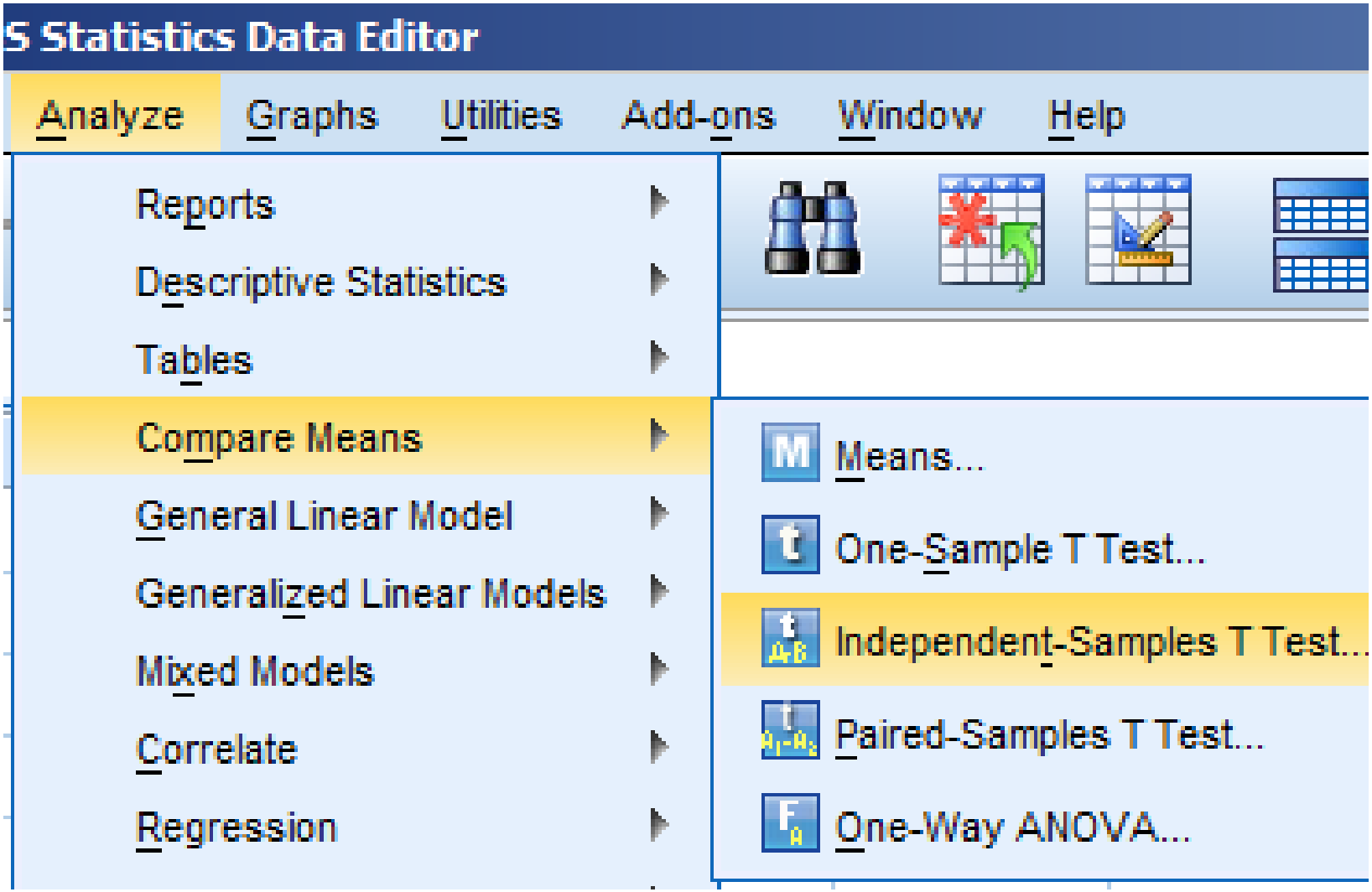
Independent t-test data needs to be all in a single column (speed). A second column is used as a ***grouping variable*** to tell SPSS which sample each car belongs to.

speed	colour	var
58.0	Red	
48.1	Red	
29.3	Red	
49.6	Red	
36.1	Red	
53.9	Blue	
51.0	Blue	
56.7	Blue	
55.7	Blue	



