Descriptive Statistics in SPSS (Practical)

The development of this E-Book has been supported by the British Academy.

Descriptive statistics for categorical variables practical

Welcome to the descriptive statistics practical in which we will look at how to investigate categorical variables in SPSS. Categorical variables can take only predefined values (or categories) and can be of two types - nominal and ordinal. For nominal variables each category has a name but there is no natural order to the categories. For example, a nominal variable might measure "fruit choice", in which it does not make sense to describe an orange as somehow coming between an apple and a banana in an ordering. Ordinal categories by contrast do have a natural ordering. For example, an ordinal variable might measure opinions on an issue with the categories Strongly disagree; Disagree; Agree; and Strongly Agree forming an ordering in terms of "strength of ageement" from low to high. (This variable is an example of a *Likert scale* .) In this practical we will simply look at summaries and plots of categorical variables that apply equally to both nominal and ordinal variables.

In this example we look at the distributions of two variables collected from 15 year-olds in England as part of the 2015 PISA study. The first, INFGGAS, is a measure of a student's scientific awareness in which they were asked to rate how informed they felt about the environmental issue of greenhouse gases in the atmosphere. The second, PAREDU, is a measure of family background - specifically the highest educational qualification held by a parent of the child. Parental education is often used in social science as a broad measure of a family's socio-economic resources.

The first categorical variable we will look at is **INFGGAS**, labelled How informed about greenhouse gases which has 4 categories: *Never heard of this*, *Heard of, but not really able to explain this*, *Know something, could explain in general* and *Familiar with this, could explain well*. The second variable is **PAREDU**, labelled Highest qualification of parent which has 3 categories: *Low: GCSE or equiv*, *Medium: A-level or equiv* and *High: University degree*.

We will begin by looking at how frequent the various categories for **INFGGAS** are by choosing the following in SPSS:

- Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
- Copy the **How informed about greenhouse gases[INFGGAS]** variable into the **Variable(s)** box.
- Click on the **OK** button to produce the tables as shown below.

Two tables will now appear in the output window. The first is really just a count of number of observations in the dataset and how many are missing for the variable, **INFGGAS**.

Statistics

How informed about greenhouse gases

N	Valid	4837
	Missing	357

In this case there are 357 missing observations for variable **INFGGAS** with 4837 valid values. In the second table we can see a list of the different categories in the data and their frequencies:

How informed about greenhouse gases

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Never heard of this	164	3.2	3.4	3.4
	Heard of, but not really able to explain this	618	11.9	12.8	16.2
	Know something, could explain in general	1700	32.7	35.1	51.3
	Familiar with this, could explain well	2355	45.3	48.7	100.0
	Total	4837	93.1	100.0	
Missing	System	357	6.9		
Total		5194	100.0		

This second table has 5 columns which we will now describe. The first column simply gives the categories for the variable, **INFGGAS** so that we can tell what each row refers to. In the second column headed Frequency we get the actual numbers of occurrences of the variable and so we see that there are 164 occurrences of *Never heard of this*, 618 occurrences of *Heard of, but not really able to explain this*, 1700 occurrences of *Know something, could explain in general* and 2355 occurrences of *Familiar with this, could explain well*. This is a useful summary as we can compare the counts within the dataset. We might however not simply be interested in this dataset in isolation and so it is often useful to convert these counts into percentages and this is done in column 3. Here we see that 3.2 percent of observations are in category *Never heard of this*, 11.9 percent of observations are in category *Heard of, but not really able to explain this*, 32.7 percent of observations are in category *Know something, could explain in general* and 45.3 percent of observations are in category *Familiar with this, could explain well*. There are 6.9 percent of observations that are missing and not included in the percentages above and so we often want to look at percentages out of valid observations only. These are given in the fourth column and we see that 3.4 percent of valid

observations are in category *Never heard of this*, 12.8 percent of valid observations are in category *Heard of, but not really able to explain this*, 35.1 percent of valid observations are in category *Know something, could explain in general* and 48.7 percent of valid observations are in category *Familiar with this, could explain well*. Finally in the fifth column we look at cumulative percentages so that we see that 3.4 percent are in the first category, 16.2 percent of valid observations in the first 2 categories, 51.3 percent of valid observations in the first 3 categories and unsurprisingly 100 percent of valid observations are in one of the 4 categories. These cumulative percentages are of most use when the variable is ordinal because, by definition, the numbering and ordering of categories of a nominal variable is arbitrary.
We can now repeat this for our second variable, **PAREDU** and to do this in SPSS we do the following:

- Select **Frequencies** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
- Remove the **How informed about greenhouse gases[INFGGAS]** variable into the **Variable(s)** box.
- Copy the **Highest qualification of parent[PAREDU]** variable into the **Variable(s)** box.
- Click on the **OK** button to produce the table as shown below.

Two tables will again appear in the output window. The first is again a count of number of observations in the dataset and how many are missing for the variable, **PAREDU**.

Statistics

Highest qualification of parent

N	Valid	4759
	Missing	435

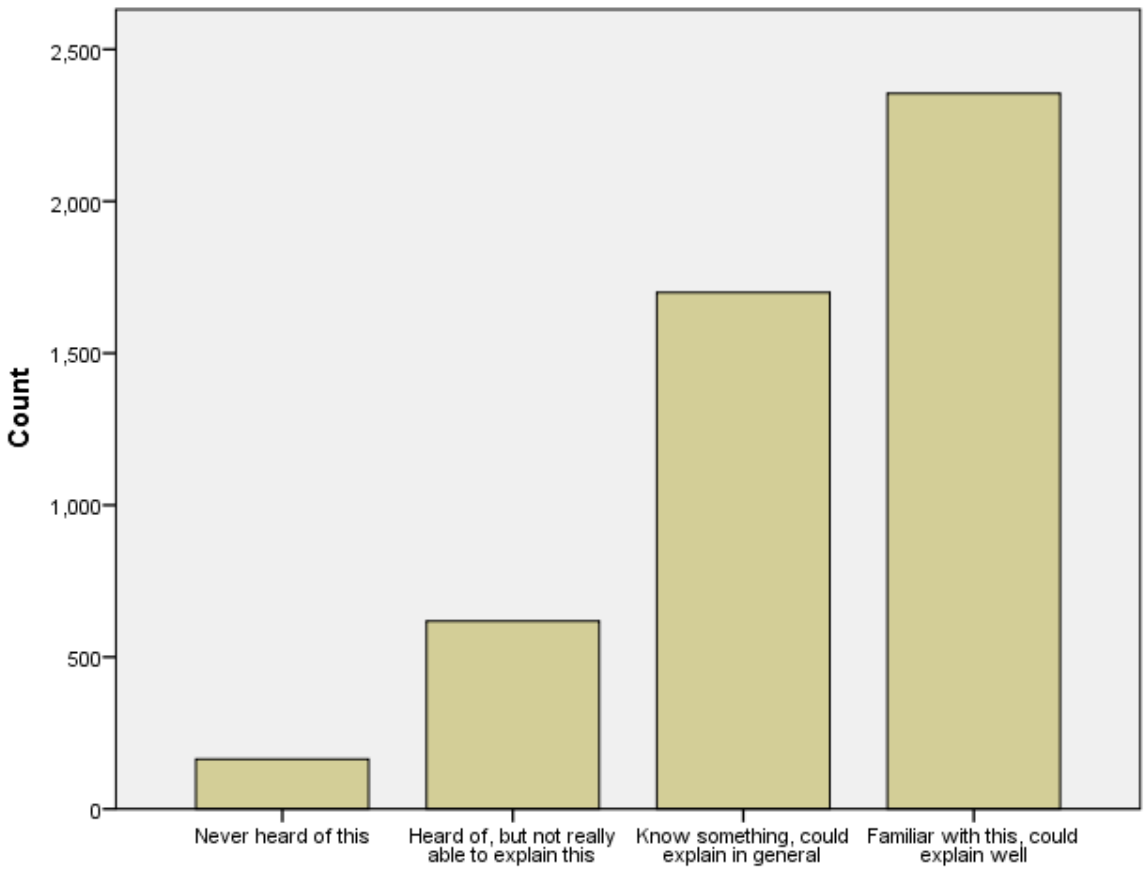
In this case there are 435 missing observations for variable **PAREDU** with 4759 valid values. In the second table we can see a list of the different categories in the data for **PAREDU** and their frequencies:

Highest qualification of parent

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Low: GCSE or equiv	904	17.4	19.0	19.0
	Medium: A-level or equiv	1701	32.7	35.7	54.7
	High: University degree	2154	41.5	45.3	100.0
	Total	4759	91.6	100.0	
Missing	System	435	8.4		
Total		5194	100.0		

Again, the first column simply gives the categories for the variable, **PAREDU** so that we can tell what each row refers to, and the second column headed Frequency gives the actual numbers of occurrences of the variable. Here there are 904 occurrences of *Low: GCSE or equiv*, 1701 occurrences of *Medium: A-level or equiv* and 2154 occurrences of *High: University degree*. Converting these counts into percentages in column 3, we see that 17.4 percent of observations are in category *Low: GCSE or equiv*, 32.7 percent of observations are in category *Medium: A-level or equiv* and 41.5 percent of observations are in category *High: University degree*. There are 8.4 percent of observations that are missing and not included in the percentages above. To see the percentages out of valid observations only we look to the fourth column, which shows that 19.0 percent of valid observations are in category *Low: GCSE or equiv*, 35.7 percent of valid observations are in category *Medium: A-level or equiv* and 45.3 percent of valid observations are in category *High: University degree*. Finally, turning to the cumulative percentages in the fifth column we see that 19.0 percent are in the first category, 54.7 percent of valid observations are in the first 2 categories and 100 percent of valid observations are in one of the 3 categories.
We will next look at how we might illustrate the distribution of the variables graphically. When we have categorical data we can plot the data as a bar graph and to do this for our first variable, **INFGGAS** in SPSS we need to do the following:

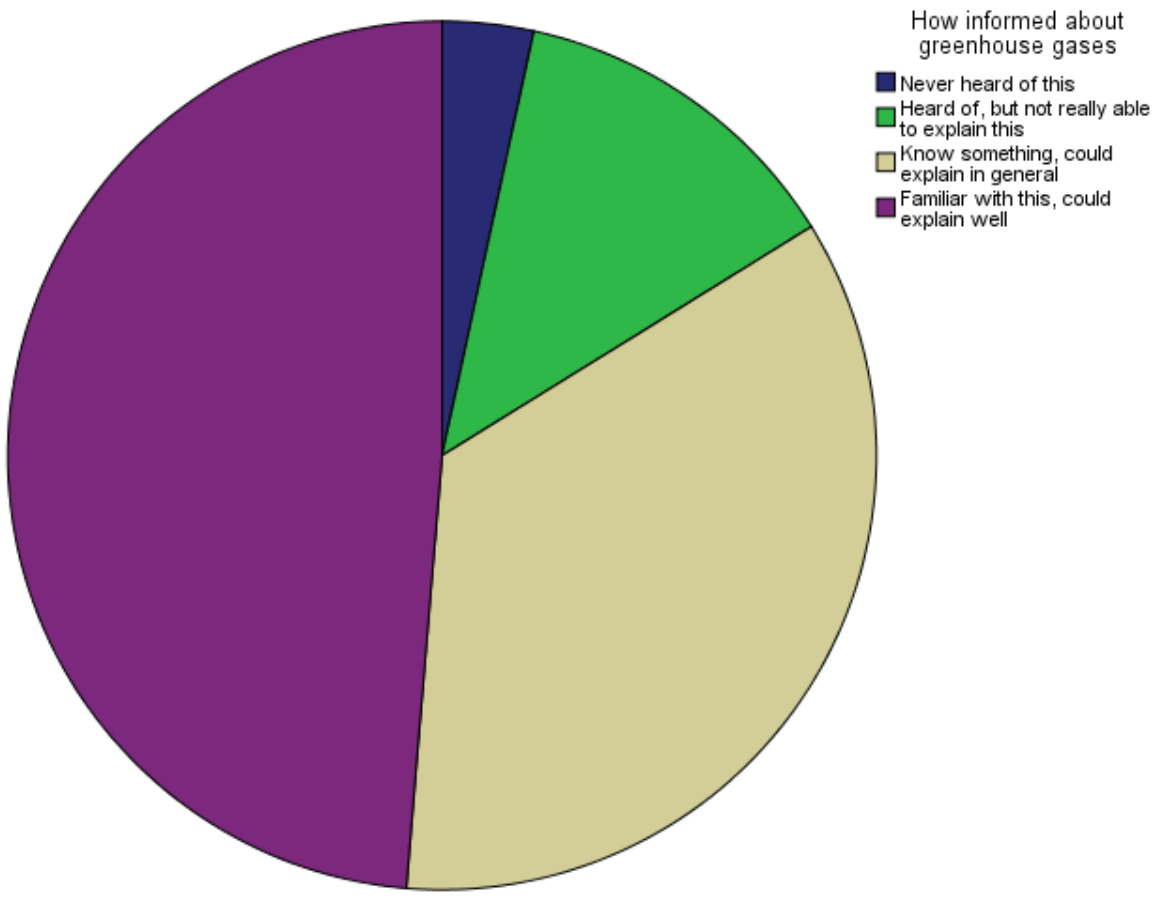
- Select **Bar...** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
- We will use the defaults which are **Simple** and **Summaries for groups of cases** and then click on **Define** .
- Copy the **How informed about greenhouse gases[INFGGAS]** variable into the **Category Axis** box.
- Click on the **OK** button to produce the graph as shown below.



