Košťál, Jaroslav

Vybrané metody vícerozměrné statistiky (se zvláštním zaměřením na kriminologický výzkum)
Selected multivariare statistics methods (with a special focus on criminological research)

ISBN: 978-80-7338-128-8

Summary

Multivariate statistics have only recently started to be used in Czech criminology while Western criminology has used them for two or even four decades. This is not just an issue of prestige as "multivariate" is in; this is also a matter of being right since one- or two-dimensional statistics (still too much relied upon) may lead to incorrect conclusions. They can, for instance, be the source of incorrectly accepted or rejected hypotheses about relationships between variables and therefore can lead to such conclusions and predictions of criminal phenomena that are not supported by real evidence.

This treatise consists of 5 parts. The first is devoted to a general characterization of multivariate statistics, the second part focuses on factor analysis, the third on cluster and correspondence analyses and the last two describe multidimensional scaling methods.

Multivariate approach is applied when we concurrently search for deeper relationships among three or more variables. Using multivariate rather than bivariate analyses generally allows to identity and describe phenomena in a more precise and at the same time more complex and reliable manner. This complexity also increases confidence on which we can base our judgment, e.g., findings with regard to criminal behaviour across time, or evaluations of anger management training in aggressive offenders.

Multivariate statistics is based on a) test techniques, then not always b) on comparative random distributions against which the significance of estimated parameters (means, counts, variances etc.) are tested and c) the concept of distances between the observed phenomena in (extrapolated) space. This text is limited to description of factor, cluster and correspondence analysis and to a certain degree also to multidimensional scaling, marginal mention of decision trees and the use of structural equation modeling. In addition, we demonstrate

application of these techniques to various data examples of Institute of Criminology (ICSP) projects and discuss possible pitfalls.

Factor analysis, addressed in the second part, is a statistical method which reduces the number of observed or manifest variables to a smaller number of latent (unobserved, construed) factors. These factors explain the variance of measured (manifest) variables. If multiplied and added, i.e., by the sum of their linear combinations, these factors equal the measured variables plus error terms (measurement error and the unexplained part of variance –the error of the model).

FA was invented by Spearman in the form of principal component analysis 100 years ago. Significant development of the method took place in the 1960s and it has since been applied across many fields of natural and social sciences. The point of departure of explorative FA is the correlation or covariance matrix of a set of the measured variables. This is the basis for statistical decision whether and which of these manifest variables are closely related to each other and therefore belong together or whether they belong to a different common factor.

Example 2 describes a survey focused on the opinion of Czech adults about the authorities involved in criminal proceedings. As expected, the factor analysis of responses to 19 proceeding items produced 4 factors: evaluations of work performed by the police, the courts, the state prosecution and by prison officers. The preliminary steps of factor analysis depend on correlation and partial correlation matrices. Examples are used to demonstrate FA requirements on input data, and desirable transformation of input correlation matrices into tetrachoric coefficients.

Following data transformation, factor analysis was executed for Example 3 with a two-factor solution and principal components (factors) dividing activities to socially or individual harmful categories. Strategies for estimating of optimal number of factors (principal components) are discussed with a stress on Horn´s parallel analysis which is an optimal yet rarely applied option. Shortcomings of other popular choices, such as Kaiser criterion and Pearson correlation matrix are discussed.

The difference between the principal component extraction method and the factor analysis method is explained. Example 5 focuses on the degree of consensus among young people in

the Czech Republic with regard tomoral principles (codes of behaviour). Data of this type are not suitable for factor analysis treatment by the maximum likelihood extraction, despite the fact that the input matrix has been adapted and meets the requirements of continuous metric data. For that reason, data are analysed using an alternative acceptable method of principal axis factoring. The scales corresponding to three individual factors are also submitted to Cronbach's Alpha reliability analysis.

Example 6 deals with young people's opinion on the conduct of crime witnesses. It demonstrates factor analysis performed by maximum likelihood estimate and its advantages. for this particular case and notes on controversy which relates to the use of factor analysis versus principle component analysis. Some scholars strongly recommend oblique rotation and the maximum likelihood extraction method as the results are not dependent on the peculiarities of the concrete data set. We stand behind this recommendation and also suggest to plan for the statistical analyses at the time of the construction of research methods (in the questionnaire, expert scale).

The last segment of the factor analysis chapter highlights structural equation modeling (SEM), which is one of the most multifaceted and flexible methods of analysis of the interrelated variables. SEM is most frequently used in so-called confirmatory factor analysis, typically applied to test a current theory or structure revealed by exploratory factor analysis. Example 6 (Young people's opinion on the conduct of crime witnesses) a result of exploratory factor analysis, is submitted to CFA. The model derived from preceding FAs with was entered to the analysis proved itself fully compatible with the data and confirmed that respondents' views on witness behaviour depend on the context, i.e., whether it is the perperrator or the victim who is in the focus.

