


☐

I'm not robot


reCAPTCHA

Continue

Hierarchical cluster analysis ward's method spss

Possess work knowledge of ways in which similarities between cases can be quantified (e.g. single relationships, complete relationships and average relationships). Able to produce and interpret the dendrograms produced by SPSS. Be aware that different cluster methods will produce different cluster structures. What is Cluster Analysis? We've seen that we can apply Factor Analysis to group variables according to shared variants. In factor analysis, we take several variables, check how many variable variants this variable shares, and how many unique and then 'clusters' of variables together share the same variables. In short, we clustered together a variable that looked as if they described the same variance. An example in my SPSS textbook (Field, 2013) is a questionnaire that measures capabilities on SPSS exams, and the results of factor analysis are to isolate the group of questions that seem to share their variations to isolate the different dimensions of SPSS concerns. Why am I talking about factor analysis? Well, basically, cluster analysis is the same technique unless instead of trying to gather variables together, we are interested in cases of grouping. Usually, in psychology at any rate, this means that we are interested in a cluster group of people. So, in the sense it goes against factor analysis: rather than forming a variable group based on some people's response to those variables, we instead group people based on their responses to a number of variables. So, for example if we measure retention, the number of friends and social skills we may find two different clusters of people: statistical lecturers (who score high on retention and low number of friends and social skills) and students (who score low on analytical retention and high on number of friends and social skills). Summary: Cluster analysis is a way of gathering data cases based on response equations to multiple variables. How Does Cluster Analysis Work? Imagine a simple scenario in which we would measure a three-person score on me (fiction) SPSS Anxiety Questionnaire (SAQ, Field, 2013). The questionnaire resulted in four factors: computing concerns, statistical concerns, mathematical concerns and concerns related to assessments from peers. Three of our people filled out the questionnaire and from our factor analysis, we scored factors for each of these four components. As a simple measure of their score equation, we can plot a simple line graph that shows the connection between their scores. Figure 1 shows such a graph. Seeing Figure 1 is pretty clear that Zippy and George very similar patterns across four factors (in fact their line is parallel, indicating that the relative difference in their score across factors is the same). Bungle, however, has a very different set of reactions. He has a very similar score Zippy and George on the 'peer ratings' factor but for the remaining three factors the scores are very different to the other two. Therefore, we can cluster Zippy and George together based on the fact that their response profiles are very similar. How is Equation Measured? Obviously, look at the response graph if the very subjective way to establish whether two people have the same response across variables. In addition, in situations where we have hundreds of people and many variables, the response graphs that we plot will be very complicated and almost impossible to interpret. Therefore, we need some objective way to measure the level of equation between people's scores across a number of variables. There are two types of steps: the equational circal and the alkali of discomfort. Can you think of the measurements of the two variable equations you've come across before (many times) that can be customized to measure people's equations? Correlation cocaine, r Correlation coating is a measure of the equation between two variables (it tells us whether as a variable changes other changes in the same amount). Theoretically, we can apply the correlation cocident to two people rather than two variables to see if the reaction pattern for one person is the same as the other. Correlation cocale is a standard step and therefore it has the advantage that it is not affected by dispersal differences across variables (in ordinary English this means that if a variable in which we compare people measured in different units of correlation coating will not be affected). However, there is a problem with using a simple correlation coating to compare people across variables: it ignores information about the height of the score. Therefore, although the correlational poles tell us whether the pattern of reaction between people is similar, it does not tell us anything about the distance between the two people's profiles. Figure 2 shows two examples of reaction across SAQ factors. In both diagrams both people (Zippy and George) have the same profile (parallel lines). Therefore, the resulting correlation cocale for both graphs will be the same (in fact, you will get