Mathematical Problems of Computer Science 22, 2001, 80–85.

Package "SIFAnalysis" on Testing Sequences Homogeneity and Computing Information Distances

Harutyun A. Shahumyan

Institute for Informatics and Automation Problems of NAS RA and YSU
E-mail: harutiun@yahoo.com

Abstract

This article presents a statistical program package "SIFAnalysis" created by the author. The package allows to check the homogeneity of numerical sequences, find change points of time series, define informative distances of different classes and etc. It is developed by STATISTICA BASIC and may be used within STATISTICA 5.0 or later. "SIFAnalysis" has been applied in some demographical and epidemiological researches and proved itself as a useful statistical computer tool.

1 Introduction

"SIFAnalysis" (Sequence & Informative Factors Analysis) makes package STATISTICA more effective by introducing new features such as sequence homogeneity testing, change point detection, informative characters selection and independence testing. These enhancements provide users with new opportunities of data researching.

"SIFAnalysis" requires that STATISTICA 5.0 or later be installed on the computer, because it uses data files and system commands of STATISTICA [2]. The package works in Windows 3x, 9x, 2000 environments. It consists of 10 separate programs and a special help system, which are developed by STATISTICA BASIC, Microsoft Visual C++ and Microsoft HTML Help Workshop.

The package enables to:

- check the homogeneity of numerical sequences,
- rank sequences in different ways,
- compute non-parametric statistics of Wilcoxon, Mood, Savage and etc.,
- define change points of time series through different slippage and a posterior methods,
- build graphics of time series and their rank statistics,
- estimate variation and Kullback-Leibler distances of different populations,
- identify the most informative characters,
- check independence of variables, etc.

The program enables easily choose the required variable from the data file, define the research interval and decide the method of analysis. "SIFAnalysis" requires the existence at least of one data file. It starts with the main panel of research methods, which consists of the following sections (figure 1).

About – gives general information about the package.

Population Comparison – contains some methods for studying of sequence homogeneity. Parametric statistics are used for comparison of two populations.

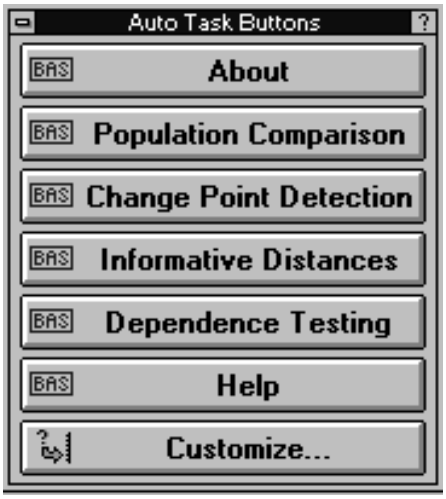

Figure 1. The main panel of "SIFAnalysis".

Change Point Detection – contains several non-parametric methods on definition of change points of time series, which can be used by slippage and a posterior procedures [6].

Informative Distances – contains methods for computing of variational and Kullback-Leibler distances between two populations.

Dependence Testing – checks hypothesis on the independence of specified variables by several parametrical statistics [5].

Help – provides information about installation and usage of the package (figure 2). It is based on the Microsoft WinHelp 4.0 and has all opportunities of Windows Help system.

