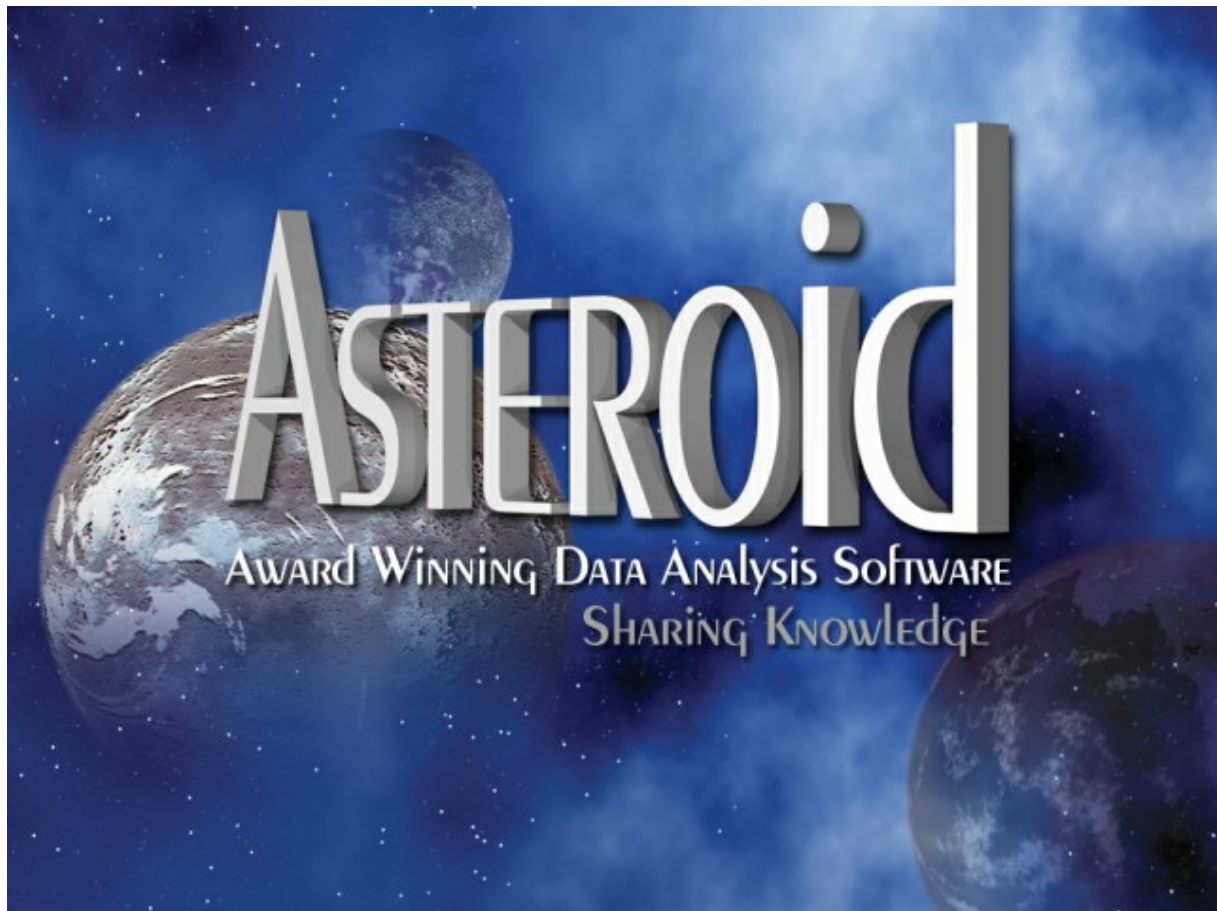


ASTEROID

Statistical Analysis Module



ASTEROID Support:

Telephone

Email

+61 3 9223 2428

asteroid.support@roymorgan.com

August 2023

Introduction

This module covers a number of statistical tools available within ASTEROID but it has been written with the non-statistically trained user in mind.

All the tools covered in this manual function independently, and so you can read the sections in whichever order suits your needs, though with the Significance Testing section please read the introduction to that section before going to a specific tool.

Course Objectives

ASTEROID provides a number of statistical tools and this guide is intended to introduce them and their applications.

At the end of this session you will be able to understand and confidently apply:

- A.I.D.
- Quantities
- Significance Tests & Standard Error
- Correspondence Analysis

Implied Knowledge

This course assumes that you will have completed the Introductory course and have a good working knowledge of¹:

- Tabulation
- Navigating ASTEROID

¹ While some references to Profiler are made in this guide it is not necessary to be familiar with this tool.

The other modules in the series are:

Introduction to ASTEROID

This is the first of the four modules in the ASTEROID training series. It is intended both as a step by step guide for new users, and a way current users can refresh, update and expand their skills.

Target Consumer Profiling

‘Profiler’ will allow you to easily compare your target Group to other Groups and examine what differentiates each Group from the others. When used with media Variables it also shows cost/reach.

Media Analysis

‘MediaPLANNER’ creates and compares media schedules, combining print, TV, radio, Cinema and website advertising.

Also Available

Software Reference Manual

The Software Reference Manual forms the online help, available through the ASTEROID Help menu, and provides more detailed information and technical specifications regarding all aspects of the software.

TABLE OF CONTENTS

A.I.D. (Automatic Interaction Detector).....	1
Definitions:.....	1
What is A.I.D.?	1
Why do we use A.I.D.?	1
When should we use A.I.D.?	1
How do we use A.I.D.?	2
The Diagram	3
An Example.....	3
Persons and %Targ	6
Understanding the output	7
Significance Threshold.....	9
Percent uncertainty explained.....	10
Correspondence Analysis	11
What is Correspondence Analysis?	11
Why do we use Correspondence Analysis?	11
When should we use Correspondence Analysis?	12
How do we use Correspondence Analysis?	12
Quantities	21
Using Quantities	23
Understanding Quantities.....	25
Significance testing.....	28
What is Significance Testing, and why do we use it?	28
Considerations When Using Significance Testing.....	29
Output Display	30
Display.....	30
Context for Markers.....	31
Outputs and Output Options.....	33
Confidence limits for Markers – the level of Significance, and the number of tails.....	33
Test Type For Markers	33
Output Display.....	34
Methodology	34
Important warnings	35
Standard Error	36
Using Standard Error.....	36
Understanding the ‘true value’	36
Cluster Analysis	37

What does Cluster Analysis do?	37
How do we use Cluster Analysis?	39
Avoiding duplication.....	41
Avoid bunched scales / variables	42
Selecting too many variables/groups.....	42
What happens within the analysis when you change the number of clusters?	43
Goodman-Kruskal:	45
Dorofeev Contribution:	46
% Explained:	46
Uniqueness:.....	46
Av. Difference:.....	46
Difference between two closest clusters:	46
Examples of the measures.....	47
Example 1.	48
Example 2	49

A.I.D. (Automatic Interaction Detector)

Definitions:

<i>Characteristic</i>	This is a behaviour (bought diet cola), attitude (“I’d like to be able to lose weight”) or attribute (living in South Australia).
<i>Target Group</i>	The group of people we are interested in.
<i>Context Group</i>	The larger population that these people belong to.
<i>Candidate Variables</i>	The characteristics we are interested in as potential predictors of Target Group members.

What is A.I.D.?

A.I.D. takes characteristics you have chosen and shows, through a tree diagram, which combination of those characteristics best defines the members of your Target Group. In simple terms A.I.D. tells you what percentage of people who have each combination of characteristics are members of your Target Group. (The higher the percentage the better a combination predicts who will be a member of the Target Group.) However, you must bear in mind that the single ‘best-combination’ group may itself contain a very small proportion of the population (and of the Target Group). You need to consider other combinations as well and form a general impression.

Why do we use A.I.D.?

We can discover a lot about a Target Group through Tabulate and (particularly) Profiler but both these tools only examine characteristics separately while A.I.D. allows us to find the best *combination* of characteristics.

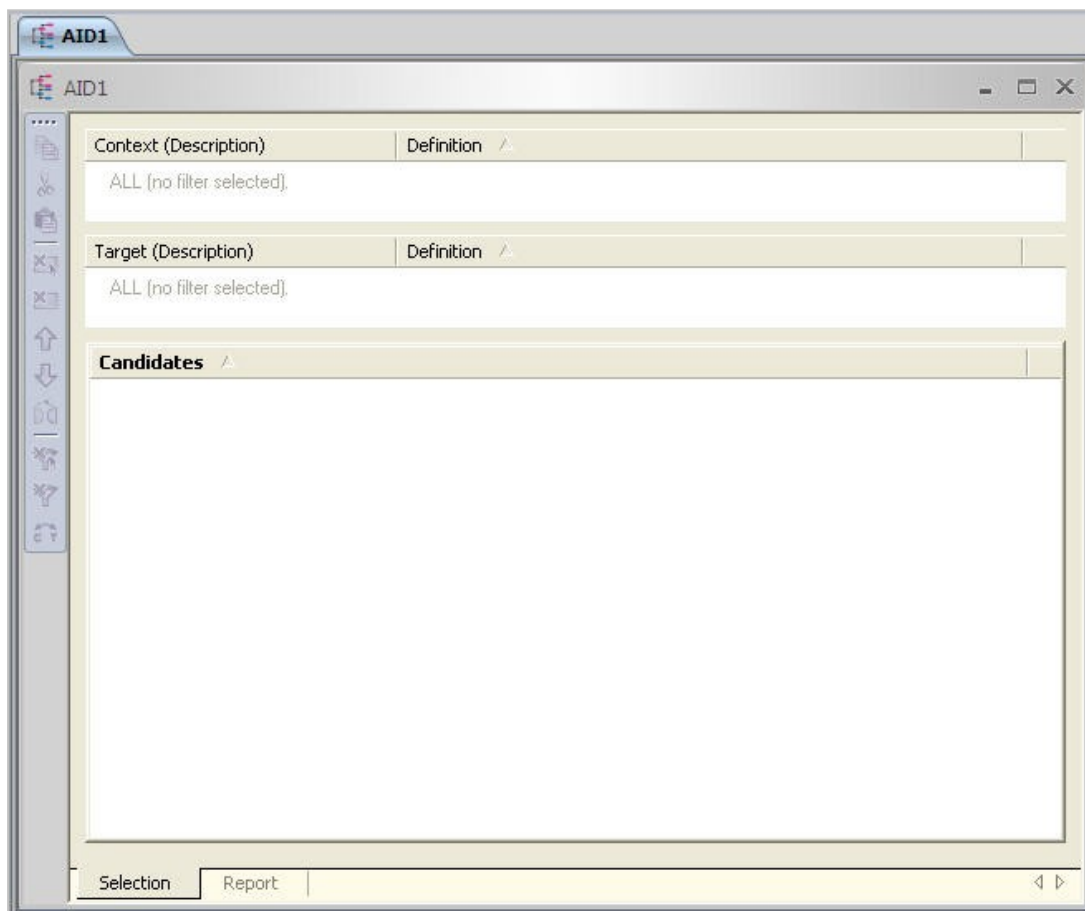
When should we use A.I.D.?

If you already know something about your Target Group you can use A.I.D. to test assumptions. For example, you might know that your Target Group fall into a ‘high’ income bracket, in which case you might assume that when they travel they would spend a certain amount, stay at certain types of hotels and look for certain things in a holiday. Using A.I.D. to look at previous and future travel information and/or attitudes, you can test your assumptions and build a more detailed picture.

Alternatively, where you know very little about your Target Group, you can use major demographics in A.I.D. as a starting point to understanding who they are. But bear in mind that ASTEROID does not automatically select the best Candidate Variables to use. You have to apply your skill and judgement to select ones that you think will be appropriate and may be helpful.

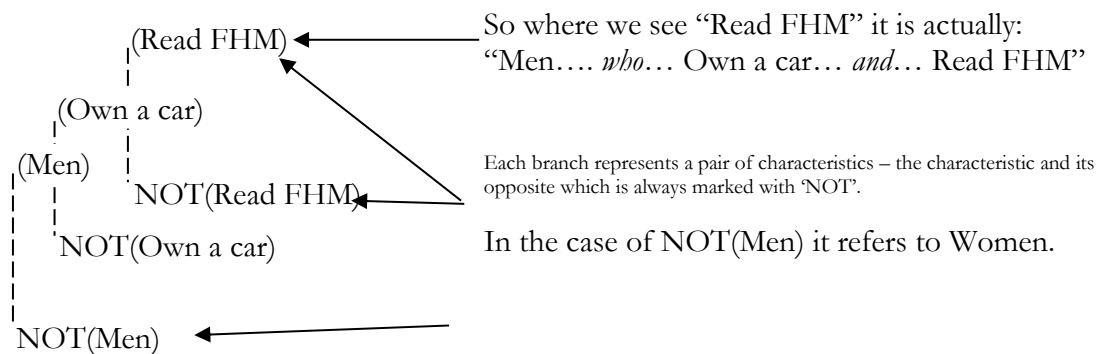
How do we use A.I.D.?