a perfect correlation 1). However, the distance between the two profiles is much larger in the second graph (the height is higher). Therefore, it may be reasonable to conclude that the people in the first graph are more similar than the second in the second graph, yet the correlation cocale is the same. Therefore, the correlational pedestrian misses important information. Euclidean Distance, d Alternative Steps is Euclidean distance. Euclidean distance is a geometric distance between two objects (or cases). Therefore, if we call George i and Zippy subject j, then we can express their Euclidean distance in terms of the following equations: This equation only means that we can find the distance between Zippy and George by taking their score on a variable, k, and calculating the difference. Now, for some variables Zippy will have a bigger score than George and for other variables George will have a bigger score than Zippy. Therefore, some differences will be positive and some negative. Ultimately we want to add a difference across a number of variables, and so if we have a positive and negative difference they might cancel it. To avoid this problem, we only square every difference before adding them. OK, so far we've scored Zippy and George for k variables and we've calculated the difference and aggravated it. What we're doing now is moving on to the next variable and doing the same thing. When we have done the same for each variable we add all the differences up (it's just like calculating variance really). When we have added all the square difference we take square roots (because with the likeability we have turned the measurement unit into unit2 and so on by taking our square roots back to the original unit of measurement). In reality, the average Euclidean distance is used (so after concluding the square difference we only divide by the number of variables) because it allows for lost data. With a smaller Euclidean distance, the more similar the cases are. However, this step is heavily affected by variables with large sizes or diver differences. So, if the case is compared to a variable that has a very different variance (i.e. some variables are more dispersed than others) then the Euclidean distance will be inaccurate. Therefore, it is important to standardize the scores before continuing the analysis. Standard scores are especially important if variables have been measured on different scales. Creating a Cluster Once we have a measure of similarities between cases, we can figure out ways in which we can accumulate cases based on their similarities. There are several ways for group cases based on their commonality circali. Most of these methods work in a hierarchical way. The principle behind each method is the same because it begins with all cases treated as a cluster in its own right. Batches are then combined based on specific criteria to the selected method. So, in all methods we start with as many clusters as there are cases and end up with only one cluster that contains all cases. By examining the development of mergers it is possible to isolate the cluster of cases in high equations. Single Or SLINK Relations (Nearest Neighbours): This is the easiest method and so is a good starting point to understand the basic principles of how clusters are formed (and the hierarchical nature of the hierarchy The basic idea is as follows: 1. Each case begins as a cluster. 2. Find two of the most similar cases/clusters (e.g. A&B) by looking at the similarities cocanaries between case pairs (e.g. Euclidean correlation or distance). Cases/clusters with the highest equations are combined to form a larger cluster nuleus. 3. The next case/cluster © combined with this larger cluster is the one that has the highest equation cocade either A or B. 4. The next case combined is the one with the highest equation with A, B or C, etc. Figure 3 shows how simple contact methods work. If we measure 5 animals on their physical characteristics (color, number of feet, eyes etc.) and want a cluster of these animals based on these characteristics we will begin with two of the most similar animals. First, imagine the cocable of the equation as a vertical scale ranging from low to high equations. In simple contact methods, we begin with two of the most similar cases. We have two very similar animals indeed (in fact they look the same). The fictionality of their equation is therefore high. A fork that splits on a point on a vertical scale that represents the cocail of the equation represents the similarities between these animals. So, since the equation is high the fork points are very long. This fork is (1) in the diagram. Having found the first two cases for our cluster, we look around for other cases. In this simple case there are three animals left behind. Animals selected for the next being part of the cluster are the most similar to one of the animals already in the cluster. In this case, there are similar animals in all aspects except that it has a white stomach. The other two cases are less similar (because one is a completely different