Helsingin yliopisto - Helsingfors universitet - University of Helsinki ID 2007-214

Tiedekunta-Fakultet-Faculty
Faculty of Social Sciences

Laitos-Institution-Department
Department of Statistics

Tekijä-Författare-Author
Kontto, Jukka

Työn nimi-Arbetets titel-Title
Visualizing large epidemiological data sets using depth and density

Oppiaine-Läroämne-Subject
Statistics

Työn laji-Arbetets art-Level
Master's thesis

Aika-Datum-Month and year
2007-05-11

Sivumäärä-Sidantal-Number of pages
84 s., 24 liites.

Tiivistelmä-Referat-Abstract
The emphasis in this work is on visualizing large data sets in epidemiology. In epidemiology, especially studies of rare events in a population require large sample size. Also, large data are collected in longitudinal studies and by national health services and other government officials. The developments in computer technology over the last ten years have increased the efficiency of processing of large data sets and this has opened new opportunities for the statistical data analysis.

Specific methods has been developed for visualizing large data sets, since most of the traditional data analyzing tools are not necessarily efficient. In this work the definition of a large data set is based on two typical characteristics of large data sets: A data set is large if plotting or computation times are long, or if plots have an extensive amount of overplotting. In this work the visualization of large data sets is discussed in terms of these two characteristics. Large data sets are discussed in a bivariate situation and added with one or two categorical variables.

The visualization of large data sets is discussed with the graphical methods based on the concepts of depth and density. Both approaches deal with overplotting by aggregating data into groups which are visualized instead of individual data points. The methods using depth define a depth value for each observation and visualize groups which are determined using the depth values. The most studied depth-based method is the bagplot whose modification, the grouped bagplot, is introduced in this work. On the other hand, the methods using density divide a two-dimensional plane into bins and analyze the data points separately in each bin. Density-based methods are more well-known and more used than depth-based methods hence in this thesis the emphasis is on depth-based methods.

The graphical methods are applied using the data sets of the MORGAM Project. The MORGAM Project is a large international follow-up study of the cardiovascular diseases and genetic risk factors. The methods are compared using an interactive web application which was developed as a part of this work for researchers of the MORGAM Project. The user of the web application selects a data set to be analyzed and graphical methods to be used. The application then shows the graphs based on selections.

It was observed during the comparison of depth- and density-based methods, that the bimodality of a data set was detected only with density-based methods. This is due to the fact that when using depth-based methods the assumption is that the underlying data set is unimodal. On the other hand, the group comparison was more efficient with depth-based methods. The comparison of the processing times of the methods showed that the methods using depth have longer processing times than the methods using density.

Avainsanat-Nyckelord-Keywords
large data sets
graphical methods
overplotting
data depth
bagplot
epidemology

Säilytyspaikka-Förvaringsställe-Where deposited

Muita tietoja-Övriga uppgifter-Additional information