- | | |
|--------------------|--|
| Step One: | Decide whether you want to define a Context Group and if so, define it |
| Step Two: | Define your Target Group |
| Step Three: | Select Candidate Variables |
| Step Four: | Run the A.I.D. |



The Diagram

The output in A.I.D. is a tree diagram, where the branches of the tree run from left to right across the screen. Each time the tree branches it represents an addition to the combination of characteristics:



An Example

To explore the use of A.I.D. we will look at finding out more about people who do yoga regularly.

1. Context Group²

For this example we won’t define the Context Group, meaning that the Context Group will represent the population covered by the database – in this case all Australians aged 14+.

2. Target Group

We will define our target group as people who do Yoga regularly.

3. Candidate Variables

² The Context and Target groups in this example are kept very simple, but you can also use And, Or & Not to create more complex definitions.

We will look at some demographics:

Age - summary

Sex

Marital Status

4. Click Run A.I.D.

(You will notice that the progress meter will flash repeatedly as the A.I.D. is generated.)

Always keep in mind that, depending on the sample size and the complexity of the A.I.D., it may take a few minutes to run.

Take a moment to look at the output from our example (shown opposite).

Above the tree diagram you will see information about what ‘went into’ the A.I.D. including the definition for both the Context and Target Groups and which Candidate Variables were selected.

We will look at the details of the tree itself in the following pages but straight away you can see that the diagram is accompanied by two columns of figures, both of which help us to work out how well the various combinations of characteristics describe our Target Group.

You might also have noticed the two following lines (one above and one below the diagram):

Significance level: $p < 0.001$

Percent uncertainty explained = 8.391

Both of these refer to the statistical principles involved in A.I.D. and will be covered toward the end of the section.

In the following examples the weighted counts (sums of weights) are intended to represent the population in thousands. Although this is common practice, it may not apply to all surveys.

Context group: ALL

Target group: Yoga

Variables:

AGE: AGE - summary

SEX: SEX

MARRIAGE: MARITAL STATUS - detailed

Significance level: $p < 0.001$

SECTION OR SUB-SECTION OF POPULATION	PERSONS	% TARG
Single [marriage]	1871	5%
NOT (65 and Over) [age]	7003	4%
NOT (Single) [marriage]	5132	4%
Women [sex]	8205	4%
65 and Over [age]	1202	1%
ALL	16190	2%
De Facto [marriage]	518	2%
NOT (Women) [sex]	7984	1%
35-49 + 50-64 [age]	828	2%
NOT (Married) [marriage]	3239	1%
NOT (35-49 + 50-64) [age]	2411	1%
NOT (De Facto) [marriage]	7467	1%
Married [marriage]	4228	1%

Percentage uncertainty explained = 6.109%

Persons and %Targ

Context group: ALL

Target group : Yoga

Variables:

AGE: AGE - summary

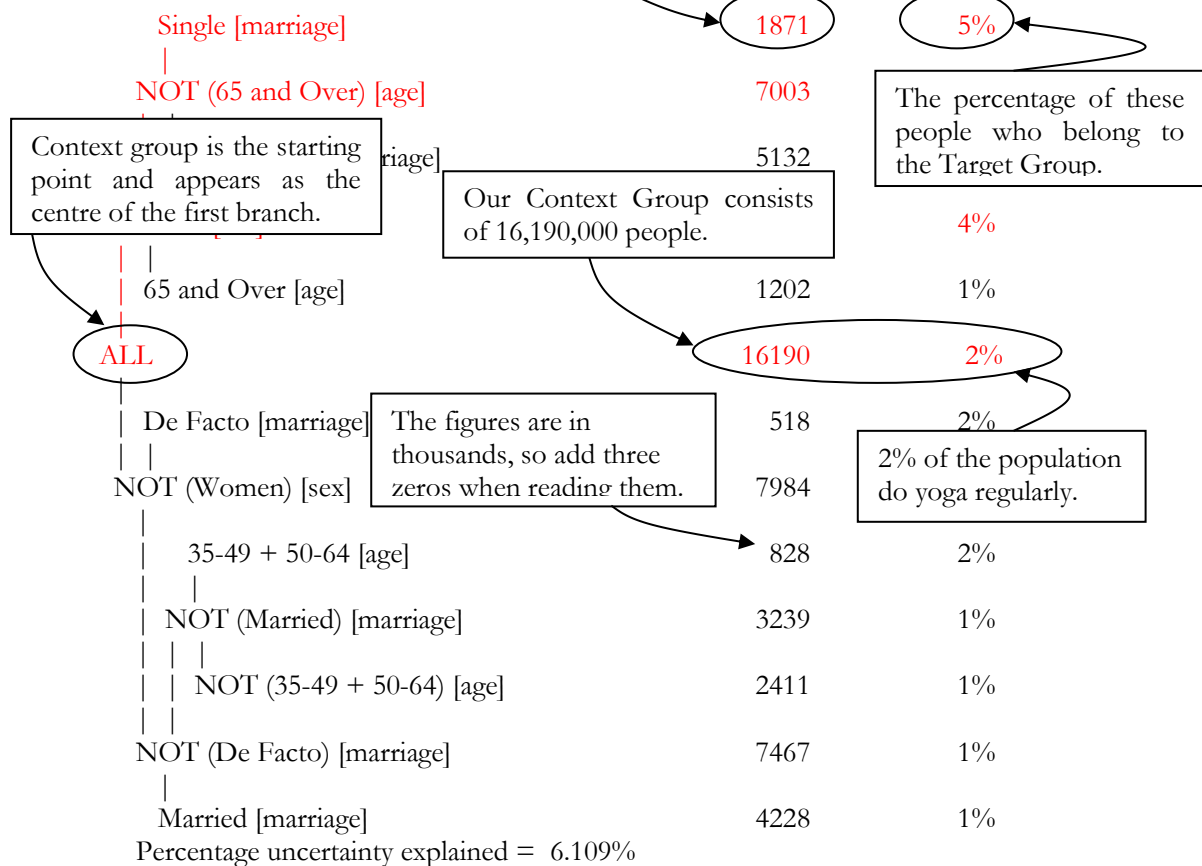
SEX: SEX

MARRIAGE: MA

Significance level: p

The number of people in the Context Group who have each combination of characteristics.

SECTION OR SUB-SECTION OF POPULATION PERSONS % TARG



Understanding the output

The basic principle of reading A.I.D. output, is that the top line of the tree shows the combination of characteristics which best predict membership of the Target Group.

So in our example:

- Out of the whole population it is single women, under the age of 65, who are most likely to do yoga regularly.
- Of all single women under 65, 5% do yoga regularly.
- There are 1,871,000 single women, under the age of 65.

SECTION OR SUB-SECTION OF POPULATION	PERSONS	% TARG
Single [marriage]	1871	5%
NOT (65 and Over) [age]	7003	4%
NOT (Single) [marriage]	5132	4% A.
Women [sex]	8205	4%
65 and Over [age]	1202	1% B.
ALL	16190	2%

(Partial output)

Looking at some of the other figures:

- A. There are 5,132,000 people who are women under the age of 65 who are “not single” and 4% of these people do yoga regularly. It is important to note here that “not single” means the respondents ticked any of the other options (eg engaged, widowed, de facto) but not single.
- B. 1,202,000 people are women aged over 65 and 1 % of them do yoga regularly.

Sharing Knowledge

If you find reading the figures difficult it may help to remember to read it in the following way, inserting the words specific to your example wherever you see the relevant brackets:

Of (Context Group), (Persons) are (combination of characteristics) and (%) of them are/do/etc
(Target Group).

As mentioned above, you should always start by reading the top line of the AID, to get the combination of characteristics that best describes your Target Group, but you should also look down the %Targ column for any other branches that have a high %Targ. A branch with a high %Targ can indicate further sets of characteristics that may be of interest. However, when looking at the %Targ you must always check the Persons column to make sure there are enough people in the group.

If the Persons column shows the sample is small, then it may not be large enough to be actionable, and it can indicate that the 'high' percentage in %Targ is due to the small sample having an exaggerating effect.

Also remember that the more often an A.I.D. branches the smaller the group of people represented by the last branch gets.

Reading the labels

Each label consists two parts:

The group/s
 ↘
(Single) [marriage]
 ↖
The variable

As described above one label in each pair also have NOT, being the opposite of the other in the pair.

Significance Threshold

When we looked at the components of the A.I.D. output we briefly touched on the following line of text:

Significance level: $p < 0.001$

This is set through Settings (Task) on the task toolbar.

It could in fact be counted as another step in running the A.I.D. but we are treating it separately here because the default setting of 0.001 will not often require changing.

A.I.D. takes all the characteristics we select through the Candidate Variables and looks at which characteristic is the best single predictor of who will be in our Target Group. For instance in our example the best single predictor was 'women' - i.e. more women do yoga regularly than men.

After A.I.D. has found this, it looks at which other characteristic combines with the first one to create an even better predictor - in our example, this was being aged 18 to 34 and female. Then A.I.D. looks for another characteristic which, in combination with the other two, becomes an even better predictor and so on. This continues on until one of two things happens:

- It runs out of selected characteristics.
- The difference between how well combinations predict who will be in the target group becomes too small to be statistically meaningful.

It is the second point where the Significance Threshold becomes important because the Significance Threshold draws the line between what is considered a large enough difference to be important and what is considered too small.

The "strictest" threshold is $p = 0.001$ and the "least strict" is $p = 0.05$.

In simple terms, the impact of the threshold setting is that the "stricter" the threshold the fewer times the tree branches. Generally it's good to avoid too many branches because, as mentioned earlier, the more branches the smaller the sample size for each branch.

Remember a small sample size (number of people) may mean that the %'s are exaggerated and/or that the group is too small to be worth considering.

Another reason that the ‘stricter’ test provides a better output, is that where a more relaxed threshold adds characteristics to the top branch that are only slightly better predictors, then you might ignore an almost equally important/useful group.

For example if we run our A.I.D. with a Significance Threshold of 0.05 and by relaxing the threshold our best predictor becomes single women aged 18 to 49 – what does this really mean?

If the 18 to 49 yr olds are really only slightly more likely to do yoga regularly than the under 18's or 50 - 64 yr olds, then are we ignoring a potentially useful part of our market if we follow this A.I.D.?

So what should you set the Significance Threshold to?

It is a good policy to leave it on 0.001 because if you find you don't get enough branches with this setting you can always run it again on 0.1 or 0.5, but if you are running complex A.I.D.'s, then you may prefer to start with a more relaxed test simply because of the time it takes to run an A.I.D. containing a lot of Candidate Variables. Just remember to look carefully at the figures.

Percent uncertainty explained

The final line in the output from our example was:

$$\text{Percent uncertainty explained} = 6.109$$

This figure indicates how well our selected characteristics describe members of our Target Group. What it does, in simple terms, is compare how much more accurately we predicted membership of the Target Group with these characteristics than we would have if we'd just taken a random sample from the Context Group.

Correspondence Analysis

What is Correspondence Analysis?

Correspondence Analysis allows us to see relationships between variables/groups and is only used where other methods for identifying relationships cannot be. It looks for relationships between all the variables/groups selected, graphs them in multi-dimensional space and then shows them in a 2D graph.

WARNING. Correspondence analysis is a visual tool and gives a general impression of the relationships to be found. It does not produce precise measures of the strength of the relationships it suggests. In all cases it is *essential* to re-examine the table from which it is derived to see how strong (how different from random) these relationships are and how large the overlap groups are that lie at the heart of a relationship. Correspondence analysis should never be used in isolation.

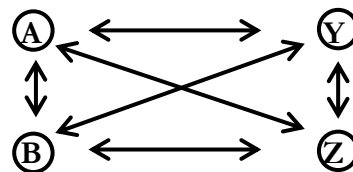
Why do we use Correspondence Analysis?

When we use Tabulation the resulting table shows the relationship of one row item to one column item at a time, but Correspondence Analysis takes all the selected items and looks at all the relationships at once. (Illustrated below.)

Tabulate looks at the relationship between two items at a time:

	Y	Z
A	xx	nm
B	yy	nn

Correspondence Analysis looks at the relationships between all items at once:



When should we use Correspondence Analysis?

