Section 6: Analysing Relationships Between Variables

- Choosing a Technique
- The Crosstabs Procedure
- The Chi Square Test
- The Means Procedure
- The Correlations Procedure

So far, any analysis we have done, has been restricted to simple frequency tables and summary statistics. In this section, we'll begin to look at techniques for analysing the relationships *between* variables. As with the section on summarising variables, we will use the concept of 'levels of measurement' to help guide us through the various options.

Choosing a Technique

There are of course a very wide variety of statistical and analytical procedures available to the data analyst in software packages like SPSS Statistics. In this section, we will introduce some of the most widely-used procedures and explain how to interpret the results. When choosing a particular analytical technique, it's important to keep in mind the level of measurement of the variables concerned. To begin with, we can concentrate on examples concerning *pairs* of variables (these kinds of analyses are known as 'bivariate'). Figure 6.1 lists some of the analytical techniques that are employed when dealing with different combinations of categorical and continuous variables.

Variable Type Combination	Analysis Techniques
Categorical by Categorical	Crosstabs, Tables of Percentages, Clustered/stacked Bar Charts, Panelled Pie Charts
Categorical by Continuous	Tables of means (or other summary measures), Stacked/Grouped Histograms, Error Bars, Box Plots
Continuous by Continuous	Correlations, Scatterplots

Figure 6.1 Examples of analytical techniques associated with different combinations of Categorical and Continuous variables

Crosstabs (sometimes called 'contingency tables') are one of the most popular and useful ways to explore interactions and relationships between categorical variables and this will be the first technique that we explore.

The Crosstabs Procedure

Crosstabulation allows us to compare the number or percentage of cases that fall into each combination of the groups created when two or more categorical variables interact. A good way to begin using crosstabs is to think about the data in question and to begin to form questions or hytpotheses relating to the categorical variables in the dataset. These might include:

Is the respondents' place of birth related to their type of employment?

Do married and unmarried people rate the ferry the service in the same way?

Are men and women equally likely to experience health problems?

Let's begin by looking at the relationship between place of birth and employment type. To request crosstabs, from the main menu click:

Analyze

Descriptive Statistics

Crosstabs

The crosstabs dialog requires at least one variable to be added to the row dimension and one added to the column dimension. From the source variable list select:

Place of birth (placeofbirth)

Send the variable into the *Row(s)* dimensions by clicking the corresponding button. Click:

Now from the source variable list, select:

Type of paid employment (job)

Send the variable into the *Column(s)* dimensions by clicking the corresponding button. Click:

You may notice that the 'OK' button becomes active as we have now specified the minimum requirements to run the Crosstabs procedure with the default options. Figure 6.2 shows the completed dialog.

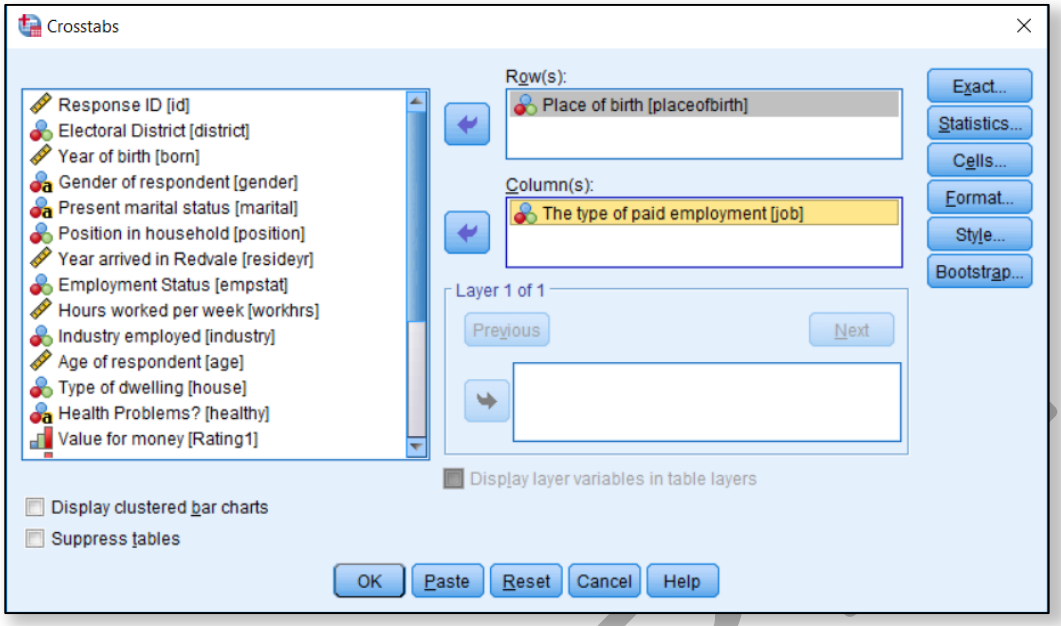


Figure 6.2 Completed Crosstabs dialog with default settings

Run the procedure by clicking:

OK

The output is shown in figure 6.3.

Crosstabs

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Place of birth * The type of paid employment	330	100.0%	0	0.0%	330	100.0%

Place of birth * The type of paid employment Crosstabulation

Count

		The type of paid employment				Total
		Permanent	Seasonal	Other	None	
Place of birth	Redvale Island	82	0	2	63	147
	Elsewhere in Ruritania	85	2	2	42	131
	Elsewhere in Europe	22	8	0	7	37
	Elsewhere in the World	5	0	0	10	15
Total		194	10	4	122	330

Figure 6.3 Basic Crosstab showing just frequency counts in the cells

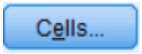
Apart from the initial 'Case Processing Summary' table which indicates that there were no missing values and that all 330 records were present, the crosstab itself simply shows how many respondents fell into the different groups regarding their place of birth and type of paid employment. The values where the frequency counts appear are referred to as the *cells* in the table. We can also see that it makes no difference which variable is in the rows or the columns as the frequency counts have no *direction.* The simplicity of the table however betrays the fact that it's hard to *compare one group to another.* This is because as the column and row totals show, the group sizes are different. For example, there are 147 people born on Redvale Island but only 37 born elsewhere in Europe. Let's re-run the crosstab and request further statistics in the cells that will enable us to compare the proportional differences more effectively. From the main menu click:

Analyze

Descriptive Statistics

Crosstabs

The dialog is retrieved with the variables still in their respective dimensions. To compare the respondents born in different locations *in terms of their employment type*, click the button marked:



This opens a sub-dialog where we can request additional statistics. You can see that *Observed Counts* are already selected. In the area marked 'Percentages', check the box marked:

Row

Figure 6.4 shows the *Cells* sub-dialog.

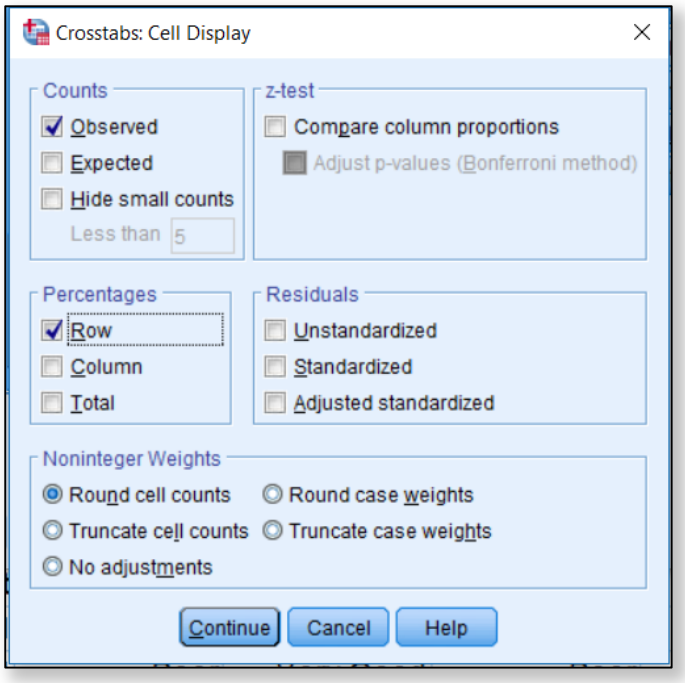


Figure 6.4 Requesting Row Percentages to be displayed in a Crosstab Cells

To run the updated crosstab, click:

Continue

OK

The Output is displayed in figure 6.5.