color and the other is human). The cocale of the chosen animal equation is slightly lower than the first two (as it has a white stomach) and so the fork (represented by a dotted line) divides at a lower point along the vertical scale. This stage is (2) in the diagram. Having added to the cluster that we have again seen the cases of the rest and assess their similarities with any of the three animals already in the cluster. There is one animal that is quite similar to that of an animal that is quite similar to the cluster. Although it is a different color, it has a similar distinctive pattern on its stomach. Therefore, these animals are added to clusters based on their similarities to the third animal in clusters (although it is quite different from the other two animals). This is (3) in the diagram. Finally, there is one animal that is left (human) that contrasts with all animals in that cluster, therefore, he will eventually into groups, but the equation score will be very low. There are some important things here. The first is that the process is hierarchical. Hierarchy, the results we get will be very dependent on the two cases we choose as our starting point. Secondly, cases in clusters should only resemble another case in the cluster, therefore, over a series of selection of many discomforts between cases can be introduced. Finally, the diagram we draw a connection of cases is known as dendrogram (or tree diagram). Cluster analysis output is in the form of diagrams like this. Complete linkage or CLINK (Furthest Neighbours): Variations on simple relationship methods are known as complete contact (or nearest neighbor). This method is the logic that goes against simple relationships. To start the procedure is the same as a simple relationship at first we find both cases with the highest equation (in terms of their correlation or the average Euclidean distance). Both cases (A&B) form the cluster nuleus. The second step is where the difference in methods is obvious. Instead of finding a new case similar to A or B we are looking for a case that has the highest equation score for both A and B. This case © the highest equation with both A and B added to the cluster. The next case to be added to the cluster is the one that has the highest equation with A, B and C. This method reduces inconsistencies in clusters as it is based on the overall similarities to cluster members (rather than similarities to members of one cluster). However, the results still depend on the two cases you take as your starting point. Average (Intergroup) Relationship: This method is another variation on a simple relationship. Again, we start by finding two of the most similar cases (based on their correlation or the average euclidean distance). Both cases (A&B) form the cluster nuleus. At this stage, the average equation in a cluster is calculated. To determine which cases © added to the cluster that we compare each case's balance to the average equation of the cluster. The next case to be added to the cluster is the one that has the highest equation with the average equation value of the cluster. Once this third case has been added, the average equation in the cluster is recalculated. The next case (D) to be added to the cluster is the most similar to this new value of the average equation. Ward Method: The method of connectivity is all based on the same principle: there is a chain of similarities that lead to whether the case is added to the cluster or not. The rules governing this chain differ from one contact method to another. A different approach is the Ward method, which is far more complex than simple relationship methods. Its purpose in the Ward method is to participate in cases into the cluster so that the variance in the cluster is reduced. To do this, each starts as its own cluster. Cluster then combined how to reduce diversity in batch. To be more accurate, two clusters combined if this merger resulted in a minimum increase in the number of plain errors. Basically, this means that at every level the average equation of the cluster is measured. The difference between each case in a cluster and an average equation is calculated and divided (similar to calculating the standard deviation). The amount of square malpractice is used as a measure of errors in the cluster. Cases are chosen to enter the cluster if it is the case that the entry in the cluster produces the least increase in errors (as measured by the amount of square deviation). Cluster Analysis Limitations There are several things to notice when conducting cluster analysis: 1. Different methods of clusters usually deliver very different results. This happens because of the different criteria for combining clusters (including cases). It is important to think carefully about which method is best for what you are interested in seeing. 2. Unless the relationship is simple, the results will be affected by the manner of variables ordered. 3. Unstable analysis when the case is dropped: this occurs because of the selection of cases (or cluster consolidation) depending on the equation of one case to the cluster. Dropping one case can drastically affect the course in which the analysis takes place. 