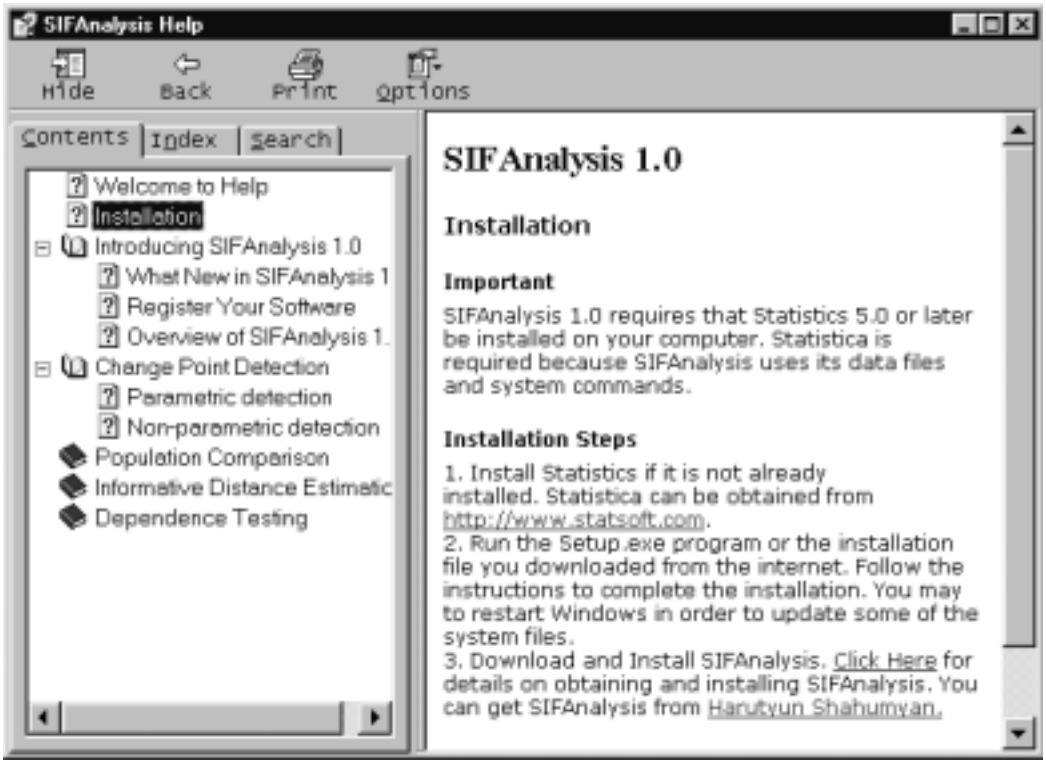

Figure 2. The Help system of "SIFAnalysis".

2 Methods on population comparison

This subprogram compares two populations for definition of the significance of their distributions difference. The parametric statistics Z_0 and T are used for that purpose.

The program starts with the selection of event and population variables [5] (figure 3).

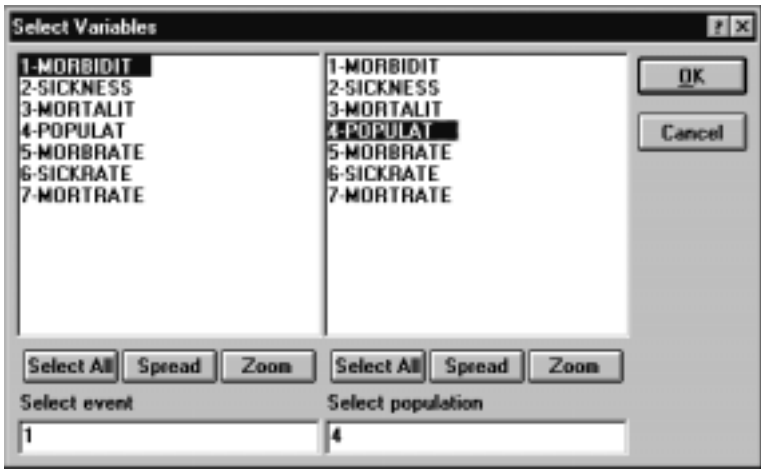

Figure 3. Selection of observing variables.

It is possible to investigate **Time series** of the specified event or select **Two specific cases** from the series and test the homogeneity of the corresponding populations.

The interval or elements of sequence which must be investigated are specified in the windows "Range of Time Series".

The results of **Two specific cases** comparison are displayed immediately. "SIFAnalysis" supports the investigation of **Time series** by some **slippage** and **a posterior** procedures. The dimension of slipping window must be specified for the slippage procedure. It has the default value 2 and can be only a positive even number.

As a result of analysis the values of Z_0 and T parametrical statistics and their critical values are displayed.

3 The methods on change point detection

The change point detection procedure is based on the sequential usage of rank statistics [6]. There are different ranking methods in the package (figure 4).

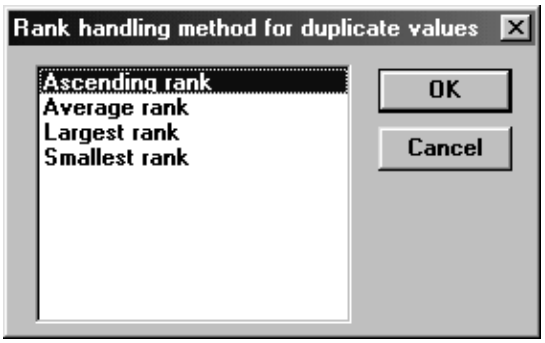

Figure 4. Ranking methods.

It is available that:

- the **Ascending rank** is assigned to each duplicate value in the specified variable,

- the **Average rank** for all duplicate values will be assigned to each duplicate value in the specified variable,
- the **Largest rank** is assigned to each duplicate value in the specified variable,
- the **Smallest rank** for all duplicate values will be assigned to each duplicate value in the specified variable.

The appropriate rank statistics must be chosen after ranking (figure 5). **Wilcoxon**, **Mood**, **Polynomial** and **Savage** statistics are available. In the case of polynomial statistics, coefficients are also specified by the user.

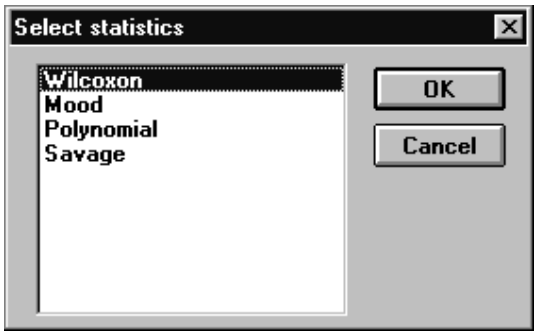

Figure 5. Selection of non-parametric rank statistics.