Place of birth * The type of paid employment Crosstabulation

			The type of paid employment				
			Permanent	Seasonal	Other	None	Total
Place of birth	Redvale Island	Count	82	0	2	63	147
		% within Place of birth	55.8%	0.0%	1.4%	42.9%	100.0%
	Elsewhere in Ruritania	Count	85	2	2	42	131
		% within Place of birth	64.9%	1.5%	1.5%	32.1%	100.0%
	Elsewhere in Europe	Count	22	8	0	7	37
		% within Place of birth	59.5%	21.6%	0.0%	18.9%	100.0%
	Elsewhere in the World	Count	5	0	0	10	15
		% within Place of birth	33.3%	0.0%	0.0%	66.7%	100.0%
Total		Count	194	10	4	122	330
		% within Place of birth	58.8%	3.0%	1.2%	37.0%	100.0%

Figure 6.5 Crosstab showing frequency counts and row percentages

So now we can see that an extra row has been added for each level of the 'Place of Birth' variable. The new row label says '% within Place if birth'. It is easier to compare these groups using percentages, as we can now say that, for example, 55.8% of those respondents born on

the island are in permanent jobs as opposed to 64.9% of those born 'Elsewhere in Ruritania'. Furthermore, the crosstab indicates that 21.6% of those born 'Elsewhere in Europe' are in 'Seasonal' employment. This is the largest proportion by far but notice that the frequency count in the cells indicates that this group is comprised of only 8 cases. So, although adding percentages does make the crosstab more interpretable, it's important to include the frequency counts as we need to bear in mind that some of the groups sizes are relatively small. Only 37 respondents in total were born 'Elsewhere in Europe' and only 15 were born 'Elsewhere in the World'.

Let's see the effect of adding more information to the cells. From the main toolbar, click the 'Recall recently used dialogs' button:

Crosstabs

Cells

In the area marked 'Percentages', check the box marked:

Column

Figure 6.6 shows the completed *Cells* sub-dialog.

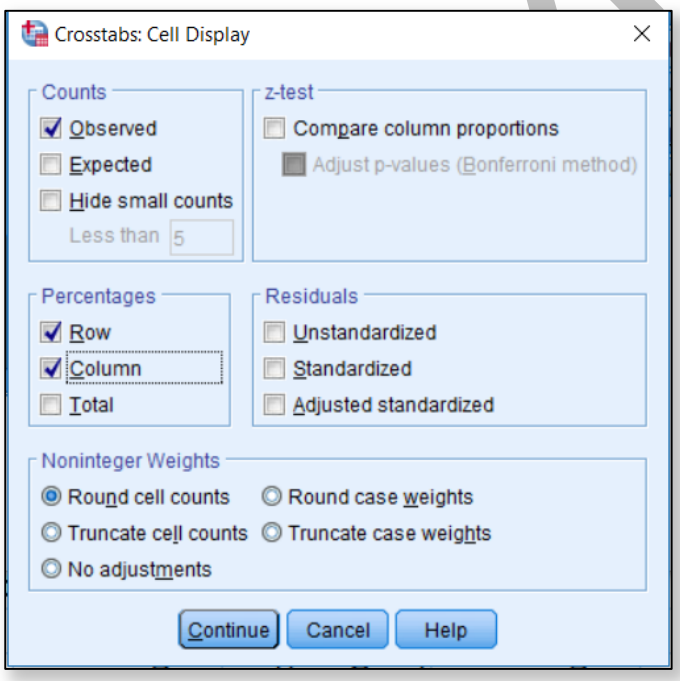


Figure 6.6 Requesting Row and Column Percentages in a Crosstab

To run the updated crosstab, click:

Continue

OK

The Output is displayed in figure 6.7.

Place of birth * The type of paid employment Crosstabulation

			The type of paid employment				
			Permanent	Seasonal	Other	None	Total
Place of birth	Redvale Island	Count	82	0	2	63	147
		% within Place of birth	55.8%	0.0%	1.4%	42.9%	100.0%
		% within The type of paid employment	42.3%	0.0%	50.0%	51.6%	44.5%
	Elsewhere in Ruritania	Count	85	2	2	42	131
		% within Place of birth	64.9%	1.5%	1.5%	32.1%	100.0%
		% within The type of paid employment	43.8%	20.0%	50.0%	34.4%	39.7%
	Elsewhere in Europe	Count	22	8	0	7	37
		% within Place of birth	59.5%	21.6%	0.0%	18.9%	100.0%
		% within The type of paid employment	11.3%	80.0%	0.0%	5.7%	11.2%
	Elsewhere in the World	Count	5	0	0	10	15
		% within Place of birth	33.3%	0.0%	0.0%	66.7%	100.0%
		% within The type of paid employment	2.6%	0.0%	0.0%	8.2%	4.5%
Total		Count	194	10	4	122	330
		% within Place of birth	58.8%	3.0%	1.2%	37.0%	100.0%
		% within The type of paid employment	100.0%	100.0%	100.0%	100.0%	100.0%

Figure 6.7 Crosstab with Row and Column percentages in the cells

The crosstab is now looking a lot larger and more complex than before. Again, an extra row of values has been added. This time, the numbers relate to the percentage within 'The type of paid employment' which is the column variable. Here we can see that although 55.8% of the respondents born on the island are in permanent employment (which is 82 out of a *row* total of 147 cases), *of those in permanent employment, 42.3% were born on the island* (which is 82 out of a *column* total of 194). We can also see that 80% of the 10 seasonal workers are people born elsewhere in Europe.

We can see that crosstabs can convey a great deal of detailed insight when comparing the interactions between categorical variables. Let's look at adding a *third* dimension to a crosstab with a new example.

From the main toolbar, click the 'Recall recently used dialogs' button:

Crosstabs

Reset

The crosstab dialog is now reset to its default settings. From the source variable list choose the variable we created in the last chapter:

Marital Status (simplified) [marital3]

...and add it to the rows dimension. Now choose the rating scale variable,

Facilities for small children (rating6)

...and add it to the columns dimension.

Request *'Row Percentages'* to be displayed in the cells and click:

OK

The output is displayed in figure 6.8.