Wherever you want to find what, if any, relationships exist between different variables/groups. This may be to help identify points of similarity between groups when creating new segmentation, or it may simply be to gain a better understanding of a target market, but it is important to follow the ‘rules’ below when considering whether to apply Correspondence Analysis to your data.

Data that *shouldn’t* be used for Correspondence Analysis:

- Variables that have a logical sequence (like age groups or income brackets).
- Variables where a respondent may be in more than one group (eg one respondent may read all the magazines in the Women’s Lifestyle Readership variable, or buy more drink more than one brand of soft drink in a period.) A small degree of overlap may be tolerated, but any overlap weakens the power of the analysis.
- Groups that are subtotals that will replicate data already selected.
- Summary grid variables.

So ask yourself these questions when considering using Correspondence Analysis:

- A. Is it ‘non-ordinal’ data (i.e. data with no logical sequence)?
- B. Is the data mutually exclusive (i.e. a respondent can only appear in one group within the variable)?
- & Remember not to include any groups that replicate data.

How do we use Correspondence Analysis?

Running a Correspondence Analysis is simple:

- Step One:** Check that the data meets the ‘rules’.
- Step Two:** Use Tabulation to create a table with the variables you are interested in. The resulting table must have a minimum of 3 rows and 3 columns.
- Step Three:** Click the Cor. An. tab at the bottom of the screen. The Correspondence Analysis then displays.

To illustrate how to use Correspondence Analysis, we will work through an example and examine the relationship of Federal voting intention to the state you live in.

1. The first thing we'll do is test whether our data meets the criteria explained above:
 - A. Is it 'non-ordinal' data? Yes – neither state nor voting intention are in any sort of sequence.
 - B. Is the data mutually exclusive? Yes – you can only vote for one party and you can only live in one state.
 & Remember not to include any groups that replicate data. – In this case (as you'll see below) we have to decide with voting intention whether we want to look at Liberal and National separately or as the Lib./Nat. coalition because to include both will replicate data.

2. We can run the following table:

Columns: States *(the whole variable)*

Rows: Federal Vote - 1st Preference:

ALP, Liberal, National Party, Greens, Australian Democrats, One Nation Party, Independent/Other

*Filter*³: Total 18 and Over

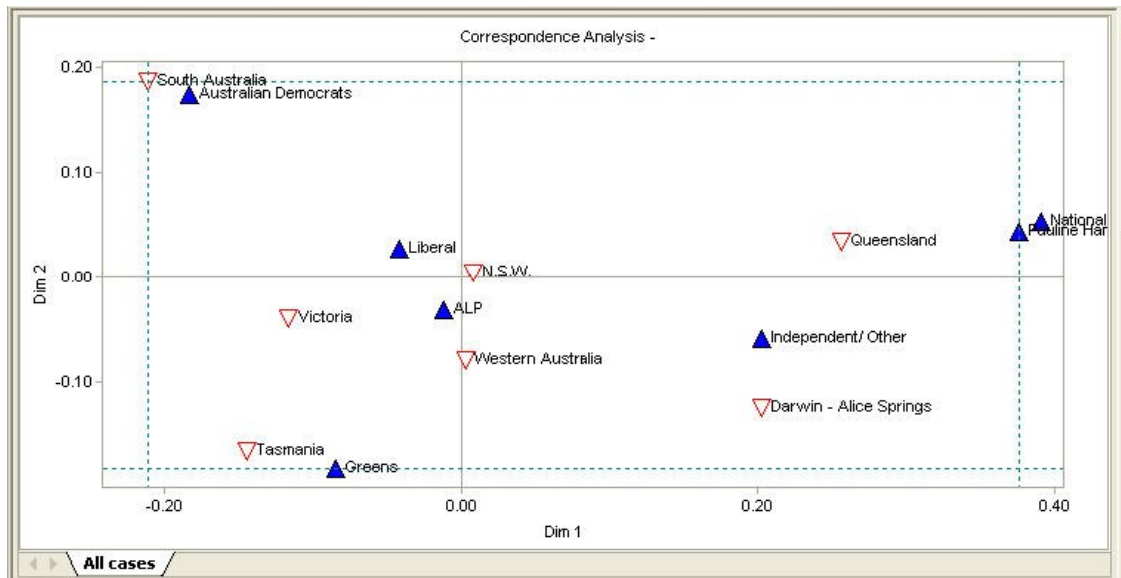
This gives us the following table:

	A	B	C	D	E	F	G	H	I	J	K
4	Filter: All cases							Projected population of Australia 14+ (in '000)			
5								No ranking			
6											
7								STATES			
8			TOTAL	N.S.W.	Victoria	Queensland	South Australia	Western Australia	Tasmania	Darwin - Alice Springs	
9											
10	(unweighted)	uc	56030	18803	13333	10157	4539	5455	3214	529	
11	(Pop'n '000)	wc	15740	5571	3969	2935	1236	1551	380	99	
12		h%	100%	35%	25%	19%	8%	10%	2%	1%	
13											
14	FEDERAL VOTE - 1ST PREFERENCE (electors)										
15											
16	ALP	uc	17681	5895	4478	3094	1340	1520	1202	152	
17		wc	4979	1738	1363	895	370	440	142	31	
18		v%	31.6%	31.2%	34.4%	30.5%	30.0%	28.3%	37.3%	30.9%	
19		h%	100%	35%	27%	18%	7%	9%	3%	1%	
20		ix	100	99	109	96	95	90	118	98	
21											
22	Liberal	uc	17268	5770	4191	2832	1658	1655	1001	161	

³ We apply a filter because only people aged 18+ are allowed to vote.

3. Click the Cor. An. Tab.

This is the resulting Correspondence Analysis:



Understanding the output

The steps to interpreting the scatter graph are:

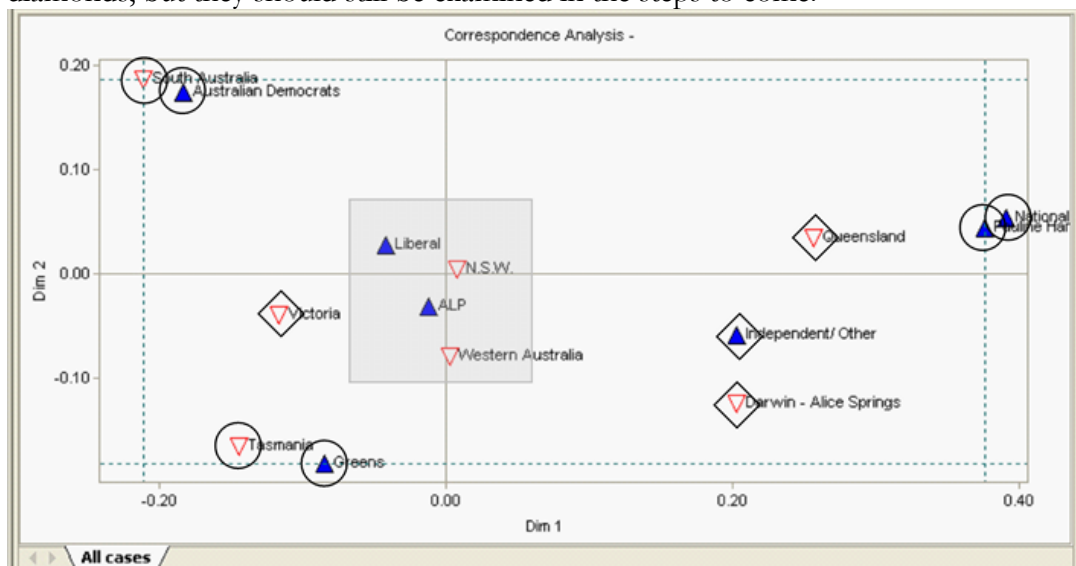
- | | |
|--------------------|---|
| Step One: | Look at 'how far' from the Origin the groups are. |
| Step Two: | Look for groups that lie in the same direction from the Origin. |
| Step Three: | Look for Outliers. |
| Step Four: | Look for groupings. |
| Step Five: | Look at the dimensions. |
| Step Six: | Review all the above information. |

1. How far are groups from the Origin?

Note: This is not about *distance* (i.e. not about saying ‘this group is X cm from the origin’) – it is an issue of the position of a group on the graph.

The reason we ask this question is because the further a group is from the Origin the more distinctive it is, or in other terms, the less like the average it is.

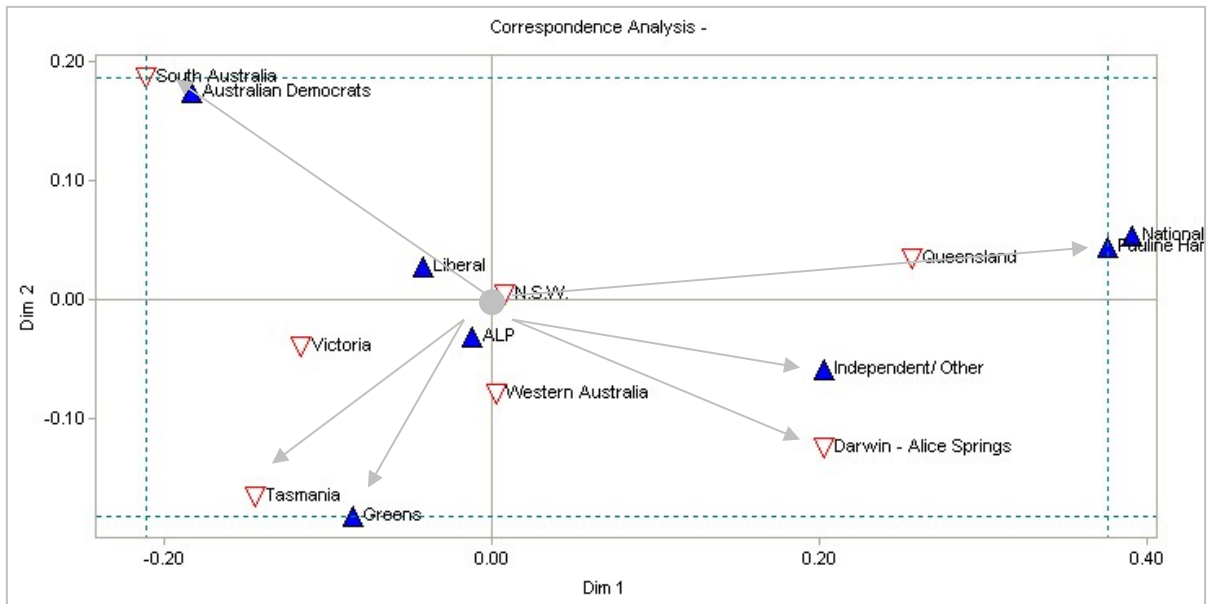
In the example below, the circled groups are far enough from the origin to be considered distinctive. We shouldn’t place too much emphasis on the distinctiveness of the groups in the diamonds, but they should still be examined in the steps to come.



If you look at the groups in the shaded area around the origin, you’ll notice they represent large proportions of the sample (eg the majority of people vote either ALP or Liberal meaning that these parties make a ‘large’ contribution to the average).

2. Do any groups lie in the same direction from the origin as other groups?

The direction of a group from the origin is not significant in itself but groups lying in the same direction indicate a relationship between those groups.



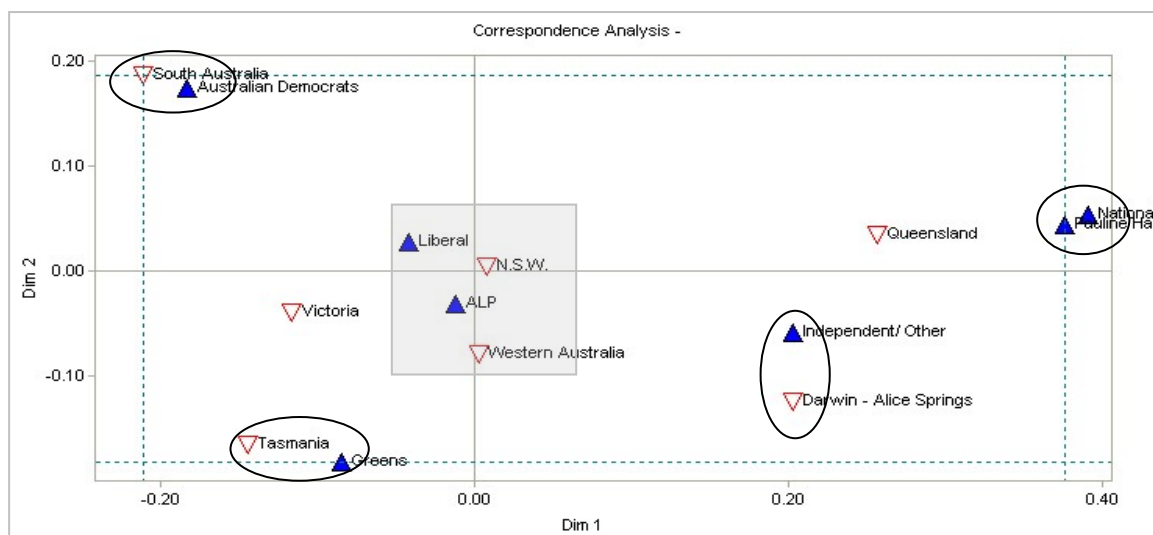
Most strikingly QLD, Pauline Hanson's One Nation and the National Party all lie in the same direction from the origin, indicating a relationship between intention to vote for these two parties and living in Queensland. (Again here the distance the groups are from each other is not the issue.)