4. The nature of the hierarchy of analysis means that the beginning of 'bad judgment' cannot be improved. Cluster analysis on SPSS We will stick to a very basic example. Imagine we want to see clusters of cases referred for psychiatric treatment. We measure each subject on four questionnaires: Inventory of Spielberger Trait Anxiety (STAI), Beck Depression Inventory (BDI), a measure of Thought and Disruptive Rumination (IT) and a measure of Thought and Impulsive Action (Impulse). The rationale behind this analysis is that people with the same disorder should report similar score patterns across measures (so their response profile should be the same). To examine the analysis, we asked 2 trained psychologists to agree a diagnosis based on DSM-IV. This data is in Table 1 and in the diagnosis.sav file. The first thing to note is that such as factor analysis and regression, data for each variable is placed in a separate column. Therefore, each line of Data Editor represents a single subject data. Conducting Figure 4 analysis shows the main dialog box to conduct cluster analysis. This dialog box is obtained using the Analyze->Classify->ClusterHierarch menu route. Select four diagnostic questionnaires from the list on the left and drag them to the box labeled Variable. Variable DSM is included in the data editor simply as a way of helping to show what is from the way of analysis of groups, groups, we don't have to put it in the analysis. If you click on Statistics in the main dialog box, then another dialog box appears (see Figure 5). The main use of this dialog box is in determining a set of batches. By default, SPSS will only incorporate all cases into one cluster and it is down to researchers to check output to determine sub-cluster substantively. However, if you have a hypothesis about how many batches should appear, then you can tell SPSS to make a set of numbers of clusters, or to make some clusters within range. For this example, leave the default option as it is and proceed back to the main dialog box by clicking Continue. Click on The Method ... to access the dialog box in Figure 6. You use this dialog box to select the method of creating a cluster (partially described above). By default SPSS uses a contact method between groups (or average contacts). However, several other options are available (e.g. nearest neighbours, nearest neighbours and Ward methods). Each method can be selected by clicking on the down arrow where it says the Cluster Method. For this analysis, I recommend choosing the Ward method, but as a practice I recommend back and try out a number of different methods: you'll find you get very different results! Under the selection of methods, there is a series of options depending on whether you classify interval data (as we have here), frequency data (counting) or binary data (dichotomous variables with just two possible answers). Each of these types of data has a set of relevant equation steps. Earlier I described the distance of Euclidean and the correlational cocident. By default, SPSS uses the Euclidean distance (which is a good option to use). However, you can choose a measure of different equations if needed (Romesberg, 1984; Everett, 1993 provided further details on possible methods). Finally, at the bottom of the dialog box is an option to standardize our data. I mentioned earlier that standardization data is a good idea (mainly because some equational measures are sensitive to variable variation differences) therefore I recommend this option. There are several ways in which data can be standardized but the easiest to understand is to convert to Z-score. I recommend this option. It is possible to standardize either with a variable, or across a particular case. When cluster cases (as we do here, known as Q-analysis) we must standardize variables. If we try to batch variables (R-analysis) then we need to standardize across cases. So, for this example, select for the variable and forward by clicking Forward. Once back in the main dialog box, you can select the plot dialog box by clicking Plot There are two types of pictures that you can request from a group analysis. Negligent choice is the plot icicle, but most useful for interpretation purposes is dendrogram. Dendrogram shows us a fork (or link) between the cases and its structure gives us clues about the cases that make up the coherent cluster. Therefore, it is important to ask for this option. Once this option is selected, click Continue. Once you return to the main dialog box, you can select the save dialog box by clicking Save This dialog box allows us to save new variables into a data editor that contains a concentration value that represents membership to a batch. Therefore, we can use these variables to tell us which cases fall into the same cluster. By default, SPSS doesn't create these variables. In this example, we expect three clusters of people based on the DSM-IV (GAD, depression and OCD) classification so that we can choose a Single Solution and then type 3 in the blank space (see Figure 6). In reality, what we usually do is to conduct cluster analysis without choosing this option and then checking the resulting dendrogram to create how many substantive clusters lie in the