How informed about greenhouse gases

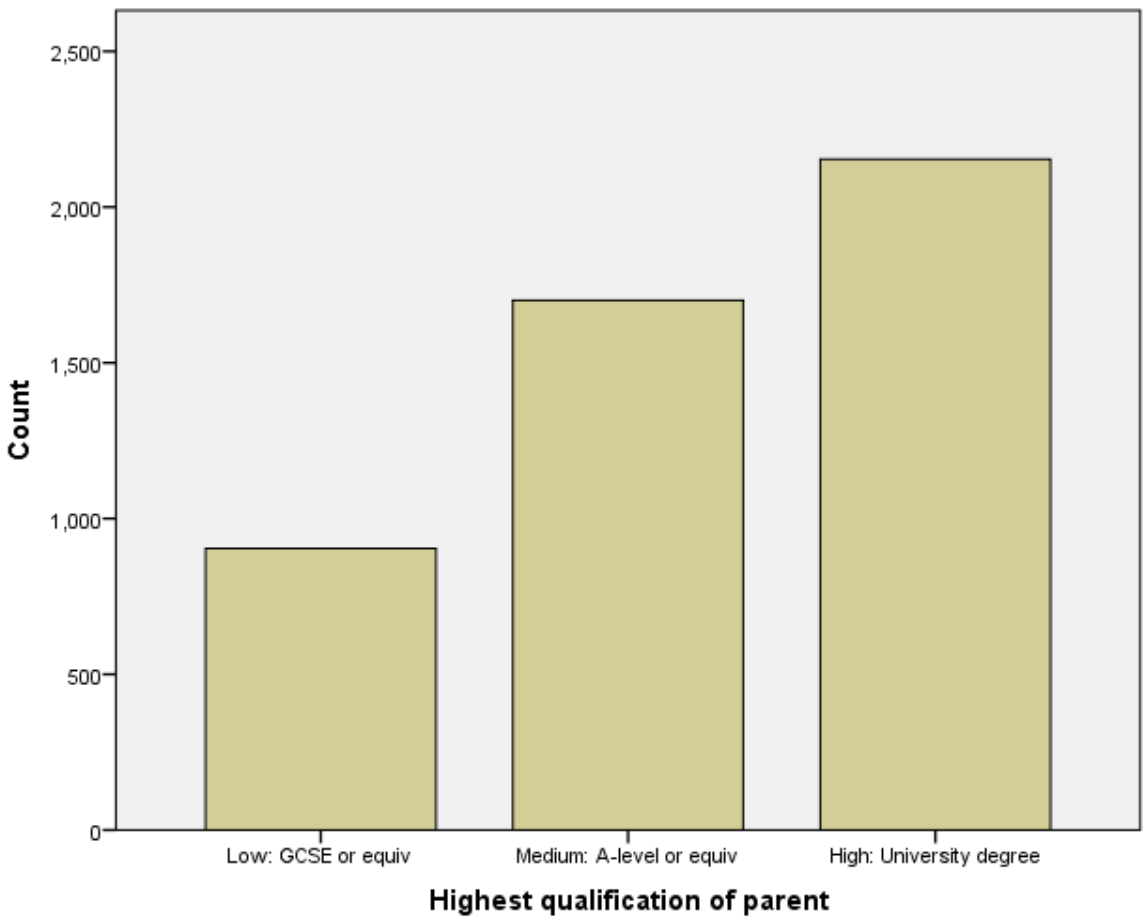
You will see that the graph contains one bar for each category and each bar is labelled with the category name. The graph plots the actual counts in each category so we can see there are for example 164 observations in category *Never heard of this*.
We can also look at the data in the form of a pie chart which is an alternative way of displaying the data (although pie charts are often criticised as providing less visual clarity than a bar chart). This can be done via the following:

- Select **Pie...** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
- Keep the choices as **Summaries for groups of cases** before clicking on **Define**.
- Add the **How informed about greenhouse gases[INFGGAS]** variable in the **Define Slices by:** box.
- Click on the **OK** button to produce the graph as shown below.



Here the pie chart represents each category in a different colour with the proportion of the total that are in each category being proportional to the size of the slices of pie. The circle consists of 360 degrees in total, so the 3.4 percent of (valid) observations that are in category *Never heard of this* are represented by a slice of angle 12 degrees, the 12.8 percent of (valid) observations that are in category *Heard of, but not really able to explain this* are represented by a slice of angle 46 degrees, the 35.1 percent of (valid) observations that are in category *Know something, could explain in general* are represented by a slice of angle 127 degrees and the 48.7 percent of (valid) observations that are in category *Familiar with this, could explain well* are represented by a slice of angle 175 degrees.
We can next plot a bar graph for the second variable, **PAREDU** by doing the following in SPSS:

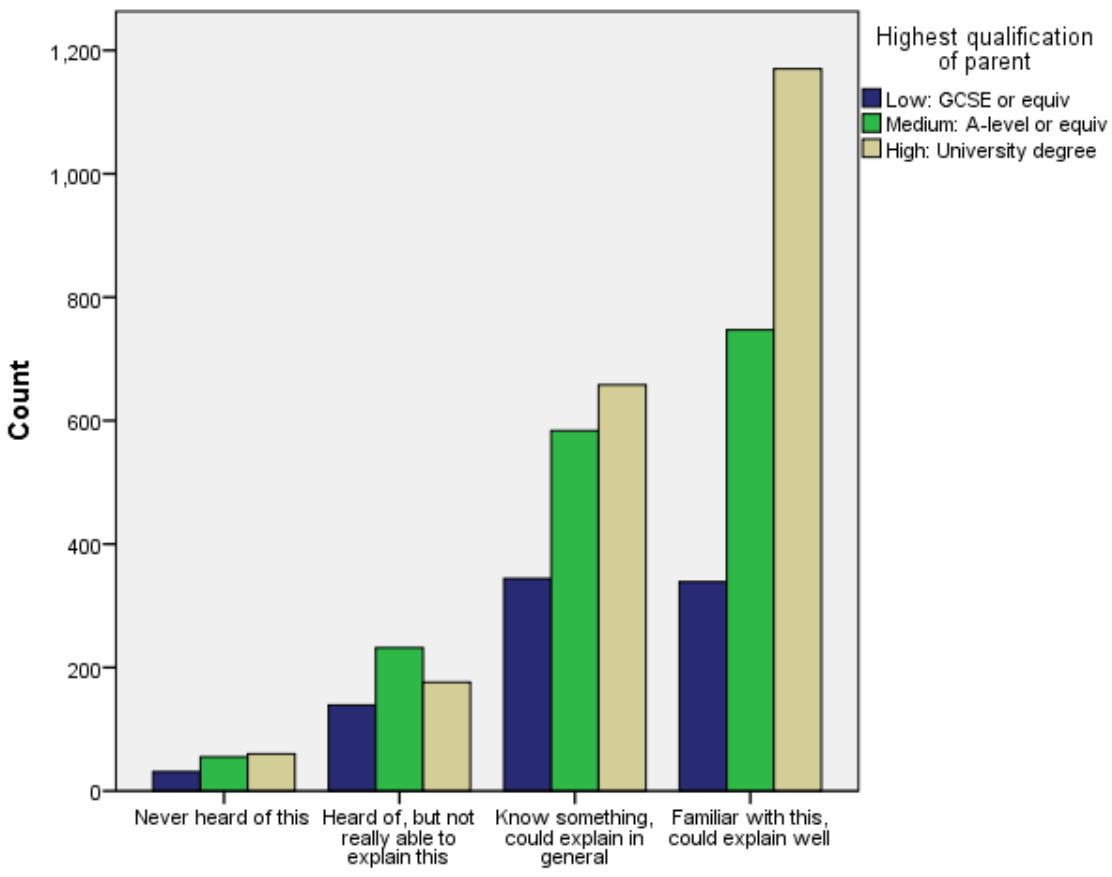
- Select again **Bar...** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
- Use the defaults which are **Simple** and **Summaries for groups of cases** and then click on **Define**.
- Remove the **How informed about greenhouse gases[INFGGAS]** variable into the **Category Axis** box.
- Copy the **Highest qualification of parent[PAREDU]** variable into the **Category Axis** box.
- Click on the **OK** button to produce the graph as shown below.



Highest qualification of parent

This graph should have a similar presentation to the first graph with this time 3 bars and with for example 904 observations in category *Low: GCSE or equiv*. Note that again each bar is labelled, there are labels for each axis of the graph and there are gaps between each bar.
We can also look at the two variables in combination. This is useful for looking at the distribution on one variable relative to the other and can be achieved by doing the following commands in SPSS:

- Select **Bar...** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
- This time we will change from the defaults to **Clustered** and **Summaries for groups of cases** before clicking on **Define** .
- Copy the **How informed about greenhouse gases[INFGGAS]** variable into the **Category Axis** box.
- Copy the **Highest qualification of parent[PAREDU]** variable into the **Define Clusters by** box.
- Click on the **OK** button to produce the graph as shown below.