Marital Status (simplified) * Facilities for Small Children? Crosstabulation

			Facilities for Small Children?			
			Excellent	Very Good	Poor	Total
Marital Status (simplified)	Married	Count	110	49	13	172
		% within Marital Status (simplified)	64.0%	28.5%	7.6%	100.0%
	Never Married	Count	56	33	8	97
		% within Marital Status (simplified)	57.7%	34.0%	8.2%	100.0%
	Previously Married	Count	28	21	5	54
		% within Marital Status (simplified)	51.9%	38.9%	9.3%	100.0%
Total		Count	194	103	26	323
		% within Marital Status (simplified)	60.1%	31.9%	8.0%	100.0%

Figure 6.8 Crosstab showing Marital Status (simplified) against rating of 'Facilities for Small children'

You can see that there are some marked differences between people of different marital statuses and their evaluation of the ferry service's facilities for small children. Exactly 64% of married respondents rate this aspect of the ferry's service as 'Excellent' compared to only 51.9% of those previously married. It would be interesting to break this relationship down further by finding out if these differences are true for male and female respondents. To do so, from the main toolbar, click the 'Recall recently used dialogs' button:

Crosstabs

Choose the variable:

Gender of respondent (gender)

...and add it to the box marked:

Layer

The completed dialog is shown in figure 6.9.

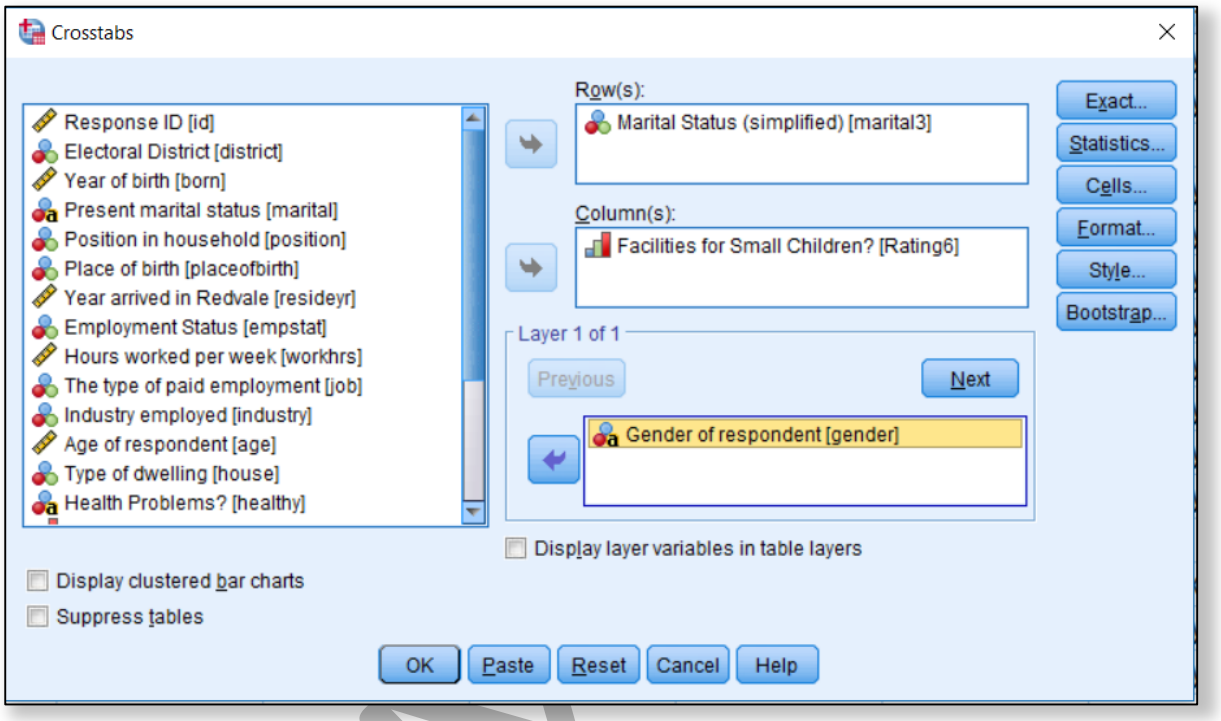


Figure 6.9 Crosstab dialog with an additional variable in the Layer box

To run the procedure, click:

Ok

The output is shown in figure 6.10

Marital Status (simplified) * Facilities for Small Children? * Gender of respondent Crosstabulation

Gender of respondent				Facilities for Small Children?			Total
				Excellent	Very Good	Poor	
female	Marital Status (simplified)	Married	Count	51	24	11	86
			% within Marital Status (simplified)	59.3%	27.9%	12.8%	100.0%
		Never Married	Count	21	21	3	45
			% within Marital Status (simplified)	46.7%	46.7%	6.7%	100.0%
		Previously Married	Count	20	14	4	38
			% within Marital Status (simplified)	52.6%	36.8%	10.5%	100.0%
	Total		Count	92	59	18	169
			% within Marital Status (simplified)	54.4%	34.9%	10.7%	100.0%
male	Marital Status (simplified)	Married	Count	59	24	2	85
			% within Marital Status (simplified)	69.4%	28.2%	2.4%	100.0%
		Never Married	Count	35	11	5	51
			% within Marital Status (simplified)	68.6%	21.6%	9.8%	100.0%
		Previously Married	Count	8	7	1	16
			% within Marital Status (simplified)	50.0%	43.8%	6.3%	100.0%
	Total		Count	102	42	8	152
			% within Marital Status (simplified)	67.1%	27.6%	5.3%	100.0%
Total	Marital Status (simplified)	Married	Count	110	48	13	171
			% within Marital Status (simplified)	64.3%	28.1%	7.6%	100.0%
		Never Married	Count	56	32	8	96
			% within Marital Status (simplified)	58.3%	33.3%	8.3%	100.0%
		Previously Married	Count	28	21	5	54
			% within Marital Status (simplified)	51.9%	38.9%	9.3%	100.0%
	Total		Count	194	101	26	321
			% within Marital Status (simplified)	60.4%	31.5%	8.1%	100.0%

Figure 6.10 Crosstab output split by layer variable 'Gender of respondent'

What is striking from the crosstab output, is that the effect of adding the gender variable as a layer field shows that although the difference between married and previously married respondents in terms of their likelihood to rate 'facilities for small children' as 'Excellent' is still present, this is *particularly true* for male respondents. In fact, close to 70% of male respondents who were classed as either married or never married rate this aspect of the service as 'Excellent'. Again, this shows the power of crosstabs to reveal complex interactions within categorical variables.

Let's look at a couple of final examples using the Crosstabs procedure, or rather let's look at an example of using crosstabs in conjunction with a test of statistical significance.

The Chi Square Test

The chi square test (also known as 'Pearson's Chi-Squared Test') was developed in 1900. It's a popular statistical test primarily because it is such a useful addition to the crosstabs procedure. Chi Square tests are often the first 'significance test' that students of statistics are introduced to. Although we don't have time to go into the statistical theory that underpins the test, we can illustrate how it is applied with a simple example.

Once again, from the main toolbar, click the 'Recall recently used dialogs' button:

Crosstabs

Clear the crosstab by clicking:

Reset

From the source variable list, choose the variable:

Gender of respondent (gender)

And add it to the 'Row(s)' box. From the source variable list, choose the variable:

Value for money (rating1)

And add it to the 'Column(s)' box

Click the 'Cells' button and request that row percentages are displayed.

Now click the button marked:

Statistics

You may notice that there are a number of statistical tests one could choose. From the sub-dialog, check the box marked:

Chi-square

Figure 6.11 shows the completed sub-dialog.

