Survey Analysis Workshop **Copyright © 2010 John F Hall**

Block 2: Analysing one variable **Interval and ratio variables**

2.2.1.1 Frequencies for interval variables

[2 December 2010]

Previous tutorials: **2.1.2.3 - 2.1.2.9 Frequencies etc. for nominal and ordinal variables**

Exemplar: Pre-course questionnaire

Task: What is the age distribution of the sample and what is its average age? What shape does the distribution have? Where are the cutting points to divide the sample into three approximately equal groups?

File: myclass3.sav

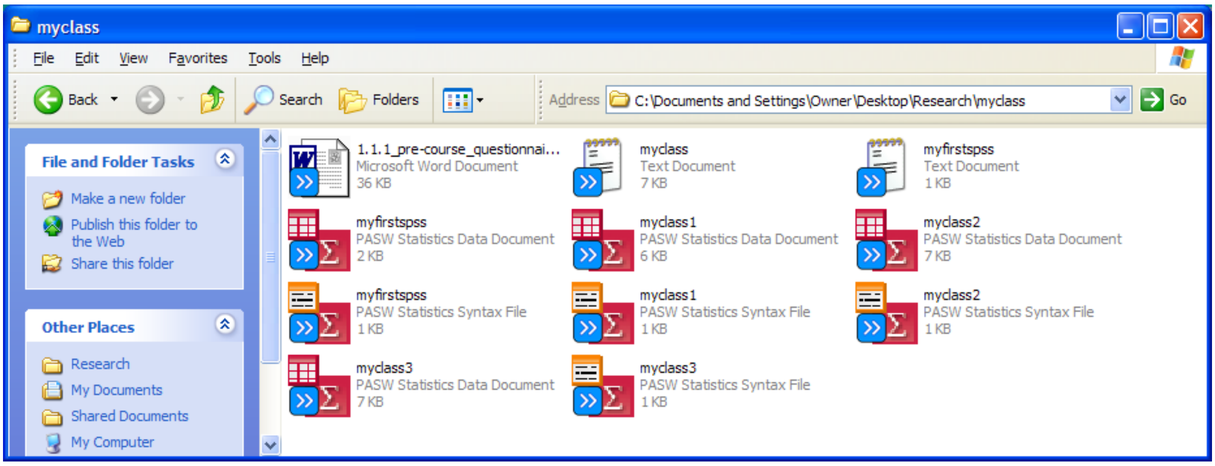
SPSS commands used: FREQUENCIES[1]

We have already seen how to use the SPSS command **FREQUENCIES** to produce tables of distributions, barcharts and pie-charts. Over and above the measurement properties of nominal and ordinal variables, variables measured on interval scales have an additional property, a known metric with fixed intervals (eg age in years, height in metres). Because of this we can legitimately calculate meaningful summary statistics such as averages and measures of spreads and shapes of distributions. If the variable has a true zero point (eg weight in kilograms) it is known as a **ratio** scale: this enables us to say that someone who weighs 90 kg is **twice** as heavy as someone who weighs 45 kg, whereas 100° Fahrenheit is **not** twice as hot as 50° Fahrenheit.

In this first work-through exercise for interval variables we shall be using our original data from the pre-course questionnaire on interests and skills.

Go to your desktop and open your folder **myclass**

If you have done the earlier exercises, the folder will not be empty: your **myclass** folder should already have some files in it.



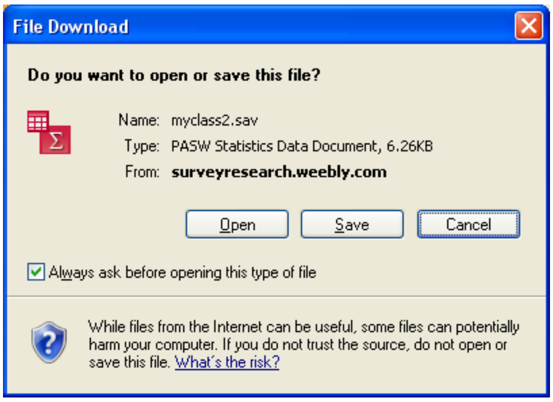
Address: C:\Documents and Settings\Owner\Desktop\Research\myclass

1.1.1_pre-course_questionnai... Microsoft Word Document 36 KB
myclass Text Document 7 KB
myfirstspss Text Document 1 KB
myfirstspss PASW Statistics Data Document 2 KB
myclass1 PASW Statistics Data Document 6 KB
myclass2 PASW Statistics Data Document 7 KB
myfirstspss PASW Statistics Syntax File 1 KB
myclass1 PASW Statistics Syntax File 1 KB
myclass2 PASW Statistics Syntax File 1 KB
myclass3 PASW Statistics Data Document 7 KB
myclass3 PASW Statistics Syntax File 1 KB

[1] **General format:**

```
FREQUENCIES <varlist>
        /STATISTICS <descriptive statistics list >
        / <graphics options>
        /  <other options>
```

The following example uses my version of the the cumulative data from previous waves of students. If you did the full set of exercises from **1.4 Extending your data dictionary**, you can just double-click on file **myclass3.sav** which you should have created in a previous session.   If not you need to download file myclass3.sav from this site and save it in your **myclass** folder.



File Download

Do you want to open or save this file?

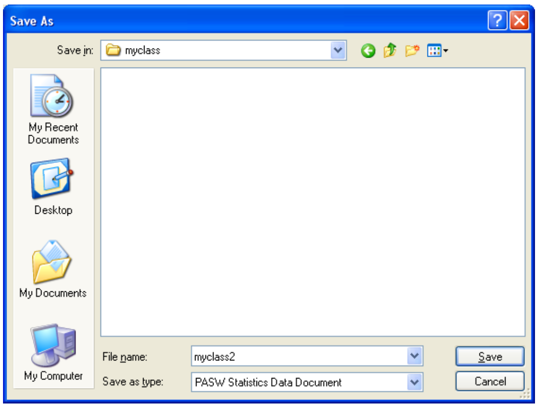
Name: myclass2.sav
Type: PASW Statistics Data Document, 6.26KB
From: **surveyresearch.weebly.com**

☑ Always ask before opening this type of file

While files from the Internet can be useful, some files can potentially harm your computer. If you do not trust the source, do not open or save this file. What's the risk?