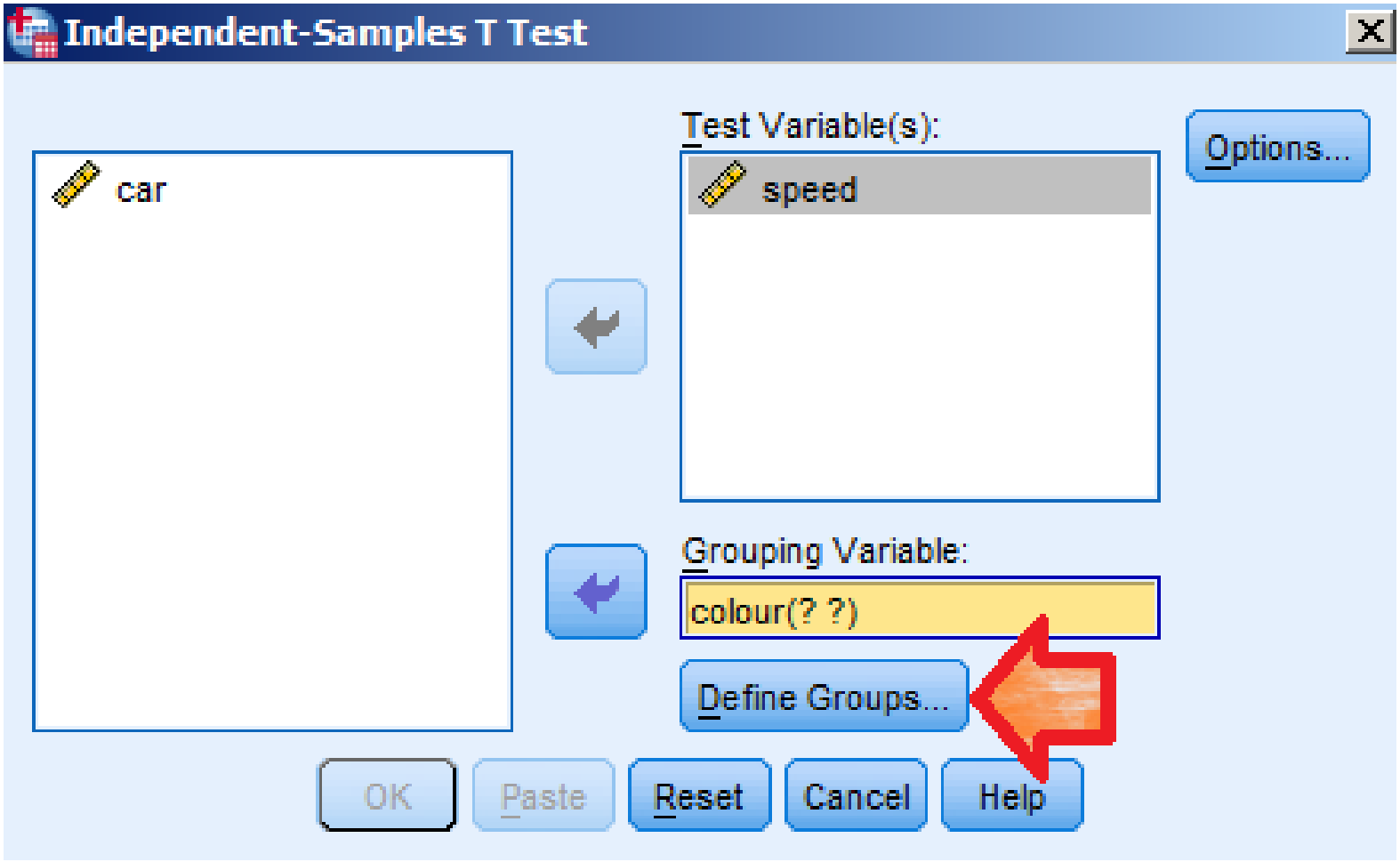


To do this test, go to ***Analyze → Compare Means → Independent Samples T-Test…***



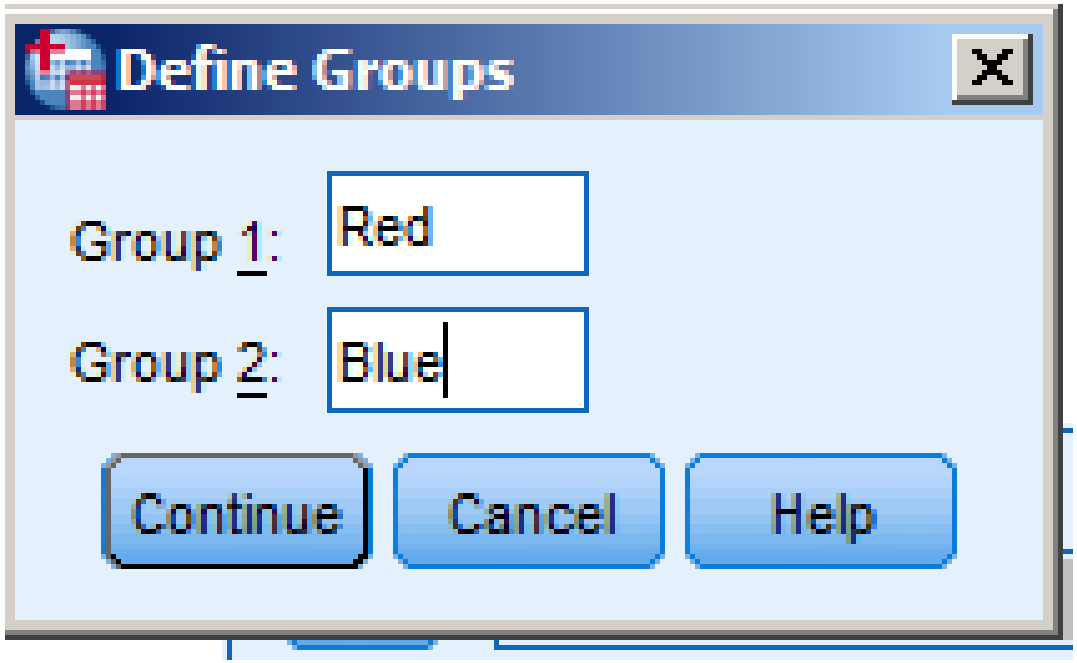
Put the response (speed) into the Test Variable(s) section.

Put the grouping variable (colour) into the Grouping Variable spot, and click ***Define Groups.***



Type “Red” into one group, and “Blue” into the other.

Be very careful of speling and cApitalization. It has to be ***exactly*** the same as the names in the grouping variable.



Then click Continue and click OK

SPSS outputs a large table. The first part is the results from testing the assumption of ***equal variance.*** This is what tells us if pooled standard deviation S_p is reasonable.

		Levene's Test for Equality of Variances	
		F	Sig.
speed	Equal variances assumed	2.269	.137
	Equal variances not assumed		

The null assumption is equal variance holds. The p-value is .137, which is large, so we'll use S_p, the top row results.

The middle part is the actual hypothesis test results.

Independent Samples Test

	t-test for Equality of Means				
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
	1.275	66	.207	3.6618	2.8722
	1.172	39.879	.248	3.6618	3.1254

T, df, and Sig. (2-tailed)are the t-score, degrees of freedom, and p-value respectively. Just like in a one-sample t-test.

Mean Difference is the difference between the sample means.

Std. Error Difference is the standard error of the difference.

Independent Samples Test

t-test for Equality of Means				
t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
1.275	66	.207	3.6618	2.8722
1.172	39.879	.248	3.6618	3.1254

The top row uses the assumption of equal variances. Note that this row has more degrees of freedom.

The last part is the confidence interval approach to the same problem.

	95% Confidence Interval of the Difference	
	Lower	Upper
	-2.0726	9.3963
	-2.6554	9.9791

This is the confidence interval of the difference between the means, the null hypothesis being a difference of 0. Note that 0 is in this confidence interval, what does that mean?

One-Sample Proportion Tests

One-sample proportion tests are used to test if a proportion is significantly different from a specified value. They are similar to t-tests, but are used when all responses are in a “Yes”/”No” or 0/1 format.

Here we test if the proportion of bearded dragons that are fancy is significantly more than 20%. **(Dragons.sav)**

Quick Reference:

Analyze → Nonparametic Tests → Legacy Dialogs → Binomial

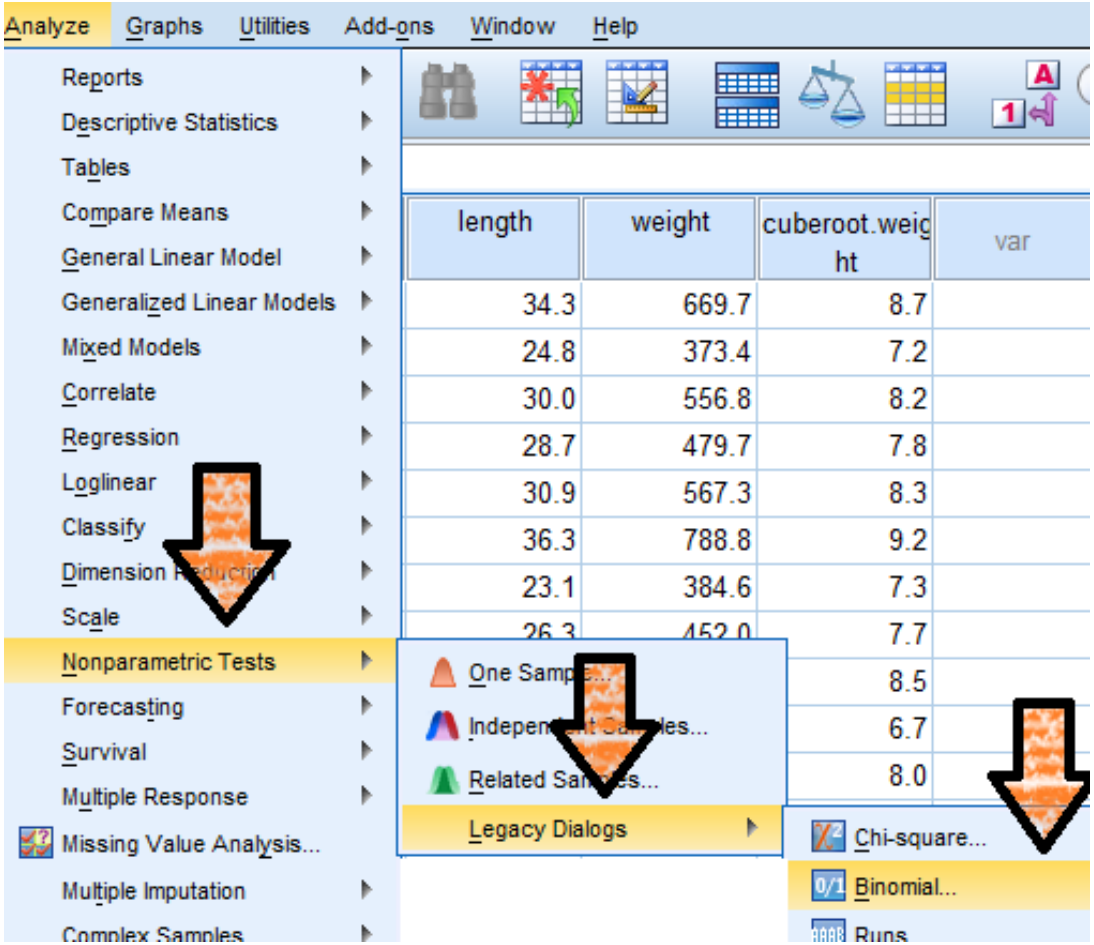
Here is the sex and colour data from 8 of the 300 bearded dragons that we have data on.

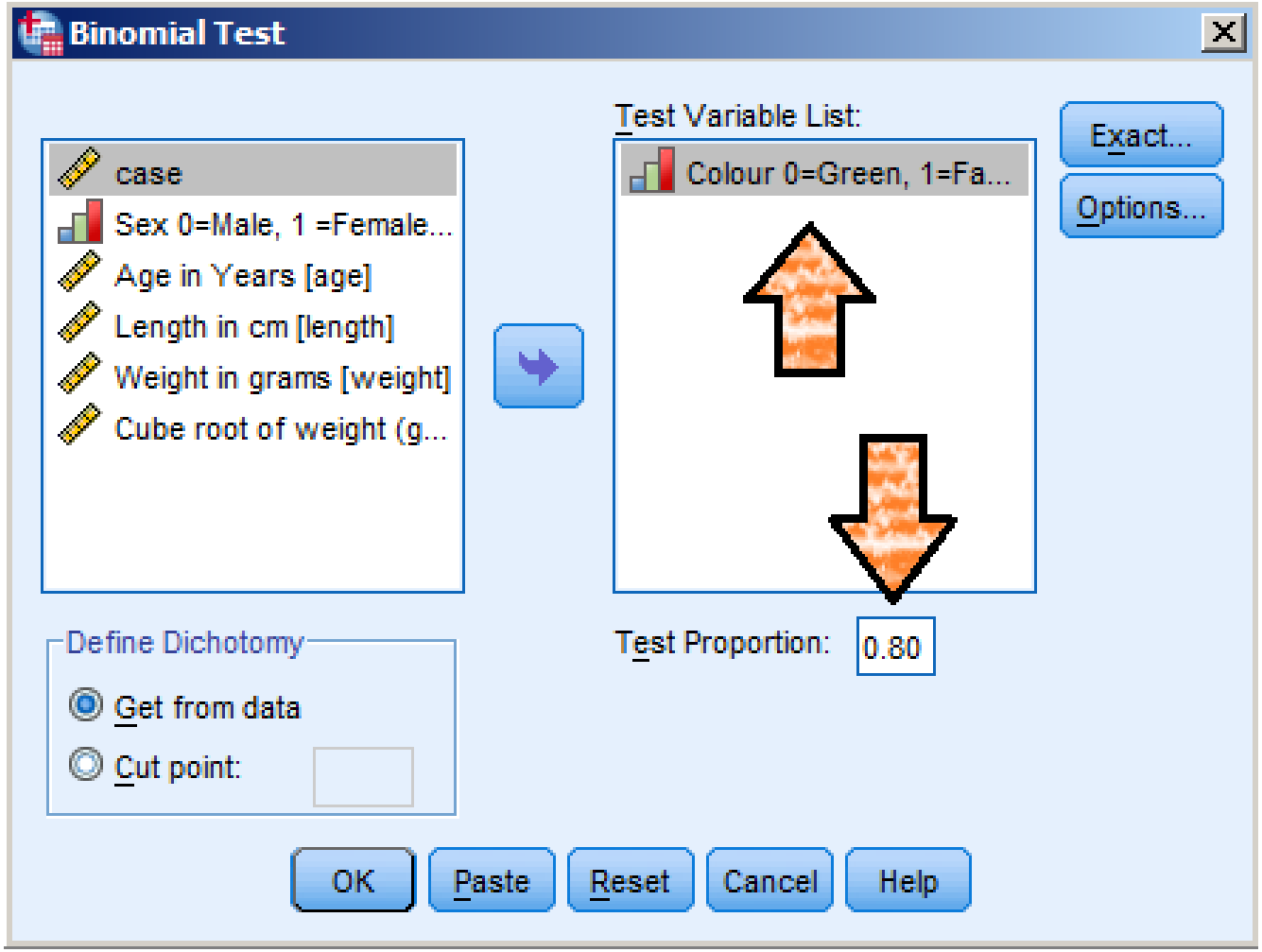
case	sex	colour
1	1	0
2	0	0
3	1	1
4	1	0
5	0	0
6	1	0
7	1	1
8	1	1