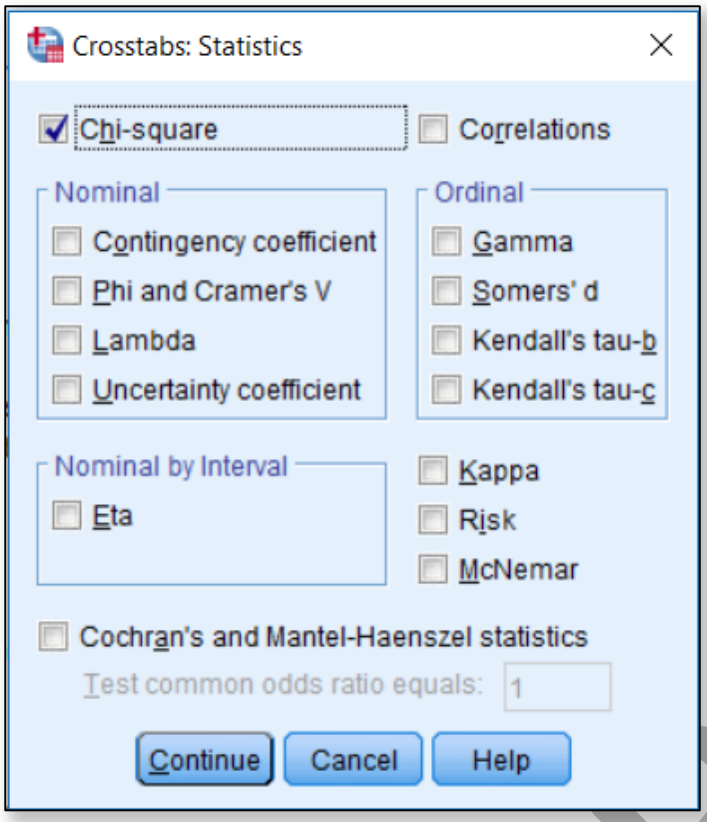


Figure 6.11 Requesting a Chi-square test

To run the procedure, click:

Continue

OK

The generated output is shown in figure 6.12.

The crosstab shows some apparent differences in the way in which male and female respondents have evaluated the ferry service in terms of 'value for money'. Note that 31.5% of female respondents have rated the service as 'Excellent' in this respect compared to 23.9% of male respondents. Considering this discrepancy between the sexes in the crosstab output, we might ask ourselves whether the magnitude of the differences is so small that we would expect to see them in many cases, or whether they are so large that we would only encounter them relatively rarely. This is what researchers mean when they talk about statistical 'significance'. Looking that the table below the crosstab, we can see the Pearson Chi-Square test itself.

Gender of respondent * Value for money Crosstabulation

			Value for money				
			Excellent	Very Good	Poor	Very Poor	Total
Gender of respondent	female	Count	53	100	6	9	168
		% within Gender of respondent	31.5%	59.5%	3.6%	5.4%	100.0%
	male	Count	37	100	12	6	155
		% within Gender of respondent	23.9%	64.5%	7.7%	3.9%	100.0%
Total		Count	90	200	18	15	323
		% within Gender of respondent	27.9%	61.9%	5.6%	4.6%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	4.929[a]	3	.177
Likelihood Ratio	4.979	3	.173
N of Valid Cases	323		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.20.

Figure 6.12 Crosstab output with associated Chi-square test of significance

The chi-square test is used to assess whether two categorical variables are independent of one another *in the population*. To put that in practical terms, our data is drawn from a sample, so it could be that the differences we're seeing between males and females here, are simply the result of sampling variation (the fact that no two data samples are likely to give *exactly* the same results). The column marked 'Asymptotic Significance (2-sided)' shows us a probability value that allows us to assess how often these sorts of differences would occur due to chance. In this case, the value for Pearson Chi-Square is 0.177 (equivalent to 17.7%). In statistical analysis, the convention follows that this is too large a value to regard this relationship is 'statistically significant' as it indicates that differences of this magnitude can occur too often due to random variation to assume that female respondents *really do* rate the ferry service in terms of 'value for money' differently from male respondents.

The most common rule of thumb is that the significance value (technically speaking, this is *the probability that the null hypothesis is true*) should be no larger than 0.05 (5%) before regarding an observed relationship as statistically significant. Let's run one last crosstab. From the main toolbar, click the 'Recall recently used dialogs' button:

Crosstabs

Swap the variable 'Value for money (rating 1)' in the columns dimension with the variable 'Facilities for small children (rating 6)'. Now click:

OK

The output is shown in figure 6.13.

Gender of respondent * Facilities for Small Children? Crosstabulation

			Facilities for Small Children?			
			Excellent	Very Good	Poor	Total
Gender of respondent	female	Count	92	59	18	169
		% within Gender of respondent	54.4%	34.9%	10.7%	100.0%
	male	Count	104	42	8	154
		% within Gender of respondent	67.5%	27.3%	5.2%	100.0%
Total		Count	196	101	26	323
		% within Gender of respondent	60.7%	31.3%	8.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	6.760[a]	2	.034
Likelihood Ratio	6.860	2	.032
N of Valid Cases	323		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.40.

Figure 6.13 Second crosstab output with associated Chi-square test of significance

Notice that 67.5% of male respondents have evaluated this aspect of the ferry service as 'Excellent' compared to 54.4% of females. Indeed, twice as many female respondents as males have indicated that they thought the facilities for small children were 'Poor'. In this case, the Chi-Square test of significance is shows a value of .034 (equivalent to 3.4%) this below the .05 (5%) threshold so we would regard this relationship as 'significant'.

Thus far we have used the crosstabs procedure as an example of a technique we can employ to examine the relationships between categorical variables. In the next section, we will see a procedure that enables us to examine the relationships between categorical and continuous variables.

The Means Procedure

A simple approach to comparing the groups within a categorical variable in terms of a continuous variable is to use a summary statistic. Summary statistics such as sums, means or median values can be used to compare one group to another. As an example, from the main menu click:

Analyze

Compare Means

Means

The Means dialog is generated. Despite its title, the Means procedure can be used to compare groups in terms of many more summary measures than arithmetic averages. You may also notice that the dialog is made up of two boxes marked 'Dependent List' and 'Layer 1 of 1'. In this case we can think of the top dependent box as the place where we send *continuous* variables (or whichever variable we wish to use a summary measure such as an average with) and the bottom *layer* box as the target location for the categorical field (or the variable containing the groups we want to compare).

From the source variable list, select the variable:

Age of respondent (Age)

And send it to the box marked *'Dependent List'*. Now, select the variable:

Health problems? (healthy)

And send it to the box marked *'Layer 1 of 1'*. The complete dialog is shown in figure 6.14.

To run the procedure, click:

OK

The resultant output is shown in figure 6.15.

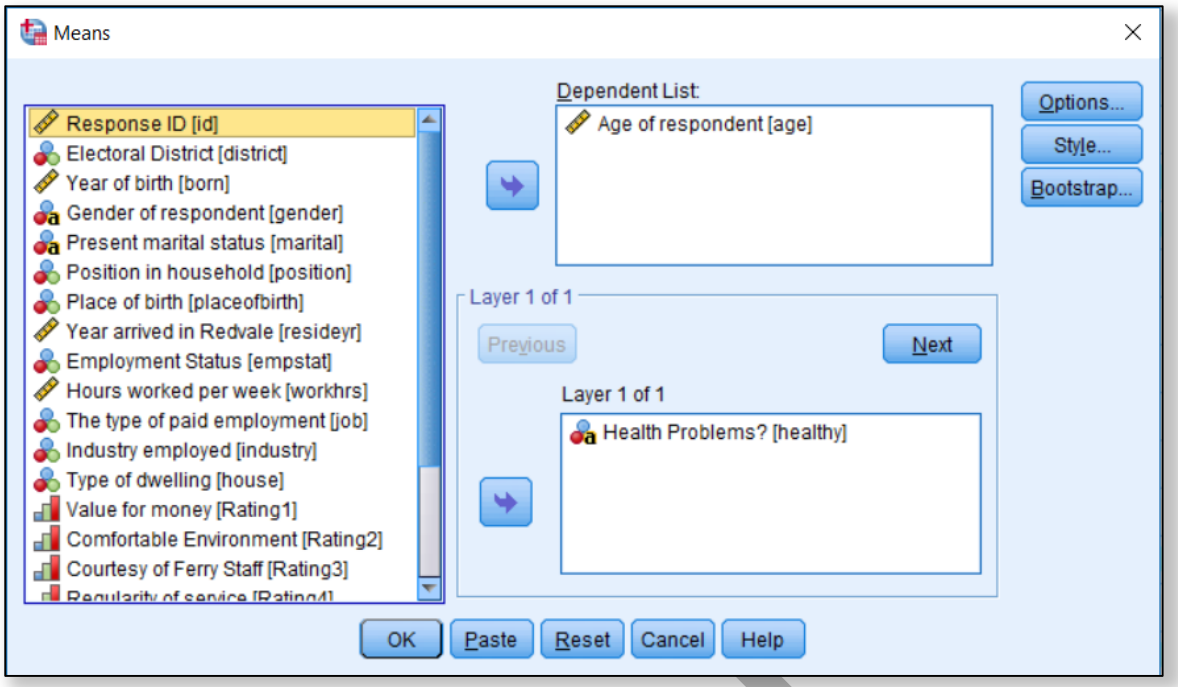


Figure 6.14 Completed Means Dialog

Means

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Age of respondent * Health Problems?	330	100.0%	0	0.0%	330	100.0%

Report

Age of respondent

Health Problems?	Mean	N	Std. Deviation
No Health Problems	43.37	295	16.841
Has Health Problems	65.60	35	21.894
Total	45.73	330	18.708

Figure 6.15 Output from the 'Means' procedure

The means output is fairly straightforward to interpret. It shows that the average age of the 295 people who report having *no* health problems is 43.37 whereas for the 35 people who indicated that they *do* have health problems, the mean age is 65.6 years. For all 330 respondents in the sample, the mean age is 45.73. You may also notice that by default the output includes the standard deviation value for each group. Standard Deviations are known as *measures of dispersion* as they indicate the degree of '*spread*' within each of the groups. In

other words, these values quantify the average amount by which the individual members of a group differ from the *overall* mean value for the group. Let us return now to the means procedure and explore it a little further. From the main toolbar, click the 'Recall recently used dialogs' button:

Means

From the returned dialog, within the Layer box where it says, 'Layer 1 of 1', click:

Next

The box now appears empty and has the title 'Layer 2 of 2'. From the source variable list, click:

Gender of respondent (gender)

Now let's enhance the procedure with additional statistical measures. Click the button marked:

Options

A sub-dialog is generated offering you the opportunity to change which measures are displayed in the means procedure. As you can see there are several additional measures. From the list choose:

Median
Minimum
Maximum

...and click the selection button each time:

The completed sub-dialog is shown in figure 6.16.

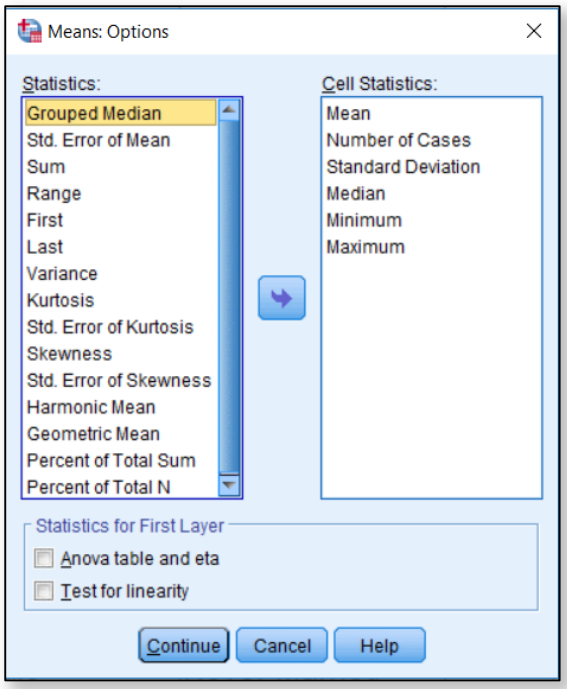


Figure 6.16 Sub-Dialog for additional measures for Means Procedure

To run the procedure, click:

Continue

OK

Figure 6.17 shows the results.

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Age of respondent * Health Problems? * Gender of respondent	328	99.4%	2	0.6%	330	100.0%

Report

Age of respondent

Health Problems?	Gender of respondent	Mean	N	Std. Deviation	Median	Minimum	Maximum
No Health Problems	female	45.20	149	17.711	45.00	17	90
	male	41.37	145	15.725	38.00	16	83
	Total	43.31	294	16.842	41.00	16	90
Has Health Problems	female	72.50	22	15.595	78.50	36	93
	male	57.00	12	25.035	57.00	26	92
	Total	67.03	34	20.499	75.50	26	93
Total	female	48.71	171	19.676	47.00	17	93
	male	42.57	157	17.023	39.00	16	92
	Total	45.77	328	18.681	42.00	16	93

Figure 6.17 Layered Means report with additional summary measures

Interestingly, the layered report shows that for those with health problems, there is a marked difference in the average age of male and female respondents (57 and 72.5 respectively). The means procedure is a simple but powerful tool for exploring group differences between combinations of continuous and categorical variables. At this point, we can briefly introduce the concept of using pivot trays with SPSS output tables. Pivoting allows us to alter the display of values and fields within tables by transposing fields, moving rows and columns or creating multidimensional layers. To activate the pivot controls:

Double click on the table of means in the output viewer

You may notice that a dotted line appears around the outside of the table. To request the pivot trays, from the main menu within the viewer window, click:

Pivot

Pivoting trays

The pivoting tray is generated. To illustrate how we can use the tray to manipulate the dimensions in the table,

Click the icon next to the word 'Statistics' and drag it to the empty slot in the top left box under the word 'Variables' (in the 'Layer' dimension)

Now:

Click the icon next to the word 'Gender' and drag it to the empty slot in the top right box previously occupied by the 'statistics' icon (in the columns dimension)

Figure 6.18 illustrates this process. The results of this process are shown in figure 6.19.