Click on Save[2]

. . . and navigate to your **myclass** folder. (If there are already any SPSS saved files in it, their icons will display)

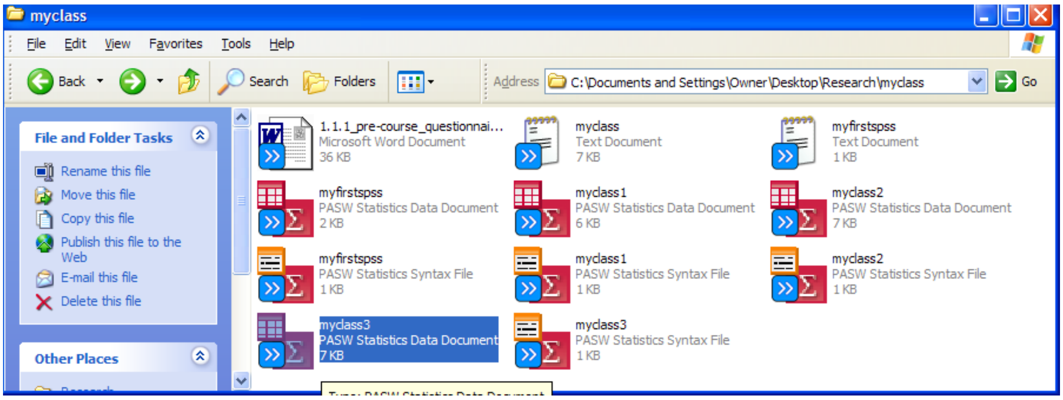


Save As

Save in: myclass

File name: myclass2
Save as type: PASW Statistics Data Document

Click on Save again. If you've followed all the previous exercises, your folder **myclass** should now look like this:



myclass

Address: C:\Documents and Settings\Owner\Desktop\Research\myclass

1.1.1_pre-course_questionnai...
Microsoft Word Document
36 KB

myclass
Text Document
7 KB

myfirstspss
Text Document
1 KB

myfirstspss
PASW Statistics Data Document
2 KB

myclass1
PASW Statistics Data Document
6 KB

myclass2
PASW Statistics Data Document
7 KB

myfirstspss
PASW Statistics Syntax File
1 KB

myclass1
PASW Statistics Syntax File
1 KB

myclass2
PASW Statistics Syntax File
1 KB

myclass3
PASW Statistics Data Document
7 KB

myclass3
PASW Statistics Syntax File
1 KB

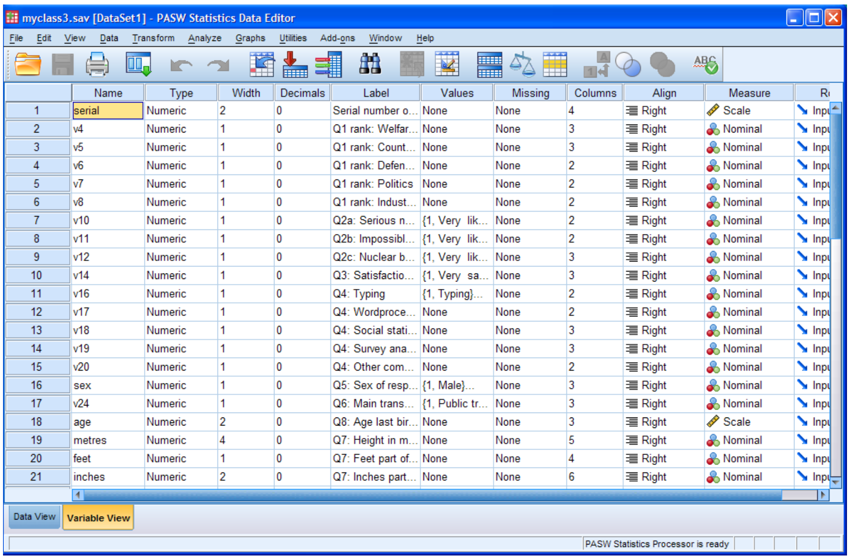
Double click on **myclass3.sav** and wait (...and wait, and wait: SPSS 15 is much quicker!).

[2] If you click on Open **myclass3.sav** will not download, but SPSS will open a blank data editor and syntax editor. It's not yet clear to me if there's a workaround for this, so stick to the instructions above

SPSS automatically generates the following syntax and displays it in the output file:

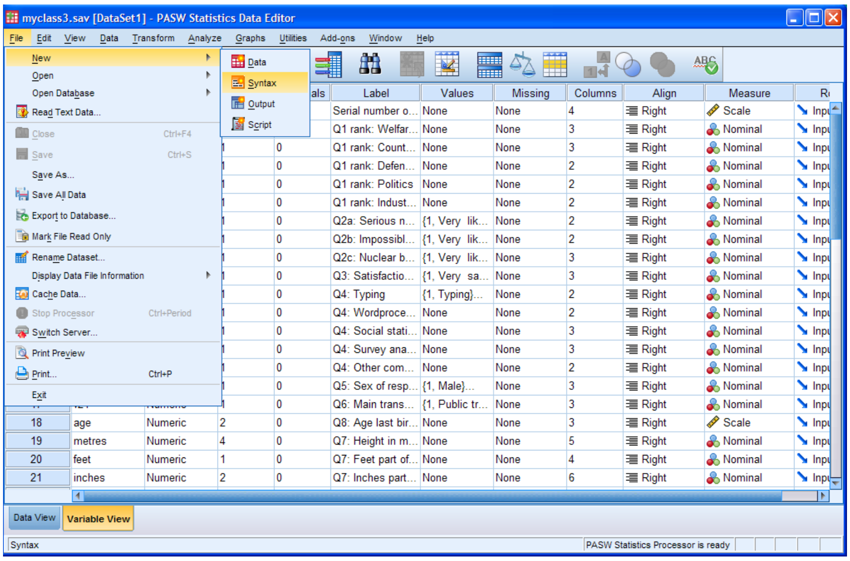
```
GET
  FILE='C:\Documents and Settings\Owner\Desktop\myclass\myclass3.sav'.
DATASET NAME DataSet1 WINDOW=FRONT.
```

If the data editor opens in Data View switch to Variable View.