Name	Type	...	...	Label
case	Numeric	3	0	
sex	Numeric	1	0	Sex 0=Male, 1 =Female
colour	Numeric	1	0	Colour 0=Green, 1=Fancy

To start a test against a hypothesized proportion go to

Analyze → Nonparametric Tests → Legacy Dialogs → Binomial




Binomial Test
case
Sex 0=Male, 1 =Female...
Age in Years [age]
Length in cm [length]
Weight in grams [weight]
Cube root of weight (g...
Test Variable List:
Colour 0=Green, 1=Fa...
Exact...
Options...
Define Dichotomy
Get from data
Cut point:
Test Proportion:
0.80
OK
Paste
Reset
Cancel
Help

The results appear in a table with the observed proportion (what was actually seen in the data), and the test proportion (the null hypothesis proportion).

As usual, a smaller p-value indicates strong evidence against that null hypothesis.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Colour 0=Green, 1=Fancy	Group 1	0	210	.7	.8	.000[a]
	Group 2	1	90	.3		
	Total		300	1.0		

a. Alternative hypothesis states that the proportion of cases in the first group < .8.

A bit unusual for SPSS is that the **one-tailed p-value** is given.

For a two-tailed test, double the p-value.

The alternative hypothesis is generated automatically to be the one that would generate the lower p-value of the two one-tailed alternative hypotheses.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Colour 0=Green, 1=Fancy	Group 1	0	210	.7	.8	.000[a]
	Group 2	1	90	.3		
	Total		300	1.0		

a. Alternative hypothesis states that the proportion of cases in the first group < .8.

Two-Sample Proportion Tests

Two-sample proportion tests are to two-sample t-tests as one-sample proportion tests are to one-sample t-tests. The common null hypothesis to check is whether or not there is a difference in the proportions of two groups.

Here we test for a difference in the proportion of car colours by the gender of driver. **(RedCars.csv)**

Quick Reference:

- Analyze → Descriptive Stats → Crosstabs

First, construct a crosstab of gender and colour in

Analyze → Descriptive Stats → Crosstabs

…as found in the crosstabs chapter, starting on p.95.

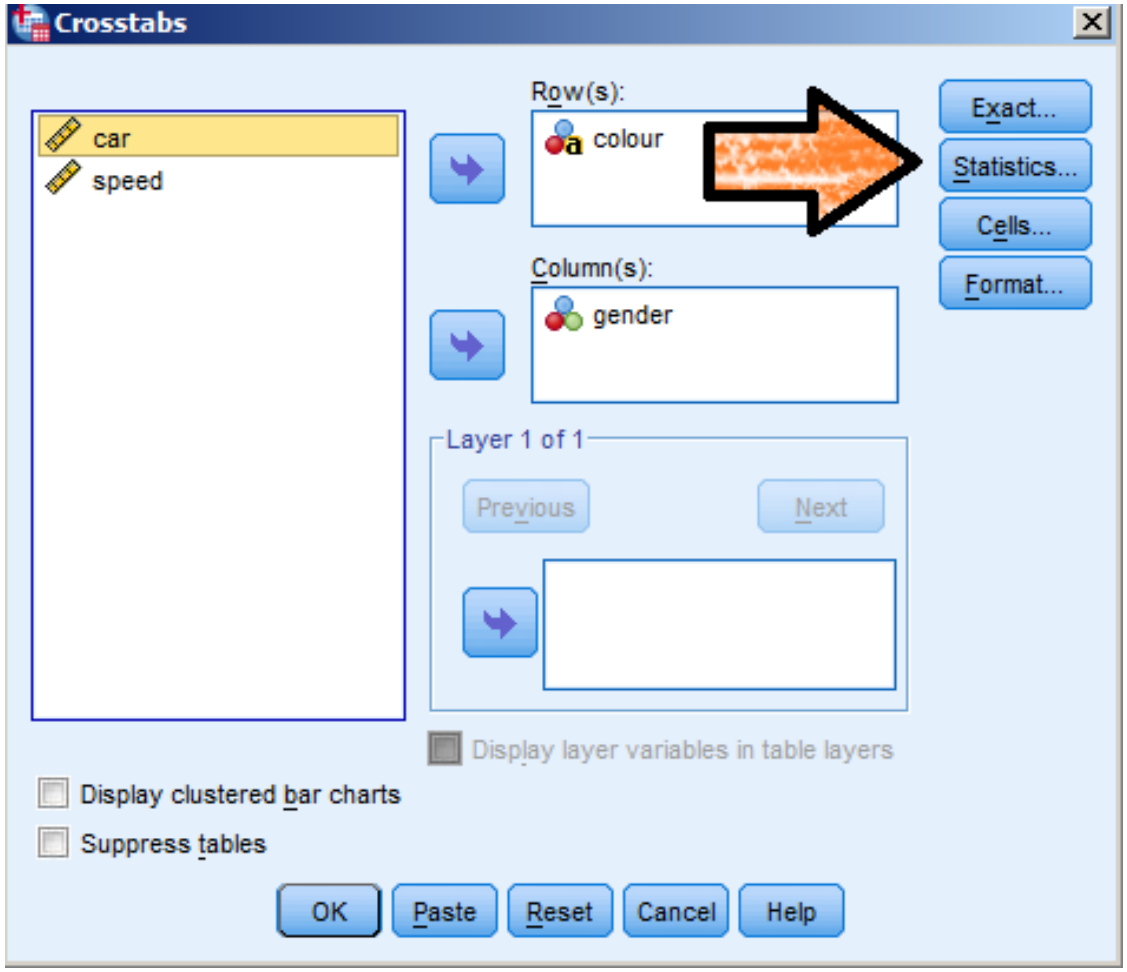
(0 is Male, 1 is Female)