The Australian Democrats and South Australia also lie in the same direction, showing a relationship between the state and voting intention.

The other arrows on the graph show that there is a mild similarity in direction with these two sets of groups but again this will need to be considered in the context of the other elements of interpreting the graph.

3. Do any of the groups form 'clusters' or 'groupings'?

Where groups appear 'close' together it indicates a relationship between those groups but this must always be considered in the context of distance and direction from the origin.



In our example we see that two groupings are quite clear – Pauline Hanson’s One Nation and the National party forming one grouping and the Australian Democrats and South Australia forming another. There will be more to say about the National Party / Pauline Hanson grouping in the section on Outliers but obviously where groups are very close together they will also be the ‘same distance’ from the origin and in the ‘same direction’.

The rectangle around the origin contains a grouping (or series of groupings) too but these groups are just too close to the origin for that to be meaningful.

More interestingly we have two other groupings where it is less obvious how meaningful they are. In the case of the Dar-Alice & Independent/Other grouping, neither group is very far from the origin nor is there a strong similarity in direction, which combined, indicates we cannot say there is a relationship here.

In the case of the Greens & Tas grouping the similarity in direction is no stronger than that of the Dar-Alice & Independent/Other grouping but here, both groups are far enough from the origin to be considered distinctive and that indicates that there is a relationship between these groups.⁴

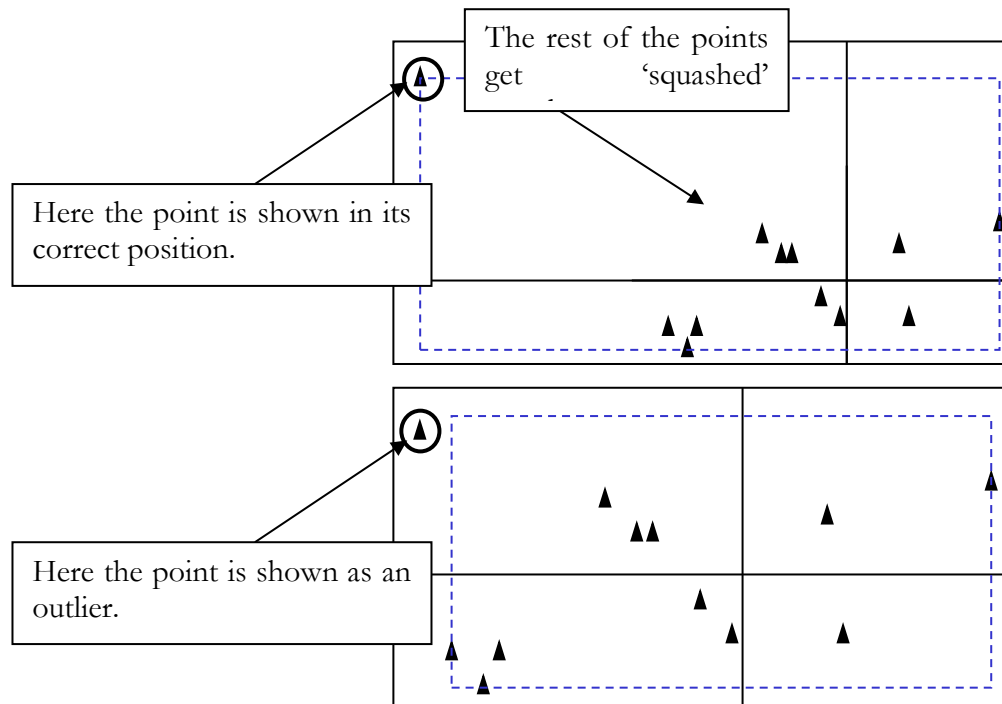
⁴ The relationship would need to be confirmed by an examination of the underlying table – more on this at the end of the section.

4. Outliers

The border on a Correspondence Analysis graph is formed by running a rectangle through the “outermost” points on the graph in each direction (i.e. the groups furthest from the origin up, down, left and right).

In our example the border is formed on the left by Pauline Hanson’s One Nation, at the top by the Greens and on the bottom and right by South Australia. An outlier is any point which is shown on or outside the border.

These points are not plotted on the graph in their true positions, because if they were the rest of the graph would be distorted as shown below:



How do outliers effect our interpretation of the graph?

Distance: The fact that a point qualifies as an outlier indicates that it is a very distinctive group because it lies so far from the origin.

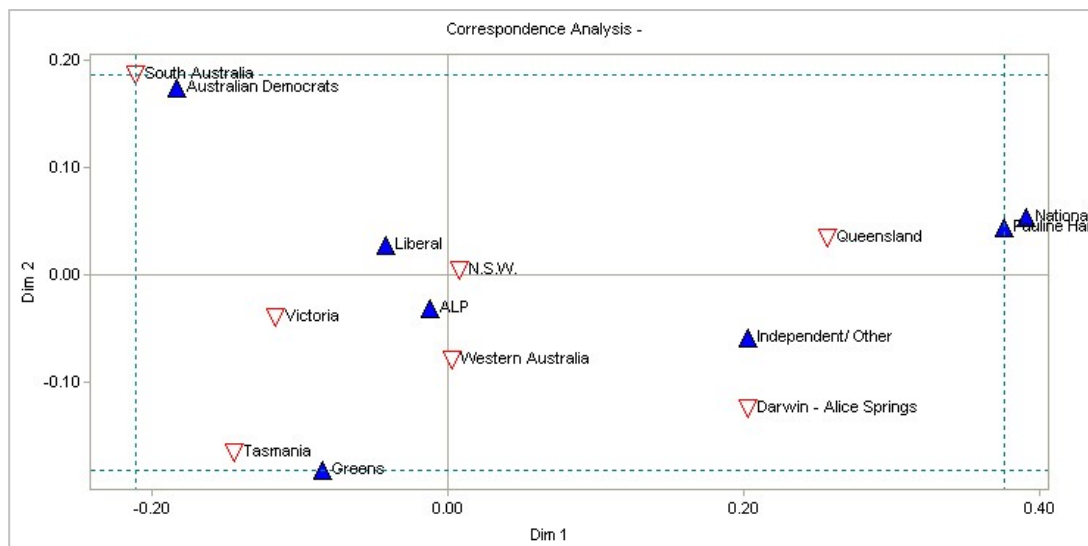
Direction: An outlier is positioned in the same direction from the origin as if it were being plotted in its proper position.

Groupings: If the groups within the grouping that includes the outlier, lie in the same direction from the origin then the grouping still indicates a relationship.

5. The dimensions

On a normal graph, the axes represent specific aspects of the data being graphed. For example, if you're graph looked at money spent monthly then one axis would represent \$ spent and the other the months. In Correspondence Analysis the axis represent two 'unnamed' dimensions and it is not always possible to determine what these dimensions are.

In our example we can gain some idea of what the dimensions represent by looking how the groups are spread along each axis. The horizontal spread indicates that we are going from parties that are extremely right wing through to parties that aren't – though we can't call the Democrats left wing, so maybe it is 'socially progressive' or similar.



Looking at the vertical spread it is probably even harder to decide what dimension represents.

Something separates the Greens and the Democrats possibly.

The significance of the dimensions is generally of more interest in the advanced application of Correspondence Analysis, but in simple terms interpreting the dimensions can tell us something further about the groups.

For instance, Victoria is too close to the Origin to show any meaningful relationships with any other groups but if we look at its position on the X axis we see that it is further to the ‘socially progressive’ end of the political spectrum than NSW.

By looking at the underlying table, we could find out why it appears at this end of the graph.

6. Review all information

As you may have gathered it can be difficult to accurately interpret the output, and that is why the final step in interpretation is looking at all the information gathered so far and carefully reviewing it. There are two things to remember when doing this review:

- Are results being caused by an external factor? An example of this is the correlation between Pauline Hanson’s One Nation and Queensland, where the external factor is that Queensland is the state from which this party originated.
- Does the underlying table support the interpretation of the Correspondence Analysis? For instance the relationship between Tasmania and the Greens is not ‘strong’ on the graph and we can check its real strength by looking at the table.

A final word on interpretation: Correspondence Analysis is a fairly advanced statistical tool and while we have endeavoured to explain its application in layman’s terms here, the tool is most ‘safely’ used by people who have a full understanding of it as a statistical tool.

A note on label position

You can move the labels next to the data points on the Correspondence Analysis simply by clicking on them and then moving them to a new position.

Quantities

Generally in ASTEROID we look at the number, or percentage, of people who have particular characteristics (e.g. the number of people who have bought a new car in the past year). However, sometimes it's useful to look at dollars spent or quantity drunk instead, and this is when you use quantities.

By adding a quantity to your tabulation, you automatically change the output from showing people to showing, for instance, dollars. Below are extracts from two tables, where one shows 'population' and the other shows 'dollars spent per person on last short trip'.

		TOTAL	Own car or 4WD
(unweighted)	uc	125295	40138
(popn. '000)	wc	15937	5310
	h%	100%	33%
Total hotel/ motel	uc	18195	12241
	wc	2394	1646
	v%	15.0%	31.0%
	h%	100%	69%
	ix	100	206

		TOTAL	Own car or 4WD
(unweighted)	uc	125295	40138
(Dollars 000s)	wc	1863336	911843
	h%	100%	49%
Total hotel/ motel	uc	18195	12241
	wc	680044	374168
	v%	36.5%	41.0%
	h%	100%	55%
	ix	100	112

You can see in the table on the left, the (popn. '000) tag, indicating the wc is in thousands and representing people. Compare that with the (Dollars 000s) tag in the second table, indicating the wc is in thousands of dollars. Note that the uc is the same in both tables – it still represents the number of people interviewed. These are some typical examples of quantities:

- Serves purchased
- Glasses drunk
- Number of visits
- Amount spent
- Value of accounts at main financial institution

By default, the wc in a quantified table refers to the quantity (e.g. Dollars). However, it is possible

to display 'wc' for population AND 'qwc' for quantity. It is recommended that you use this option, particularly if using quantities for the first time (this can be set as the default via Settings). Go to Settings (Task or Global) and Show Numbers. On this screen, tick the option for 'Show both quantified and unquantified weighted counts, when a quantity is applied':

Side Tags	Description	Decimals
<input checked="" type="checkbox"/> uc	Unweighted Counts	
<input checked="" type="checkbox"/> wc	Weighted Counts	<input type="text" value="0"/>
<input checked="" type="checkbox"/> v%	Vertical Percents	<input type="text" value="0"/>
<input checked="" type="checkbox"/> h%	Horizontal Percents	<input type="text" value="0"/>
<input checked="" type="checkbox"/> z%	Layer Percents	<input type="text" value="1"/>
<input type="checkbox"/> ix	Index Figures	<input type="text" value="0"/>
<input checked="" type="checkbox"/> mn	Mean Scores	<input type="text" value="2"/>

----> ☐ Show both quantified and unquantified weighted counts, when a quantity is applied

Show Side Tags

☐ Auto ☒ On ☐ Off

The table on the left shows how quantities were displayed previously. The table on the right shows the uc and wc for population, as well as qwc giving the Dollars weighted count:

		TOTAL	Own car or 4WD
(unweighted) (Dollars 000s)	uc	125295	40138
	wc	1863336	911843
	h%	100%	49%
Total hotel/ motel	uc	18195	12241
	wc	680044	374168
	v%	36.5%	41.0%
	h%	100%	55%
	ix	100	112

		TOTAL	Own car or 4WD
(unweighted) (Dollars 000s)	uc	125295	40138
	wc	1863336	911843
	h%	100%	49%
Total hotel/ motel	uc	18195	12241
	wc	2394	1646
	qwc	680044	374168
	v%	36.5%	41.0%
	h%	100%	55%
	ix	100	112

Using Quantities

To see whether or not your database contains quantities, look at the tabs at the bottom of the Data Dictionary. Where quantities have been included there will be a Quantities tab.



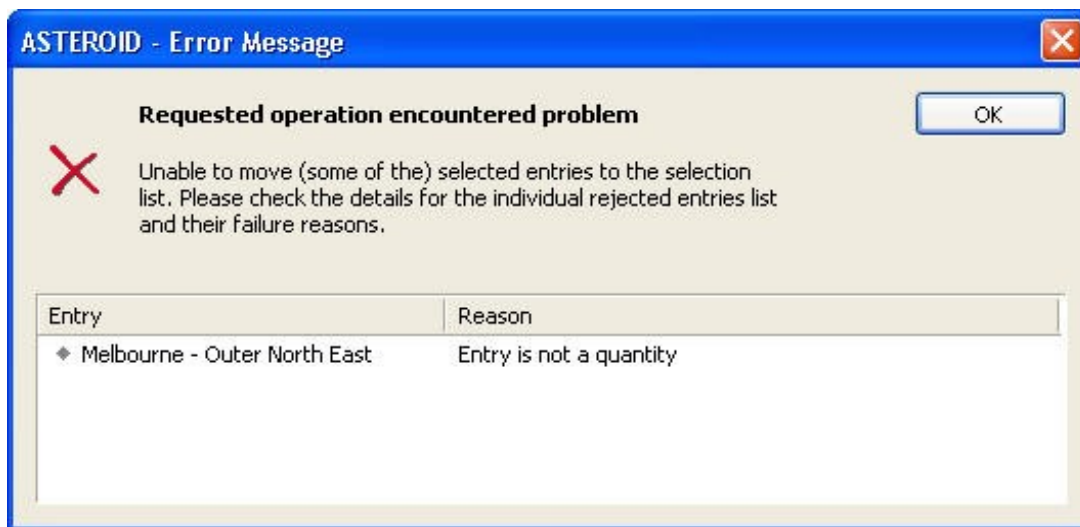
To apply a quantity to a table, you must select it from the Quantities tab. You can either double-click or drag-and-drop it to the Quantity section at the bottom of the selection screen.



You can see that quantity items have a different icon to variables within the database, which makes them easy to identify.

When working with quantity data the mean and median values can be displayed from the Setting menu, as shown below:

Note: ASTEROID will not allow you to place a quantity into another section of the selection screen, e.g. Rows or Columns, or to place a variable/group (from the All Items tab of the Data Dictionary) into the Quantity area. If you try this, the following message will appear:



Understanding Quantities

The following example relates to the last short trip taken, where we have mode of transport used in our columns, type of accommodation used in our rows and a quantity of Cost per person on last short trip:

Filter: All cases					
Quantity: Cost per person on Last Short Trip					
MODES OF TRANSPORT USED ON LAST HOLIDAY OF 1-2 NIGHTS					
		TOTAL	Own car or 4WD	Friend's/ Relative's car or 4WD	Hire car or 4WD
(unweighted)	uc	125295	40138	7641	2039
Dollars (000s)	wc	1863336	911843	214649	133018
	h%	100%	49%	12%	7%
TYPES OF ACCOMMODATION USED ON LAST HOLIDAY OF 1-2 NIGHTS					
Total hotel/ motel	uc	18195	12241	1909	922
	wc	680044	374168	70408	70890
	mn	284	227	291	529
	cl	+/- 8	+/- 6	+/- 25	+/- 65
	v%	36.5%	41.0%	32.8%	53.3%
	h%	100%	55%	10%	10%
	ix	100	112	90	146

Note that this table is in the default format, without displaying the wc AND qwc.

Sharing Knowledge

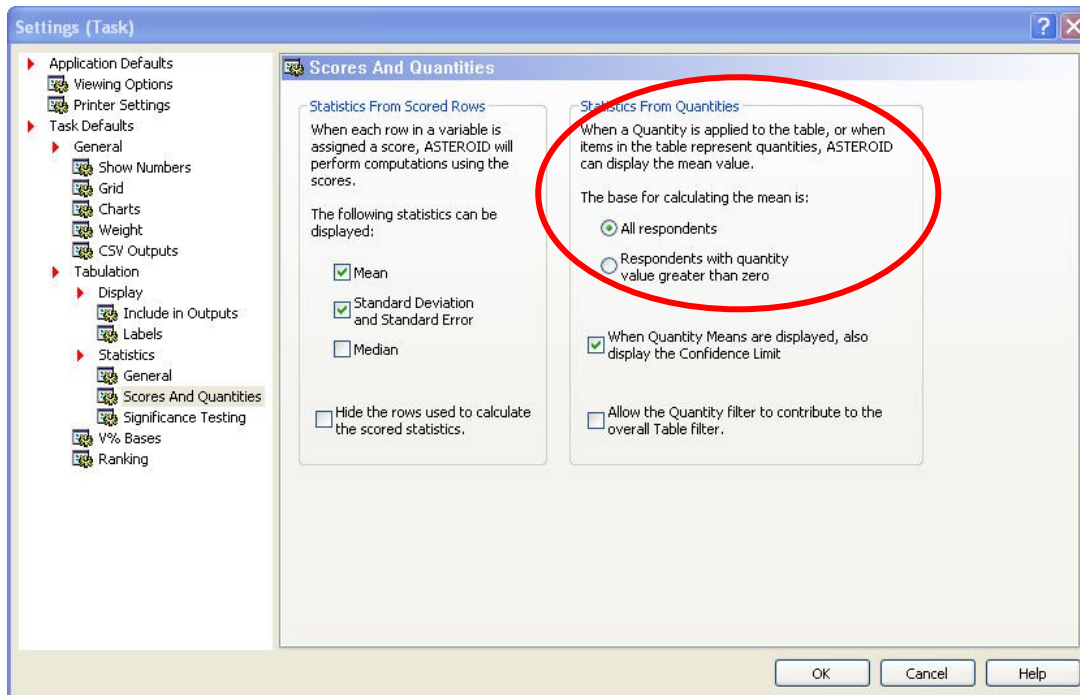
The table tells us that:

uc	There were 18,195 respondents to this question in the survey, who stayed in a hotel/motel on their last holiday of 1-2 nights.
wc	The total expenditure of all people on their last short trip, by those who used their own car or 4WD on their last holiday of 1-2 nights, came to \$374,168,000
mn	Those who stayed in a hotel/motel and drove their own car/4WD on their last holiday of 1-2 nights, spent an average of \$227 per person on their last short trip.
cl	For those who stayed in a hotel/motel on their last holiday of 1-2 nights, and travelled by own car or 4WD, 95% of the time their average will range from \$221 and \$233
v ⁰ %	Of those people who drove their own car/4WD on their last 1-2 night holiday, 41% of their expenditure was by those who stayed in a hotel/motel on their last short holiday.
h%	55% of expenditure by people who stayed in a hotel/motel on their last 1-2 night holiday, was by people who drove their own car/4WD on their last short holiday

To account for sampling error, the confidence limits⁵ (cl) for the mean are displayed. This tells us the range by which the mean might vary 95% of the time. This allows us to more accurately interpret the mean amount spent per person.

⁵ This is the 95% confidence limits (cl) of the estimated mean.

When reading mean values, be aware of two different settings that can be used:



‘All respondents’ will include those who are not included in the Quantity which is applied to the table, whereas ‘Respondents with quantity >0’ will filter to only those who have contributed to the quantity total.

Significance testing

What is Significance Testing, and why do we use it?

When we make an estimate based on a sample, we are aware that another sample, selected independently but in exactly the same way, would give a different estimate. If a very large number of similarly-drawn samples were used we would see a range of different estimates. Most estimates would be bunched around a central value and a few would be outliers. Extreme values are rare, but still possible. This is known as sampling variation.

Significance testing is a way of examining differences that are observed within the data, or between a sample-based estimate and some assumed value, to see how likely it is that these differences can be ascribed to natural sampling variation.

When we look at a tabulation, we often compare the result in one column to that in another column. While these results differ, the fact that the data source is a sample of the population means that some or all of this difference may be due to sampling variation.

Significance testing allows us to measure the likelihood that this is so, and conversely, the likelihood that an observed difference is representative of an actual difference in the real world.

A statistical analysis of the sample allows us to express statistical confidence in the difference as a percentage, which is displayed in ASTEROID as markers.

The significance testing available in ASTEROID provides two ways of assessing the differences in the data.

- Testing individual columns against the total.
- Testing individual columns against other columns.

Considerations When Using Significance Testing

In interpreting a significance test, there are three critical factors to consider:

- A 'significant' difference is not necessarily meaningful. You must always consider first whether the size of a difference is big enough to be of interest / concern or to be actionable.

For example; a study found that the difference in the average length of the wool on the left and right sides of a sheep was statistically significant, but the actual difference in length was so small that it had no meaningful impact on wool yields.

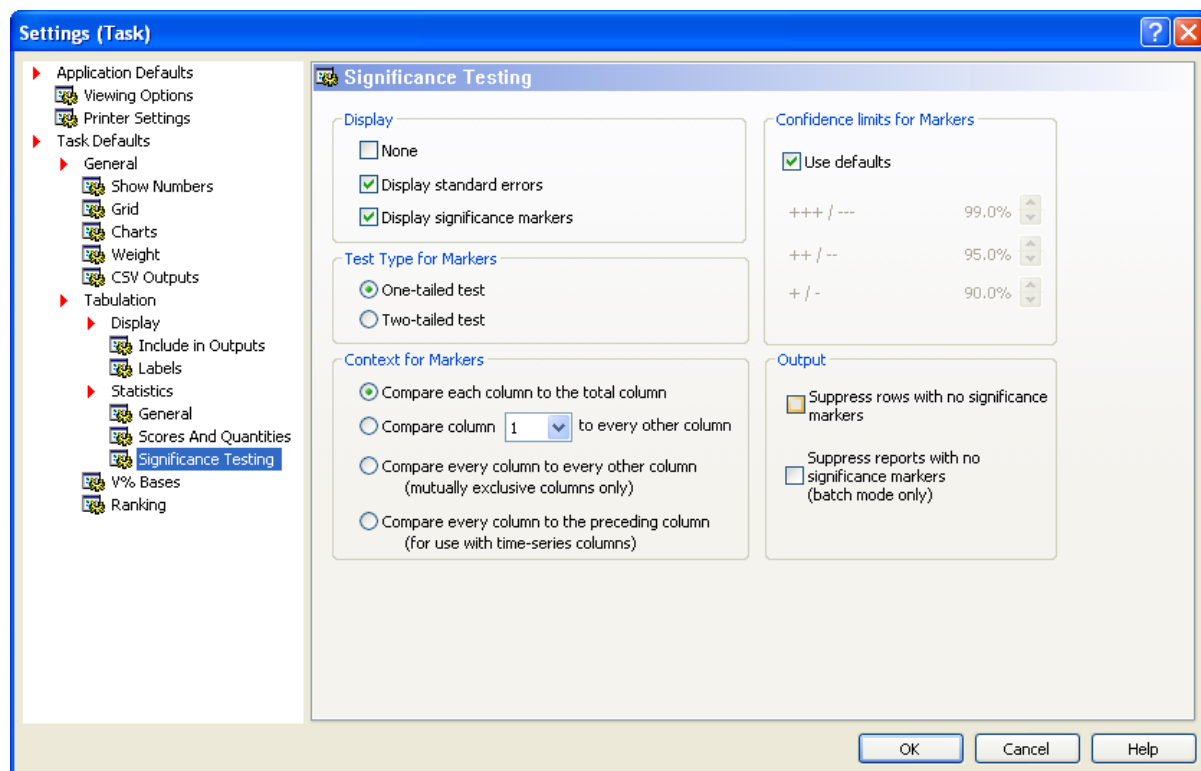
- The significance of a difference is related to an assumed difference of zero. If we would reasonably expect the real-life difference to be other than zero, the significance test must be interpreted appropriately.

For example; we would reasonably expect a lower percentage of people who do factory work to have university degrees, than people working in scientific research. So, when we expect to see this difference between them, a significance test confirming this doesn't really tell us anything.

- The confidence level at which the difference is significant is not necessarily a measure of the size of the difference. A small difference can be very statistically significant, when drawn from a large sample, and a large difference can be insignificant, if drawn from a small sample.

ASTEROID significance testing takes account of the increase in variance of percentages caused by any weighting applied, but there may be other factors affecting them that are not taken into account. You should expect that ASTEROID will tend to slightly overstate the significance of differences.

Output Display



Display

There are three options for displaying Significance Testing, the first of which is **None** – this is the default setting.

Displaying the Standard Error

When this setting is selected, the approximate standard error of the vertical percentage is shown as an absolute percentage below each cell. This is useful in analysing the relative confidence in the accuracy of the percentages, but is not in itself a test for statistical significance. In fact, it is important to note that this is not the only measure from which the significance is calculated, as it only relates to one of the two groups being compared.

It does, however, allow you to assess the likely confidence interval of the 'v%' figure, so that you can apply your own test to any difference based on specific assumptions relative to the circumstances. It thus offers a high level of flexibility. See Standard Errors on page 36 for more detail.

Display significance markers

By selecting this option, Confidence Limit Markers are displayed on the table when you run it. The following settings relate to the way in which the markers are calculated, and what they represent.

Context for Markers

Significance testing based on confidence limits can be reported graphically, based on one of four settings. In each of these four settings, each significance marker is derived by comparing one result against another – the selection of the context determines how these pairs of results are selected.

Please read and be aware of the “Important warnings

” on page 35 when considering significance tests and confidence limit estimates.

Compare each column to the total column

"Which of my columns differ significantly from the norm?"

Each cell is compared with the corresponding row total and an indicator is shown below the cell to show the approximate level of statistical significance of the difference between the cell total and the row total cell (in effect comparing the vertical percentages).

Note that this means the *complement set* is compared, so that the comparison is based on mutually exclusive sets. In other words, **a marker indicates the significance of the respondent set in the column compared to the set of respondents that do not qualify for the column.**

For instance, a column defined as respondents "Aged 18-25" would be compared to respondents outside that age group. This is useful where you are not starting out with an assumption that groups will differ from the total column.

Compare column 'n' to every other column

"Which columns are significantly different to this one?"

This is similar to the process above, except that instead of comparing to the total column, each column is compared to a nominated column.

An important difference is that while in the circumstances described above the column is always a subset of the row total, in this case there is no such restriction. After resolving any overlap (see subsequent note) **each marker indicates the significance of the difference between the respondent set in the column compared to the respondent set in the nominated comparison column.**

As above, an indicator is shown below the cell to show the approximate level of statistical significance. Because the direction of the change depends on the interpretation of the two test groups, the symbols used are always asterisks.

Compare every column to every other column (mutually exclusive columns only)

"Which columns are significantly different to each other?"

For those columns in the table that are mutually exclusive, each pair of columns will be tested for significant differences. Mutually exclusive columns are identified with a letter in the column heading. For each column, the column to its right for which a 'significant' difference is found are indicated by corresponding letters shown beneath the cell. Therefore, **each marker indicates the significance of the difference between the respondent set in the column with the marker, and the respondent set in the column the marker refers to.**

Upper-case letters are used where the difference is judged to be above the '***' significance level (by default, the 95% confidence interval); for others judged at least '**' (by default, the 99% confidence interval) lower-case letters are used.

Note that only 26 columns in a table can be distinguished using letters.

Compare every column to the preceding column

"Which periods represent a significant change from the previous period?"

This test is designed for, but not limited to, use with time-series columns. Where the columns are a time series set, the expectation is that each column is a later period than the one to its left, and so **each marker indicates the significance of the respondent set in the one period compared to the respondent set in the previous period.**

It is important to note that when displaying this type of test, the likelihood of significant differences may be reduced by the granularity of the periods. For instance, a measure may display a statistically significant change across the course of a year, but each month or week might not show a significant change, due to the smaller change and lesser sample size. It is also worth

considering using the second context option, and comparing all preceding time periods to the latest.

Although designed for time series columns, the test is equally appropriate for continuous variables (for instance, age groups), where there is a particular order to the series of columns.

Outputs and Output Options

Confidence limits for Markers – the level of Significance, and the number of tails.

‘Significance’ markers show the approximate level of confidence which may be placed in the difference between cell and row total. ASTEROID expresses the result as being at a particular significance level (e.g. at the 99% significance level) and the higher the significance level, the more confident you can be that the differences you’re examining weren’t caused by variation in the sample.

A 99% significance level tells you there is a 1/100 probability that the difference is a chance event, and so on.

By default, ‘+’ symbols are used where the cell figure is higher and ‘-’ where the cell is lower.

The confidence levels can be adjusted, but the default values are:

+++	- - -	99%
++	- -	95%
+	-	90%
=		<90%.

As discussed, a +++ marker does not necessarily describe a larger change than a + marker, merely a difference for which we can be more confident that it is not due to sampling variation.

Test Type For Markers

The choice between a ‘one-tailed’ and a ‘two-tailed’ significance test is left to the user to decide in the light of the specific circumstances. Generally, where it is possible to include a statement of the direction of the difference, a one-tailed test is more appropriate. For instance if we are asking the question “What is the probability that the ‘true’ difference between two estimates would be zero or less?” we are concerned with only one end of the distribution of possibilities, not with the possibility that the difference might really be much greater, so a one-tailed test would be preferred. In the case of a two-tailed test, which is essentially non-directional, one, two or three asterisks replace the ‘+’ or ‘-’ symbols.

Output Display

The Output options allow for suppression where there are no significance markers:

- You can opt for rows to not be displayed if there are no markers
- Reports may be suppressed if they have no significant differences – this is applicable in batch mode only. This option is useful when auditing a database, or as a data mining tool.

Please read the “Important warnings

” and “Considerations When Using Significance Testing” sections before using these options.

Methodology

Every significance marker is the result of a test between two independent sets of respondents.

Where the two sets (test group and context group) are already mutually exclusive, they are recognised as being independent, but in other situations ASTEROID must identify the mutually exclusive components.

- Where there is a partial overlap, the overlapping set of respondents is excluded from the statistical comparison.
- Where one group is a subset of the other, the subset group is compared to the complement.

This is an important concept to remember. For instance, comparing a January-June result with a February-July result is effectively comparing January to July, as the overlapping respondents are removed from the analysis. It is not a test of how much a large sample has changed within one month, but how much a smaller sample has changed within six months. To include the over-lapping respondents in both groups would invalidate the statistical test – you cannot analyse the difference between two sets of data when they are the same data.

The test applied is by default a simple large-sample one-tailed \bar{x} -test (a two-tailed test can be used if appropriate). In the case of proportions this is equivalent to a 2×2 χ^2 (chi-squared) test. Each test is a comparison of the results for a pair of respondent sets, tested independently – there is no test of the distribution of any variable as a whole as there is no control over overlap between groups.

Important warnings

Roy Morgan Research accepts no responsibility for the interpretation of significance testing or confidence limits made by users.

1. The calculation of 'significance' takes account of the increase in variance of estimates caused by any weighting applied. It does not, however, make any allowance for the effect on variance estimates of any factors affecting sample design or execution including (but not limited to) stratification, post-stratification, clustering and finite population corrections. Where sample designs are complex, users must make due allowance for this in the interpretation of significance tests. In the case of quota or other non-probability samples, it is not possible to calculate a design effect as there is no evidence within the data of departures from equal probability of selection of the sample units.
2. Significance tests rest on the hypothesis that an observed value should not be different from an 'expected' or 'null-hypothesis' value. In automating significance testing the only general standard that can be applied is that the 'null hypothesis' be taken as 'no difference between cell and row total', or 'no difference between the two cells being compared'. In many cases this may not be valid or appropriate. Users must make their own judgement about the appropriateness of this null hypothesis in each case. Users must similarly decide whether a one-tailed or two-tailed test is the more appropriate for the circumstances.
3. Standard errors and confidence limits are measures of the *precision* of sample-based estimates, not of their *accuracy*. They relate to the reproducibility of results, that is, to the likely range of estimates which would be yielded by multiple samples of the same nature, not with the absence of bias. An individual result may differ from the 'true' value because of natural sampling variation but also because of a range of factors to do with the sampling method and its execution and also because of non-sampling factors, for instance in sample surveys question wording and order, measurement or memory errors or psychological pressures on respondents

Standard Error⁶

The Standard Error provides a way of judging how much confidence you can have in an individual $v\%$.

If we were able to conduct the same survey multiple times in exactly the same way the results would vary from sample to sample. The Standard Error is an estimate of the amount of variation we would expect to see in the estimates.

We recommend that you do not use Standard Errors unless you are reasonably familiar with the theory and reasoning behind them. The following simplified summary is not a substitute for a proper basic statistics course.

Using Standard Error

For survey data, it is unlikely that the $v\%$ matches the ‘true’ value - that is, the value that a very much larger sample would have yielded - with exact precision. We use statistical tables to project a range for which we can be confident. For instance, if we assume a standard distribution, the range which we can be 95% confident encompasses that ‘true’ percentage is calculated as $\pm(1.96 \times \text{Standard Error})$.

It is important to understand that this allows us to project confidence against such a ‘true’ value, and that when comparing two figures from surveyed data, the sampling variation for both needs to be accounted for. This is where the significance testing tools come in.

It is also important to understand that there is a relationship between the size of the sample and the size of standard error. You would expect to see that the larger the sample, the smaller the standard error.

Understanding the ‘true value’

The standard error is a measure of precision or reproducibility, not of accuracy or external validity. It is a measure of how much variation there would be between a large number of samples of the same size selected in the same way.

What you are comparing an estimate with, when using a standard error, is the more reliable estimate you would get from a *very* large sample or the mean you would expect from a large number of similar samples. However, as you (normally) have only one sample to work with you are trying to estimate how far such a more reliable estimate might be from what you have got, or the chances that it would differ from the estimate you have by more than a certain amount. The

⁶ Note that this is not Relative Standard Error.

‘real world value’ may be different because of a range of other sampling and non-sampling errors, especially biases, not just sampling variation.

Cluster Analysis

Cluster Analysis is a statistical tool allowing you to divide a specified population into groups that share characteristics and behaviours. It can be used both to gain a deeper understanding of a population or target group, and also as an initial step in creating segmentation.

It is important to understand that Cluster Analysis is an exploratory tool – not a one click, one answer tool – and creating good clusters is a multi-stage activity.

Note: Cluster Analysis ‘segments’ (divides up) a population, but this isn’t the same thing as creating the ‘segmentation’ we usually talk about in market research.

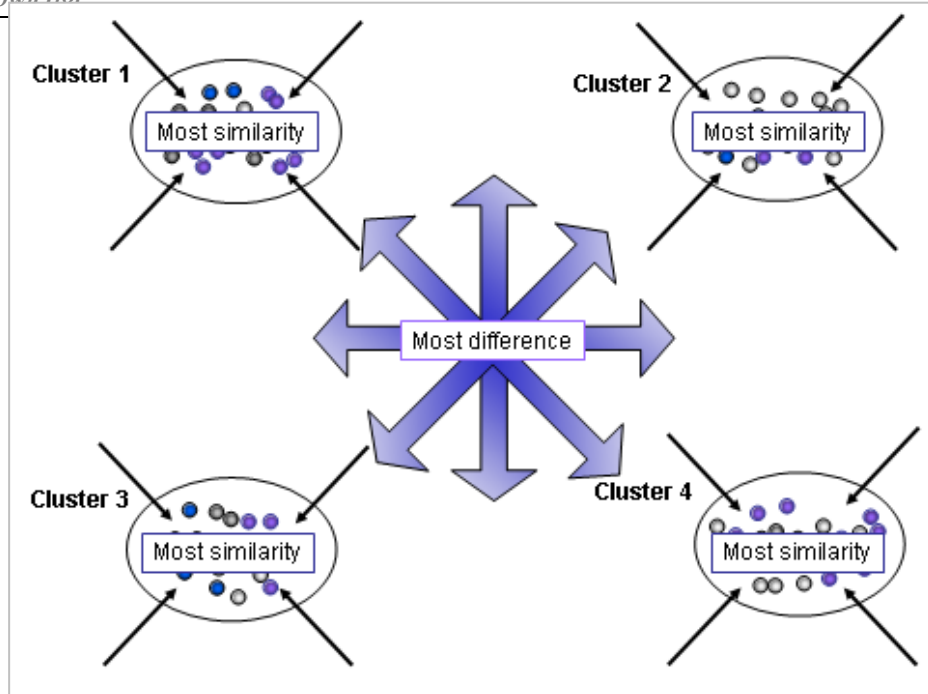
- With segmentation you develop rules that determine the people who belong to a particular segment.
- With Cluster you are using statistical analysis to divide a specific population into a preset number of sub-groups.

Generally, Cluster Analysis is most useful where you already have a good understanding of your data and wish to look into it further. For example, you may know what makes consumers of a product stand out from the rest of the population, but be interested in what sub-groups exist within that group of consumers.

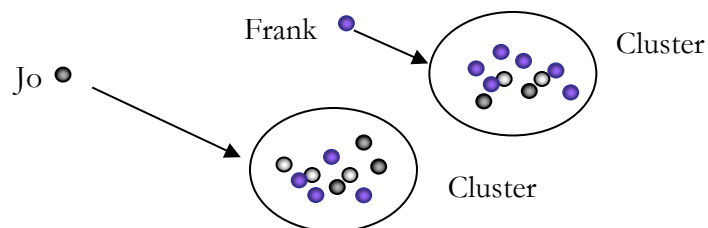
What does Cluster Analysis do?

Cluster Analysis takes the population you specify and divides it into clusters based on the variables you ask it to examine.

More specifically, it takes people’s values of the Candidate Variables you choose (such as ‘Intention to buy a new car.’, ‘Going interstate on last holiday.’ etc), it examines each person to determine how similar/different they are to the other people, and then allocates them to a cluster.

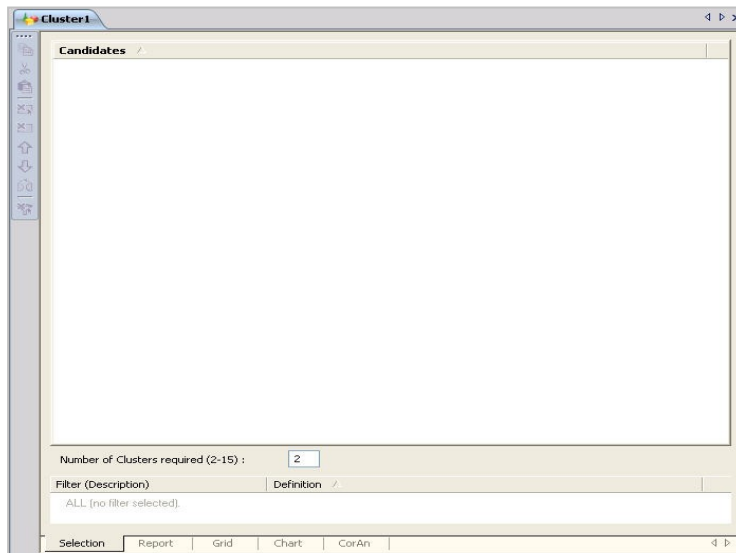


The intention is to allocate each respondent to a cluster such that each cluster is internally as homogenous as possible and externally as different as possible from the other clusters.



The maths used to create the clusters, groups people who are most statistically similar to each other and most different from everyone else. In other words, the clusters are as different as possible from each other, but within each cluster the people are as similar as possible.

How do we use Cluster Analysis?



To generate a Cluster Analysis you select your Candidate Variables, specify the number of clusters you want and then decide whether you want to apply an overall filter. It is choosing which variables to use and how many clusters you want that takes an understanding of the data and a degree of trial-and-error, to find a useful cluster solution.

The output from Cluster Analysis – the cluster solution – provides you with a cross-tabulation. Looking at the v⁰s you can work out which variables define the clusters.

Sometimes they will divide so strongly on a variable that you will be able to see it at a glance – other times it will take careful examination. To assist with this, indexes over 20% above or below the index are marked in colour.

A number of statistical measures have been included at the bottom of the output, to help you decide on how successful the cluster solution has been. These will be discussed in detail from page 46.

In the following pages we will look at the steps involved in running a successful Cluster Analysis.

- | | |
|--------------------|-------------------------------|
| Step One: | Know your data. |
| Step Two: | Selecting variables. |
| Step Three: | How many clusters? |
| Step Four: | Applying a filter. |
| Step Five: | Examine the output. |
| Step Six: | Re-run or Save. |
| Step Seven: | Explore the clusters further. |

Step One – Know your data

To be able to use Cluster Analysis effectively you must have a good understanding of the data you are working with. You must be able to make informed decisions about what to include and how to interpret the results.

Step Two – Selecting variables

When selecting variables for Cluster Analysis you must follow these guidelines:

- Variables must be ‘continuous’ or single-choice
- Avoid demographics
- Avoid duplication
- Avoid ‘bunched’ scales/variables
- Avoid selecting too many variables / avoid variables containing too many groups.

Single choice variables

Because so much market research data is multi-choice (where a respondent can choose more than one answer to the question), ASTEROID has been designed to convert them into single choice variables in Cluster Analysis; treating each group as a variable. (You will notice when selecting multi-choice variables that they appear as individual ‘groups’ in the Candidate Variable window.)

Variables used in cluster analysis must have some numerical sense or contain a logical progression (e.g. ‘age’, ‘number of children’, ‘education level’). Single-choice variables with no such ordering (e.g. ‘State’) can be used, but by treating each category as a separate variable, in effect a separate ‘yes/no’ variable.

Avoiding demographics

Demographics are often our first choice when choosing characteristics on which a population can be divided, because they have such strong correlations with both other demographics and many other characteristics.

For example, many products are purchased predominately by people of a particular age, sex or income level, and there are also links between demographics such as age, income, education level, socio-economic group etc.

There is no real reason why demographic variables should *not* be used at all, except that their effect may already be well-known and cluster analysis is used to explore the areas we don't know about. A division along demographic lines might mask far more interesting information, but equally, if in a mix of demographic and other variables the demographics dominate the clustering, it presumably means that other variables don't have so much of a role to play.

As a basic guide, however, it is recommended to avoid using demographics in the Cluster Analysis, and then cross-tabulate the clusters against demographics later to get the demographic profile.

Avoiding duplication

Many variables – particularly those containing attitudinal statements – contain a number of groups that essentially say the same things, or diametrically opposite things, in different ways. The inclusion of such variables will bias the division of the clusters in a particular direction.

For example, in the variable below, there are 3 attitudes regarding the environment/eco-tourism (see in bold). If we include them all in a Cluster Analysis, we might expect a cluster to form around them.

I'm always very active on holidays.

On holidays I like to do as little as possible.

I prefer the bright lights and big cities when I travel.

I prefer to holiday where I can see nature or be in a natural setting.

I usually book and arrange all my holiday travel details myself.

I sometimes organise holidays on behalf of my family and friends.

I usually leave holiday arrangements to someone else.

I avoid staying at accommodation that does not have genuine environmental policies.

For my next holiday, I'd really like a total ecotourism experience.

I'd like to holiday where I can experience the local culture.

This problem can be avoided by taking only some of the groups and selecting them carefully. Obviously if you aren't sure how strong a relationship there is between the groups you should cross-tabulate them or even try running a cluster solution to see if they do form a cluster.

Avoid bunched scales / variables

Bunched scales – where most people fall at one point on a scale, will contribute weakly as they do not separate people adequately

High/Low incidence – groups within a variable that have particularly high or low incidence will similarly influence the formation of the clusters less.

Selecting too many variables/groups

The more variables you include, the less likely it is that they will all contribute to the Cluster Analysis which means they will simply create ‘noise’ and make it harder to see the useful information. This means it will take you longer to refine your clusters and also creates greater potential for misinterpretation.

If a variable contains a lot of groups then it is likely that you will either run into the problem of high/low incidence mentioned above, or they will all have such low incidence that it’s unlikely they will make any useful contribution to the Cluster Analysis.

Step Three – how many clusters?

The next step is to decide how many clusters to ask ASTEROID to create. You are looking for clusters that make sense – a rule of thumb for this is whether you can give each cluster a name.

Where there is no obvious starting point, start with 2 clusters, then run 3, 4 etc and see how the output changes.

If there is something in the variables that would create a two-way split then you can try three or four clusters to see what happens. (If you still get two cluster splitting on those lines, you can try removing that variable/group.)

Remember that you must try different things – you might find that you get equally useable solutions when you do 3 clusters and 8 clusters but they will be different types of segments.

Most importantly, remember that:

There is no magic number, and there is no one solution.

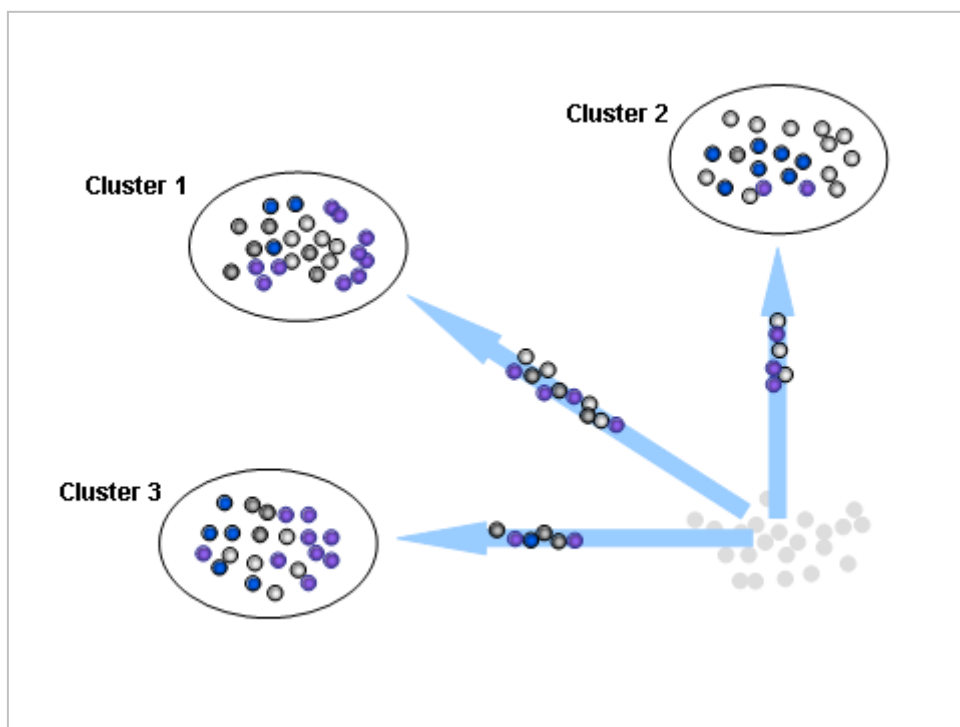
The best cluster solution is the one that provides the most insight and new information.

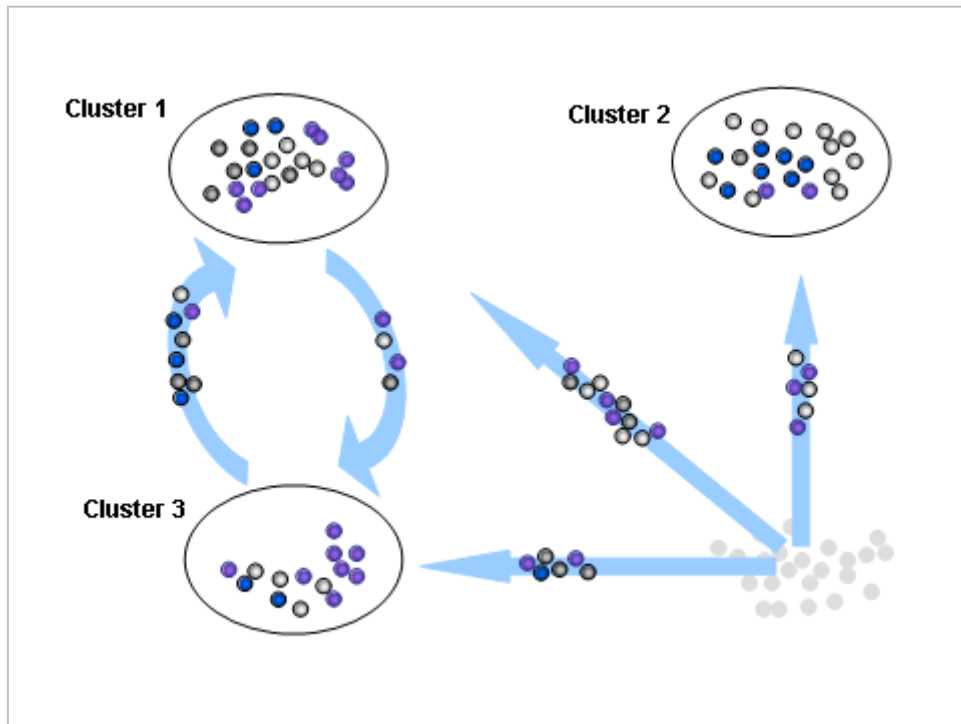
What happens within the analysis when you change the number of clusters?

When you increase or decrease the number of clusters you will usually find:

- They are largely re-arranged (forming very different clusters).
- One or two clusters split up / re-join leaving the other clusters relatively unchanged.

One cluster is re-distributed:





All / some clusters re-arrange themselves:

If you want to see how your clusters have split/merged you can save both cluster solutions and then cross-tabulate them against each other. This will show the movement of people between them.

Step four – applying a filter

You create the filter just like in any other tool in ASTEROID, and this then becomes the population examined in the Cluster Analysis.

Filter (Description)	Definition <small>▲</small>
ALL (no filter selected).	

Page

Step Five – Examine the output; is the cluster solution useful / useable?

In examining the output for a useable cluster solution, there are two things to consider:

- Do the clusters make sense? You should be able to see how the clusters are forming around the variables you've included and, as mentioned earlier, if you can put a name to your clusters, then you are on the way to a useable and hopefully useful cluster solution.
- Are the clusters 'strong'? It is important to remember you are using a statistical analysis tool and so the cluster solution must be statistically 'sound'.

Some cluster solutions will make sense but be statistically 'weak' and others may be statistically 'strong' but of no use to you – so you must always consider both aspects when looking at a cluster solution.

The first stage in determining the usefulness of a cluster solution is to examine the v% to see how the clusters have formed. If the nature of the clusters (all or some) isn't clear, then you need to change the variables included or alter the number of clusters. Before you decide on this, always look at the statistical measures at the bottom of the output.

The statistical measures provide two broad types of information: How important to the solution each variable is, and how effective ('strong') the solution is overall.

Each measure is described here in layman's terms and tries to indicate how to use them practically.

Goodman-Kruskal:

This measure indicates which variables are 'driving' the formation of the clusters – these will be the variables with the higher %. Variables that have a 0% are not contributing to the Cluster Analysis and you should consider removing them, however you might prefer to try altering the number of clusters before removing them, as their level of contribution may change.

There are no rules as to what is a 'good' % or a 'bad' %.

Dorofeev Contribution:

With a value between 0 and 1, this tells you whether the distribution of that variable in the current cluster solution is similar (low value) or different (high value) to its distribution in the population as a whole. You should consider removing the variable where the value is low but again, there isn't really a fixed 'low' / 'high' definition.

% Explained:

Practically, if the %s are low or negative for any cluster in the solution then you shouldn't use that solution – a change to the number of clusters or to the variables selected is probably required.

There is no rule as to what is a 'good' % but generally the higher the better.

Uniqueness:

Also a value between 0 and 1, this measure tells you the degree of difference between the cluster and the population. If some of the clusters have a low number then the cluster solution is probably not very useful.

Av. Difference:

Also a value between 0 and 1, this provides an indication how much each cluster differs from the other clusters. As Cluster Analysis tries to form clusters where the greatest possible distance exists between each cluster, the greater the average distance the 'better' the cluster solution from a statistical point of view.

Practically this is probably of most use where you have more than one cluster solution that looks viable and you need help in deciding between them - the one with the higher average distances is the 'better' of the two.

Difference between two closest clusters:

If the difference between the closest clusters is 'high' then obviously all the other distances will be high also, indicating it is a 'good' cluster solution. There is no rule as to what is high (again the number will be between 0 and 1).

Examples of the measures

On the following pages are examples of the statistical measures for two different cluster solutions. (Note that these are partial outputs and have been altered slightly for use in this manual.)

Example 1 shows a lot of difference between the contributions the variables are making to the clusters, varying from 4% to 27%, with one making a zero contribution. The total Goodman Kruscal score is 7.6%. The Dorofeev Contribution indicates that the distribution of the variables isn't differing much from the total population. The % explained varies from a very low 0.99% for cluster 2, up to 16.25% for cluster 4.

Even without looking at the Uniqueness and Av. Difference we would say that this cluster solution isn't particularly 'good' because the measures are low overall. In this case we could increase the number of clusters and see if that strengthened or weakened the figures.

In contrast, Example 2 has higher values across the measures. There is still one variable making a zero contribution but all the others fall between 10% and 18% with a total Goodman Kruscal of 12.9%. The % explained varies between the clusters but in this case the lowest score is 14% and the highest 39%. The other measures, while not stunningly high, are stronger than in our previous example.

The overall impression from the measures is that this cluster solution is 'good' and, assuming the clusters make sense to us as described earlier, we could be confident saving it and using it.

Example 1.

	Goodman Kruskal	Dorofeev Contribution			
I don't read the ads in newspapers and magazines	4.52%	0.10			
I can't help noticing advertising on buses	7.01%	0.13			
I enjoy buying magazines	0.00%	0.06			
I can't miss seeing those big billboard signs	14.29%	0.18			
I find TV advertising interesting	24.10%	0.22			
TV advertising often gives me something to talk about	19.67%	0.18			
Nearly all TV advertising annoys me	26.59%	0.23			
Some TV advertising is devious	2.47%	0.04			
Quite often I find TV advertising more entertaining than the programs	9.45%	0.09			
I often take advantage of the special offers on the back of my supermarket shopping docket	14.92%	0.19			
I usually notice the advertisements on shopping trolleys when I go grocery shopping	8.53%	0.09			
I often notice the advertisements on the tops and backs of taxis	11.07%	0.12			
Advertising posters in shopping centres and malls don't interest me	12.56%	0.14			
Total	7.63%	(Clust 1)	(Clust 2)	(Clust 3)	(Clust 4)
% explained (cluster)		6.88%	0.99%	10.16%	16.25%
Uniqueness	0.11	0.12	0.10	0.10	0.12
Av. Difference	0.18	0.19	0.18	0.16	0.18
Closest clusters are 3 and 4 with difference 0.14					

Example 2

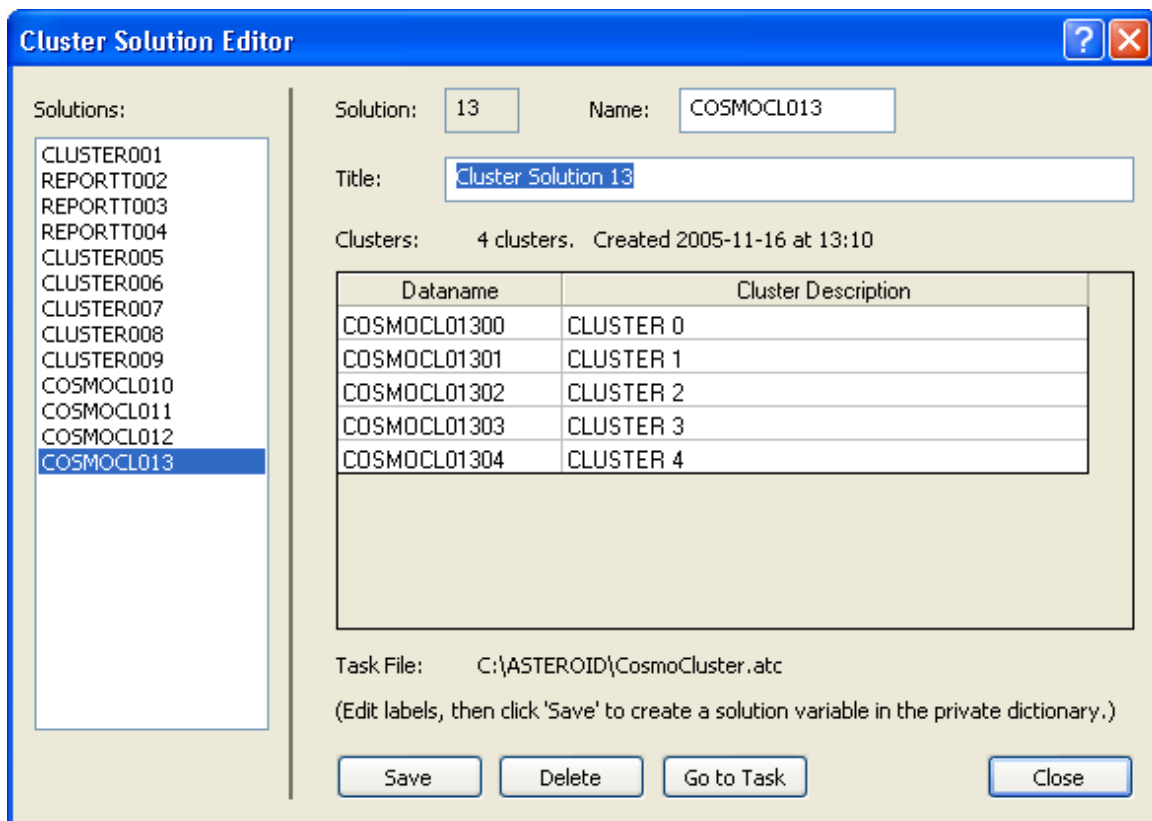
	Goodman Kruskal	Dorofeev Contribution				
I always think of the number of calories in the food I'm eating	14.70%	0.20				
A low fat diet is a way of life for me	17.55%	0.24				
Health food is not necessary if you eat properly	0.00%	0.01				
I try to buy additive free food	10.63%	0.28				
I love to do as many sports as possible	10.91%	0.13				
I'm eating less red meat these days	13.09%	0.26				
I would like to be able to lose weight	17.36%	0.26				
I'm concerned about my cholesterol level	13.86%	0.35				
I try to get enough calcium in my diet	18.03%	0.22				
Total	12.90%	(Clust 1)	(Clust 2)	(Clust 3)	(Clust 4)	(Clust 5)
% explained (cluster)		14.02%	21.86%	39.37%	30.31%	33.04%
Uniqueness	0.21	0.31	0.18	0.21	0.19	0.18
Av. Difference	0.32	0.40	0.28	0.33	0.31	0.30
Closest clusters are 3 and 4 with difference 0.21						

Step six – Re-run or save?

Decide whether you need to alter the number of clusters and/or remove the low contributors (remembering that a variable making a low contribution on a 2 cluster solution may make a higher contribution on a 3 or 4 cluster solution).

When you have a useable cluster solution, then you can save it into your Private Dictionary, through the button on the toolbar.

The Cluster Solution Editor dialog box lists all the cluster solutions run in the current database. It lists the clusters for the highlighted solution, and includes a cluster 0 which represents the people not included in the cluster solution.



The Cluster Solution Editor dialog box is shown. It has a title bar with a question mark and a close button. The dialog is divided into two main sections. On the left, under the heading 'Solutions:', there is a list box containing the following items: CLUSTER001, REPORTT002, REPORTT003, REPORTT004, CLUSTER005, CLUSTER006, CLUSTER007, CLUSTER008, CLUSTER009, COSMOCL010, COSMOCL011, COSMOCL012, and COSMOCL013. COSMOCL013 is selected and highlighted in blue. On the right, there are several fields and a table. The 'Solution:' field contains the number '13'. The 'Name:' field contains 'COSMOCL013'. The 'Title:' field contains 'Cluster Solution 13'. Below these, it says 'Clusters: 4 clusters. Created 2005-11-16 at 13:10'. Below this is a table with two columns: 'Dataname' and 'Cluster Description'. The table contains five rows: COSMOCL01300 (CLUSTER 0), COSMOCL01301 (CLUSTER 1), COSMOCL01302 (CLUSTER 2), COSMOCL01303 (CLUSTER 3), and COSMOCL01304 (CLUSTER 4). Below the table, the 'Task File:' is 'C:\ASTEROID\CosmoCluster.atc'. Below that is a note: '(Edit labels, then click 'Save' to create a solution variable in the private dictionary.)'. At the bottom, there are four buttons: 'Save', 'Delete', 'Go to Task', and 'Close'.

Dataname	Cluster Description
COSMOCL01300	CLUSTER 0
COSMOCL01301	CLUSTER 1
COSMOCL01302	CLUSTER 2
COSMOCL01303	CLUSTER 3
COSMOCL01304	CLUSTER 4

The Title acts as the name of the cluster ‘variable’ – it appears in the Private Dictionary as a variable with the clusters forming the groups. You should give it a meaningful Title and give each cluster a meaningful name before clicking the Save button.

Use the Close button to exit the Cluster Solution Editor when you’re done.

Step Seven – Explore the clusters further.

Once you have saved a cluster solution you can cross-tabulate or Profile it against demographics or other data to further develop your understanding of those clusters.

Note: The Cluster Analysis is based on the population of the current database (entire or filtered portion) and so the results of a Cluster Analysis will not apply to other databases.