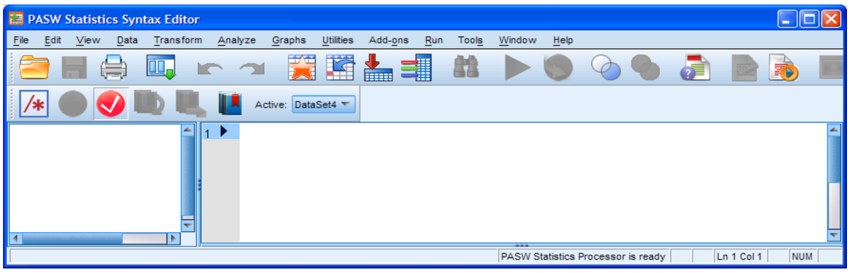

myclass3.sav [DataSet1] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Data View
Variable View
PASW Statistics Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	R
1	serial	Numeric	2	0	Serial number o...	None	None	4	Right	Scale	Inpu
2	v4	Numeric	1	0	Q1 rank: Welfar...	None	None	3	Right	Nominal	Inpu
3	v5	Numeric	1	0	Q1 rank: Count...	None	None	3	Right	Nominal	Inpu
4	v6	Numeric	1	0	Q1 rank: Defen...	None	None	2	Right	Nominal	Inpu
5	v7	Numeric	1	0	Q1 rank: Politics	None	None	2	Right	Nominal	Inpu
6	v8	Numeric	1	0	Q1 rank: Indust...	None	None	2	Right	Nominal	Inpu
7	v10	Numeric	1	0	Q2a: Serious n...	{1, Very lik...	None	2	Right	Nominal	Inpu
8	v11	Numeric	1	0	Q2b: Impossibl...	{1, Very lik...	None	2	Right	Nominal	Inpu
9	v12	Numeric	1	0	Q2c: Nuclear b...	{1, Very lik...	None	3	Right	Nominal	Inpu
10	v14	Numeric	1	0	Q3: Satisfactio...	{1, Very sa...	None	3	Right	Nominal	Inpu
11	v16	Numeric	1	0	Q4: Typing	{1, Typing}...	None	2	Right	Nominal	Inpu
12	v17	Numeric	1	0	Q4: Wordproce...	None	None	2	Right	Nominal	Inpu
13	v18	Numeric	1	0	Q4: Social stati...	None	None	3	Right	Nominal	Inpu
14	v19	Numeric	1	0	Q4: Survey ana...	None	None	3	Right	Nominal	Inpu
15	v20	Numeric	1	0	Q4: Other com...	None	None	2	Right	Nominal	Inpu
16	sex	Numeric	1	0	Q5: Sex of resp...	{1, Male}...	None	3	Right	Nominal	Inpu
17	v24	Numeric	1	0	Q6: Main trans...	{1, Public tr...	None	3	Right	Nominal	Inpu
18	age	Numeric	2	0	Q8: Age last bir...	None	None	3	Right	Scale	Inpu
19	metres	Numeric	4	0	Q7: Height in m...	None	None	5	Right	Nominal	Inpu
20	feet	Numeric	1	0	Q7: Feet part of...	None	None	4	Right	Nominal	Inpu
21	inches	Numeric	2	0	Q7: Inches part...	None	None	6	Right	Nominal	Inpu

Click on File > New > Syntax



. . . to open a blank syntax editor:

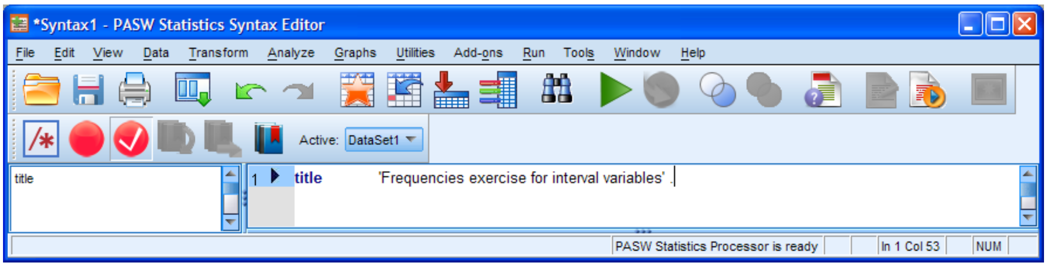

PASW Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet4
1
PASW Statistics Processor is ready
Ln 1 Col 1
NUM

Frequencies work-through exercise:

In my examples I use **UPPER CASE** for general syntax formats and **lower case** for actual **SPSS syntax** needed to run the analyses. I also use the same colour-coding as SPSS for **commands**, **sub-commands** and **keywords**. For text you have to type in yourself I use **light brown**.

Write a **TITLE** command (any text, but don't forget the primes). Start in column 1 and don't forget the full stop (period).

title 'Frequencies exercise for interval variables' .


*Syntax1 - PASW Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet1
title

```
title         'Frequencies exercise for interval variables' .
```

PASW Statistics Processor is ready | In 1 Col 53 | NUM

Click on the green ▶ (or press **[CTRL] + R**) and SPSS displays:

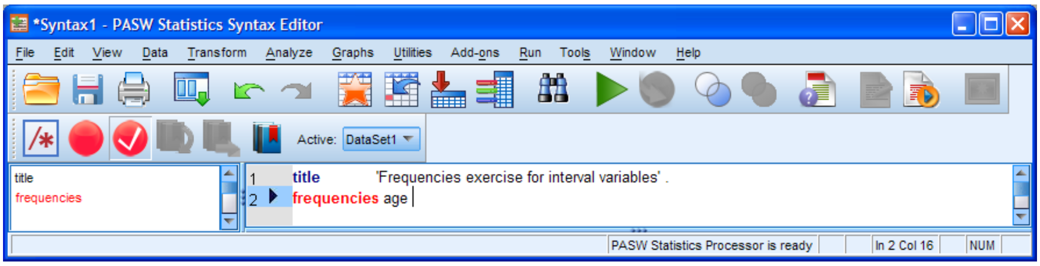
FREQUENCIES

The simplest form of the **FREQUENCIES** command is:

FREQUENCIES <varlist>

Type the following on the next line:

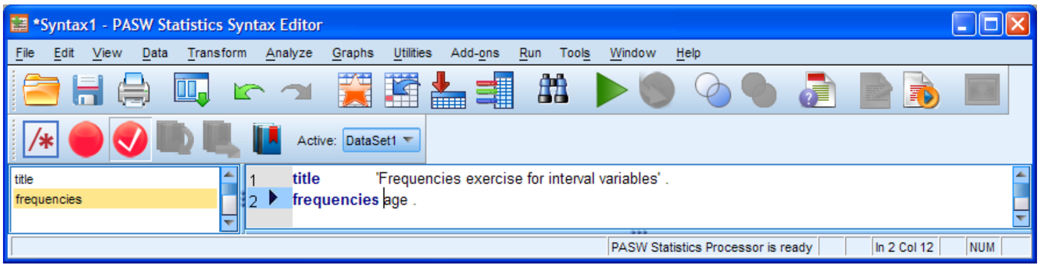
frequencies age .


*Syntax1 - PASW Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet1
title
frequencies

```
title         'Frequencies exercise for interval variables' .
frequencies age
```

PASW Statistics Processor is ready | In 2 Col 16 | NUM

Oops! Forgot the full stop (period)!


*Syntax1 - PASW Statistics Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Run Tools Window Help
Active: DataSet1
title
frequencies

```
title         'Frequencies exercise for interval variables' .
frequencies age .
```

PASW Statistics Processor is ready | In 2 Col 12 | NUM

Click on the green ▶ (or press **[CTRL] + R**) to get:

Statistics

age

N	Valid	166
	Missing	3

age

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	18	2	1.2	1.2	1.2
	19	5	3.0	3.0	4.2
	20	5	3.0	3.0	7.2
	21	10	5.9	6.0	13.3
	22	7	4.1	4.2	17.5
	23	10	5.9	6.0	23.5
	24	5	3.0	3.0	26.5
	25	7	4.1	4.2	30.7
	26	9	5.3	5.4	36.1
	27	6	3.6	3.6	39.8
	28	15	8.9	9.0	48.8
	29	3	1.8	1.8	50.6
	30	11	6.5	6.6	57.2
	31	9	5.3	5.4	62.7
	32	4	2.4	2.4	65.1
	33	3	1.8	1.8	66.9
	34	3	1.8	1.8	68.7
	35	5	3.0	3.0	71.7
	36	7	4.1	4.2	75.9
	37	6	3.6	3.6	79.5
	38	6	3.6	3.6	83.1
	39	4	2.4	2.4	85.5
	40	8	4.7	4.8	90.4
	42	1	.6	.6	91.0
	43	4	2.4	2.4	93.4
	44	3	1.8	1.8	95.2
	46	1	.6	.6	95.8
	48	1	.6	.6	96.4
	49	1	.6	.6	97.0
	50	1	.6	.6	97.6
	52	3	1.8	1.8	99.4
	69	1	.6	.6	100.0
	Total	166	98.2	100.0	
Missing	System	3	1.8		
Total		169	100.0		

There are 169 cases in the file, but only 166 have given their ages. The youngest is 18 and the oldest 69, but there is a gap of 17 years between this case and the next youngest at 52. Cases like this are known as **outliers** and can have a disproportionate effect on statistical analysis. Sometimes it's better to leave them out of the analysis.

Note also that SPSS **has not printed rows** for ages that are not represented in the sample: there are no students aged 41, 45, 47, 51 or 53 to 68.

Descriptive statistics

To get **descriptive statistics** for scale (interval and ratio) variables, in addition to **mode**, **minimum** and **maximum**, there is a range of other statistics to measure centrality, spread and shape.

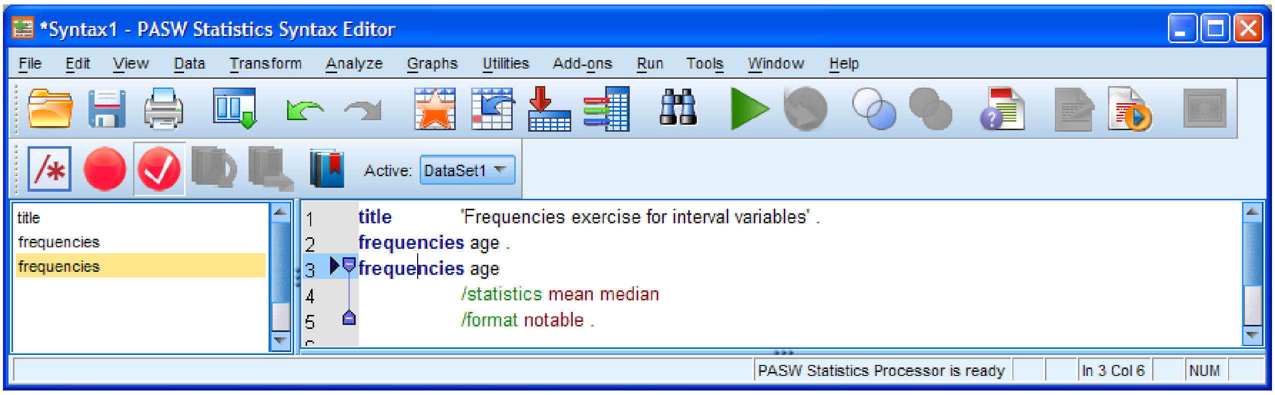
Format:

```
FREQUENCIES <varlist>
        /STATISTICS  followed by one or more of

        MEAN STDDEV MINIMUM MAXIMUM
        (default if /statistics used, but none specified)

        SEMEAN VARIANCE RANGE MODE MEDIAN SKEWNESS

        or ALL
        (This is very naughty if you don't know what they are, or what they're for!)
```

I'll explain what these are later in the course, but for now let's stick to finding the average age of the sample. Go back to the syntax editor and type in:

```
frequencies age
        /statistics mean median
        /format notable .
```

(If you don't suppress the table, SPSS prints that as well, but we already have it above)



```
title        'Frequencies exercise for interval variables' .
frequencies age .
frequencies age
             /statistics mean median
             /format notable .
```

PASW Statistics Processor is ready | In 3 Col 6 | NUM

Click on the green ▶ (or press **[CTRL] + R**) to get:

Statistics

age

N	Valid	166
	Missing	3
Mean		30.59
Median		29.00
Skewness		1.031
Minimum		18
Maximum		69

[**NB:** The calculated mean is misleading: **age** is recorded as age last birthday, so 20 is anywhere between 20 and 21, 35 anywhere between 35 and 36.   Best guess at central value is the mid-point of the interval, so we need to adjust it by adding 0.5 (6 months) to yield a more accurate figure of 31.09.]