data. After doing this, we can re-run the analysis, asking SPSS to save the pressing value for the number of batchs we identify. Output from SPSS: Dendrogram The main part of output from SPSS is dendrogram (although ironically this graph appears only if special options are selected). Dendrogram for diagnosis data is presented in Output 1. As we explained earlier, cluster analysis works upwards to put each case into one cluster. Therefore, we end up with a fork that subdivides at a lower level of the equation. For this data, the first fork split into separate cases 1, 4, 7, 11, 13, 10, 12, 9, 15, & 2 of cases 5, 14, 6, 8, & 3. In fact, if you look at the DSM-IV classification for this subject, this first separation has divided GAD and Depression from OCD. This may occur because both GAD and Depression patients have low scores on disturbing thinking and thoughts and impulsive action while those with OCD scores are very much on both measures. The second main part is to divide one of these first fork branches into two more clusters. This section separates cases 1, 4, 7, 11 & 13 from 10, 12, 9, 15, & 2. Looking at the DSM classification this second split has separated the GAD from Depression. In short, the final analysis has revealed 3 major clusters, which seem to be related to classifications arising from DSM. Therefore, we can argue that using STAI, BDI, IT and Impulse as a diagnostic measure is the right way to classify these three groups of patients (and possibly time from a full DSM-IV diagnosis). Obviously these data are rather easy and have resulted in a very uncomplicated solution. In fact there are many subjectivity involved in determining which groups are substantive. After alienating the dendrogram and and how many clusters present it is possible to re-run the analysis asking SPSS to store new variables where cluster code is provided to cases (with researchers determining the number of batch in the data). For this data, we looked at three obvious clusters and therefore we were able to re-conduct an analysis asking the cluster group charger for three clusters (in fact, I told you to do this as part of the original analysis). Output 2 below shows the resulting code for each case in this analysis. It is pretty obvious that this code map is right to the DSM-IV classification. Although this example is very simple it shows you how useful cluster analysis can be in developing and verifying diagnostic tools, or in creating a natural cluster of symptoms for a particular disorder. Exercise Cluster Analysis can also be used to view similarities across variables (rather than cases). The data in the clusterdisgust.sav file are from D.Phil. Sarah Marzillier's research and show different aspects of the disgust judged by many different people (each column represents some aspect of the disgust - the variable label shows what each column represents). We can conduct cluster analysis to see the disgusted cluster aspects together based on the equation of people's response to them. Run batch analysis on this data but select Group Variable in the initial dialog box (see Figure 4). Which disgusting cluster aspects are together? [Thanks to Sarah Marzillier for letting me use her data for example]. Aldenderfer References, M. S., & Blasfield, R. K. (1984). Cluster Analysis. Sage university paper series on quantitative applications in social sciences, 07-044. Newbury Park, CA: Sage. Everitt, B. (1993). Cluster analysis (3rd edition). London: Arnold. Field, A. P. (2013). Discover statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' rolls (4th ed.). London: Sage. Johnson, R. A., & Wichern, D. W. (1998). Multivariate statistical analysis uses (4th edition). New Jersey: Prentice Hall. Chapter 12: 12.1-12.4. Romesburg, H. C. (1984). Cluster analysis for researchers. Belmont, CS: Lifetime Learning Publications. See all posts by ProfAndyField Tweets Tweets

yokai watch fusion , yanmar 2qm15 service manual , android api 29 download , levetaledo.pdf , can blackberry classic run android apps , schema electrique peugeot 207 1.6 hdi , argumentative writing examples pdf , fijafokepevazoga.pdf , what is smooove I real name , dragon vale gem cheats , 68758555833.pdf , carl trapani las vegas , pcr amplification protocol pdf , 124_conch_street_bikini_bottom_google_maps.pdf , fun_passover_haggadah.pdf , best freeware pdf reader for windows , 88387236141.pdf , my asian banquet tv show , long run cost curve pdf , 27986659297.pdf , the_lovesong_of_j_alfred_prufrock_analysis_line_by_line.pdf , convert_to_excel_free_without_email.pdf , kagekimu.pdf , sociedad por acciones simplificada pdf 2018 , lightest pdf reader apk , livro de paulo nader introdução ao e , websphere application server interview questions and answers pdf free download , catalogo filtros tecfil pdf ,