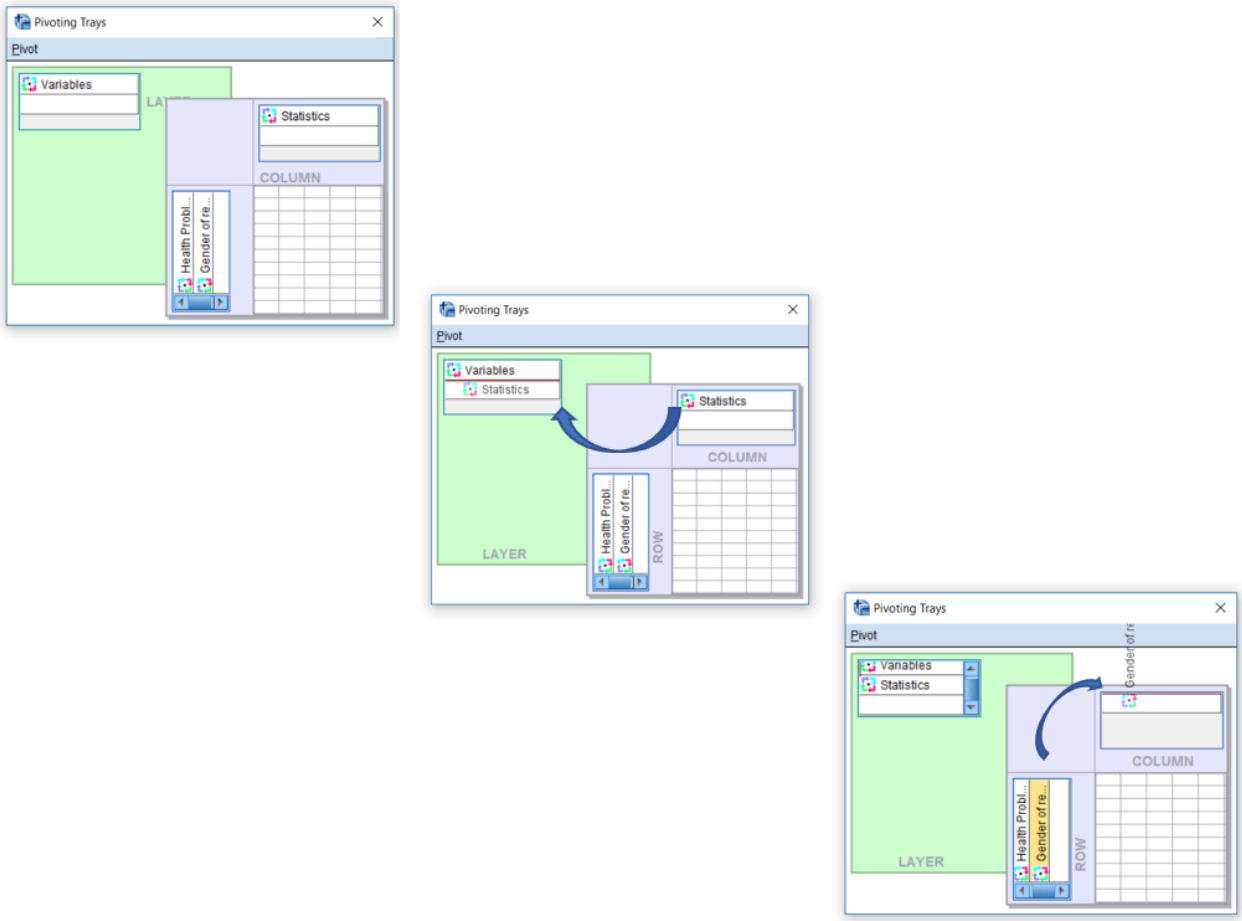


Figure 6.18 Using pivot trays to change table appearances

Report

Variables: Age of respondent

Statistics: Mean

No Health Problems	female	45.20
	male	41.37
	Total	43.31
Has Health Problems	female	72.50
	male	57.00
	Total	67.03
Total	female	48.71
	male	42.57
	Total	45.77

Figure 6.19 Pivoted Means table with statistics measures layered

As the statistical measures have now been placed in the layer dimension, we can interact with them and choose to display one at a time. As shown in figure 6.20, using the drop-down menu we can now choose to display a different summary measure such as a median.

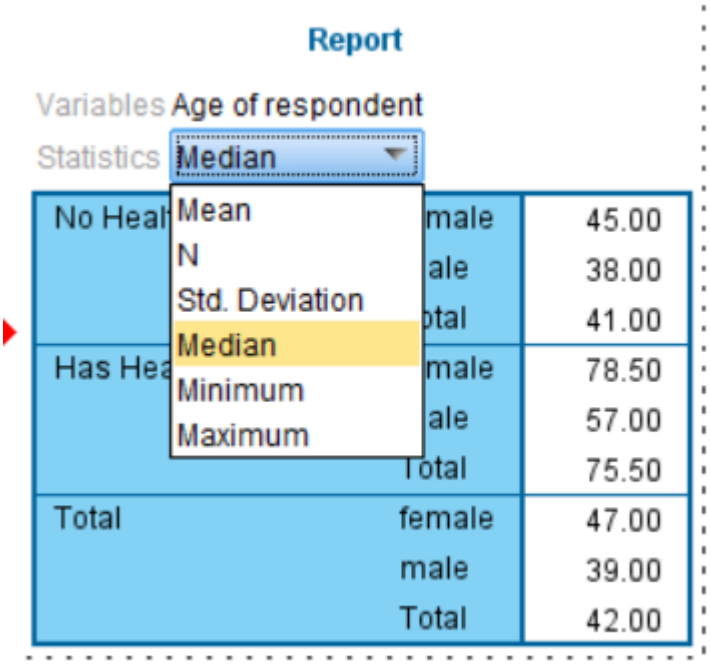


Figure 6.20 Pivoted Means table interacting with the layered statistics

Later in the course we will take another look at interacting with SPSS Statistics output. For now, we can turn our attention to analysing the relationships between pairs of continuous variables.

The Correlations Procedure

In this section of the course, we have used two procedures to help us investigate the relationships between exclusively categorical variables (crosstabs) or combinations of categorical and continuous fields (the 'Means' procedure). In this next example, we will see how we can employ correlation coefficients to examine the relationship between pairs of continuous fields. The best way to illustrate how correlation measures work is to think of how we normally visualize the relationships between pairs of continuous variables. Scatterplots are the graphical equivalent of correlation coefficients. A correlation coefficients is a single value that indicates how strong the *linear* relationship is between the two variables and whether the relationship is positive of negative. If we look at the charts in figure 6.21 we can see different scatterplots with various correlation values. It is important to understand that correlation coefficients always range from -1 to +1. If we were to request a correlation of a continuous variable against a *copy of itself* we would find that the correlation value was 1.00 indicating that the variable is perfectly correlated with its copy. A positive value simply indicates that high values in one variable are associated with high values in another. A negative correlation indicates that as one value rises the other tends to fall (such as alcohol intake and reflex reaction time).

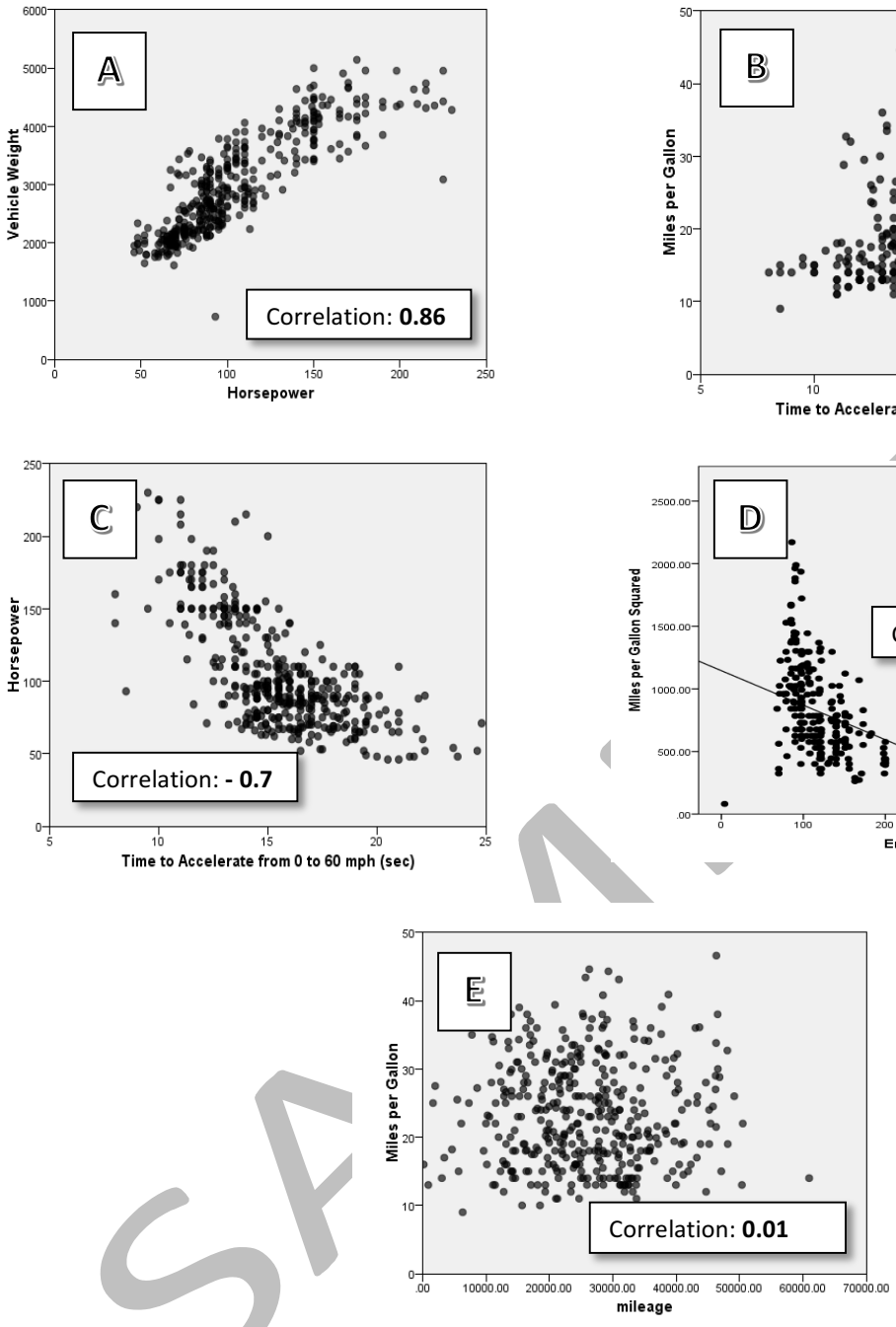


Figure 6.21 Illustration of various correlation values and their associated scatterplots

Figure 6.21 contains a number of scatterplots of pairs of continuous variables. Each measure relates to an aspect of a car's performance or history. In fact, each point in the scatterplots represents the individual make and model of an automobile. Chart A shows a strong positive correlation between horsepower and vehicle weight (0.86). This simply illustrates that on average, the weight of a car is highly correlated with its horsepower. In fact, we can see that this relationship is linear in nature as it tends to follow a straight line. Chart B on the other hand illustrates a positive relationship between the time to accelerate from 0 to 60 mph and

the miles per gallon of gasoline consumption. In other words, the longer a vehicle takes to reach 60 mph from a standing start, the more efficient its overall fuel consumption is. We can see from the chart that the relationship is a little more random (or 'noisy') and as such, the correlation value, although positive, is weaker (0.43) than the previous example. Chart C shows a strong *negative* relationship (- 0.7) between time to accelerate and horsepower. Here we can see that cars with a lot of horsepower take *less time* to reach 60 mph that those with less horsepower: as one value increases the other tends to decrease. Chart D illustrates the limitations of correlations although it shows a strong negative relationship (- 0.67) it is not well expressed by a *straight line.* We can conclude from this that weak correlations don't necessarily indicate that there is no relationship, just that there is no *linear* relationship. In this example, the correlation value would be even stronger if the relationship was less *curvilinear*. Finally chart E shows that when the relationship is effectively random, the correlation value is close to zero.

To generate our own correlations for continuous variables, from the main menu click:

Analyze

Correlate

Bivariate

This generates the correlation dialog. From the source variable list choose the following variables:

Age of respondent (age)
Hours worked per week (workhrs)
Year arrived in Redvale (resideyr)

Notice that the type of correlation is indicated by the check box marked 'Pearson'.

The completed dialog is shown in figure 6.22. To run the procedure, click:

OK

The output is shown in figure 6.23.

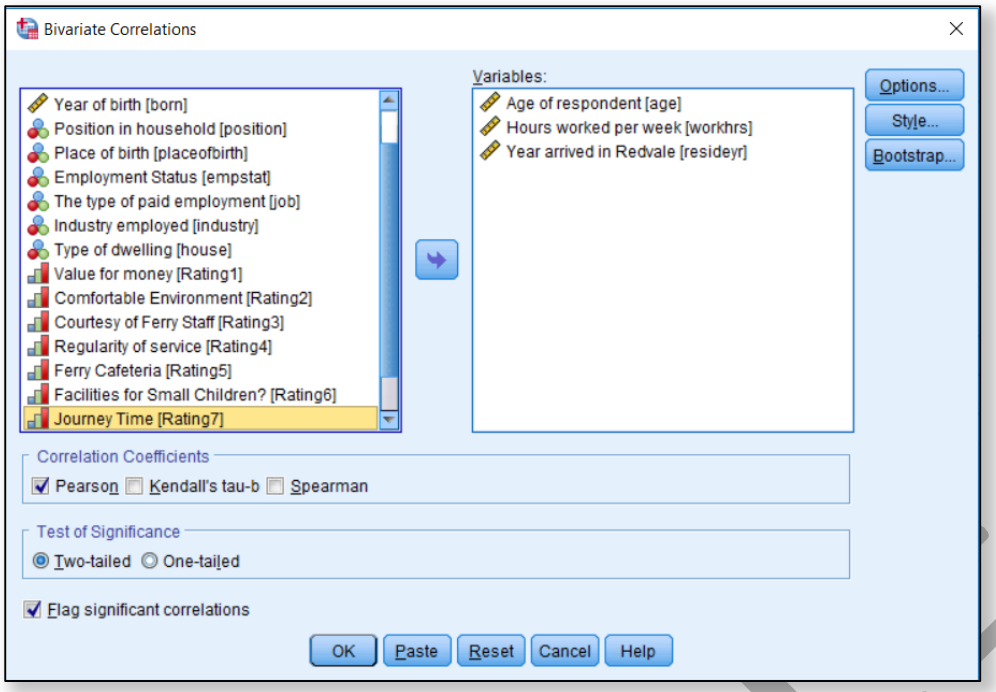


Figure 6.22 Completed Correlations Dialog

Correlations

Correlations

		Age of respondent	Hours worked per week	Year arrived in Redvale
Age of respondent	Pearson Correlation	1	-.209**	-.674**
	Sig. (2-tailed)		.002	.000
	N	330	208	189
Hours worked per week	Pearson Correlation	-.209**	1	.394**
	Sig. (2-tailed)	.002		.000
	N	208	208	126
Year arrived in Redvale	Pearson Correlation	-.674**	.394**	1
	Sig. (2-tailed)	.000	.000	
	N	189	126	189

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

Figure 6.23 Correlations output

We've only asked for correlations for three variables, yet the output shows a table with three values each in nine cells. Of course, the cells running diagonally from left to right show a correlation of 1 as each variable is being correlated with itself. However, if we look at the middle cell in the first column, we can see that this is the correlation of 'Hours worked per week' against 'Age of respondent'. Here the value is -.209 indicating a weak negative correlation (older people tend to work fewer hours). We can also see that the middle value corresponds with the row label 'Sig. (2-tailed)'. This is a test of significance that tells us that although the correlation is weak, it is unlikely to be the result of chance (technically speaking

this is testing the null hypothesis that the correlation is actually equal to zero in the population). The double asterisk next to the correlation value itself simply tells us that the significance value is below .01 (the 1% level). Generally, the significance value is not very interesting in correlation analysis since it doesn't indicate that the correlation is particularly strong, merely that the value is large enough to not be regarded as a random fluctuation. The last value in each of the cells is the number of cases that were used to calculate the correlation. Although our strongest correlation is a negative relationship between the year the respondent arrived in Redvale and their age (a correlation value of -.674) we can see that only 189 out of 330 cases was used in the calculation. This of course is because a large proportion of people were *born* on the island and they are coded as 'Not Applicable' (missing). The third correlation is a reasonably strong positive one with a value of .394. This is between year of arrival on the island and hours worked per week, indicating that the more recent the arrival of the respondent the more hours they work.

Earlier in the course we discussed the fact that sometimes ordinal data are treated as if the variables were continuous in nature. This is particularly true when researchers wish to calculate average scores with rating scales. In fact, within statistical research, there are a number of procedures that make less stringent assumptions about the nature of the data that you may be working with. These procedures are referred to as *non-parametric* techniques. The correlations procedure in SPSS Statistics includes a non-parametric technique known as Spearman's correlation that is often used when analysing correlations between pairs of *ordinal* variables. To demonstrate this, from the main toolbar, click:

Bivariate Correlations

Reset

The dialog is now cleared back to its default setting, from the source variable list choose the following ordinal variables:

Value for money (Rating1)
Comfortable Environment (Rating2)
Courtesy of Ferry Staff (Rating3)
Regularity of service (Rating4)
Ferry Cafeteria (Rating5)
Facilities for Small Children (Rating6)
Journey Time (Rating7)

Within the dialog…

Uncheck the box marked Pearson

Check the box marked Spearman

The completed dialog is shown in figure 6.24.

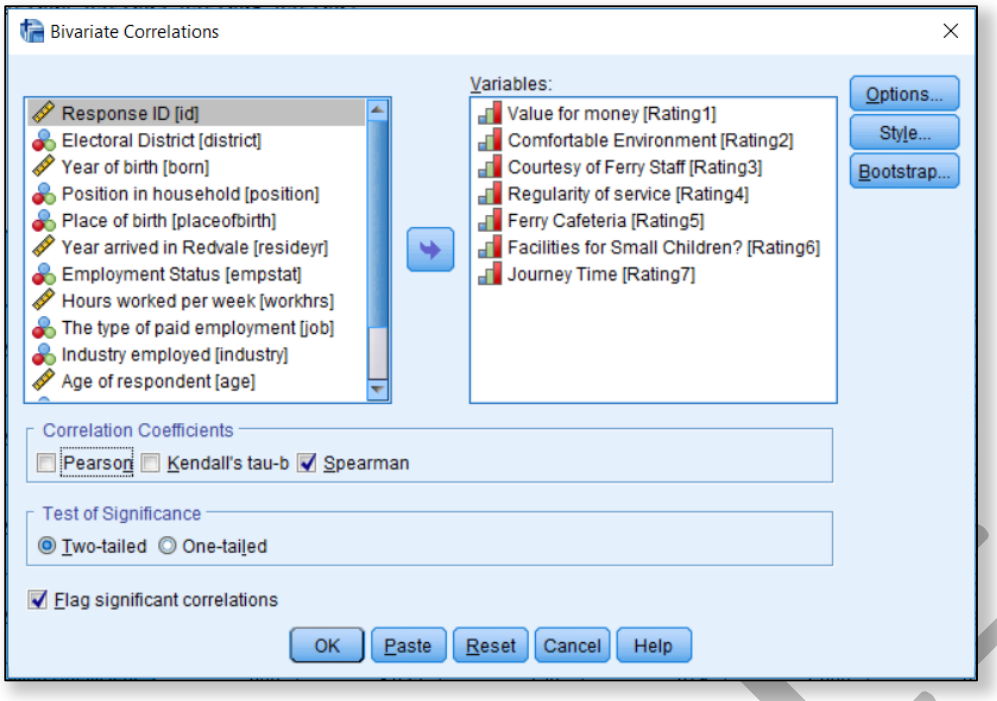


Figure 6.24 Completed Correlations dialog with Spearman correlation selected

To run the procedure, click:

OK

The initial output is shown in figure 6.25.

Correlations

			Value for money	Comfortable Environment	Courtesy of Ferry Staff	Regularity of service	Ferry Cafeteria	Facilities for Small Children?	Journey Time
Spearman's rho	Value for money	Correlation Coefficient	1.000	.015	.090	.102	.000	.177**	.195**
		Sig. (2-tailed)	.	.794	.111	.070	.997	.001	.000
		N	325	320	317	315	298	321	324
	Comfortable Environment	Correlation Coefficient	.015	1.000	-.085	.085	-.035	.069	.096
		Sig. (2-tailed)	.794	.	.130	.131	.549	.218	.086
		N	320	324	317	317	300	320	323
	Courtesy of Ferry Staff	Correlation Coefficient	.090	-.085	1.000	.105	.150*	.088	-.043
		Sig. (2-tailed)	.111	.130	.	.065	.010	.119	.443
		N	317	317	321	311	294	317	320
	Regularity of service	Correlation Coefficient	.102	.085	.105	1.000	.079	.008	.132*
		Sig. (2-tailed)	.070	.131	.065	.	.172	.884	.018
		N	315	317	311	319	297	315	318
	Ferry Cafeteria	Correlation Coefficient	.000	-.035	.150*	.079	1.000	.010	-.066
		Sig. (2-tailed)	.997	.549	.010	.172	.	.859	.255
		N	298	300	294	297	302	298	301
	Facilities for Small Children?	Correlation Coefficient	.177**	.069	.088	.008	.010	1.000	.145**
		Sig. (2-tailed)	.001	.218	.119	.884	.859	.	.009
		N	321	320	317	315	298	325	324
	Journey Time	Correlation Coefficient	.195**	.096	-.043	.132*	-.066	.145**	1.000
		Sig. (2-tailed)	.000	.086	.443	.018	.255	.009	.
		N	324	323	320	318	301	324	328

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

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

Figure 6.25 Spearman correlation matrix

As we can see, the procedure has created a large table with lot of values. To make it easier to focus on the Spearman correlation values themselves, double-click on the output table and

use the pivoting tray to send the statistics to the layers dimension so that the output looks like figure 6.26.

Nonparametric Correlations

Correlations

Correlation Coefficient

		Value for money	Comfortable Environment	Courtesy of Ferry Staff	Regularity of service	Ferry Cafeteria	Facilities for Small Children?	Journey Time
Spearman's rho	Value for money	1.000	.015	.090	.102	.000	.177**	.195**
	Comfortable Environment	.015	1.000	-.085	.085	-.035	.069	.096
	Courtesy of Ferry Staff	.090	-.085	1.000	.105	.150*	.088	-.043
	Regularity of service	.102	.085	.105	1.000	.079	.008	.132*
	Ferry Cafeteria	.000	-.035	.150*	.079	1.000	.010	-.066
	Facilities for Small Children?	.177**	.069	.088	.008	.010	1.000	.145**
	Journey Time	.195**	.096	-.043	.132*	-.066	.145**	1.000

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

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

Figure 6.26 Pivoted Spearman Correlation table

The non-parametric correlation table, shows that allow there are many statistically significant correlations, the values themselves are very weak so we shouldn't spend too much time concerning ourselves with them.

In this section, we have looked at three key methods for analysing relationships between variables with different combinations of measurement level. In the next section, we will return to the 'Transform' menu to look at how we can further enhance the analysis process by creating new variables.