Notice that the median is lower than the mean by around two years. This is because the distribution is slightly skewed: the tail is pulled out to the right (positive skew) mainly by the 69-

year-old, but also by all the other cases further away from the middle, which exert more influence and drag the value of the arithmetic mean upwards (think of a child almost on the far end of a see-saw with a parent on the other side closer to the centre, in perfect balance). If the child moves further away, the seesaw will tip towards the child: to restore balance, the fulcrum (balancing point) needs to be moved towards the child. In such cases the **median** (the 50/50 point at which half the cases lie above, and half below) may be a better measure than the **mean** (the balancing point) of where the middle is.

You can obtain cutting points other than 50/50 by asking for **percentiles** (**/PERCENTILES**) These can be as many as you like and anywhere from 0.1% to 99.9%. To get **percentiles** (for interval scale variables only)

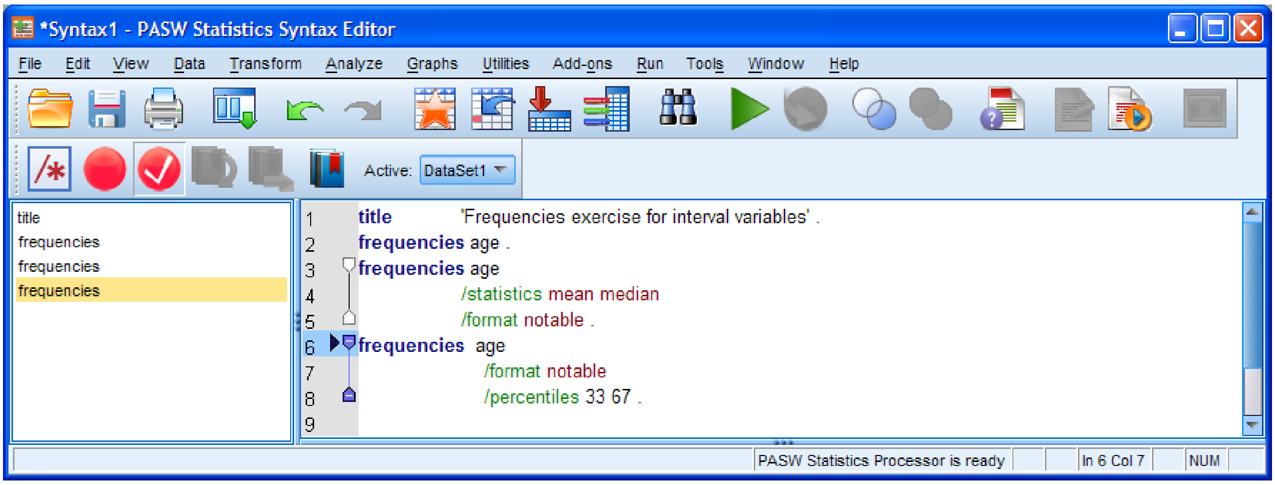
Format:

```
        FREQUENCIES <varlist>
                /PERCENTILES = <cutting point(s)>
                /FORMAT ~ ~ ~
```

To find two cutting points to divide the sample into three age groups of approximately equal size, go back to the syntax editor and type in:

```
      frequencies  age
                /format notable
                /percentiles 33 67 .
```

(If you don't suppress the table, SPSS prints that as well, but we already have it above)



```
title          'Frequencies exercise for interval variables' .
frequencies age .
frequencies age
               /statistics mean median
               /format notable .
frequencies  age
                 /format notable
                 /percentiles 33 67 .
```

Statistics

age

N	Valid	166
	Missing	3
Percentiles	33	26.00
	67	33.89

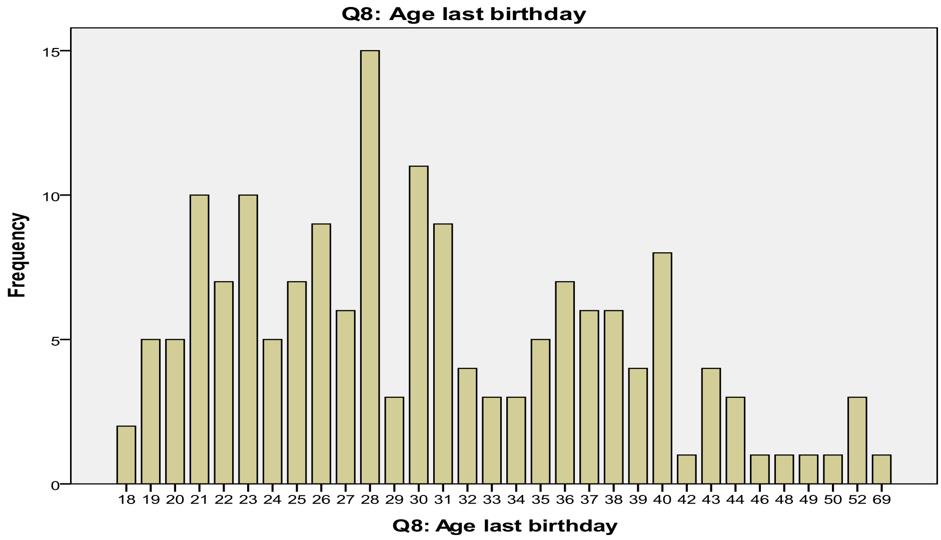
..which can be done just as easily and more usefully by looking at the **Cumulative Percent** column in the original table and deciding to make the first cut between values 25 and 26 and the second between 32 and 33. It's always a good rule of thumb to use common sense rather than mathematical precision, but in this case the cutting points would be the same whichever method you use.