Cluster analysis (CA) is a subject of the third part. In example 8 (Opinion of young people in the Czech Republic on how much problematic certain negative phenomena are), we explain the "correlation profile" and the essentials of CA. In example 9 (Rating the degree of severity of negative social phenomena by the Czech adults), we illustrate the necessity to identify the clusters'content and subsequently to characterise them as actually existing groups of people with distinct inclinations and behaviour. Hierarchical cluster analysis is described by example 10 (Analysis of crime rates over the past decades). CA is suitable for metric data sets that are small and easy to keep in mind. Some periods of criminal statistics may gradually

(hierarchically) merge into units with similar pattern. Icicle charts and a dendrograms may identify and explain their change and common features.

K-means is a method of cluster analysis used for large samples with at least ordinal, and ideally metric data. Example 11 (European Values Study and justifiability of behaviour) shows how clusters emerge during gradual analysis (iteration steps) and how properly name them. Importantly, the number of clusters is decided by the researcher himself/herself and he/she must carefully consider and select those which are stable, balanced in number and which provide the best explanation of the subject matter. The suitability of clusters for international comparisons is discussed (particularly the same understanding of test batteries). It is highlighted, that the same configuration and metrics (invariance) of the battery across various countries may be tested by SEM.

Well identified clusters tend to meaningfully correlate with other research variables. Example 8 (Opinion of young people in the Czech Republic on how much problematic certain negative phenomena are) divided respondents into groups who voiced problems are serious, semi-serious and those who minimized their importance. These groups significantly differed by their criminal sensitivity indicated by 43 items. These groups also differed by their anamnestic data (e.g., grades of conduct at school).

The two-step cluster analysis (SPSS) provides a compromise between hierarchical analysis and k-means. Its particularly suitable for work with large samples and can process both categorical and metric data. Example 4 (Czech adults' experience with psychotropic substances) demonstrates a three-cluster solution where non-users were differentiated from people with casual, isolated usage and from habitual drug users.

A shorter segment is devoted to decision trees, the TREE procedure in SPSS, suitable for processing and clear presentation of any type of data.Tree procedure was applied in case of example 4 (Czech adults' experience with psychotropic substances) in order to assess the significance of social, cognitive (drug awareness) and socio-demographic variables (such as age, gender, education, income, size of village or town etc.) for drug usage differences. And whether such differences in drug user career are affected by awareness of the effects and countermeasures taken in this area.

Part four is dedicated to correspondence analysis. Its origins are linked with linguistics. Example 2 (The opinion of Czech adults about the authorities involved in criminal proceedings) was used to organize data which otherwise appeared to be chaotic. Graphic seizure provided by correspondence analysis allowed insight with respect to rating of individual authorities in various regions of the Czech Republic. Conveniently correspondence analysis works with any type of data, it does not require the data to comply with such assumptions as normality of distribution. Example 12 (Life principles of potentially problematic and unproblematic inhabitants of the Czech Republic) categorised respondents from the point of view of their social mal/adaptation. Using correspondence analysis we assigned those groups significantly different principles of conduct and opinions from negative extreme of a street gang moral code to the other pole in well socially adjusted attitudes to media, police and politics.

The concluding fifth part deals with multidimensional scaling, above all with the ALSCAL technique. ALSCAL is particularly suitable for metric data (e.g., criminal statistics) or data gained from content analysis of documents or correspondence. These data, transformed to proximity or dissimilarity matrices, are subsequently processed into so-called perception maps such as in example 4 (Czech adults' experience with psychotropic substances). The maps clarify the distances and dimensionality of individual phenomena, such as soft and hard drug usage. We demonstrate how to identify the optimum number of dimensions. This is important because it is the researcher who must determine their number and enter it into SPSS. The three-dimensional solution to example 4 revealed more than factor analysis in which a comparable initial matrix had been used (the Pearson correlation). In addition to the assessment of different characteristics irelated to usage of hard or soft drugs, ALSCAL revealed also the dynamics of proneness to hard drugs (e.g., that the facilitaing agents tend to be tobacco and alcohol rather than soft drugs), and the intensity and type of dependency. Analysis thus may generate further hypotheses which can be tested by other multivariate techniques. And this is also the main raison d'être for this text—not just to illustrate the solutions in criminological research but to encourage creativity, new questions and hypotheses.

Translated by: Presto