At the end of described process a special control panel allows to display the results of investigation (figure 6). It has the following construction.

Report – shows the main results of analysis. The report includes information about observed variables, interval of investigation, ranking methods, applied statistics, critical values, detected change points and etc.

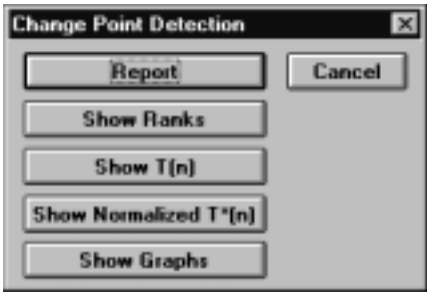

Figure 6. Results control panel.

Show Ranks – displays the ranks of observed sequence.

Show T(n) – displays the values of applied non-parametric statistics.

Show Normalized T*(n) – displays the values of corresponding normalized statistics.

Show Graphs – opens graphs panel, which allows to create graphics of the **original sequence**, **T(n)** and **T*(n)** statistics or display all mentioned graphs in the same window.

Graphics created in "SIFAnalysis" can be modified by the rich graphical tools of STATISTICA and be saved in "bmp" or "wmf" formats.

4 Computation of information distances

The informative distances are effective tools for reduction of analyzed variables space dimensionality and for selection of the most informative indices [5].

For calculation of variational and Kullback-Leibler distances of observing groups, the appropriate character(s) and grouping variable must be specified. As a result the values of

information distances are displayed. They can be used for selection of the most informative characters: the greater the distance the more the chosen character is informative.

5 The practical usage of the package

"SIFAnalysis" has been successfully applied in some demographical, epidemiological and medical researches. It has been used in the solution of the following practical problems:

- estimation of mortality rate alteration in Armenia from 1990 to 2000,
- investigation of homogeneity of tuberculosis epidemiological characteristics in Armenia from 1980 to 2000,
- detection of change points of time series of tuberculosis morbidity, prevalence and mortality in Armenia,
- selection of the most informative symptoms of some gastroenterological diseases [9].

References

[1] Afifi A. A., Azen S. P. Statistical analysis: A computer oriented approach, Academic Press, New York, 1979.

[2] Borovikov V. P., Borovikov I. P. STATISTICA – Statistical analysis and data handling in Windows, (in Russian), Filin, Moscow, 1998.

[3] Duarte S., Stam A., Nonparametric two-group classification: Concepts and a SAS-based software package, American Statistician, 25, pp. 185-198, 1998.

[4] Dyuk V. Data handling by PC in examples, (in Russian), Piter Publishing, St. Peterburg, 1997.

[5] Haroutunian E., Kazanchyan T., Mesropian N., Asatryan D., Harutyunian M., Sahakyan M., Shahumyan H. Probability and applied statistics, (in Armenian), Gitutiun, Yerevan, 2000.

[6] Myles H., Douglas W. Nonparametric statistical methods, John Wiley & Sons, New York, 1973.

[7] Petrosyan P. A. Program realization of the algorithm on time series change point detection, (in Russian), Mathematical problems of computer science, XIX, pp. 32-39, Yerevan, 1998.

[8] Statistical Methods for digital computers, edited by Enslein K. and Ralston A., Wilf H., (in Russian), Nauka, Moscow, 1986.

[9] Shahumyan H. A., Baghdasaryan N. G., Haroutunian E. A. On the Most Informative Indices Selection Method, Transactions of the Institute for Informatics and Automation Problems of NAS RA and YSU: Mathematical Problems of Computer Science, XXII, Yerevan, 2001.

[10] Upton G., Cook I. Understanding Statistics, Oxford University Press, London, 1996.

Հաջորդականությունների համասեռության ստուգման և տեղեկատու հատկանիշների ընտրության SIFAnalysis ծրագրաշարը

Հ. Ա. Շահումյան

Ամփոփում

Հոդվածում ներկայացվում է հեղինակի ստեղծած SIFAnalysis ծրագրաշարը: Այն հնարավորություն է ընձեռում իրականացնել հաջորդականությունների համասեռության ստուգման, ժամանակային շարքերի հատկությունների փոփոխման պահերի հայտնաբերման, տեղեկատու հատկանիշների ընտրության և հատկանիշների անկախության ստուգման մի շարք յուրօրինակ եղանակներ: Ծրագիրը հաջողությամբ կիրառվել է ժողովրդագրական, համաճարակաբանական և բժշկական որոշակի խնդիրների լուծման գործընթացներում և նկատելիորեն ընդլայնել է այդ բնագավառներում STATISTICA հաշվարային միջավայրի հնարավորությունները: