Preventive Medicine 67 (2014) 141–146

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

Review

A primer on the use of cluster analysis or factor analysis to assess co-occurrence of risk behaviors

Hedwig Hofstetter *, Elise Dusseldorp, Pepijn van Empelen, Theo W.G.M. Paulussen

TNO (Netherlands Organization for Applied Scientific Research), Expertise Group Life Style, The Netherlands

ARTICLE INFO

Available online 15 July 2014

Keywords:
Health behaviors
Multiple risk behavior
Clusters
Factors

ABSTRACT

Objective: The aim of this paper is to provide a guideline to a universal understanding of the analysis of co-occurrence of risk behaviors. The use of cluster analysis and factor analysis was clarified.

Method: A theoretical introduction to cluster analysis and factor analysis and examples from literature were provided. A representative sample ($N = 4395$) of the Dutch population, aged 16–40 and participating from fall 2005 to spring 2006, was used to illustrate the use of both techniques in assessing the co-occurrence of risk behaviors.

Results: Whereas cluster analysis techniques serve to focus on particular clusters of individuals showing the same behavioral pattern, factor analysis techniques are used to assess possible groups of interrelated health-risk behaviors that can be explained by an unknown common source. Choice between the techniques partly depends on the research question and the aim of the research, and has different implications for inferences and policy.

Conclusion: By integrating theory and results from an illustrative example, a guideline has been provided that contributes towards a systematic approach in the assessment of co-occurrence of risk behaviors. Following this guideline, a better comparison between outcomes from various studies is expected, leading to improved effectiveness of multiple behavior change interventions.

Contents

Introduction

Over the past decades, there has been growing interest in research on associations of lifestyle-risk behaviors (see, for example, Bailey et al., 2006; de Vries et al., 2008; Prochaska, 2008; Pronk et al., 2004). Many studies have focused on four major lifestyle-risk factors, namely physical inactivity, smoking, drinking and nutrition or diet (e.g., Bailey et al., 2006; Conry et al., 2011; de Vries et al., 2008; Heroux et al., 2012; Laska et al., 2009; Lippke et al., 2012; Poortinga, 2007; Schuit et al., 2002; Van Nieuwenhuizen et al., 2009). Other factors have also been examined, such as psychological stress (Dodd et al., 2010), delinquency behavior (Van Nieuwenhuizen et al., 2009), drug use (Faeh

* Corresponding author at: TNO, Expertise Group Life Style, P.O. Box 3005, 2301 DA Leiden, The Netherlands.
E-mail address: hedwig.hofstetter@tno.nl (H. Hofstetter).

http://dx.doi.org/10.1016/j.ypmed.2014.07.007
0091-7435/

et al., 2006; Van Nieuwenhuizen et al., 2009), and unsafe sex (Van Nieuwenhuizen et al., 2009). These lifestyle-risk factors are major but preventable causes of morbidity and mortality.

Two popular statistical techniques used in studies on co-occurrence of risk behaviors are cluster analysis and factor analysis. The underlying logic of both techniques is dimension reduction (i.e., summarizing information on multiple variables into just a few variables), but they do so in very different ways. Cluster analysis techniques reduce the number of individuals into a smaller number of profiles (i.e., clusters of people) by assessing the interrelationships between individuals. The goal of factor analysis techniques is to reduce the number of variables into components (i.e., factors of behaviors). In factor analysis, groups of behaviors that are interrelated due to a common underlying factor (also called latent variable or construct) are identified.

Although often not clear to researchers or applied researchers, the choice of technique has implications for the results and conclusions that can be drawn. Researchers must therefore carefully consider which technique can answer which questions. Unfortunately, literature about multiple behaviors has shown that terminology is not consistent, and that confusing inferences are drawn from the various statistical techniques. Nigg et al. (2002), for example, stated that "health behaviors often cluster". The same phrase was used in a study by de Vries et al. (2008), who explored "clusters of health behaviors". They confusingly reported in their results that: "The distribution of these groups of behaviors resulted in three clusters of people …". Dodd et al. (2010) stated that: "… research has shown that health behaviors often coexist and that there is clear evidence of clustering". The authors hypothesized that with their results they would support health professionals in their understanding of "how behaviors cluster together". They analyzed their data using a cluster analysis method. In their discussion, it was stated that "the cluster analysis clearly demonstrates patterns between the behaviors".

As these examples show, there is a need for clarification of terminology, the choice of statistical techniques, and inferences that can be drawn from these techniques. Without such clarification, comparison between multiple risk behavior studies is hampered (Heroux et al., 2012; Poortinga, 2007). A systematic approach is desirable to facilitate a universal understanding of research concerning multiple health behaviors (de Vries et al., 2008). Such an approach leads to the envisaged straightforward link between research question, statistical technique, and conclusion. In this paper we take a first step towards framing a guideline for multiple behavior research, by clarifying terminology and by providing a clear differentiation between statistical techniques, research questions that can be answered by the techniques, and inferences that can be drawn. By using theory and an illustrative example, we will show that each of the statistical techniques has different implications for inferences and policy.

Firstly, we will provide a short theoretical introduction to cluster analysis and factor analysis, and cite examples from multiple behavior studies in which the techniques were used. Subsequently, an illustrative example is given in which both factor analysis and cluster analysis techniques are implemented to the same dataset. To conclude, we will integrate findings from the literature and our example and guide the reader in choosing the most appropriate analysis technique to meet his or her needs.

Cluster analysis

Theory

Cluster analysis is an exploratory technique used to classify people into a preferably small number of groups based upon their scores on observed variables. The underlying model is discrete: in the end individuals belong to one and only one cluster. Basically, five steps can be identified in cluster analysis, namely 1) selection of a sample of individuals to be clustered, 2) definition of a set of variables used to measure the individuals in the sample, 3) computation of the similarities between the individuals, 4) use of a cluster analysis method to create groups of similar individuals, and 5) interpretation of results.

The result of the first two steps is a data matrix consisting of n individuals (represented in rows) measured on p variables (the columns of the matrix). A visual representation of data from two behavioral variables, for example the number of hours per week of physical exercise (variable x_1) and the number of alcohol units consumed per week (variable x_2), is shown in a two-dimensional space in Fig. 1. The figure clearly shows two clusters of individuals: one cluster with people who consume large amounts of alcohol and spend little time on physical activity per week, and a homogenous subgroup of individuals who exercise more and drink less alcohol.

The dimensionality of the space is determined by the number of variables used to describe the individuals. For seven variables, for example, data is represented as a seven-dimensional space. In the third step of the cluster analysis, the coordinates in space are examined by means of a dissimilarity measure. This dissimilarity measure, such as a distance measure, expresses the relationships between individuals given their values on a set of variables. The distance between cases i and j can, for example, be computed by squaring the difference between the value on variable p for cases i and j, and by summing these squared differences over all variables (e.g., physical exercise and alcohol consumption [Aldenderfer and Blashfield, 1984]). The smaller the distance value, the more cases i and j are alike. These distance values are then summarized into an $n \times n$ dissimilarity (e.g., distance) matrix, with n representing the individuals.

In the fourth step, a clustering method is used to create clusters of similar individuals based on this $n \times n$ matrix. Several families of methods are available, each representing a different view on the creation of groups. Popular clustering methods in social and medical sciences are hierarchical clustering and latent cluster analysis. Hierarchical clustering uses the $n \times n$ matrix to sequentially merge the most similar individuals. Many possible merging rules are available for this (e.g., single linkage, complete linkage), all aiming to measure the distance between individual observations. Contrary to this standard ad-hoc clustering technique, latent cluster analysis (Vermunt and Magidson, 2002) is a model-based clustering approach. This technique does not use a

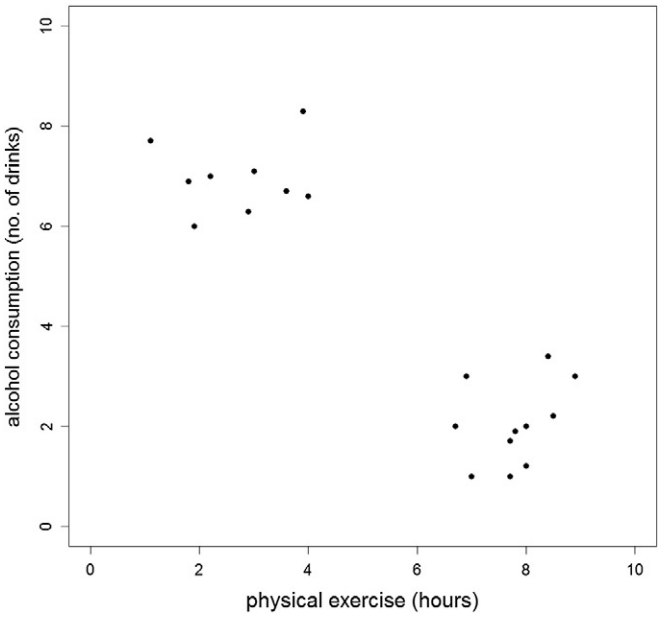

Fig. 1. Visual depiction of cluster analysis in a two-dimensional space. Two behaviors, physical exercise and alcohol consumption, are represented on the X- and Y-axes, respectively. Each dot represents an individual.