Whilst many statistical routines require maximum precision in variables like age and height, it is also useful to generate groups for tabulation purposes.

Thus, using information from the cumulative frequencies or percentiles above, we can use the SPSS data transformation command **RECODE**[3] to group the values for **age** into a new variable[4] **agegrp3** with three categories of approximately equal size, 18 - 25, 26 - 33, and 34 - 69. Grouped age can then be used in tabulation against other variables.

Charts

As well as frequency tables and descriptive statistics, you can also specify graphic displays. It is legitimate to use a **barchart** (**/BARCHART**) with interval variables, but the chart can be **misleading** if there are **empty cells** as in the table above. The barchart should have gaps to indicate ages which are not present in the data.



Histograms

With scale variables it is better to use a **histogram** (**/HISTOGRAM**). Histograms can only be used when the underlying measure has a **known and fixed interval** (and the width of the bars therefore has meaning). To get a histogram of a distribution

Format:

```
FREQUENCIES  <varlist>
/HISTOGRAM .
```

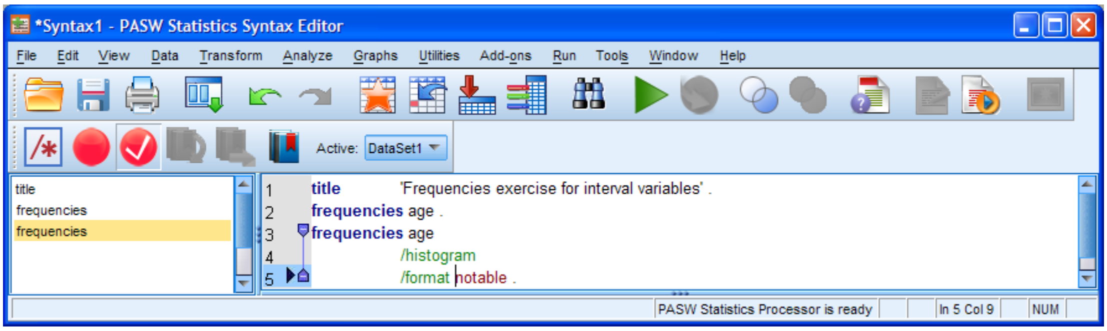
If we only want the chart or the statistics (as here, because we've already got the table) we can suppress the frequency table by **/FORMAT NOTABLE**. (Remember, SPSS colour-codes syntax as you type it into the syntax window.) Go back to the syntax editor and type in the following:

```
frequencies age
        /histogram
        /format notable .
```

[3] See 2.3.1.1 Tutorial - Data transformations

[4] **recode age (18 thru 25 = 1)(26 thru 33 = 2)(34 thru 69 = 3)(else = sysmis) into agegrp3** (See also 2.3.1.1 Data transformations)

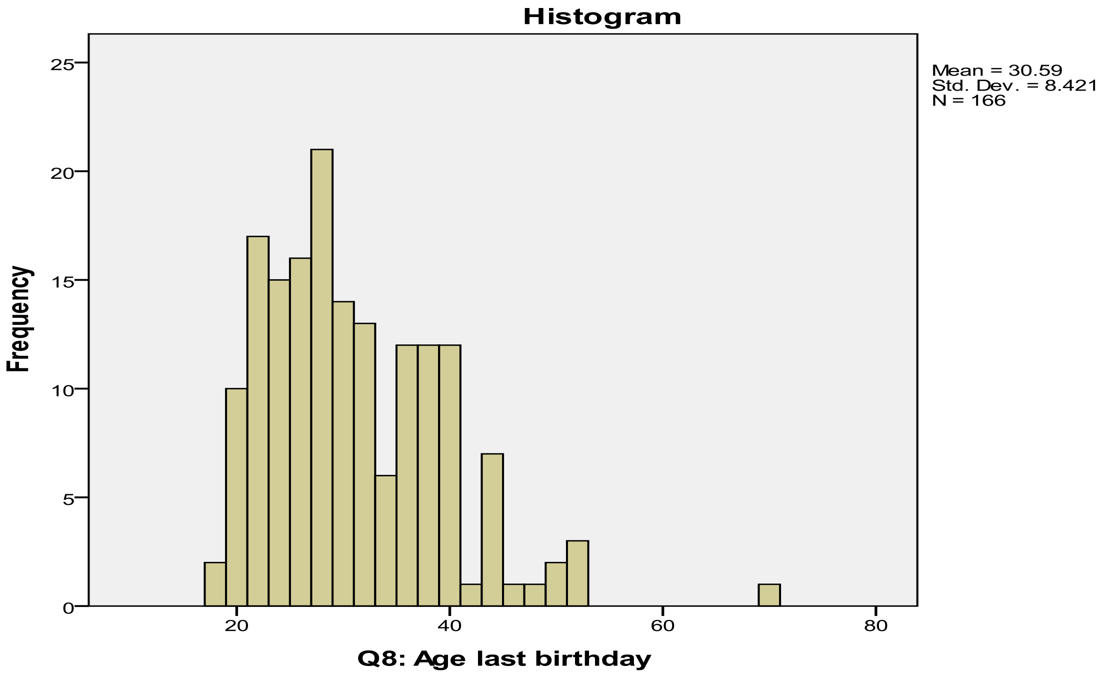
As you type, SPSS will prompt you with menus of commands, sub-commands and keywords for you to select by clicking. You can use these if you like, but your syntax will all be in upper case. I just ignore them and carry on typing: the syntax gets colour-coded anyway.



```
title          'Frequencies exercise for interval variables' .
frequencies age .
frequencies age
               /histogram
               /format notable .
```

PASW Statistics Processor is ready | In 5 Col 9 | NUM

Click on the green ► (or press **[CTRL] + R**) to get:



This shows graphically that the distribution is skewed as the tail is pulled to the right (positive skew) by the older students, particularly the 69-year-old. Note that the values on the horizontal axis represent the mid-point of the intervals: because the values have been grouped, the shape accurately reflects the distribution even though some ages were not present in the data. The outlier aged 69 is clearly visible and the large gap is because there were no ages in between 52 and 69. Note also that **histogram** automatically displays the mean, standard deviation and valid N in a small table at top right.

