

## Survey Analysis Workshop

### Block 3: Analysing two variables (and sometimes three)

---

#### Section 3.2: Three (or more) variables

#### Sub-section 3.2.1 Elaboration

© Copyright 2019 [John F Hall](#)

[New tutorial 30 April 2019: **Draft only**]

#### 3.2.1.5 Earnings differences 2009: Download and check file

(Replication, using 2009 data, of elaboration exercise [3.2.1.1 Earnings differences – Elaboration](#) )

Data source: [British Social Attitudes Survey, 2009](#)<sup>1</sup> (UKDS SN 6695)

#### Page Contents:

2	<a href="#">Model</a>
3	<a href="#">Variables to be extracted</a>
4	<a href="#">Downloading the data file from the UK Data Service (UKDS)</a>
9	<a href="#">Checking contents of the downloaded file</a>
13	<a href="#">Frequency counts for dependent variables</a>

#### Introduction

This set of tutorials will use data from the 2009 British Social Attitudes Survey to explore the following research questions.

- 1: Is there a difference between men and women in their earnings (from paid work)?
- 2: What other variables might account for differences in earnings?
- 3: What effect do these other variables have by themselves?
- 4: What happens to differences between men and women in their earnings when controlling for these other variables?

Tutorial [3.2.4.1 Income differences - Elaboration](#) used the 1989 British Social Attitudes Survey (BSAS) to analyse differences between men and women in their earnings of from paid work. This tutorial replicates that exercise on data from the 2009 survey.

I have retained the original variable names, but other dictionary attributes such as missing values, variable and value labels, measurement levels and formats have been added if they are absent or edited if they are confusing, incomplete or incorrect.

For many variables missing values are displayed as **None**. For other variables (**lo thru -1**) is specified. However (**-1, -2**) is specified for many variables which have other negative values in the range **-3 to -6**, which also need to be treated as missing. Yet other variables have positive values such as **7** or **97** "Refused", **8** "Can't choose", **9** "Not answered" or **98** "Don't know" which also need to be treated as missing, but are not. For statistical analysis these values need to be treated as missing.

---

<sup>1</sup> National Centre for Social Research. (2011). *British Social Attitudes Survey, 2009*. [data collection]. UK Data Service. SN: 6695, <http://doi.org/10.5255/UKDA-SN-6695-1>

### 3.2.1.5 [Earnings differences 2009] Download and check file

For example [EJbHrsX] " Respondent: Is job full or part-time? :Q1007

-4 = "Self-employed"  
-3 = "Not currently employed"  
-1 = "Never had a job"  
98 = "Don't know"  
99 = "Refusal"

#### Model

This exercise will download the data for BSAS 2009, extract a dependent variable [REarn] (gross earnings from paid work) an independent variable [Rsex] (sex) and a selection of work-related and demographic test variables to analyse the following elaboration<sup>2</sup> model:

$X \rightarrow Y . T$  (the effect of  $X$  on  $Y$  controlling for  $T$ ) where:

Y = Dependent variable  
X = Independent variable  
T = Test variable(s)

Y (Dependent)	X (Independent)	T <sub>n</sub> (Test or control)
Gross earnings from paid work	Sex	T <sub>1</sub> Working full time or part time T <sub>2</sub> Employee or self employed T <sub>3</sub> Economic sector T <sub>4</sub> Socio-economic grade of work T <sub>5</sub> Level of education T <sub>6</sub> Qualifications T <sub>7</sub> Age T <sub>8</sub> Geographical region

#### Previous research questions:

1: Is there a difference between the earnings (from paid work) of men and women?

See sessions: [2.3.1.6.2: Specimen answer for tasks 3 and 4](#)  
[3.1.4.1 Income differences work-through](#)

2: What other variables might account for differences in earnings?

See sessions: [3.1.4.2 Income differences - Build working file](#)  
[3.1.4.3 Income differences for test variables](#)  
[3.1.4.4 Income differences - Choose test variables and cutting points](#)

3: What effect do they have by themselves?

See session: [3.1.4.5 Income differences for derived test variables](#)

#### Further research question:

What happens to differences in earnings between men and women when controlling for these other variables?

---

<sup>2</sup> (See [Elaboration](#) (extract from Jim Ring's [Statistical Notes](#) specially written for this course)

### 3.2.1.5 [Earnings differences 2009] Download and check file

#### Variables to be extracted

**Dependent variable**                    **[REarn]** "Respondent's gross earnings from paid work" [if working]

Variable **[REarn]** "Grouped gross earnings" has valid values ranging from **1** to **20** denoting grouped earnings per calendar month. Value **-1** "Item not applicable" is declared as missing: values **97** "Refused information", **98** "Don't know" and **99** "Refused" are **not declared as missing**.

There is also a derived variable **[REarnQ]** "Quartile groups of R's gross earnings" which groups earnings into four categories. This helps to keep contingency tables small and manageable. Values **7** "Refused information" and **8** "Don't know" are **not declared as missing**.

**Independent variable**                    **[Rsex]** "Sex of respondent"

**[Rsex]** "SEX OF respondent? :Q356" [sic] is coded **1** "Male" **2** "Female" and has no missing values. Users may prefer to rename it as **[sex]** or **[gender]** according to their preferences.

**Weighting**       Some analyses may also require the weighting factor **[Wtfactor]**

Y (Dependent)	X (Independent)	T <sub>n</sub> (Test or control)	
Gross earnings from paid work	Sex	T <sub>1</sub>	Working full time or part time
		T <sub>2</sub>	Employee or self employed
		T <sub>3</sub>	Economic sector
		T <sub>4</sub>	Socio-economic grade of work
		T <sub>5</sub>	Level of education
		T <sub>6</sub>	Qualifications
		T <sub>7</sub>	Age
		T <sub>8</sub>	Geographical region

#### Test variables: (Work-related)

If the respondent is working, several work-related variables are available:

There is no single variable for working part-time (Under 30 hours a week) and working full-time (30 or more hours a week). There are two separate derived variables, one for employees and another for the self-employed:

**[EjbHrCal]** "Hours R works per week, including overtime [employee].

**[SJbHrCal]** "Hours R works per week, including overtime [self-employed].

These variables are mutually exclusive, so will have to be combined into a single variable.

**[Remploye]** "Employee/self-employed",

**[rocsect2]** "Economic sector" and

**[rnsoccl]** "Social class of job"

### 3.2.1.5 [Earnings differences 2009] Download and check file

#### Demographics

##### Age

**[Rage]** "Respondent's age last birthday"

There are two existing groupings for age, but it may be preferable to create completely new groups

**[Ragecat]** Age of respondent (grouped into 7 categories"  
**[RAgeCat2]** Age of respondent (grouped into 6 categories

##### Education

**[Tea]** "Age completed full-time education"  
**[hedqual2]** "Highest educational qualification"

##### Geographical

**[GOR2]** "Region within UK"  
**[country]** "Country within UK".

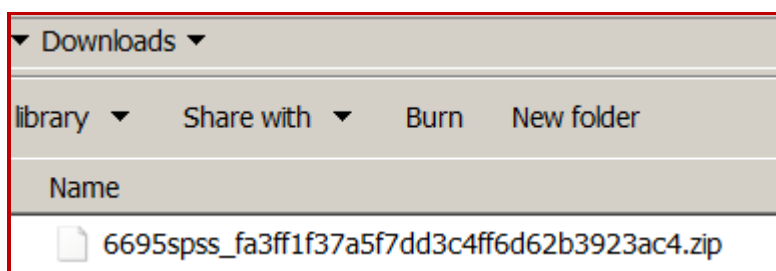
#### Downloading the data file from the UK Data Service (UKDS)

If you wish to perform the exercises yourself, you need to be a **registered user** with UKDS, be **logged in** and have **Depositor Authorisation** to download and use the data<sup>3</sup>. However, even without the actual data and without access to SPSS, you should be able to understand and follow this exercise.

The documentation and data for the 2009 survey are on [BSAS 2009](#) at UKDS.

The files arrive in your **Downloads** folder in a \*.zip folder:

**6695spss\_fa3ff1f37a5f7dd3c4ff6d62b3923ac4.zip**



**Either** 1) Create a new folder **BSAS 2009** on Desktop:

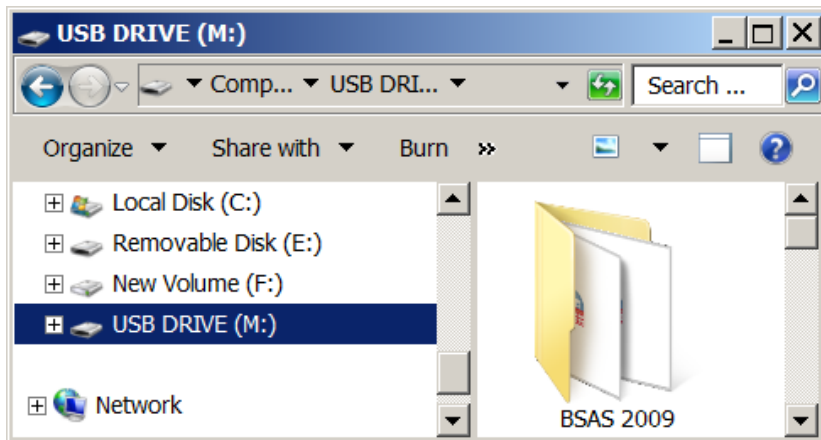


<sup>3</sup> For downloading and access see: [Downloading British Social Attitudes Survey \(BSAS\) data from the UK Data Service](#)

### 3.2.1.5 [Earnings differences 2009] Download and check file

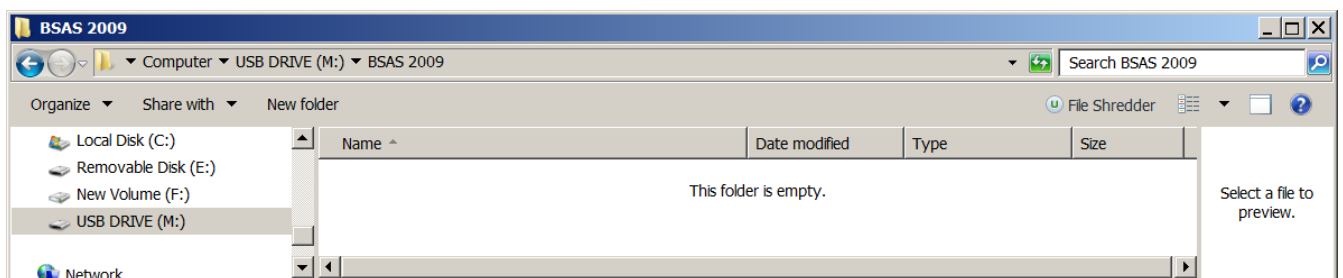
or [preferable for users needing to move between machines]

2) Create a new folder **BSAS 2009** on a USB stick (here drive **M:**)

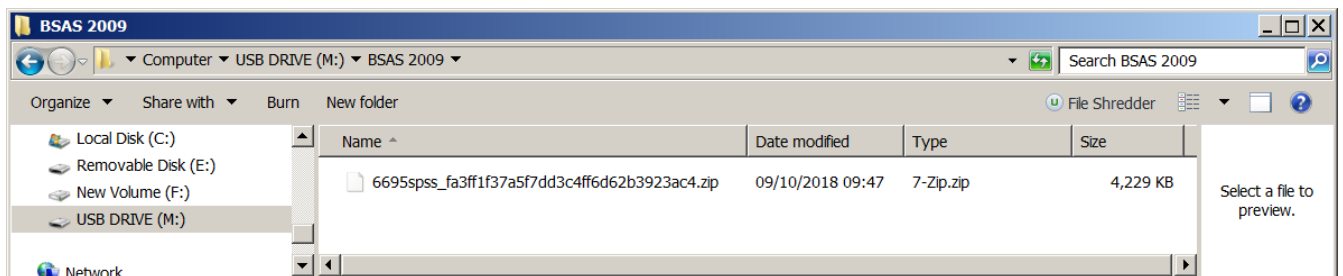


**USB DRIVE (M:)** is used in this and the following sessions.

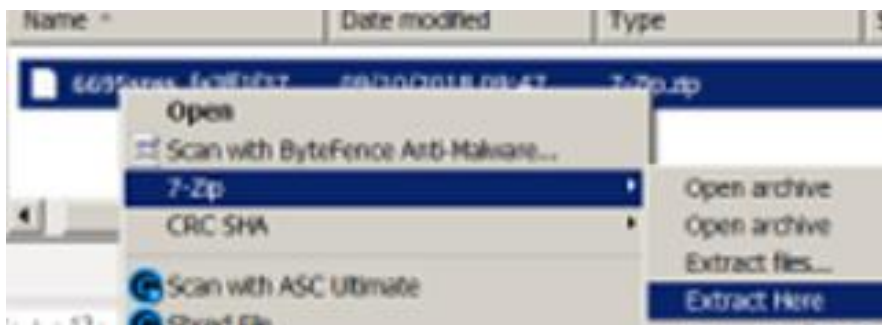
Open folder **BSAS 2009**:



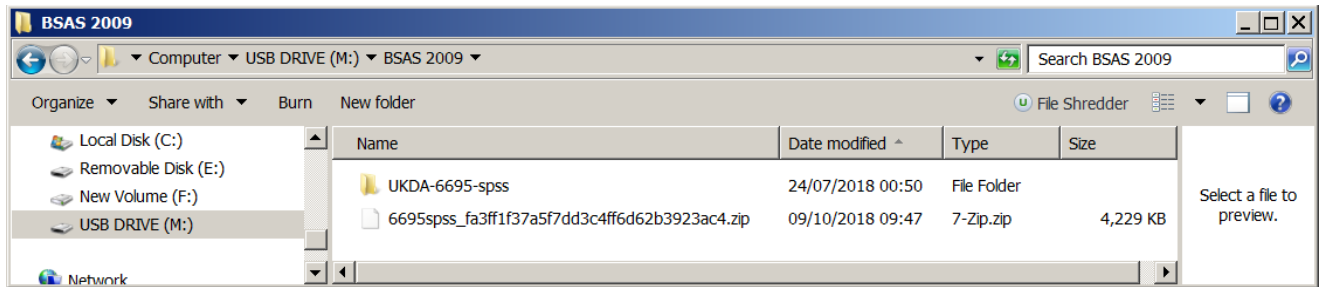
Copy the zip file from your **Download** folder to  **BSAS 2009**



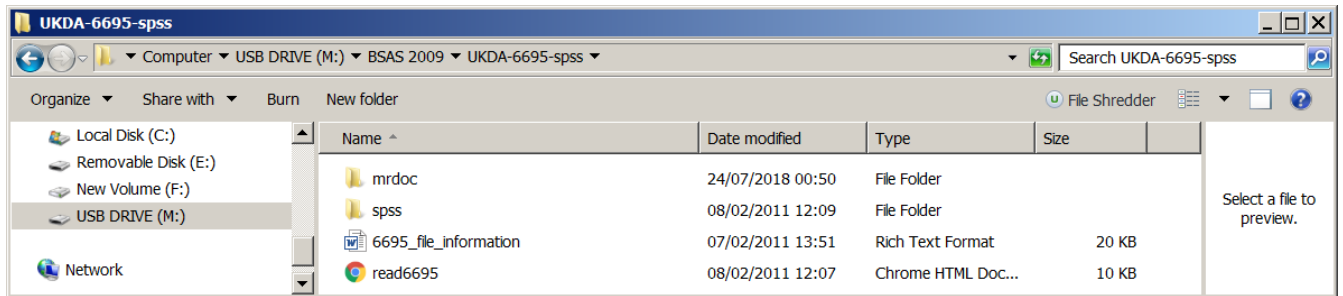
**Right click** the zip folder >> **left click 7-zip** >> **left click Extract here**:



### 3.2.1.5 [Earnings differences 2009] Download and check file

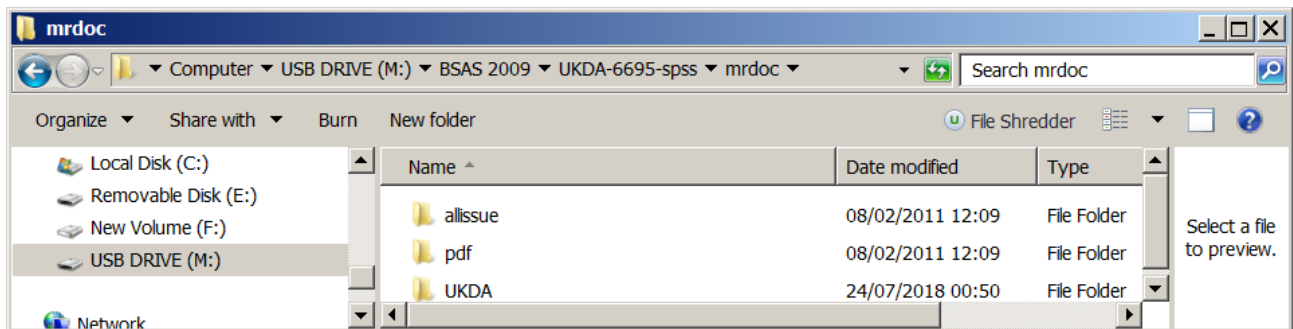


Double click  UKDA-6695-spss

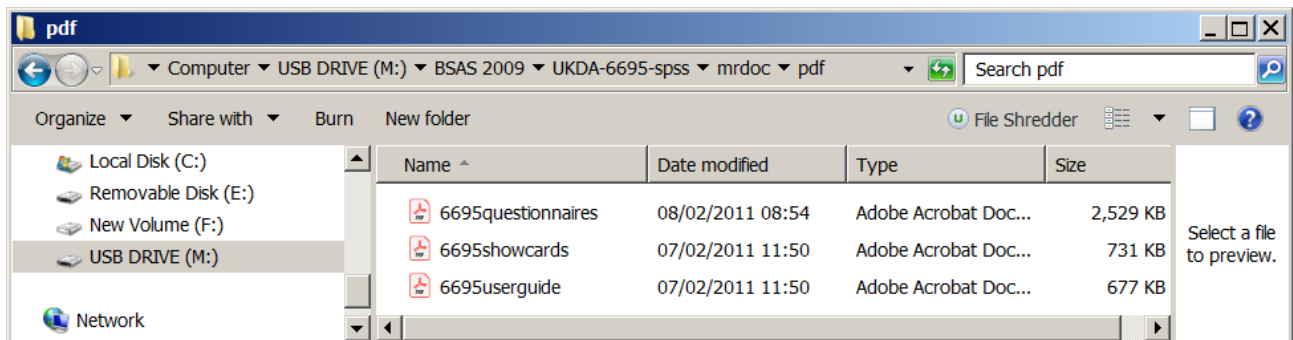


To see the documentation:


Double click  mrdoc

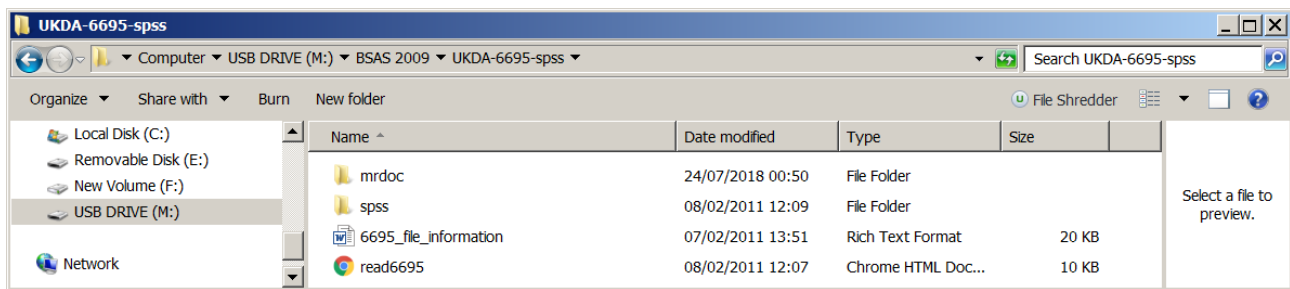


Double click  pdf



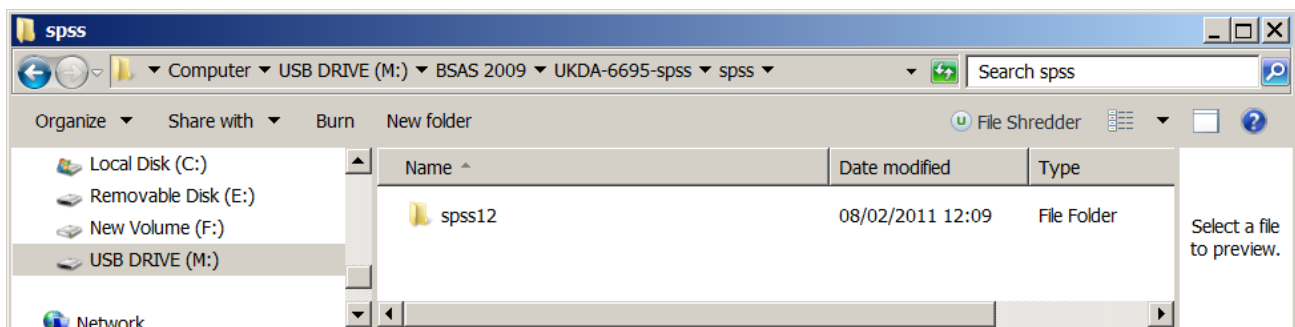
### 3.2.1.5 [Earnings differences 2009] Download and check file


We don't need to see these at this point, so go back to  **BSAS 2009**



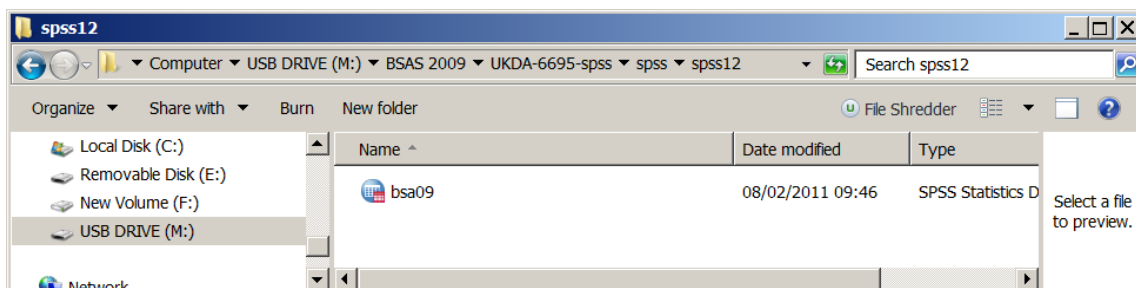
To open the SPSS saved file:

Double click  **spss**



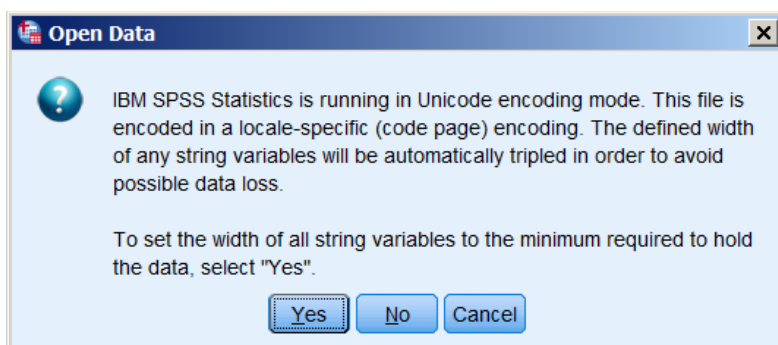
[NB: The  **spss12** icon indicates that the file was created using SPSS release 12]

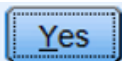

Double click  **spss12**



Double click  **bsa09**

[If you are using SPSS 20 or later, this notification will appear]



Click  to open  **bsa09**



### 3.2.1.5 [Earnings differences 2009] Download and check file

#### Beginning of Data Editor in Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Serial	Numeric	6	0	Serial Number...	None	None	10	Right	Scale	Input
2	SPoint	Numeric	3	0	Sample point:...	None	None	7	Right	Scale	Input
3	StratID	Numeric	4	0	Stratification I...	None	None	9	Right	Nominal	Input
4	PopBand	Numeric	1	0	Population De...	{1, 0-2.789...	None	9	Right	Nominal	Input
5	GOR2	Numeric	2	0	Government of...	{1, North E...	None	5	Right	Nominal	Input
6	WtFactor	Numeric	10	4	final BSA weig...	None	None	12	Right	Nominal	Input

Scroll down to the end of the file (the slider at the right-hand edge is quicker):

#### End of file in Variable View

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
843	Censor	Numeric	2	0	Censorship fil...	{-1, Skip, n...	{-1, -2	8	Right	Nominal	Input
844	leftrigh	Numeric	6	4	Left-right scale...	{-1.0000, S...	{-1.0000, -2...	10	Right	Scale	Input
845	libauth	Numeric	6	4	Libertarian-aut...	{-1.0000, S...	{-1.0000, -2...	9	Right	Scale	Input
846	welfare2	Numeric	6	4	welfare scale B	{-2.0000, S...	{-1.0000, -2...	10	Right	Scale	Input
847	Qtime	Numeric	2	0	How long to co...	{-1, Skip, n...	{-1, -2	7	Right	Nominal	Input
848											

The last non-empty row is 847: the file contains **847 variables**.

Variables **[Serial]** **[SPoint]** and the ages of respondent and other household members are declared as **Scale**, as are the derived variables **[leftrigh]** **[libauth]** and **[welfare2]**: all other variables are specified as **Nominal**. These and other anomalies will be dealt with later.

Click on **Data View**

	Serial	SPoint	StratID	PopBand	GOR2	WtFactor	OldWt	ABCVer	Country
1	210005	123	12	2	12	0.8339	1.1089	2	2
2	210008	307	104	4	9	0.6650	0.5545	2	1
3	210009	312	106	4	8	1.4435	1.1089	1	1

Scroll down to the end of the file (the slider at the right-hand edge is quicker):

	Serial	SPoint	StratID	PopBand	GOR2	WtFactor	OldWt	ABCVer	Country
3420	216776	180	40	2	3	0.7031	0.5545	3	1
3421	216778	219	60	3	4	0.4229	0.5545	3	1
3422									

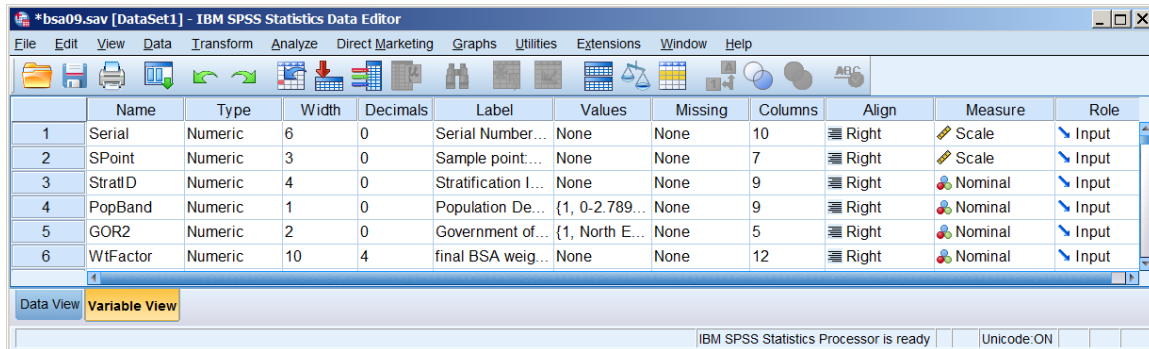
The last non-empty row is 3421: the file contains **3421 cases**.