colour * gender Crosstabulation

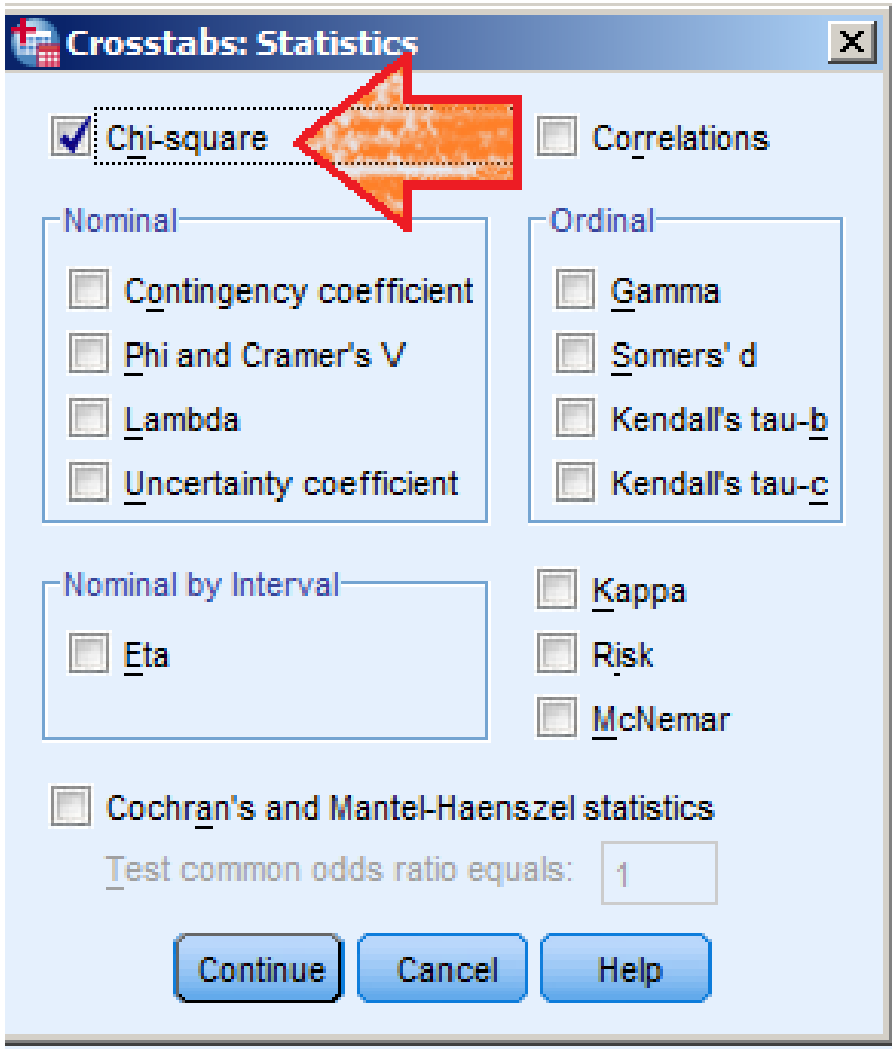
Count

		gender		Total
		0	1	
colour	Blue	22	20	42
	Red	14	12	26
Total		36	32	68

When making a crosstab, in the main crosstabs dialog, you can also calculate the chi-squared statistic by clicking on the ***Statistics*** button.



Then, put a check next to ***Chi-Square*** in the upper left.



Then click Continue, then OK.

The **Pearson Chi-Squared** is the value that matters here.

The Z-score is the square root of the Pearson Chi-Squared (**ONLY IN THIS 2x2 CASE**), so

$$Z = \text{sqrt}(0.014) = 0.118.$$

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.014[a]	1	.906		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.014	1	.906		
Fisher's Exact Test				1.000	.553
N of Valid Cases	68				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.24.
b. Computed only for a 2x2 table

Z = 0.118.

If you are using a Z-table, you can verify that about 0.453 of the probability mass is above Z = 0.12. Since the test is two-sided, that's 0.453 on each side, to make 0.906 in total.

Just like the p-value here.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	.014[a]	1	.906		
Continuity Correction[b]	.000	1	1.000		
Likelihood Ratio	.014	1	.906		
Fisher's Exact Test				1.000	.553
N of Valid Cases	68				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 12.24.
b. Computed only for a 2x2 table

Weights

In problem 23.28 and others in Chapter 23, you deal with count data, like the **smokecess.por** dataset as shown below.

	treatmen	smoking	count
1	Chantix	No	155
2	Bupropion	No	97
3	Placebo	No	61
4	Chantix	Yes	197
5	Bupropion	Yes	232
6	Placebo	Yes	283

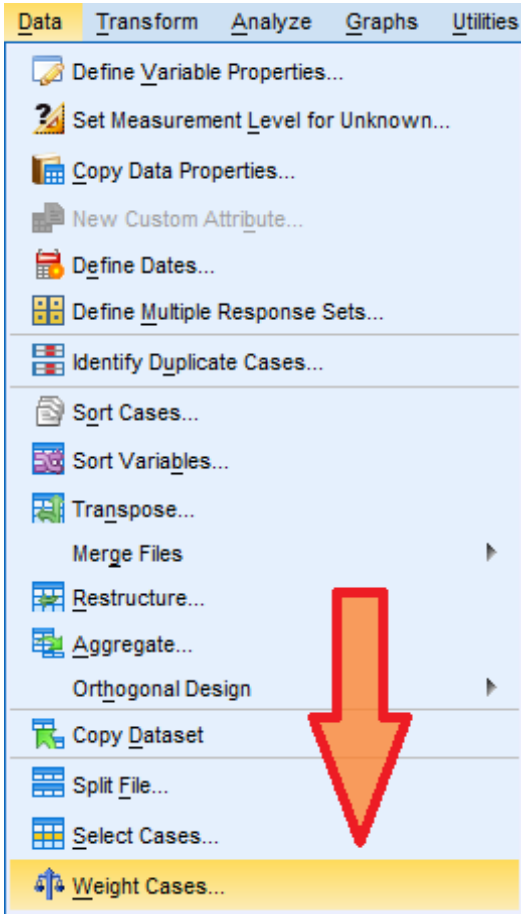
Most of the data you've dealt with up until this point has been in long format, meaning that one row is one observation.

However, in the **smokecess.por** dataset, the first row represents 155 observations (i.e. the count) of non-smokers on Chantrix, the second row represents 97 observations and so on.

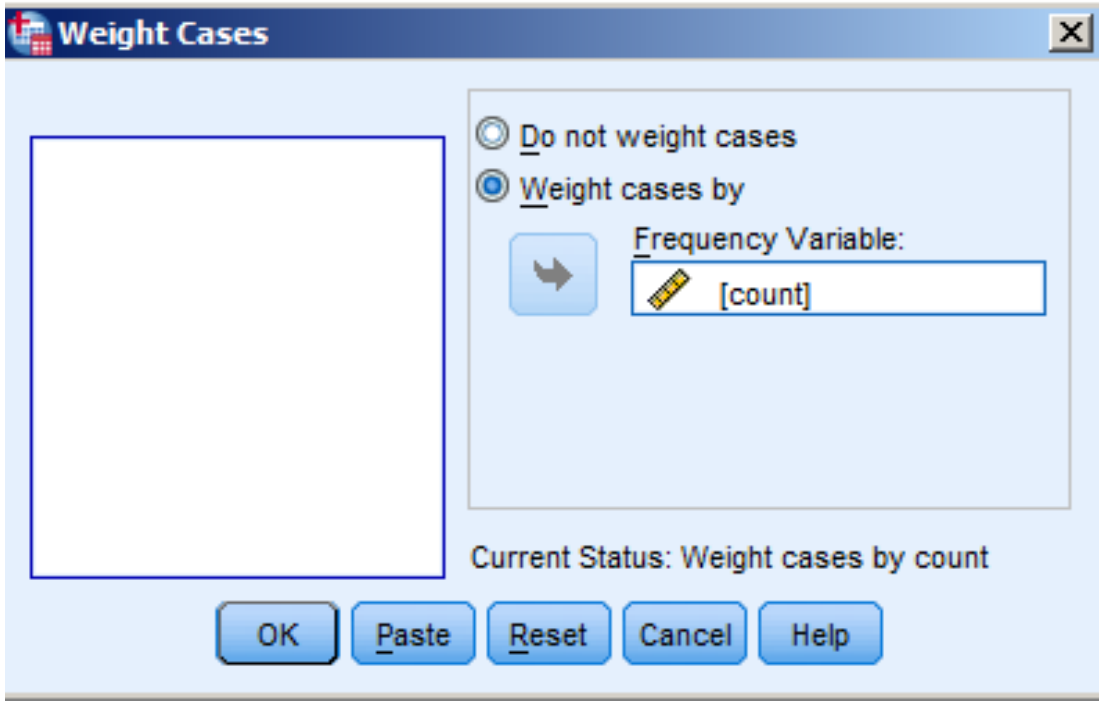
	treatmen	smoking	count
1	Chantix	No	155
2	Bupropion	No	97
3	Placebo	No	61
4	Chantix	Yes	197
5	Bupropion	Yes	232
6	Placebo	Yes	283

We need to tell SPSS that each row is more than a single observation. In short, we need to ***weight*** the observations.

To set observation weights, go to ***Data → Weight Cases***



In the dialog that opens, put the variable you wish to have determine the weights (In the case of 23.28, this is [count]) in the Frequency Variable field. Then click OK.



Now you can continue with your analysis and SPSS will read the data correctly.