Normal distribution overlay

For variables like scores on attitude scales, height or weight, it is sometimes useful to be able to overlay the histogram with a normal curve[5]. Age is not really a suitable variable for this as, in the population at large, it tapers from 100% at birth to an eventual 0%. However age at enrolment on this course has a different distribution and is a reasonable candidate for demonstration purposes.

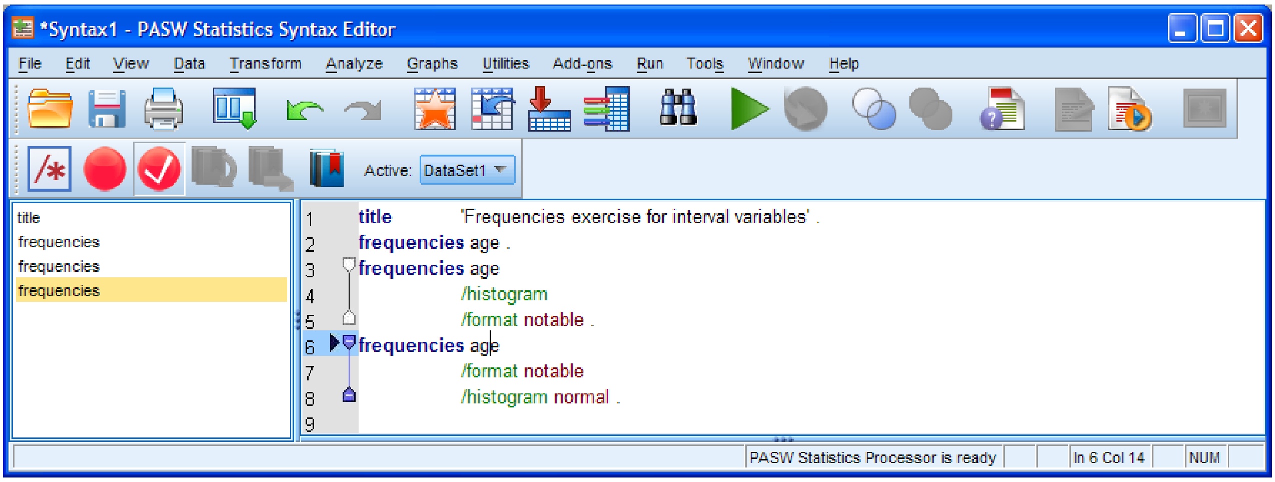
[5] I"ll explain statistical terms elsewhere. For now you are referred to the Statistics notes to accompany course, the recommended textbooks or to the explanations available from the SPSS menus via Help > Statistics Coach

To get a histpogram overlaid with a normal curve, we use **/HISTOGRAM NORMAL .**

```
FREQUENCIES  <varlist>
        /HISTOGRAM NORMAL
        /FORMAT NOTABLE .
```

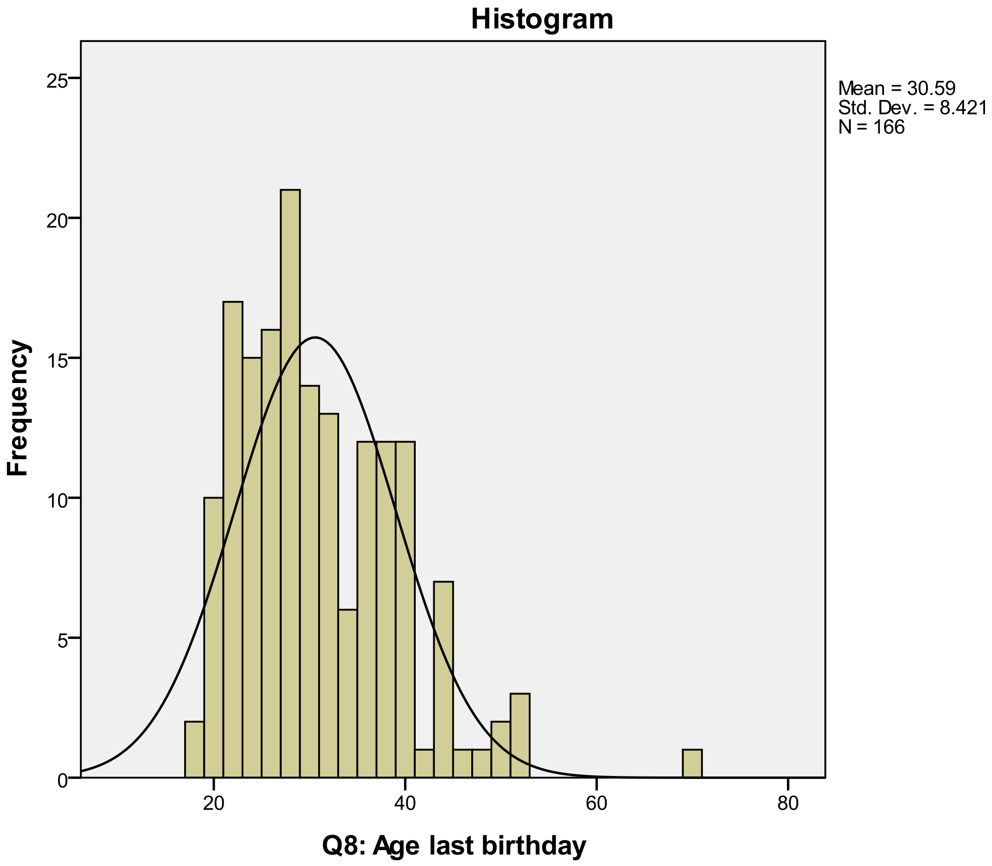
Go back to the syntax editor and type in the following:

```
frequencies age
        /format notable
        /histogram normal .
```



```
title          'Frequencies exercise for interval variables' .
frequencies age .
frequencies age
               /histogram
               /format notable .
frequencies age
               /format notable
               /histogram normal .
```

Click on the green ► (or press **[CTRL] + R**) to get:



As we have already seen, SPSS is case insensitive for syntax, and I prefer to work with abbreviated syntax in lower case as this saves time and is much easier on the eye. However, if you start typing direct into the syntax window, SPSS prompts with menu suggestions. If you click on these, the full command appears in colour-coded **UPPER CASE** text. If the command is incomplete or SPSS can't make sense of the syntax, then all or part of the text will appear in **red**. If you use abbreviated syntax, SPSS may not colour-code all or part of the command. The previous examples follow the colour-coding of SPSS and, for now, I typed them out in full.

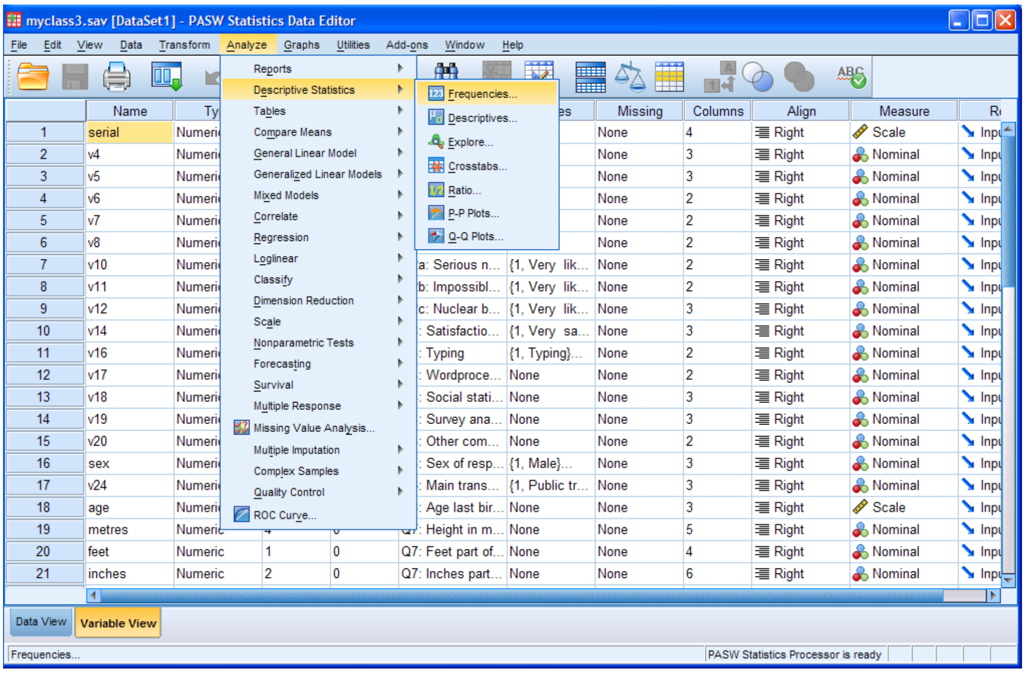
Once you get used to writing SPSS syntax, you can begin to use the abbreviated forms. SPSS only reads the first 3 or 4 characters of commands and keywords anyway, so the following will work just as well, but not everything will be colour-coded in the syntax window. Variable names always have to be written in full. You may also find it useful to put a blank line between commands to keep them visually separate, and also before the full stops as this makes it easier to check that they are present at the end of each command (or not, a common source of errors).

Thus the following examples will also work (but SPSS won't colour-code the abbreviations).

```
freq age   /his nor  /for not .
freq  age  /sta mea std .
freq age   /per 33 67 .
```

Try it yourself and see. Whilst you're at it, why not play around with the **FREQUENCIES** command on other variables in the file? Also, if you're only looking at one variable at a time, why don't you experiment with the drop-down menus using:

Analyze **>** Descriptive Statistics **>** Frequencies



	Name	Ty				es	Missing	Columns	Align	Measure	R
1	serial	Numeri					None	4	Right	Scale	Inpu
2	v4	Numeri					None	3	Right	Nominal	Inpu
3	v5	Numeri					None	3	Right	Nominal	Inpu
4	v6	Numeri					None	2	Right	Nominal	Inpu
5	v7	Numeri					None	2	Right	Nominal	Inpu
6	v8	Numeri					None	2	Right	Nominal	Inpu
7	v10	Numeri			a: Serious n...	{1, Very lik...	None	2	Right	Nominal	Inpu
8	v11	Numeri			b: Impossibl...	{1, Very lik...	None	2	Right	Nominal	Inpu
9	v12	Numeri			c: Nuclear b...	{1, Very lik...	None	3	Right	Nominal	Inpu
10	v14	Numeri			: Satisfactio...	{1, Very sa...	None	3	Right	Nominal	Inpu
11	v16	Numeri			: Typing	{1, Typing}...	None	2	Right	Nominal	Inpu
12	v17	Numeri			: Wordproce...	None	None	2	Right	Nominal	Inpu
13	v18	Numeri			: Social stati...	None	None	3	Right	Nominal	Inpu
14	v19	Numeri			: Survey ana...	None	None	3	Right	Nominal	Inpu
15	v20	Numeri			: Other com...	None	None	2	Right	Nominal	Inpu
16	sex	Numeri			: Sex of resp...	{1, Male}...	None	3	Right	Nominal	Inpu
17	v24	Numeri			: Main trans...	{1, Public tr...	None	3	Right	Nominal	Inpu
18	age	Numeri			: Age last bir...	None	None	3	Right	Scale	Inpu
19	metres	Numeric	4	0	Q7: Height in m...	None	None	5	Right	Nominal	Inpu
20	feet	Numeric	1	0	Q7: Feet part of...	None	None	4	Right	Nominal	Inpu
21	inches	Numeric	2	0	Q7: Inches part...	None	None	6	Right	Nominal	Inpu

However, if you try to analyse a lot of variables this way, you'll begin to understand why I prefer to use syntax rather than the drop-down menus.

Next session: **2.2.1.2: [BSA86] Exercise - Frequencies**

[Back to Block 2 menu]