### 3.2.1.5 [Earnings differences 2009] Download and check file

#### Checking contents of the downloaded file

Go back to **Variable View**

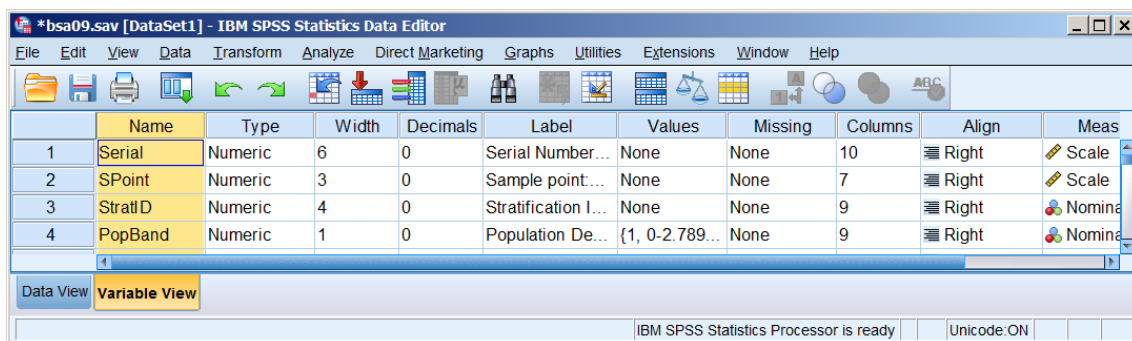


#### Dependent variables

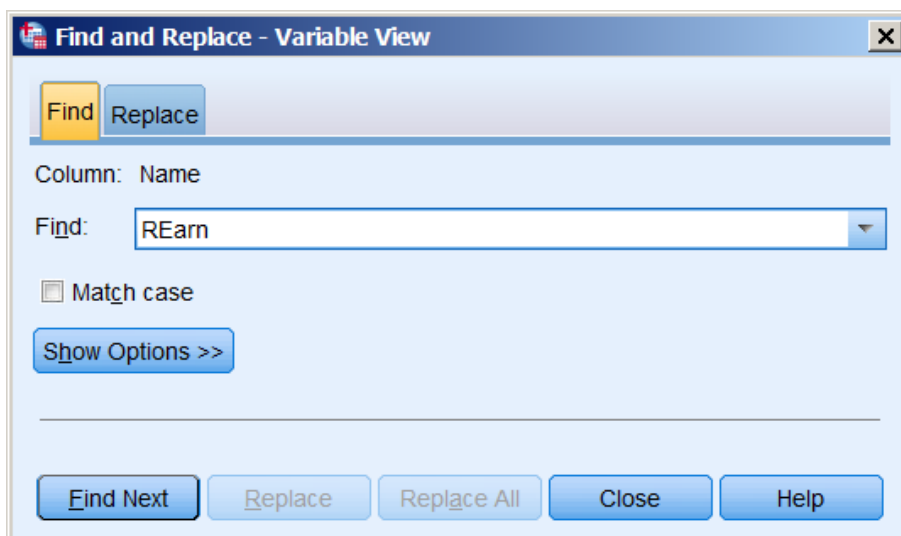
[REarn] " R's own gross or total earnings,before income tax+national insurance?:Q1376"

[REarnQ] " respondent earnings quartiles (dv):Q1377"

To find variable [REarn] click the column header for **Name** to highlight the whole column:



Press **Ctrl+F** and write "REarn" in the **Find:** box



Click **Find Next**

### 3.2.1.5 [Earnings differences 2009] Download and check file

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
615	MainInc4	Numeric	2	0	What is the m...	{1, Earning...	None	10	Right	Nominal	Input
616	HHIncome	Numeric	2	0	total income of...	{1, Q: less t...	None	10	Right	Nominal	Input
617	HHIncQ	Numeric	1	0	Household inc...	{1, less tha...	None	8	Right	Nominal	Input
618	REarn	Numeric	2	0	R's own gross ...	{-1, Skip, n...	None	7	Right	Nominal	Input

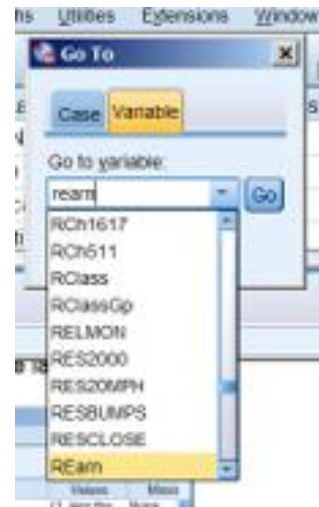
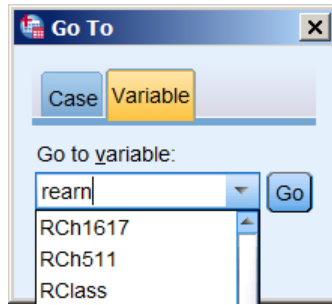
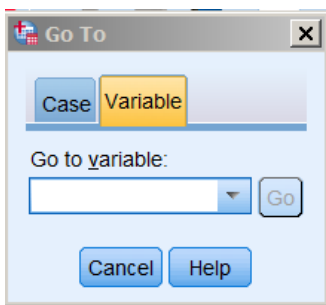
Variable **[REarn]** is on row **618** of the **Data Editor**

A quicker way to find **[REarn]** is:

In the **Data Editor** <sup>4</sup>

**Edit** >> **Go to variable:**

. .write "rearn" <sup>5</sup> in the box and click



In the **Data Editor** **REarn** is highlighted on row 618:

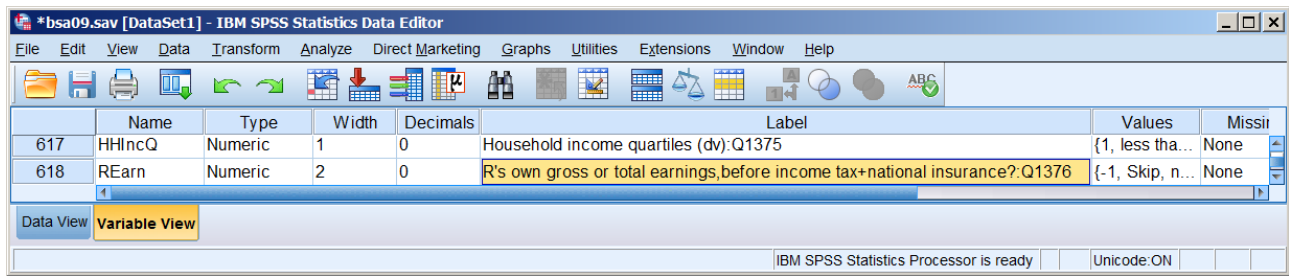
	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
615	MainInc4	Numeric	2	0	What is the m...	{1, Earning...	None	10	Right	Nominal	Input
616	HHIncome	Numeric	2	0	total income of...	{1, Q: less t...	None	10	Right	Nominal	Input
617	HHIncQ	Numeric	1	0	Household inc...	{1, less tha...	None	8	Right	Nominal	Input
618	REarn	Numeric	2	0	R's own gross ...	{-1, Skip, n...	None	7	Right	Nominal	Input

<sup>4</sup> To the author that's another new one!

<sup>5</sup> Variable names in SPSS are case insensitive. The author's preference is to use lower case.

### 3.2.1.5 [Earnings differences 2009] Download and check file

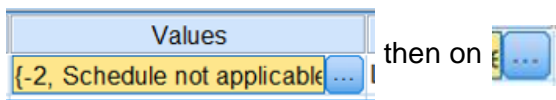
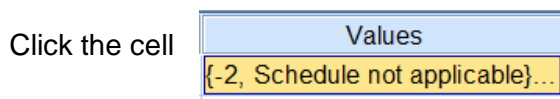
To see the full variable label, drag the right edge of the **Label** column to the right.



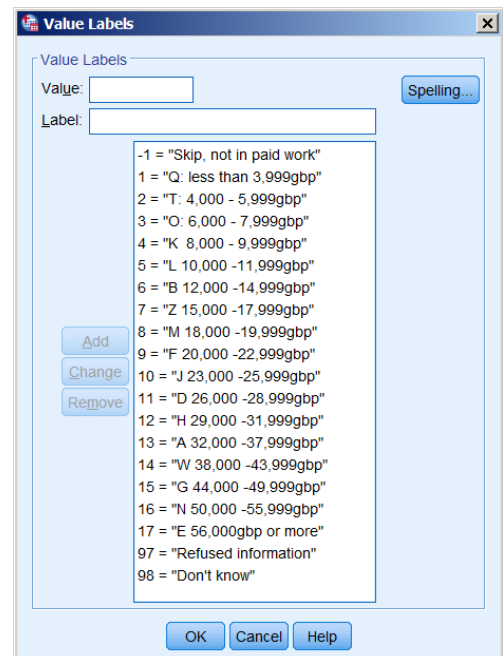
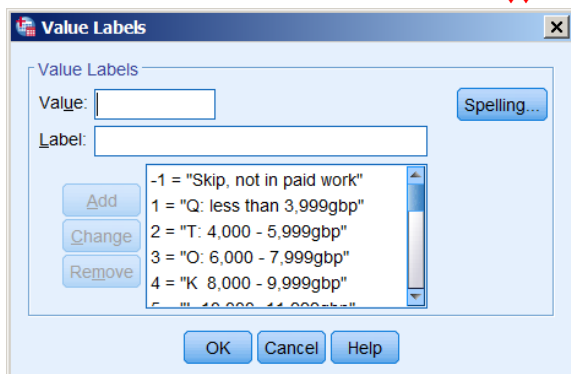
[It reads, "R's own gross or total earnings,before income tax+national insurance?:Q1376"]

To see the **value labels**, stay on the same line.

In the **Values** column:



then on



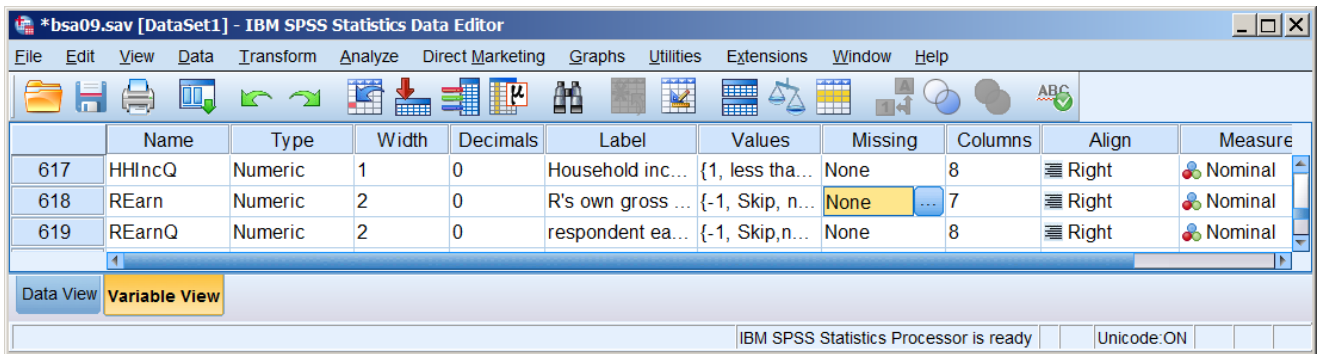
Drag the bottom edge down to see all the labels: → → →

Values **97** "Refused information" and **98** "Don't know" are **not declared as missing**.

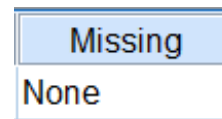
Click **OK** to return to the **Data Editor**

### 3.2.1.5 [Earnings differences 2009] Download and check file

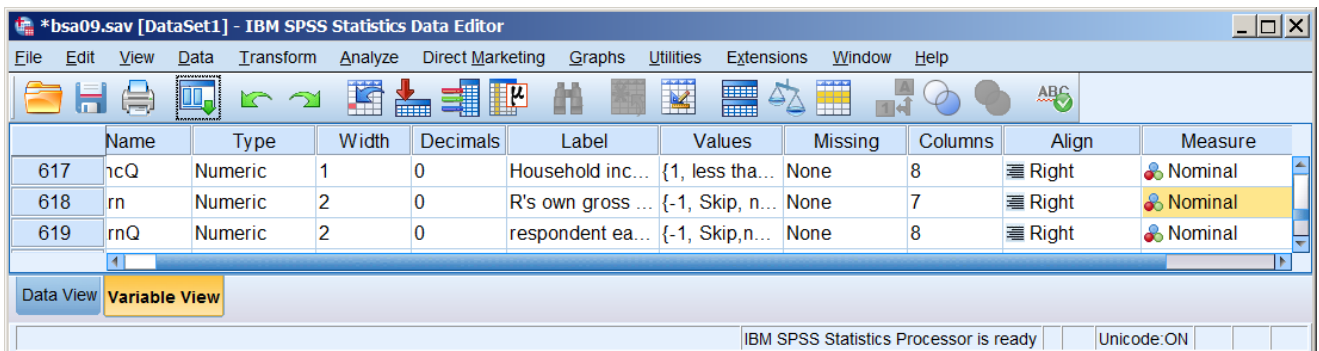
In the **Missing** column:



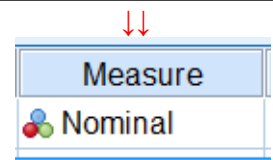
.. no missing values are declared



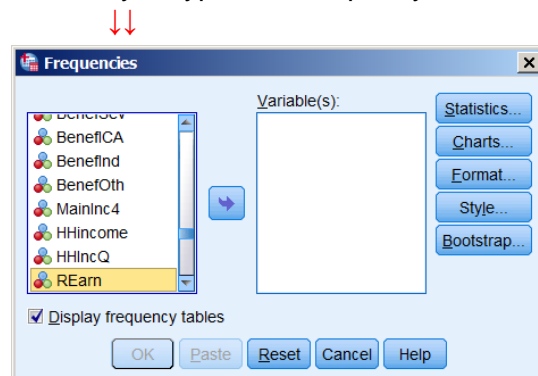
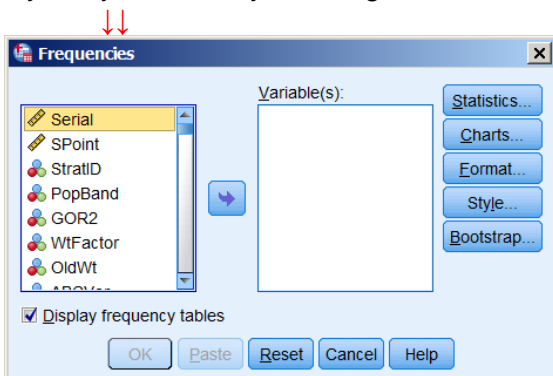
In the **Measure** column:




The level of measurement is declared as **Nominal**, but it should be **Ordinal**



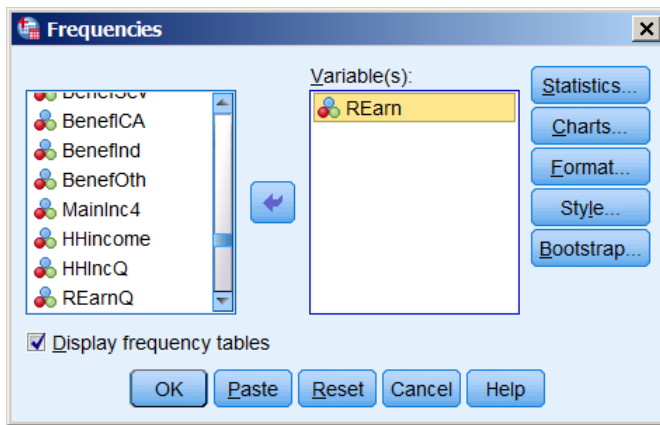
Checking the frequency count for **[REarn]** using the Graphic User Interface (GUI) will take forever if you try to find it by scrolling .. but if you type "rearn" quickly in the left pane <sup>6</sup>



Click on the blue arrow  to transfer **[REarn]** to the **Variable(s)** box:

<sup>6</sup> That's a new trick, even for the author!

### 3.2.1.5 [Earnings differences 2009] Download and check file



Click on **OK** to obtain the frequency count.

#### Frequency counts for dependent variable

**Table 1: Frequency count for [REarn] (no values declared as missing)**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	<b>Skip, not in paid work</b>	<b>1558</b>	<b>45.5</b>	<b>45.5</b>	<b>45.5</b>
	Q: less than 3,999gbp	42	1.2	1.2	46.8
	T: 4,000 - 5,999gbp	74	2.2	2.2	48.9
	O: 6,000 - 7,999gbp	70	2.0	2.0	51.0
	K 8,000 - 9,999gbp	76	2.2	2.2	53.2
	L 10,000 -11,999gbp	133	3.9	3.9	57.1
	B 12,000 -14,999gbp	160	4.7	4.7	61.8
	Z 15,000 -17,999gbp	163	4.8	4.8	66.5
	M 18,000 -19,999gbp	91	2.7	2.7	69.2
	F 20,000 -22,999gbp	117	3.4	3.4	72.6
	J 23,000 -25,999gbp	135	3.9	3.9	76.6
	D 26,000 -28,999gbp	130	3.8	3.8	80.4
	H 29,000 -31,999gbp	85	2.5	2.5	82.8
	A 32,000 -37,999gbp	116	3.4	3.4	86.2
	W 38,000 -43,999gbp	89	2.6	2.6	88.8
	G 44,000 -49,999gbp	60	1.8	1.8	90.6
	N 50,000 -55,999gbp	33	1.0	1.0	91.6
	E 56,000gbp or more	115	3.4	3.4	94.9
	<b>Refused information</b>	<b>143</b>	<b>4.2</b>	<b>4.2</b>	<b>99.1</b>
	<b>Don't know</b>	<b>27</b>	<b>0.8</b>	<b>0.8</b>	<b>99.9</b>
	<b>99</b>	<b>4</b>	<b>0.1</b>	<b>0.1</b>	<b>100.0</b>
	<b>Total</b>	<b>3421</b>	<b>100.0</b>	<b>100.0</b>	

[NB: For statistical analysis, categories " **Skip, not in paid work**", "**Refused information**", "**Don't know**" and "**99**" clearly need to be treated as missing, but are **not declared as missing values**.

Consequently, the figures could be misleading for **Valid Percent** and are inaccurate for **Cumulative Percent**.]

Clicking on **Paste** inserts the following syntax into the current **Syntax Editor**

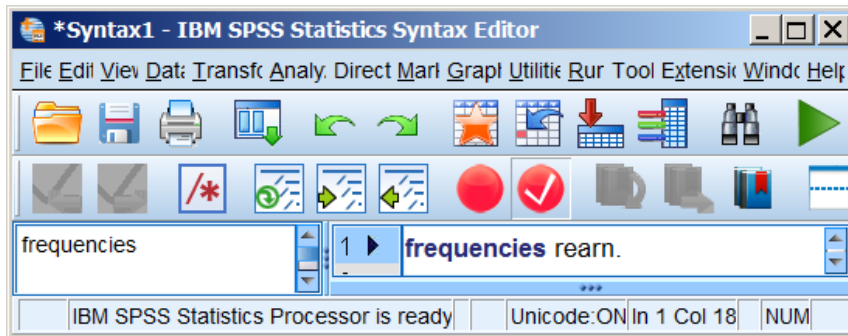
### 3.2.1.5 [Earnings differences 2009] Download and check file


```

DATASET ACTIVATE DataSet1.
FREQUENCIES VARIABLES=R_Earn
/ORDER=ANALYSIS.