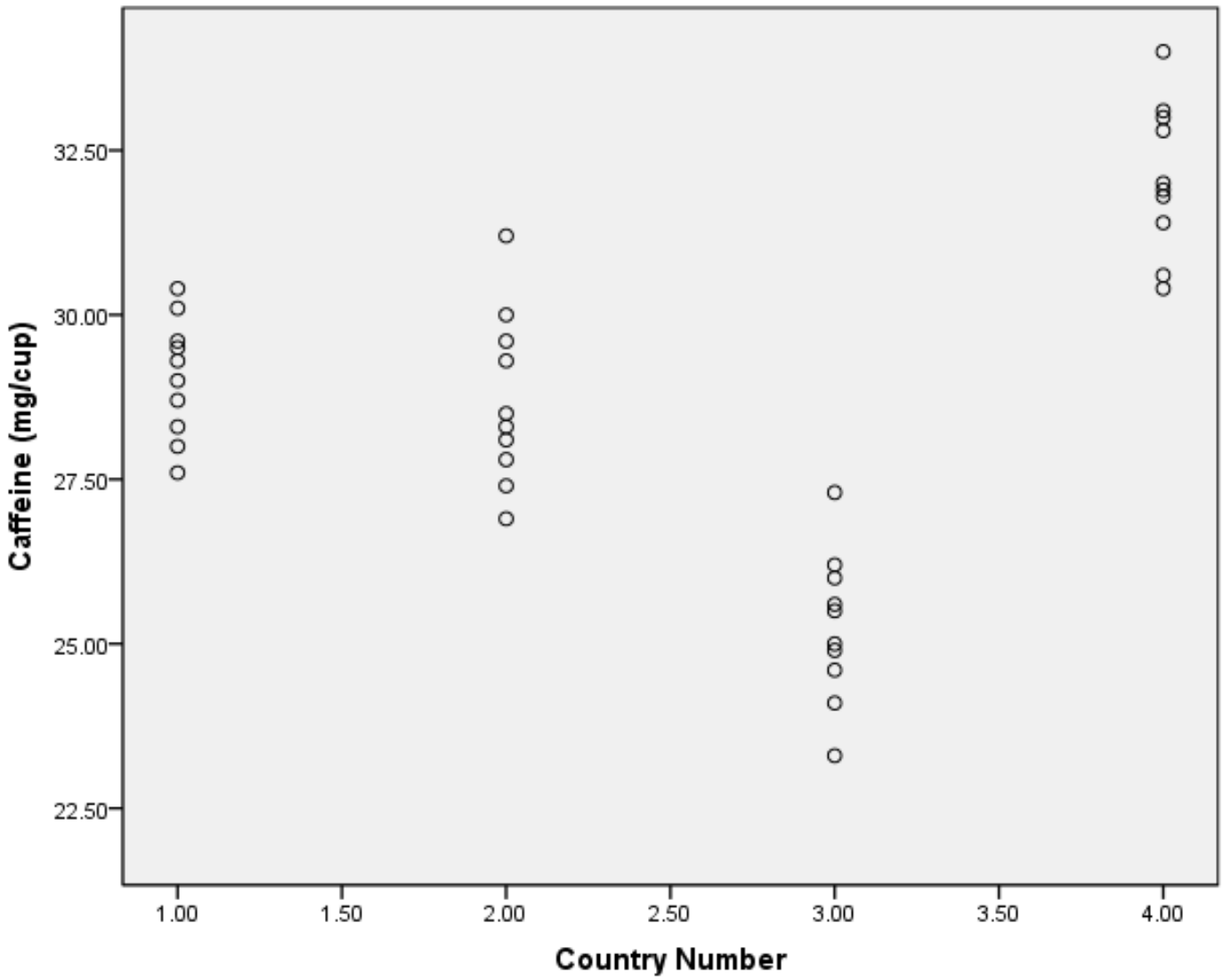
One-Way ANOVA

ANOVA is for comparing more than two means to see if the population means could be all the same. One-Way ANOVA is the last of the compare means tools in this guide.

Here we do an Analysis of Variance (ANOVA) on the caffeine content of black teas from four countries. **(Caffeine.sav)**

Quick Reference:

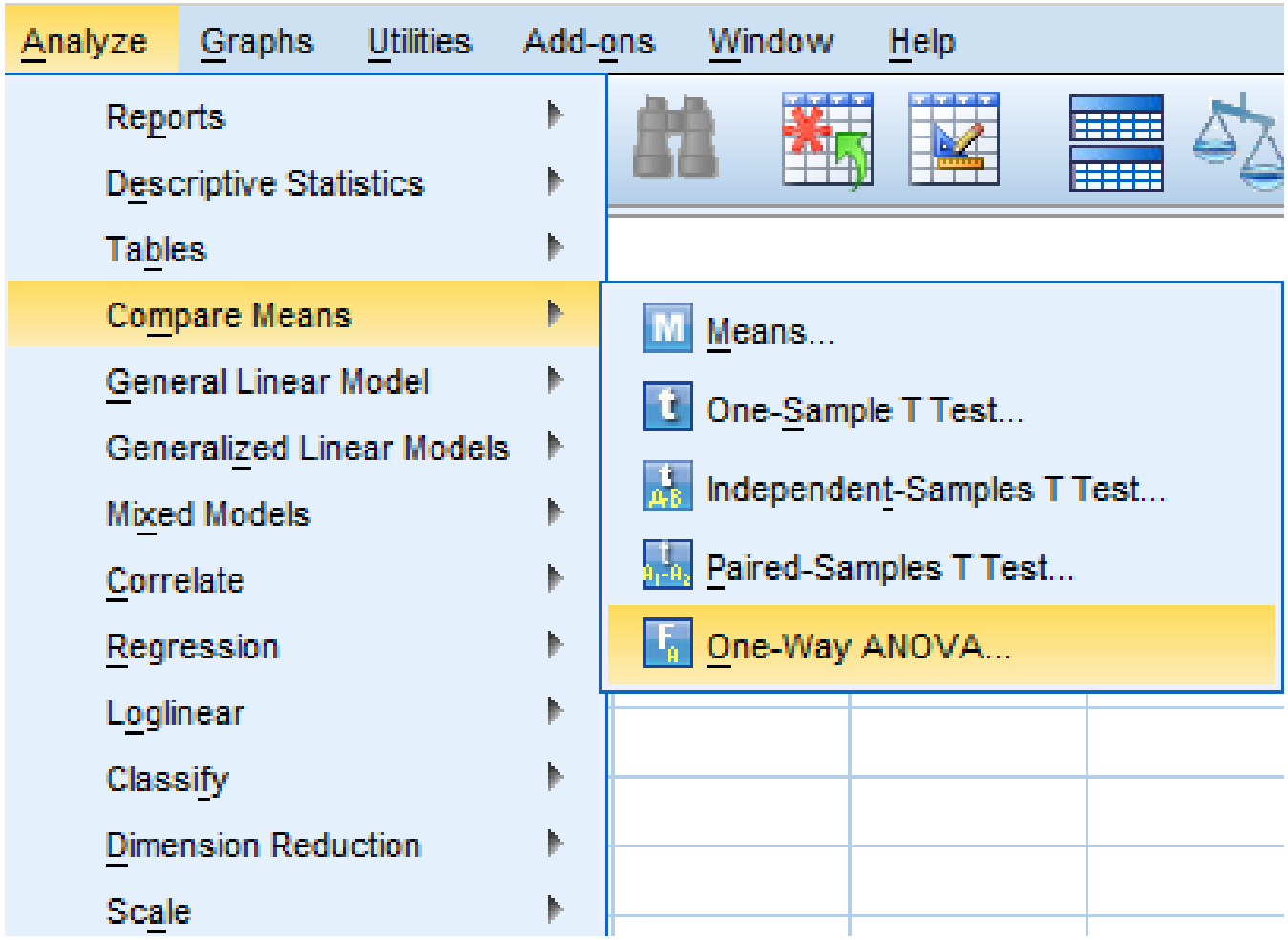
- Analyze → Compare Means → One-Way ANOVA



Here is the caffeine content of teas from our four countries. A sample of size 10 from each. (Built with a scatterplot)

First, One-Way ANOVA is in

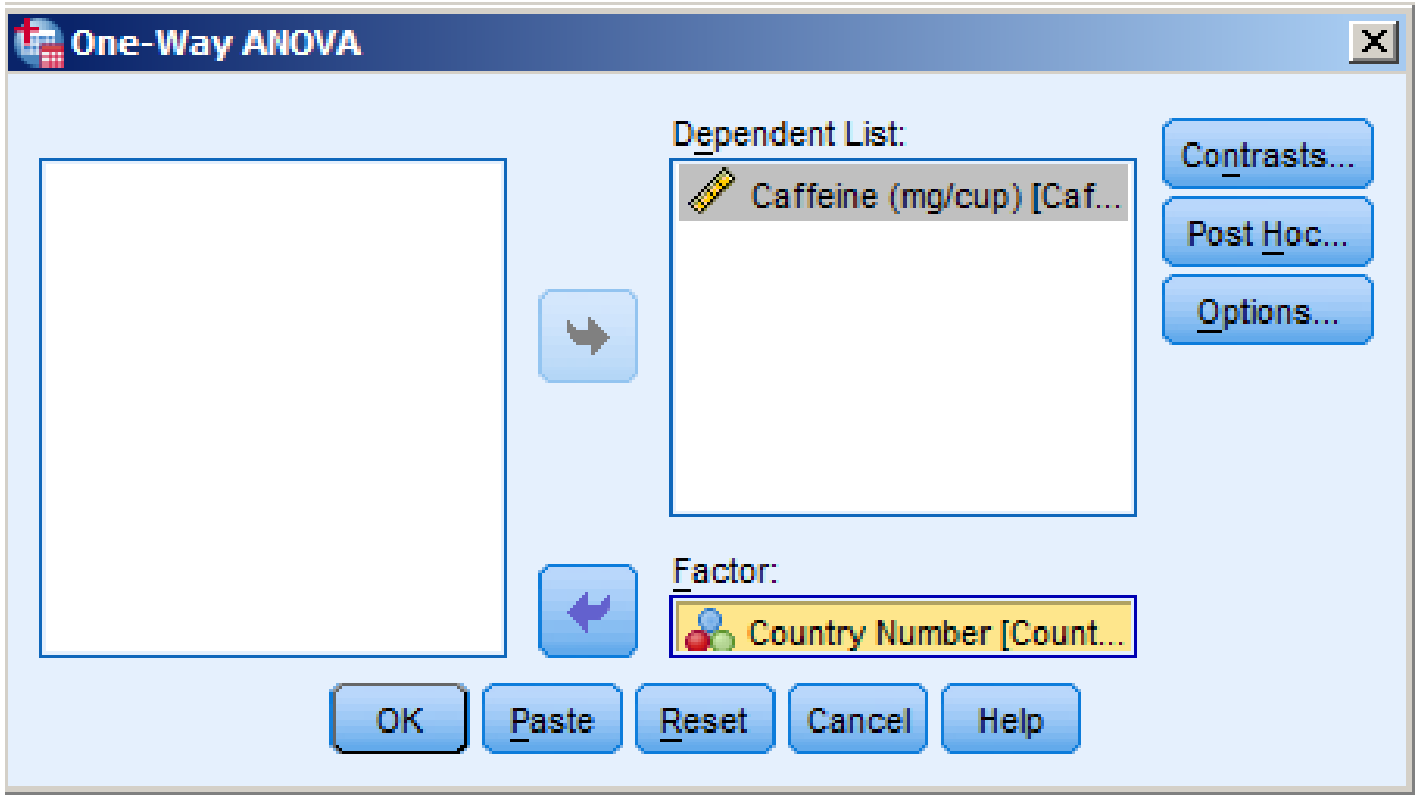
Analyze → Compare Means → One-Way ANOVA



We want to know how caffeine varies as country changes.

Put Caffeine in the ***Dependent List***, and

Country Number in the ***Factor***



ANOVA requires the factor to be numeric, hence country number instead of country name.

After clicking OK, we get this in the output.

ANOVA

Caffeine (mg/cup)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	235.611	3	78.537	61.385	.000
Within Groups	46.059	36	1.279		
Total	281.670	39			

Sig. is the p-value of the test that all the population means are the same. There is strong evidence that they aren't.

The F-Stat (61.385) and the numerator df (3) and denominator df (36) are also available for you test by table.

ANOVA

Caffeine (mg/cup)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	235.611	3	78.537	61.385	.000
Within Groups	46.059	36	1.279		
Total	281.670	39			

Also, the proportion of variance in caffeine explained by country can be found from ***Between Groups / Total***

235.611 / 281.67 = 0.836

So 83.6% of the variation in caffeine content is explained by the differences between country means.

A useful exercise: Open the redcars.sav data do an ANOVA with

- Speed of cars as dependent, and
- Colour as the factor.

Compare the ANOVA results to those of the independent samples t-test of the same data. Assume equal variances.

The p-values and df of the ANOVA and the T-test should be identical.

Why? The t-test is testing the null that **"the means of these two groups the same."**

The F-test in ANOVA is testing the null that **"the means of all the groups are the same."**

Logically, these mean the same thing when there are only two groups.

ANOVA assumes equal variance, so you need to assume equal variance with the t-test as well to make a fair comparison.

Acknowledgements, Other Resources

Thanks to Dr. Tim Swartz of Simon Fraser University and RobMagus of Reddit for the input that made this guide better.

Proportion test chapters adapted from guides at

http://www.stat.vcu.edu/help/SPSS/

at the Statistical Sciences and Operations Research department at Virginia Commonwealth University,

For tutorials of more advanced functions, try the videos for Discovering Statistics using SPSS, by Dr. Andy Field, at http://www.sagepub.com/field3e/SPSSstudentmovies.htm

For additional theoretical help, I recommend:

The Online Stat Book led by David Lane, Rice University

http://www.onlinestatbook.com/2/

The Open Learning Initiative by Carnegie Mellon University

http://oli.cmu.edu/