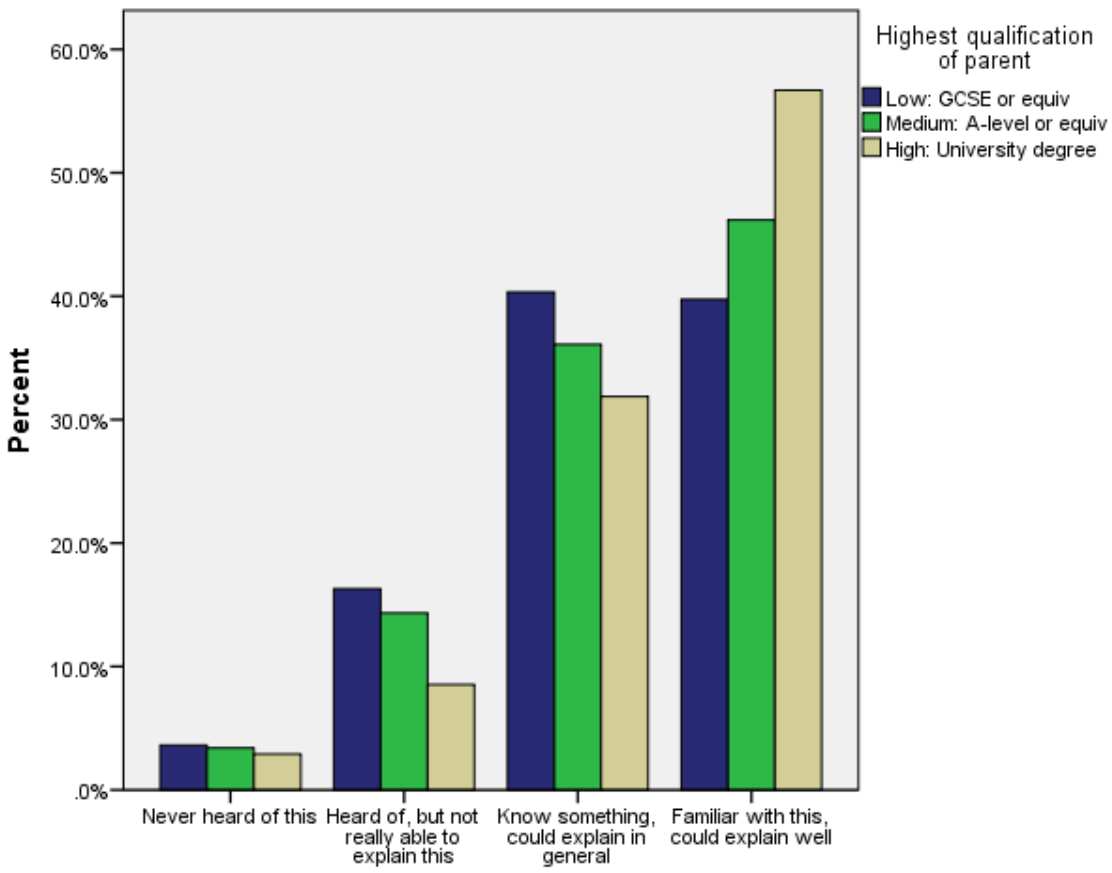


How informed about greenhouse gases

Here we see that each category of **PAREDU** is given a bar of a different colour so that one can, by eye, compare each category of **PAREDU** for different categories of **INFGGAS**. For example we see when How informed about greenhouse gases takes value *Never heard of this*, then there are 31 observations where Highest qualification of parent takes value *Low: GCSE or equiv*, 55 observations where Highest qualification of parent takes value *Medium: A-level or equiv* and 60 observations where Highest qualification of parent takes value *High: University degree*.
When the number of observations in the clustering variable categories are unequal, the plotting of counts makes it difficult to compare the relative composition of the categories plotted on the horizontal axis. As with the frequency tables we saw earlier, conversion to percentages can aid comparisons. To illustrate, to explore whether the distribution of responses to **INFGGAS** is the same for different categories of the variable **PAREDU** we can do the following in SPSS:

- Once again select **Bar...** from the **Legacy Dialogs** submenu available from the **Graphs** menu.
- Keep the choices as **Clustered** and **Summaries for groups of cases** before clicking on **Define**.
- Keep the **How informed about greenhouse gases[INFGGAS]** variable in the **Category Axis** box.

- Keep the **Highest qualification of parent[PAREDU]** variable in the **Define Clusters by** box.
- Select **% of cases** in the **Bars Represent** choices.
- Click on the **OK** button to produce the graph as shown below.



Here we see that now the bars represent the percentage of each category of **INFGGAS** that are found in each category of **PAREDU**. So for example if we look again at the cases where Highest qualification of parent takes value Low: GCSE or equiv, (the blue bars) then 3.6% percent of observations have **INFGGAS** taking value Never heard of this, 16.3% percent of observations have **INFGGAS** taking value Heard of, but not really able to explain this, 40.3% percent of observations have **INFGGAS** taking value Know something, could explain in general and 39.7% percent of observations have **INFGGAS** taking value Familiar with this, could explain well. The chart shows that, compared with this last number, 8.5% of the group for whom **INFGGAS** = Heard of, but not really able to explain this have **PAREDU** = High: University degree.

We have seen that nearly half of all 15 year-olds in England feel that they could explain greenhouse gas issues well. This is more likely among those from backgrounds with highly educated, as opposed to less educated, parents, and children from the least educated parental backgrounds make up nearly one-fifth of the sample.

This ends our practical.