```

However, to get Table 1 above, it's both **quicker** and **easier** to write " **frequencies** rearn." in the **Syntax Editor**:



.. and press the green arrow .

When missing values for **[R\_Earn]** are specified as **(-1, 97 thru 99)** the figures for **Valid Percent** are accurate.

**missing values** R\_Earn (-1, 97 thru 99).  
**frequencies** rearn.

**Table 2: Frequency count for [R\_Earn] (missing values declared)**

R's own gross or total earnings,before income tax+national insurance?:Q1376					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Q: less than 3,999gbp	42	1.2	2.5	2.5
	T: 4,000 - 5,999gbp	74	2.2	4.4	6.9
	O: 6,000 - 7,999gbp	70	2.0	4.1	11.0
	K 8,000 - 9,999gbp	76	2.2	4.5	15.5
	L 10,000 -11,999gbp	133	3.9	7.9	23.4
	B 12,000 -14,999gbp	160	4.7	9.5	32.9
	Z 15,000 -17,999gbp	163	4.8	9.7	42.5
	M 18,000 -19,999gbp	91	2.7	5.4	47.9
	F 20,000 -22,999gbp	117	3.4	6.9	54.8
	J 23,000 -25,999gbp	135	3.9	8.0	62.8
	D 26,000 -28,999gbp	130	3.8	7.7	70.5
	H 29,000 -31,999gbp	85	2.5	5.0	75.5
	A 32,000 -37,999gbp	116	3.4	6.9	82.4
	W 38,000 -43,999gbp	89	2.6	5.3	87.7
	G 44,000 -49,999gbp	60	1.8	3.6	91.2
	N 50,000 -55,999gbp	33	1.0	2.0	93.2
	E 56,000gbp or more	115	3.4	6.8	100.0
	Total	1689	49.4	100.0	
Missing	Skip, not in paid work	1558	45.5		
	Refused information	143	4.2		
	Don't know	27	0.8		
	99	4	0.1		
	Total	1732	50.6		
Total		3421	100.0		

### 3.2.1.5 [Earnings differences 2009] Download and check file

This is one of the few times that **Cumulative Percent** is useful: it helps to locate cutting points for percentiles.

Contingency tables with 17 income groups would be unwieldy: the approximate quartile cutting points indicated in **violet** above can be used to create four earnings groups of approximately equal size.

The file contains a derived variable **[REarnQ]** which actually does this:

619	<b>REarnQ</b>	Numeric	2	0	respondent earnings quartiles (dv):Q1377	{
-----	---------------	---------	---	---	--	---

The values of **[REarnQ]** range from **-1** to **8**

No missing values are declared

Missing
97 - 99, -1
None
None

**frequencies** rearnq .

**Table 3: Frequency count for **[REarnQ]** (no missing values declared)**

REarnQ		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Skip,not in paid work	1558	45.5	45.5	45.5
	less than 11999	395	11.5	11.5	57.1
	12000- 19999	414	12.1	12.1	69.2
	20000- 31999	467	13.7	13.7	82.8
	32000 or more	413	12.1	12.1	94.9
	Refused information	147	4.3	4.3	99.2
	Don't know	27	0.8	0.8	100.0
	Total	3421	100.0	100.0	



### 3.2.1.5 [Earnings differences 2009] Download and check file

[NB: For statistical analysis, categories "Skip, not in paid work", "Refused information", "Don't know" clearly need to be treated as missing, but are **not declared as missing values**. Consequently, the figures could be misleading for **Valid Percent** and are inaccurate for **Cumulative Percent**.] When values **-1**, **7** and **8** are declared as missing:

**missing values** rearnq (-1, 7, 8).  
**frequencies** rearnq .

**Table 4: Frequency count for [REarnQ] (missing values declared)**

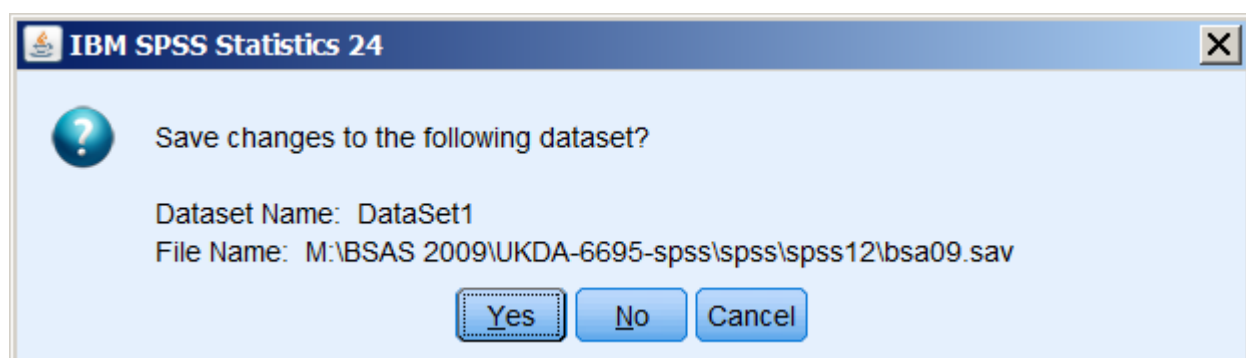
REarnQ		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	less than 11999	395	11.5	23.4	23.4
	12000- 19999	414	12.1	24.5	47.9
	20000- 31999	467	13.7	27.6	75.5
	32000 or more	413	12.1	24.5	100.0
	Total	1689	49.4	100.0	
Missing	Skip,not in paid work	1558	45.5		
	Refused information	147	4.3		
	Don't know	27	0.8		
	Total	1732	50.6		
Total		3421	100.0		

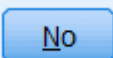
A rule of thumb for percentages is that base **n** should not be less than 40 because, with base 40, moving a single case from one category (-2.5%) to another (+2.5%) makes a net difference of five percentage points.

Elaboration compares percentages of categories in the dependent variable falling within categories of the independent and test variables. The above table has around 400 cases in each non-missing category, but as we progress through zero- order, 1<sup>st</sup> - order, 2<sup>nd</sup> - order tables, controlling for test variables, the base for percentages (n=100%) will get progressively smaller.

This grouping is much easier to use for elaboration because there are approximately equal numbers in each category. In the elaboration exercises to follow it is therefore preferable to use **[REarnQ]** as the dependent variable.

Close  bsa09



Click  (Changes will be made in later sessions, but **on a copy**, **not the original**)

**End of session:** 3.2.1.5 Earnings differences 2009: Download and check file

**Forward to:** 3.2.1.6 Earnings differences 2009: Extracting and saving selected variables

**Back to:** [Block 3 : Analysing two variables \(and sometimes three\)](#)