Eberle, Toutenburg:

Handling of missing values in statistical software packages for windows

Sonderforschungsbereich 386, Paper 170 (1999)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Handling of missing values in statistical software packages for windows

W. Eberle H. Toutenburg

September 23, 1999

Abstract

The problem of estimating parameters of distributions by an incomplete data set is theoretically considered, but in practice the implementation of the developed methods in commercial statistical software packages varies from program to program. None of the examined software offers all possible methods. In some programs the user has no choice concerning the use of a method, or no methods at all are available. However, the most popular programs have not always the largest variety of methods. Hence some work is still waiting for the producer of statistical software.

1 Introduction

In most cases, the theory of statistical methods gives answers about what to do if there is a complete data set. On the other hand, more often than we'd like to the observations are incomplete. In the last years several people dealt with the question of incomplete data and missing values. Indeed, Little and Rubin (1987) give an extensive description about the theory of missing values and how to solve this problem. In practice the statistical calculations are made by software packages. While examining an incomplete data set it is important to know how the program treats missing values and what tools are offered. This paper presents the main results of an experience of several software packages regarding the missing value problem. The results were obtained by a seminar at the Institute of Statistics at the Ludwig-Maximilians-University Munich in summer 1999. In the first section of this report the examined software packages are mentioned. The items which are considered are discussed in the second section. And the third section presents the result of this investigation. The last section contains a summary of the comparison.

2 Object of Analysis

In this investigation, the election of the statistical software packages was more intuitive than calculated. The intention was to take that into the project which is widespread in use but also more unknown programs as well. All the software packages are able to work with the operating system MS-Windows. As mentioned above the title of this paper was theme of a seminar in which each student had to examine one package. Therefore the number of elected software

MINITAB Release 12.2	SPSS Release 8.0
SYSTAT 7.0 for Windows	SAS 6.12
STATISTICA/w 5.1.	StatXact Version 4.0
Stata 6.0	LogXact Version 2.1
S-PLUS 4.0	JMP Version 3.15

Table 2.1: examined statistical software packages

was restricted by ten. The statistical software packages in the study are printed in table 2.1.

3 Aspects of the investigation

The aim of this investigation was not only to answer the question of what the program does when the data set is incomplete - that means the presentation of missing values and offered methods for treating them etc. - but also how these methods are documented in the online help and in the manuals. In the following the central items of interest will be described.

The first item concerns the missing data code. Each software package that shall work with missing values needs a code to identify them. For the handling of incomplete data sets it is necessary for the user to know how missing values are coded. Some problems may arise if the code is unknown. Especially if the code of the program and the code of the imported data set are not the same. Then mistakes may arise while reading in the data.
Imagine the input data set uses a point as a symbol for the missing value, but the program uses a blank. Then the missing value will not be recognized. As a result, an error will arise if the variable is numeral or the column will be recognized as alphanumeral if there is a missing value in the first observation. In every case, the datasets in the program and the original are different.
Another mistake arises if the original dataset uses a number, -99 for example, as a code for a missing value. The mentioned program reads the data without problems, but when calculating some statistics the results will be wrong, because the original missing data code will be recognized as a number and not as missing value and therefore will be included in the calculation. So the second item concerns the existence of a desirable option that enables one to say how missing values are coded in the original dataset and the program transforms it.

Sometimes there are some reasons for a missing value. This happens when an individual refuses to answer or it passes some questions because of a certain answer to a previous question (e.g. some questions are only for woman other are only for man) a.s.o.. Now these different reasons shall be distinguished. Then they need different codes but all these codes must be recognized as a missing value by the program. The questions are: is it possible to define more than one value as a code for missing values?, or even better: does an option exist to define a whole area of values as missing value codes?

The fourth item concerns the representation of missing values in tables and

graphics. It is examined whether it is possible to create them when the data is incomplete and how missing values are taken into account. In some cases it can be choosen whether a new category shall be created. In other cases not even the number of ignored observations is shown. Especially in time series several methods can be applied.

Several statistical methods for incomplete datasets assume that the values are missing completely at random (MCAR). That means the observed values are as well as the missing values a random subsample of the sample set. So it is necessary to test before using the method whether this assumption is fulfilled or not. Therefore the following item deals with the question if the program offers a test to assure MCAR.

A further item of the investigation concerns the calculation of descriptive statistics in presence of missing values. Descriptive statistics means: mode, median, arithmetic mean, variance, standard deviation, skewness, kurtosis, standard error and quartiles. Here, the main interest are not the methods the program offers because there is only one. It is examined if the program refuses one of these actions because of missing values.

The calculation of covariance and correlation is seperated from the descriptive statistics because here is more than one variable involved. There are two possibilities to handle missing values. The first ignores all observations having missing values in at least one of the variables. This is called complete case analysis. The second takes only variables which are involved in the next calculation and ignores the observations with missing values. That is called available case analysis. In the last case a problem sometimes arises when calculating a correlationmatrix. Each element of this matrix represents a correlation of two variables. If a dataset is incomplete the number of incomplete observations in each pair of variables may not be the same and a different amount of observations is excluded from the calculation of the matrixelements. As a result, the amount of observations for the calculation of the different correlations may not be equal. Under this circumstances, sometimes correlations higher than one or lower than minus one may arise. On the other hand, in some cases not enough observations are left for a calculation by using the complete case analysis. Therefore it is desirable to have the choice between this two alternatives. This question is dealt with by the seventh item.

The next item treats the offered options to calculate tests and confidence areas when there are missing values in the data and, of course, if it is allowed to carry out these actions with incomplete data.

The handling of missing values by applying higher statistical methods such as regression analysis, analysis of variance, cluster analysis, discriminance analysis and time series is also considered in this investigation. The results are put into the ninth item. It is not the intention to explain here exactly the theoretical background of the used methods. Therefore several statistic literature exists like Little and Rubin (1987) (as mentioned above), Rao and Toutenburg (1999), Toutenburg (1992), etc. The used methods are EM-algorithm, interpolation, extrapolation, imputation and some more.

Sometimes not all of the possible statistical methods are offered in every software package, so it would be a good thing to have a programming language to create macros for this special use. The next examined question concerns the existence of a language to program macros.

All the questions above ask for the offered possibilities to deal with incomplete data sets, but it is also important to know how the program works. The user should be informed about possibilities and options the program offers and about the assumptions that may be satisfied to use a tool in the right way. This information should be obtained by reading the manuals and the online help as well. To judge their quality several items were subject of the investigation. First of all it is examined whether and how the algorithms are explained. The second item concerns the representation of the theoretical background. The manuals and the online help should reveal the statistical background in a short way and the user is able to refresh his knowledge about the method. If he wants a detailed information there should be given a list of further literatur. This is the third item. The last item concerns examples which shall help to understand the use of the offered methods and how these can be called. It must be remarked that this judgement is not at all representative because each user may have other expectations to manuals and online help.

The above mentioned items, which are considered at each statistical software package in table 2.1, are listed in table 3.1. The results of the investigation are represented in the following section.

4 Results of the Investigation

The first considered software package is MINITAB Release 12.2. It is used in science, industrie and economy in many countries all over the world. All available procedures can be called by mouse click on the pull down menu or by writing the command in the *Session Window*.

This program codes missing values in numeral and date/time variables with a star as well. In alphanumeral variables the code is a blank. If a nondefined value arises while calculating a new variable the value is set as missing, too.

MINITAB is able to read datasets from external files which are created in MS-Excel, Quatro Pro, Lotus 1-2-3 and dBASE but also text or data files. There are no difficulties while reading in data with a different code for missing values from these files. MINITAB offers the option to change the code during or after the reading in procedure. A probleme will only arise if a missing value in a numeral variable is coded with a blank in the original data set. In this case MINITAB is not able to differ whether it is a missing value or a separator. As a result MINITAB ignores the missing value and reads the next value. Therefore the MINITAB data set differs from the original because there are no missing values and hence less observations. Here it is necessary to change the missing value code before reading or use the option *Import Special Text*. There it is possible to define the format. The alternative for reading in an external file is this: It is possible to enter the data via the command line editor in the session window. There a missing value in a numeral variable must be writen with a

1. Coding of missing values
   (a) in numeral variables.
   (b) in alphanumeral variables.
   (c) in data/time variables.
2. Existence of an option to change the code while reading an external dataset.
3. Possibility of changing the code of missing values or defining several values or even an area as missing.
4. Representation of missing values in
   (a) tables and
   (b) graphics.
5. Test on MCAR offered?
6. Possibility of calculating descriptive statistics in presence of missing values.
7. Offered options at calculating covariances and correlations.
8. Offered methods by applying tests and confinence intervals.
9. Offered methods by applying
   (a) regression analysis
   (b) analysis of variance.
   (c) cluster analysis.
   (d) discriminant analysis.
   (e) time series.
10. Possibility of programming macros.
11. Quality of manuals and online help:
   (a) Explanation of algorithm?
   (b) Presentation of the theoretical background?
   (c) List of further literature?
   (d) Quality and presence of examples?

Table 3.1: list of examined items

single quotation mark and a double in an alphanumeral variable. It is even possible to edit the data in a worksheet directly. This is the easier way because each cell is set as missing if no entry is made. MINITAB decides the type of the variable while reading the first row. If there is a missing value in a numeral variable coded with a dot MINITAB puts it as alphanumeral.

An advantage of MINITAB concerns item three in table 3.1. This program has an option to define several codes as a code for missing values and even several areas. The information about the number of observations and missing values can be obtained for each variable in the *Info Window*.

In frequency tables the number of observations and missing values is presented if this action is called by the *tally* command with the option *count*. Otherwise it isn't mentioned. The option *Cross Tabulation* offers to choose whether missing values shall be included or not or just for specific variables. If they are included MINITAB creates an additional category for the missing values for each variable that has one.
In this software package are two kinds of graphs: high-resoluted graphs (core graphs, 3D graphs, speciality graphs) and character graphs. The first offers graphics with high quality and the possibility to make some changes. The second has the advantage that its graphics can be printed with every printer but by far not as exact as high-resoluted graphs. If a character graph of a variable with missing values is called, these will be ignored. High-resoluted graphs treat this problem as follows. If the variable is categorial a new category for missing values will be created. If it is a metric variable, points with at least one missing value won't be plotted. In time series plots the point on both sides of a missing value will be connected with a straight line. Therefore the scale remains the same. In every case, the number of ignored observations are written in the *Session Window*.

A test that asures MCAR in the data set is not offered, but it is possible to program a macro.

The calculation of descriptive statistics (as mentioned above) is possible with two commands. The first (*describe*) informs about the number of observations and missing values, mean, minimum and maximum, median, standard deviation and quartiles. The second (*%describe*) is a macro which gives no information about the number of missing values, but it calculates additional variance, skewness, kurtosis and confidence intervals a.s.o. and plots some descriptive graphics. Several measurements can be called via the *stats* command or the menu bar. The treatment of incomplete data is here very easy. All observations with a missing value are excluded from these calculations.

This software package uses the available case analysis or the pairwise deletion to treat missing values while calculating covariances and correlations as well. There is no other possibility offered. Therefore it is possible to receive invalid values in a correlation matrix as mentioned in the previous section.

For calculating inductive statistics such as tests and confidence intervals MINITAB uses only complete observations. That is observation with missing values are excluded from the calculation. Only the χ^2-test of independence refuses incomplete variables.

In the regression analysis and the logistic regression incomplete observations will be excluded from the calculations. In addition the regression analysis offers the option to calculate fitted values for the response if the independent variables of the observation is complete. The output contains the number of excluded ob-

servations.
The problem of incomplete data in the analysis of variance treats MINITAB as follows. Incomplete observations will be excluded from the calculation. Unfortunately the number of excluded or included observations won't be put out. Sometimes a balanced design changes to an unbalanced. In this case the two-way-ANOVA produces an error message and no calculations will be made.
The cluster analysis in MINITAB offers two options. One tries to unite observations the other tries to group variables. In the first case variables with missing values can't be chosen in the dialog box. In the second case incomplete observations are excluded from the analysis and no information about the number of included or excluded observations will be given.
If a discriminant analysis is called to an incomplete data set MINITAB excludes all observations with missing values from the calculation. Here the number of ignored observations is given in the output.
MINITAB has several procedures for calculating time series. These are moving average, trend analysis, decomposition and single and double exponential smoothing. In each procedure it is possible to receive forecasts, but only the first three accept variables with missing values. These procedures do not ignore incomplete observations - because this would cause a change of the time scale - but still don't plot it. The output of each procedure contains the information about the number of incomplete observations.

Of course, there are not all possibilities offered to handle incomplete datasets but fortunately MINITAB enables the user to program macros. Therefore he can write programs for his requirements.

To judge the quality of the manuals two books were considered. This is the *MINITAB User's Guide*, which contains a clear overview of the structur and usage of MINITAB, and the *MINITAB Reference Manual*, which informs about all the possibilities of statistical calculations offered by MINITAB. Other manuals are available such as *MINITAB Quick Reference* and *MINITAB Mini Manual* which are only summaries of the two mentioned. The algorithms of general statistical methods and methods for missing data are explained comprehensibly. The online help contains the same information. In addition to that, the online help deals with the problem of reading in incomplete data sets from external files. The statistical theory in the manuals is not as spread out as in a school book, but users who have a certain knowledge of the statistical background have a summary and repetition. It is especially explained when the use of a special method is indicated. The theory of missing values is not given but in many sections it is said what MINITAB does if missing values enter a procedure. At the end of each chapter a list of literature for further information about the theory is given, but there is no literature found for incomplete data. For each procedure several clear and well explained examples are given, but none for missing values.
The online help has no list of literature at all except the MINITAB documentation. It contains clear examples, but also such which explain how MINITAB treats missing values. Therefore it possible to learn to work with MINITAB without using the manuals. So the manuals explain MINITAB and its abilities in a very clear way, but the information about the missing data problem is very small. On the other hand there are only a view possibilities offerd to treat missing values. The online help explications of the usage of procedures is more extensive than that in the manuals but the theory is shorter. The manual has

no section about missing values. Therefore this subject is scattered, but all in all it is a good reference book for the procedures.

The second examined software package is STATISTICA Version 5 Edition '97. This is a version in german. All available procedures can be called by the menu, but also via command lines. The data sheet is always visible. Here the data can be entered directly. It is possible to write numbers and words into the cells as well. STATISTICA defines for each word a number starting with 100. Date and time values will be recoded in real numbers. So STATISTICA has only to treat variables which are numeral. The code for a missing value is -9999, but can be chosen between -9999 and 9999. In the data sheet this number will not be shown, that means the cell is empty. In addition to that, the user is able to define this code for each variable separately.

The data can be read from different external files. These can be files which are created with MS-Excel, Lotus 1-2-3, Symphony, Quattro, dBase, Paradox, SPSS, SAS, Oracle, Sybase or ASCII-files. While reading in the data a modul *Datenmanagement* recognizes their structure and converts all logical and text variables and labes and empty cells into the STATISTICA format as well. Additional STATISTICA offers an option to change the code of variables or calculate new variables. It is also possible to enter the data via clip board into the data sheet, but in this case the data sheet must be extended at least to the size of the data set which shall be imported. Otherwise only the first ten observations will be read in.

This software package has not the ability to define more than one value or an area as missing value. Therefore the user must be aware that in case of reading in a data set with more than one code for missing value will cause a problem. Either all missing codes will be recognized as misssing by STATISTICA - then a distinction of different reasons for no value is impossible - or only one code will be accepted and the other must be recoded afterwards via the *Datamanagement*. This modul offers an option to replace missing values. There are two possibilities for the replacement: mean imputation and weighted mean imputation where the weights come from another variable. It is also possible to choose the observations which shall be used for the calculation.

While creating tables, such as frequency or contingence tables, STATISTICA enables the user to decide whether there should be a category for missing values or not. This decision is not offered if graphics will be applied. Here the program ignores incomplete cases and does not mention this. There is one exception: the option *Missing Data/Ausreisser-Plots* produces a graphic where data points for missing values are plotted. In addition to that, thresholds can be set and values above and below them values are considered as outliers.

STATISTICA calculates each descriptive statistic without problems. The number of observations, which are entered for the calculation, is shown. To call for covariance or correlation matrices there exists the choice between casewise and pairwise deletion.

If it is asked for tests and confidence areas STATISTICA applies the available case analysis.

In the regression calculations the dialogbox allows to choose between a listwise deletion and a mean imputation. Additional if multiple regression is called a pairwise deletion can be elected. It is possible to use weigthed mean imputation, too. Therefore it is necessary to manipulate the dataset with the data

management tool before. For the cluster analysis and discriminance analysis it is the same. Only the analysis of variance offers no choice. Here is always the complete case analysis used.
For the analysis of time series it is necessary to have a complete dataset. If the data contains missing values at the beginning or at the end of the time series STATISTICA excludes these cases from further calculations. The remaining holes can be filled by mean imputation or by the arithmetic mean of $2N$ neighbours where N can be chosen. If N exceed the time serie an error message appears. Then the user has to elect a smaller N. STATISTICA offers the option to fill the missing values with the median of $2N$ neighbours. In addition to that the program enables the user to elect regression imputation and linear interpolation as method for calculating estimations for the missing values.

This software package has an own programming language at its disposal. With STATISTICA BASIC the user is able to write his own macros.

The only delivered manual is *STATISTICA Benutzerhandbuch*. It has three parts. The first introduces how to use the program. It contains an index where it is possible to find a section about missing data. There are the treatments of incomplete data listed, but not explained and it is not said which method is used at each procedure. Here it is refered to the online help. The second part gives an overview about the statistical methods. Here it is somewhat difficult to find a place where treatments of incomplete data is mentioned. In fact, only in the section of correlation matrices the problem of missing values is discussed. The last part contains several example, but examples of dealing with missing values are sparse. The manual has no bibliography and some literature is only given at a few places.
In comparison with the manual the online help is much more extensive, concerning the part one and two, and deals mainly with the use of STATISTICA and the statistical theory. There is much more literature listed where further information about the theory can be found. In addition internet links can be called and connect StatSoft. In the home page of Statsoft one can find additional macros. The algorhythms are given neither in the manuals nor in the online help.

SYSTAT 7.0 enables the user to start the procedures either via pull down menu or via icons or via the *Command Editor*. It distinguishes between two types of data: numeral and alphanumeral (strings). SYSTAT marks missing values in numeral variables with a dot and in alphanumeral with a blank.
This program is able to open data sets from different files. These are SPSS, spreadsheet, database or ASCII files. It is possible to import all rows and columns or just a range by entering the number of the first and the last case or column. In ASCII-files missing numerical data is flagged by a dot and missing character data is marked by a blank which is enclosed within quotation marks. If this is forgotten SYSTAT cannot recognize the missing values and errors will arise. SYSTAT interprets each line as a row. Therefore the next observed value will be put at the place of the missing and the case has empty cells at the end of the row. Furthermore, it is not possible to change the coding while reading in the data. The user must change it after or before the import. The last is indicated when missing values are coded with a character, but in general, no problems appear while importing data from external files. Besides neither it is possible to define several values, an area nor another value as code for a missing

value. The code is fix.

A table can be extended with an additional category for missing values by using the command 'Include missing values'. In graphics incomplete observations will be ignored and no information is given about this action. It is not even mentioned how many observations are included for the graph. Only if a new category for missing values is defined, before they will be plotted.

Calculating descriptive statistics for one variable is carried out by ignoring missing values. To calculate Pearson's correlation coefficient, the covariance or the sum of squares of the cross-products of deviation (SSCP), the user can choose from one of the following methods: EM-algorithm (for metric variables), listwise or pairwise deletion. When the pairwise deletion is chosen to calculate the SSCP matrix each result is weighted with the quotient of the number of rows and the number of observations which enter the calculation for each matrix element. If the EM-Algorithmn is used the output contains information about the number of iterations and the missing pattern. Furthermore, it is possible to control the number of iterations, the convergence criteria and the influence of outlier. In addition estimations of mean and correlation matrix is given and a test on MCAR is carried out.

Tests and confidence intervals are calculated by ignoring incomplete observations. There is one exception. The χ^2-test for independence offers the option to include missing values in an additional category.

The regression analysis of SYSTAT treats missing values as follows. It doesn't matter if the missings are in the independent or dependent variables each incomplete case will be omitted. The output informs the user about the ignored cases. In the analysis of variances, the cluster analysis and the discriminant analysis, SYSTAT treats this problem treats in the same way. In the cluster analysis the manual recommends to create a new category for missing values with a binary coding. Then the option 'Join' can be used to clearify whether there is a missing data system. In time series SYSTAT has two options for treating missing values. Either they will be omitted or they will be estimated by a distance-weighted least square interpolation (DWLS-interpolation).
"DWLS interpolates by locally quadratic approximating curves that are weighted by the distance to each nonmissing point in the series. With this algorithm, all nonmissing values in the series contribute to the missing data estimates, and thus complex local features can be modelled by the interpolant" (SYSTAT Statistics).
In each case incomplete observations at the beginning or the end of the serie will be ignored.
The 'Delete'-option works as follows:"Retain only the leading nonmissing values for analysis. In series that begin with one or more missing values, the series is deleted from the first missing value following one or more nonmissing values. This option enables you to forecast missing values from nonmissing subsection of the series" (SYSTAT Statistics).
These forecasts can be inserted into the series before repeating the procedure later.

SYSTAT enables to write small programs, but no macros. These programs must be imported into the command window and submitted.

There are five books which document the use of SYSTAT. These are *Data, Graphics, Statistics, New Statistics* and *Command Reference*. *Data* contains introductory information about SYSTAT. All books except the *Command Ref-*

erence have in each chapter an introduction and a table of contents. At the end, an extensive list of literatur is given. The manuals are clear. Key words are placed on the margin and different fonts are used. The theory is only explained in a short way. It is assumed that the user already has the knowledge of the procedures which must be recalled. Clear examples support the explanation of the procedures. Methods for missing data especially the EM-Algorithm is explained extensively and problems are mentioned. On the other hand, not all of the procedure descriptions mention the treatment of missing values. The online help is clear, because of the different fonts, but is not as extensive as the manual. A list of further literatur is not given. The topic 'missing values' is only mentioned in time series and correlation calculation.

Stata 6.0 is one of the less known software packages in Germany, but in english speaking countries its use is more spread out. According to the statements on the homepage of Stata, the advantages of this program are high speed calculation and easy handling even for statistic beginners. This release is designed as a window program but its graphical surface is heavy reduced. There are no dialog boxes, icons or pull down menus to call a statistical function. Every command must be entered into the *Stata-Command-Window*. Several other windows exist: the *Data-Result-Window* containing the output, the *Variables-Window* containing a list of all variables, the *Review-Window* containing the executed commands, the *Data-Window* for showing the data set, the *Stata-Editor* for editing the data set, the *Do-File-Editor* for programming procedures and the *Graph-Window*, which contains the graphical output, but only showing one graphic. A new graphic deletes the old one. Some graphics having an ASCII format are less exact (e.g the histogram). Stata offers a wide range of tests and estimation methods with several options but therefore the commands are sometimes pretty long.

Stata marks missing values in the data sheet with a dot in numeral and a blank in alphanumeral variables. If the data will be entered directly into the data sheet the cell which is not edited in a variable is set as missing. If an ASCII file is read in a missing value in a numeral variable must be marked with a dot and in alphanumeral variables with two quotation marks. In addition to that, all "things that are not understood ... are mentioned and stored as missing values" (Getting started with Stata for windows (1999)). Missing values are coded with the highest number. As a result, if the number of individuals with an income of more than 5,000 is called, observations without entry in the corresponding variable will be counted, too. This is important if categories will be created. If one knows this it is possible to exclude missing values from creating categories. Not defined calculations and calculations with missing values lead to a missing value. In this case Stata gives a message that missing values are generated. It is not possible to define individual areas or several values as code for missing values. There are only a few types of variables available with a specific amount of numbers (e.g. byte, int, long, float, double). Each value beneath the chosen amount will be recognized as a missing value.

In tables the number of missing values is not shown except if using the command *inspect* which enables the user to see the number of missing values. Furthermore, the use of the *inspect*-command results in a mini histogramm in the text mode in which it is possible to see a rough guess of the amount of the missing values. Even the number of observed values is only sometimes

mentioned. Graphs ignore missing values. The only way to make these visible is to define a new category. In time series it is necessary to estimate the missing values to avoid holes in the graph.

Stata offers no test on MCAR, but it has an option to check the data set for two kinds of dependences. Firstly, if there is a missing value in variable a then there is a missing value in variable b, too, and vice versa. Secondly, if a is missing then variable c has a missing, but not vice versa.

Descriptive statistics will be calculated by omitting missing values. Only the correlation can be calculated by a complete case or an available case analysis.

Stata offers three methods to fill up an incomplete dataset. Firstly, it is possible to use the regression imputation. Here are at least 31 complete observations necessary. In addition to that, the variance of the estimation will be calculated. Compared to STATISTICA the completed variable is stored in a new variable, therefore the old variable is still visible. Secondly, it is possible to fill up a missing value by linear interpolation and third by linear inter- and extrapolation. In tests, regression analysis, analysis of variance, cluster analysis, time series and others, there are these three methods in addition to the complete case analysis possible. A discriminant analysis is not available. It is possible that in time series the moving average leads to missing values caused by incomplete data. This can be supressed by the *nonmiss*-command but it is not explained how it works.

This software package has a language to program macros and procedures which can be stored as *Ado-Files*. The explanations in the manuals are often not enough for understanding. Furthermore, it is possible to put them into the homepage of stata and to get some out of it.

For this examination there are seven manuals considered. These are *Getting started with Stata*, the *Stata Graphics Manual*, the *Stata User'S Guide* and four *Stata Reference Manuals*. The algorithms are explained in most cases and important formulas for tests are given. The statistical theory is given for important functions at least. Each chapter has a list of further literatur where the user can find the statistical background which is left out in the manuals. Examples are given to all procedures, but it is very often assumed that the data set is complete. Therefore there are only a few examples for the treatment of missing values. Stata offers two possibilities to receive help. Firstly, after entering a command and calling the help an extensive explanation of this command is given. Secondly, if the online help is called a word can be searched, but the result is only a list of chapters in the manuals. In addition the user can get further information via internet. Either he calls the homepage of Stata or he mails his problem. The homepage offers many files for downloading and many texts about Stata.

The software package S-PLUS is based on the S-Language which has some elements of C/C++ and even of other programming languages. It is possible to modify existing procedures. Furthermore, a menu bar is given with which actions can be started via dialogboxes. In S-PLUS, several windows exist. There is an *Object Browser* which contains all objects of the current working directory. Here the objects can be edited. The functions and commands will be entered into the *Command Window* which containing the text output, too. Graphs appear in seperat or in the same *Graphsheet*, which is optional. All submitted commands are shown in the *History Window*, so they can be repeated. A *Report*

Window will be opened if a function is called via menu bar and contains the text output. A missing data is coded with 'NA' in numerals and data/time variables. Alphanumeral variables are not allowed. In this case numbers must be entered and labels must be defined. Data can be imported from dBASE, Excel, FoxPro, MS-Access, Paradox and text files but also files from SPSS and SAS. Missing data will be recognized and recoded. If there is a self defined code for missing values it is possible to change it into the S-PLUS format while reading the data. A possibility to define several values or an interval as code for a missing data is not given. An advantage of S-PLUS are the functions for missing values. There are commands to create a vector to a variable which shows by boolean whether a value is missing or not. Furthermore, a vector can by created out of a variable by omitting all missing entries. And the observation number can be demanded where a cell of a variable is empty. There is also a function which counts the incomplete cells of a variable. In addition to that a functions exists which creates missing values within a complete vector according to a equal distribution on $[0; 1]$. Here the user can select how much observation shall become missings. S-PLUS has an option with which can be chosen whether missing values shall be placed on the top, on the bottom or be omitted in the sort or order function.

While frequency tables are requested S-PLUS offers two possibilities to treat missing values. Firstly, it ignores observations with missing values in the demanded variable. Secondly, a new category for missing values will be created. In cross tabulations is one more option available: the action will be refused. Additional it is possible to program new options. S-PLUS offers several types of graphs. Missing values in simple line plots results in a broken graph. In histograms missing values will be omitted. There is no option to choose. Pie charts will only be plotted for complete variables. The only way to receive a graphic is defining missing values as an own category. There are two functions to create normal QQ-Plots. The first (*qqnorm*) ignores missing values. The second (*qqline*) refuses the action. A test on MCAR doesn't exist.

If descriptive statistics are required S-PLUS omits missing values and it is possible to show their number. In calculations of covariance and correlation four option are available. Firstly, a complete case analysis will be performed. Secondly, an available case analysis can be chosen. Third the action will be refused, because of missing values. And fourth in matrices elements which will be calculated by incomplete variables are set as missing.

The calculation of tests and confidence intervalls excludes missing values. The only exception is the χ^2-test for independence. It refuses the action if a variable is incomplete.

The treatment of incomplete data in regression analysis is the same as in the analysis of variances. Here an error message appears, but this is optional. It is also possible to omit missing values or replace them with the arithmetic mean. The regression analysis replaces only missings in the independent variables. Observations with missing values in the response are excluded from the calculation. The cluster analysis and the discriminant analysis allow no missing values. In Classification and Regression Trees (CART) all observations with missings values in the response will be excluded. Furthermore, there exists an option with which a new factor variable can be created with an own class for missings. Another option transforms metric variables into factors. In survival analysis missing values will be excluded or an error message appears which is

optional. Missing values in time series are just allowed at the beginning and at the end. If a value is missing in the center an error message appears. In ARIMA models missing values will be treated with methods based on Kalman-filter.

It is remarkable that the manuals *Guide to Statistics*, *User's Guide* and *Programmer's Guide* are available in the online help and an entry to the MathSoft and S-PLUS hompage as well. The *Language Reference* offers an extensive help to the available functions for missing data. There are detailed information about their abilities, warnings and references to similar functions and further literature. Some examples are given, too. Though it is not mentioned in which statistical function they can be used. Here it is necessary to consult the help for the statistical methods. There are the optional arguments for treating missing values and their function is explained in a short way. But there is no hint that there is a special help page for missing value functions where the arguments will be explained extensively. The examples are mostly with complete data sets, so there are in general no examples with missing values in the help for the statistical methods. In addition some possibilities to treat incomplete data sets are not mentioned. For example the online help for linear models has no tip to the function which enables a mean imputation. These information can be found in the *Guide to Statistics*. Other arguments for treating missing values are simply given in the help page for the function in which they can be used. There the explanation is clear and extensive. The *Guide to Statistics* mentions arguments for missing data treatment in the statistical functions in a short way. Only "Classification and Regression Trees", "Survival Analysis" and "Time Series Analysis" have an own section for this problem. Additional, in the last algorithm and methods are explained in a short way. The *S-PLUS-Help/User's Guide* mentions the missing data problem simply with one sentence refering to the page of the function, which is interesting, in the *Language Reference*. The only german book for S-PLUS in this survey *Einfuehrung in S und S-PLUS* has sections for missing data and their coding, but these are not at all extensive. In general, examples are not explained in the online help.

The following software package is one of the most widespread in Europe. It offers the user a comfortable analysis of his data via pull down menu and icons, but also via command editor. SPSS has a modular feature. That means that it consists of a ground version and several additional moduls. One of them is the MVA-modul (missing value analysis). In the version 8.0 it is included. For the earlier version it must be bought seperately.

First of all the ground version distinguishes two kinds of missing values. On the one hand, the System Missing Value which is coded with the decimal sign of the country, using a dot in the english version and a comma in the german, in numeral and date/time variables. A System Missing Value exists if no valid entry is made in a cell of the data sheet. System missing values in alphanumeral variables don't exist. Even a blank is a valid sign. On the other hand, the user defined missing value. There are three possibilities to define missing value codes. Either the user defines up to three discrete numbers, one area or one area and one discrete number. This can be made for each variable separately. It can be defined missing values in alphanumeral variables, too. Here up to three words are allowed to be chosen as missing value codes.

SPSS is able to import data sets from several types of files. These are bBASE, Excel, FoxPro, MS-Access Paradox and text files. It is also possible

to read in ASCII files. Here can be chosen between the options 'Freefield' and 'Fixed Columns'. In both cases no missing value can be recoded. That means the user can say what type of variable there is in the data set and what the name of the variable is, but all values which do not match with the type of variable will be set as System Missing Value and the user receives a message. Recoding values in a variable is possible when the data set is read in. There the user can choose if the result shall be written in a new variable or overwrite the old.

The ground version of SPSS distinguishes two kinds of missing data treatments. The treatment before and while calling a specific procedure. Treatments before the analysis means either to delete a variable if it is not in the main interest of the study and it contains many missings, or the other method, which is also mentioned in the manuals, is to impute guessed values. The user can decide between five options. Firstly, there is the mean imputation. Secondly, the imputated value can be calculated from the mean or thirdly the median of the next $2n$ observed values where n can be specified. It must be remarked that a system missing value is set if not enough neighbours exist. Fourthly, the guessed value is received from linear interpolation or fifthly from the linear trend. The last works as follows: For all observed values a linear trend line will be calculated. The missing value will be replaced by the value of this trend line at its place. Of course, for the last four methods the data set must be put in an order. It can be decided if the filled variable shall replace the old or form a new one. In the first case the original variable is lost in the temporary data set and the danger of deleting this while saving it, exists. It should be mentioned that the system missings and the user defined missing values will be replaced as well.

Frequency tables show the frequency of all values that means all values inclusive the user defined missing values and the system missing values. Additional the percentage is given for all values und for the valid values, too. The cummulative percentage is only given for the valid cases. In crosstabulation the listwise deletion is used. In all kinds of graphs there are the options to select listwise or pairwise deletion. Furthermore, the user can decide whether missing values shall appear in an extra category or not. Line plots offer also an option to interpolate missing values and therefore repair the line. If a frequency statistic is called it is possible to view a graph. Three types are available. The bar chart and the histogram omits missing values. The latter is only for numeric variables. The pie chart counts user defined values twice. On the first place it is united in missing values and on the second a piece for each value is plotted. It must be remarked that user defined intervals for missing values are not counted twice.

SPSS has no problem to calculate descriptive statistics of variables with missing values. The output contains allways the information about the number of excluded observations. This software package offers three kinds of correlation the bivariate, the partial and the distances correlation. In all it is possible to delete listwise deletion and in the two first mentioned the user can select the pairwise deletion, too. The MVA-modul calculates the number of missing and nonmissing values, mean, standard deviation and extreme values. Means, covarince matrix and correlation matrixwill be estimatied using listwise, pairwise, EM or regression methods.

In all kinds of tests it can be chosen between listwise and pairwise dele-

tion. The regression analysis enables the user to elect either listwise or pairwise deletion or mean imputation to deal with incomplete data. The used method is mentioned in the output. The analysis of variance allows listwise and pairwise deletion and so it is in cluster analysis. The numbering of cases in the cluster analysis is somehow confusing, because it is not identical to that of the datasheet. The reason is: incomplete observations will be omitted and the complete cases get new serial numbers. The discriminat analysis allows only listwise deletion or mean imptutation. Missing values at the beginning or the end of time series are allowed and deleted from the analysis. If a empty cell exists in the center of it the calculation stops at this point and the time serie is only plotted till there.

As mentioned above SPSS has a modul to analyse missing values. The Missing Value procedure performs three primary functions. Firstly, to describe patterns of missing data. This includes the answers to the following questions. Where are the missing values located? How extensive are they? Tend pairs of variables to have values missing in different cases? Are data values extreme? And are values missing randomly? Secondly, to estimate means, standard deviations, covariances and correlations using a listwise, pairwise, regression or expectation-maximization method (EM method). The pairwise method also displays counts of pairwise complete cases. Thirdly, to fill in missing values with estimated values, which will be obtained by using regression or EM methods (SPSS Missing Value Analysis 7.5).
For examining the missing data pattern of a data set the modul offers three types of pattern tables. The 'Tabulated cases' shows the frequency of each missing value pattern. Counts and variables are both sorted by similarity of patterns. In addition to that, an option is given to eliminate patterns that occur in less than a chosen percentage of cases. The output table contains an additional column in which the user can see how the number of complete observations would increase if a specific variable is deleted. The option 'Cases with missing values' show case-by-variable patterns of missing and extreme values for cases that have missing values. Cases and variables are both sorted by similarity of patterns. The 'All cases' option displays for each case the pattern of missing and extreme values. Here the missing values are distinguished in system missings and the different user defined missing values. It is possible to sort them according to a specified variable. The criteria for an extreme value is the same as for boxplots. Univariate statistics can be calculated. That means for each variable the number of nonmissing values, the number and percentage of missing values, and the count and percentage of missing values are displayed. Additional, for metric variables the mean, the standard deviation and the counts of extreme high and low values are shown. Three options are offered to examine possible missing data pattern. Therefore SPSS creates internal for each variable a missing indicator variable that indicates whether the value of a variable is present or not. The 'Percent mismatch' option creates a table in which for each pair of variables the percentage of cases with one variable having and the other having not a missing value. Each diagonal element contains the percentage of missing values for a single variable. The second option compares for each quantitative variable the means of two groups using Student's t statistic. The t statistic, degrees of freedom, counts of missing and nonmissing values and means of the two groups are displayed. It is also possible to "display any two-tailed probabilities associated with the t statistics, although interpretation of these probabilities

can be problematic" as the manual *SPSS Missing Value Analysis 7.5* mentions. With the t test can be decided whether values are missing randomly or not. The MVA-module offers "Little's χ^2-statistic for testing whether values are missing completely at random"(SPSS Missing Value Analysis 7.5). This will be only calculated with the EM methods and output with the EM matrices. The third option displays a crosstabulation of categorical and indicator variables, that means a table for each categorical variable in which for each category (columns) the frequency and percentage of nonmissing values for the other variables (rows) is shown and the percentage of each type of missing value, too.
The means, the standard deviations, the covariances and the correlations can be caluculated via listwise and pairwise deletion, regression estimation and EM algorithm as well. In case of the pairwise deletion a table of frequency of missing values in pairs of variables is shown. All variables are listed and the number of pairwise complete cases are shown. In case of the regression estimation missing values are estimated using multiple linear regression. The means, the covariance matrix and the correlation matrix of the predicted variables are displayed. Here it is possible to add a random component to regression estimates. It is possible to choose between residuals, normal variates, Student's t variates or no adjustment. A maximum number of predictor variables can be set, too. In case of the EM method several assumptions can be made for the distribution of the data. These are normal, mixed normal and Student's t distribution. For the mixed normal assumption the proportion and the standard deviation ratio can be specified. For the Student's t distribution the degrees of freedom must be set. Furthermore, the number of maximum iterations can be specified after which the calculation stops (it doesn't matter if it has converged or not). In both cases of filled up data sets these can be saved via an additional option. If the data is missing completely at random all four methods provide consistent and unbiased estimates for the covariances and correlations. Besides, the MVA-module offers a summary of means and standard deviations which are calculated by different methods. Here the results can be compared. The manuals recommend to calculate scatter plots of the original variables and the completed for proofing whether the estimated values fit to the observed values.

SPSS has an own programming language. The commands can be entered into the *Syntax Editor* and submitted or a macro can be programmed. Using the syntax commands the user has some more possibilities to analyse data sets. Indeed this is different within the MVA-module. Here all commands can be applied via dialog boxes, too.

There is an extensive literature available for SPSS and its additional moduls. These are generally written in English but some of them are translated into German, French, Italian and Spanish. The manuals have clearly structured chapters. Each chapter explains the use of a specific procedure and offered options. Its use is shown by concrete examples. At the beginning of a chapter the goal of the described method is told. Afterwards the statistical background is mentioned in a short way and shown by an example. Then the way these methods and its options can be called in SPSS is described and shown by clear examples. Additionally, the examples contain an interpretation of the results. Unfortunately, the algorithms are not explained. Further literature can be found in the bibliography at the end of the manual. The books are mentioned in the text. Therefore if the user wants on overview about the literature to the topic of a chapter he must read the whole chapter and mark the mentioned books.

The online help offers the user only some short informations about the goal of a specific procedure, but there are no information about algorithm and statistical background. On the other hand it is very clearly explained how the procedure can be called. The syntax of the commands are given but unfortunately the options are not explained. An additional item in the online help leads the user to the web page of SPSS where informations can be read or questions can be posed. There are also web pages in different languages, but these contain only short information and refer to the american page.

The next considered software package is SAS in its release 6.12 . SAS is also a rather popular program especially in medical researches. If a new medicine is developed its effect must be proofed and whether there are any undesired side effects. Finally, the study results are sent to an institution which decides if a medicine may be sold. A leading institution is the american Food and Drug Administration. This wants the data in an SAS format. Therefore the use of SAS is assured.

SAS is a program whose user surface is not as comfortable as that of SPSS for example. Besides, it offers a wide range of methods to analyse data sets and create reports. If SAS is called three windows will be opened. The *Log - Window*, the *Output - Window* and the *Program Editor*. Procedures can be carried out by tipping the commands into the *Program Editor* and to submit them. The icons in the SAS window enable neither a statistical analysis nor the creating of graphs. One icon activates the *SAS/Assist*. It is a less comfortable tool to carry out actions via dialog boxes as the surface of SPSS. The user has to spend some time to find out how he can run the chosen procedures which is a disadvantage. Finally the program should be as simple in use as possible. The user might work with it without consulting manuals.

SAS distinguishes two kinds of variable types: character and numeric. The numeric type contains also date-,time- and some more formats. Several formats are united to the character type, too. A missing value in a charachter variable is coded with a blank. The code in numeric variables is a dot. Nonnumerical cell entries in a numeric variable will be set as missing. Results of calculations which are not defined are set as missing, too. The code is fix and cannot be changed. SAS allows neither the definition of more than one value nor of an area of values as missing values.

SAS can import data from text files. Here it is possible to choose a given file format, that means data files in which values are separated by commas or by tab deliminater, or a user defined file format. Equal which format is elected it is not possible to change the code of a value while reading in the data if the action is called by dialog boxes. Therefore missing values have to be recoded before or after reading the data. If the data is read in by syntax commands the user has more options. Usually the variables are of a special type. If this is defined no other type can be entered into the cells of the variable. That means a numeric variable allows no characters. SAS offers an option with which characters are allowed in numeric variables. But in calculations these characters will be regarded as missing values. This can be used to distinguish several reasons for nonresponding. SAS is also able to read data sets from MS Excel and BMDP files.

Frequency tables and crosstabulations can be called by the *PROC FREQ* procedure. The default adjustment ignores missing values, but writes the num-

ber of excluded observations in the output. Two alternatives are offered. Firstly, a new category for missing values will be created for each variable with missing values, but only their number is printed in the table. Missing values are not included in culculations of statistics. The second alternative includes additionally the missing values in statistical calculations.
This software packages supports different kinds of graphics which will be calculated by several procedures. Graphs for categorical values, these are pie charts, block charts, bar charts, etc., enables to create an additional category for missing values which is plotted in the graph. Other procedures omit incomplete observations. In time series two possibilities are given. Either the plot stops at the first missing value after an observed or missing values will be interpolated by four methods. The cubic spline fits the data inside the first and the last observed values. Additionally, the spline is extended by adding linear segments at the beginning and the end. The linear interpolation connects the observed value before and after the missing value. "The STEP method fits a discontinuous piecewise-constant curve. For point-in-time input data, the resulting step function is equal to the most recent input value. For interval total or average data, the step function is equal to the average value for the interval." "The aggregate method performs simple aggregation of time series without interpolation of missing values. If the input data are totals or averages, the results are the sums or averages, respectively, of the input values for observations corresponding to the output observations. If the input data are point-in-time values, the result value of each output observation equals the input value for a selected input observation."(SAS online help).

For calculating tests or confidence intervals SAS ignores incomplete cases. The regression analysis in SAS omits all incomplete observations. In general linear models that can be calculated by the *PROC GLM*-procedure the treatment of missing values depends on the type of analysis. If an univariate model is elected observations with missing values are omitted equal the value is missing in the response or in the independent variables. In case of multivariat models two possibilities are availabe. Either an observation will be excluded of the whole calculation if a value is missing in at least one of the response variables or it is excluded from the calculation of the considered variable if it has no valid value in this variable. The number of used or excluded observations will not be given in the output. If a probit model is calculated by the *PROC PROBIT*-procedure missing values in the response are treated as zero. The observation will be excluded if the independent variables have at least one value missing. The analysis of variances ignores any observation with missing values.
SAS distinguishes two cases in cluster analysis. "If the data are coordinates, observations with missing values are excluded from the analysis. If the data are distances, missing values are not allowed in the lower triangle of the distance matrix. The uppertriangle is ignored."(*SAS/STAT User's Guide*).
Missing values in discriminant analysis are treaten as follows. "Observations with missing values for variables in the analysis are excluded from the development of the classification criterion. When the values of the classification variable are missing, the observation is excluded from the classification criterion, but if no other variables in the analysis have missing values for that observation, the observation is classified and printed with the classification result."(*SAS/STAT User's Guide*).
In time series any missing value at the beginning of the data set will be skipped.

An option can be specified, then "the first continuous set of data with no missing values is used; otherwise, all data with nonmissing values for the independent and dependent variables are used. Note, however, that the observations containing missing values are still needed to maintain the correct spacing in the time series. For output data sets, *PROC AUTOREG* can generate predicted values when the dependent variable is missing."(SAS online help). Another procedure is available to fit "cubic spline curves to the nonmissing values of variables to form continuous-time approximations of the input series. The procedure can also estimate first derivatives of time series with respect to time, computed by differentiating the interpolated spline curve."(SAS online help).

The concept of this program bases on syntax commands as entirely mentioned. Therefore it should be easy for an experienced SAS user to program his own procedures or macros.

The quality of manuals is considered by the *SAS/STAT User' Guide* and the *SAS/GRAPH Software*. The first chapter of the User's Guide describes the idea and the theory of statistical methods in a short way. Sometimes an example is given to make the things clearer. In these chapters the user can find all available procedures of this context with a short describtion of its work. Additionally, the chapter in which the procedures are explained is named. In addition to that, a list of further literature is given. The following chapters describe the procedures extensively. The abilities of the procedures are mentioned. The syntax and the options are explained clearly. Several examples are given with a short explanation of the output. In addition to that, each of these chapters name the treatments of missing values in the procedures. Algorithms are not explicitely mentioned. The user can only guess it by reading the theory.
The online help is not as extensive as the manuals. Here neither the theoretical background nor the algoritms are mentioned. Only the idea and the syntax of the commands and its use is described. Sometimes an example is given which are not as clear as those in the manuals. Furthermore, it is not easy to find something specific, because there are several references to one item and each of the new opened pages contains another information. Sometimes it would be easier for the user if all information of one item is united in one page.

The next software package is made for exact nonparametric inference. "The goal of StatXact is to enable statisticans and data analysts to make reliable inferences by exact and Monte Carlo methods when their data are sparce, heaviliy tied, or skewed, and the accuracy of the corresponding large sample theory is in doubt. ... If a data set is too large for the exact alhorithms, StatXact computes Monte Carlo estimates of the exact p-values to any desired accuracy. If the data set is too large for both ..., it is almost certainly large enough for asymptotic theory to work accurately."(StatXact User Manual). In general this program doesn't expect missing values. There is only one command which assigns to an observation of a variable the number one, if this observation is missing and zero otherwise. It is always expecting complete data sets. Indeed the data can be incomplete. Missing values will be coded with a dot in numeral and alphanumeral variables as well. But incomplete observations will be excluded from all calculations.

StatXact allows to import data files created by other software programs. It can read in data that are in ASCII Data, BMDP Data, BMDP New system, EGRET Data, EXCEL Data, dBASE Data, LOTUS 1-2-3 Data, SAS Transport

Data, SPSS Data, STATISTICA Data or SYSTAT Data format. Missing values will be recognized and recoded into the StatXact code. It is not planed to change any variable while reading in the data. Of course, it is possible to transform variables afterwards. Furthermore, the user cannot define missing values neither one or more values nor a whole area of values. As mentioned above, StatXact is a program for inferences. Therefore graphics cannot be made, but the results can be stored in a file in a format that other programs can create a graph out of it. In tables the number of excluded observations is never mentioned. Only the number of used observations is mostly said in the output. The calculation of descriptive statistics for incomplete data sets is no problem for StatXact, missing values will be omitted. If descriptive statistics will be calculated the option 'count' carries out the number of included cases. This program offers no test on MCAR, but many other in which the available case analysis is used. Other statistics like regression analysis, cluster analysis a.s.o. are not offered. This software package has own commands to read in data, to store results or to transform variables, but it is not possible to write procedures. The syntax is easy to learn by the manuals.

The manual *StatXact For Windows: User Manual* is additionally stored in a file. If the user uses the index to find something about missing values he will be refered to one place in which only the treatment in the data editor is described. Besides, while reading the book one may find some explications about how StatXact deals with missing values in descriptive statistics and variable transformations. The manual discusses the theory extensively sometimes more than statistical literature. Examples are given to make it clearer. References are given but only at the end of the book and none at the end of each chapter. Caused by the sparce abilities to deal with missing values this problem is hardly mentioned. The online help is actually no help concerning missing values, because the search for these words leads to no result. In general the online help explains the use of the dialog boxes and the offered options. The search for algorithms, statistical theory or even literature fails. Furthermore, examples to the commands are not to find. All in all it is easy to navigate.

The following software package offers only exact methods for binary logistic regression analysis. LogXact is made from the same corporation which programmed StatXact. It "performs unconditional maximum likelihood inference, conditional maximum likelihood inference, and conditional exact on the parameters of the logistic regression model." (LogXact User Manual). In case of larger data sets asymptotical methods can be elected. Usually missing values are coded with a dot equal if it is within a numeral or alphanumeral variable. It is interesting that this program in compare with StatXact is not able to read character variables from text files. It is also impossible to change the code while reading in the data. These must be done before or after importing. Each character in numeral variable is set as missing value. In addition to that, missing values in SYSTAT data files will not be recognized or recoded to the system missing value code of LogXact. This program provides to import data set from ASCII Data, BMDP Portable Data, BMDP New System Data, EGRET Data, SAS Transport Data, SPSS PC+ Data and SYSTAT Data files.

In the global options dialog box the user can define a number within the interval $[-1 * 10^{25}; 1 * 10^{25}]$ as missing value code. This definition is valid in the whole data set. It is not allowed to define more than one value or even an

area of values. If a data set is imported after defining this missing value code all values with this value will come in as a missing value.

In tables of the output the number of used observations is outlined. Graphics can not be made. As mentioned above only logistic regression analysis can be performed. Therefore other analysis will not be treated. In fact not even descriptive statistics are offerd. Only a cross tabulation can be performed. In each case LogXact uses the available case analysis to deal with incomplete data.

As StatXact, LogXact has some commands to carry out some data management actions, but it is not possible to create own procedures.
The manual *LogXact User Manual* explaines the statistical theory as extensive as that of StatXact. Some examples help to make it more clear. Algotithms are not mentioned. Furthermore, references can be found in the end of the manual. The online help is as well as that of StatXact. The only difference between them is the fact that LogXact's online help knows the expression 'missing value'. The user will then be refered to the missing value codes which can be elected. A disadvantage are the hardly given examples.

Finally the software package JMP in its version 3.15 was considered. JMP is a program from SAS Institute Inc. and was produced for analysing smaller data sets. It offers the main methods for graphic plots and inductive statistics. For extensive analysis of the data the manual recommends to use SAS. The advantage of JMP compared to SAS is easy handling and easy learning, because of the comfortable dialog boxes. On the other hand it has no programming language to write procedures.

JMP codes missing values in numeral variables with a question mark. In alphanumeral variables it uses a blank. The narrow connection to SAS can be seen if one looks for external files to import. Only text or SAS transport files can be read in. In text files the type of a variable will be recognized by the first row. In numeral variables all nonnumeral signs will be set as missing value. Only blanks will not be noticed. In this case the value of the next column will be taken. Therefore the whole row is moved and the last variables have missings. In alphanumeral values only blanks will be set as missing values. JMP offers no tool to recode values while reading in data. In addition to that, it is impossible to change the code for missing values or to define several values or an area of values as missing code.

In frequency tables an own row for missing values is printed. In cross tabulations no category for missing values is given. The number of excluded observations must be calculated from the difference of the number of all observations and the included. In graphics missing values will be omitted without a message. A test on MCAR is not offered.

For calculations of descriptive statistics JMP omits missing values. If correlations and covariances are called the user can choose between complete case and available case analysis.

JMP excludes all incomplete observations concerning the interesting variables from the calculations of tests or confidence intervals. Regression, cluster and discriminant analysis and analysis of variance use the available case analysis for dealing with incomplete data sets as well. In time series only observed values are plotted and connected with a straight line.

The considered literature is resticted to three manuals: the *Introductory Guide* which "is a collection of tutorials designed to help " learning JMP strate-

Table 4.1: Explanation of abbreviations used in table 4.2

c	complete case analysis (listwise or case wise deletion)
a	available case analysis (pairwise deletion)
cc	change of missing value code possible
sv	several values can be defined as code for missing values
ar	an area of values can be defined as code for misssing values
nCat	a category for missing values will be created
lInt	linear Interpolation
lExt	linear Extrapolation
mI	mean imputation
wmI	weighted mean imputation
regI	regression imputation
MedoN	median of $2N$ neighbours imputation
MoN	mean of $2N$ neighbours imputation
MC	Monte Carlo methods
Kalman	methods based on Kalman filter
cSpl	cubic splines
DWLS	distance weighted least squares interpolation

gies. The *User's Guide* which has a "complete documentation of all JMP menus, an explanation of data manipulation, and a description of the calculator." The *Statistics and Graphics Guide* "documents statistical platforms, discusses statistical methods, and describes all report windows and options". The statistical theory is described and explained by clear examples. The interested reader may find further literatur in the references at the end of the *Statistics and Graphics Guide*. Algorithms are not mentioned.
The online help gives no answers to algorithms. The statistical theory is not described. Sometimes the idea is mentioned. The user will not find any literature. Fortunately the use of the online help is rather clear and comfortable. On the other hand only a few topics are given to missing values. But this is caused by the few possibilities JMP offers for treating this problem.

item	MINITAB	STATISTICA	SYSTAT	Stata	S-PLUS	SPSS	SAS	StatXact	LogXact	JMP
1 (a)	*	-9999	.	.	NA	.	.	.	.	?
(b)	blank	-9999	blank	blank		blank	blank	.	.	blank
(c)	*	-9999			NA	.	.			
2	yes	no	no	no	yes	no	yes	no	no	no
3	sv, ar	cc	no	no	cc	sv,ar ar+value	sv	no	cc[18]	no
4 (a)	c,nCat	c,nCat	c,nCat	c	c,nCat,refuse	c[13], nCat[14]	c,nCat[17]	c		nCat
(b)	c,nCat[2]	a	c	c	c,refuse	c,nCat, a,lInt	c,nCat			c
5	no	no	yes	no	no	yes[15]	no	no	no	no
6	yes	yes	yes	yes	yes	yes	yes	yes		yes
7	a	c,a	c,a, EM[8]	c,a	c,a, refuse	c,a, EM[15],regI[15]	c,a	c		c,a
8	c[3]	a	c[9]	c,regI,lInt,lExt	c[3]	c,a	c	a,MC		c
9 (a)	c	c,a[6],mI,wmI[7]	c	c,regI,lInt,lExt	c,mI,refuse	c,a,mI	c		a	a
(b)	c[4]	c	c	c,regI,lInt,lExt	c,mI,refuse	c,a	c			a
(c)	c[5]	c,mI,wmI)[7]	c	c,regI,lInt,lExt	refuse	c,a	c			a
(d)	c	c,mI,wmI[7]	c		refuse	c,mI	c			a
(e)	lInt	mI,lInt,regI, MoN,MedoN	c,DWLS	c,regI,lInt,lExt	c,refuse[11], Kalman[12]	c,stop[11]	c,lInt,lExt, stop[11],cSpl			a
10	yes	yes	no[10]	yes	yes	yes	yes	no	no	no
11[1](a)	+ / +	− / −	+ / 0	+ / −	0 / 0	0[16] / −	0 / −	− / −	− / −	− / −
(b)	+ / +	0 / +	0 / 0	+ / −	0 / +	+ / −	++ / −	++ / −	++ / −	+ / -
(c)	yes / no	no / yes	yes / no	yes / no	yes / yes	yes / no	yes / no	yes / no	yes / yes	yes / no
(d)	++ / ++	++ / 0	++ / +	++ / −	+ / 0	++ / ++	+ / 0	++ / -	++ / -	++ / +

Table 4.2: Comparison of statistical software packages. For explanation of the items and abbreviations see table 3.1 and 4.1 respectively.

5 Summary

This survey has shown that the missing value problem is treated very differently even in this small selection of statistical software packages. Some of the smaller programs as JMP, StatXact and LogXact have no idea to deal with missing values, except to omit incomplete observations. Here is not even a progamming language offered to remove the lack of methods by the user. LogXact enables only to change the missing value code. JMP merely creates a new category for missing values in tables. LogXact and StatXact allow to choose the code for missing values out of star, dot and question mark while exporting a data set. SYSTAT is somehow more comfortable than these three, because here the user has sometimes a choice how incomplete data shall be handeled. Amazingly, it offers an EM-Algorithm for calculating correlation matrices, but easier methods like imputation methods are not at all considered. Additionally, the programming of macros is not provided. Therefore it is not possible to program any of the missing methods. The explanation of the EM-Algorithm is quite extensive and SYSTAT is able to carry out a test on MCAR what is not at all taken for granted. Only SPSS offers this test in its additional MVA module. Apart from that, the statistical background is only mentioned in the manuals in a short way. All the other software packages have a programming language on its disposal to write macros and remove the lack of methods for missing value treatment. The more popular program MINITAB enables the user to define his own missing value codes, either as several discrete values or as an area of values. In addition to that, SPSS offers the user to define an area and one discrete value. MINITAB offers only complete case and available case analysis and in time series linear interpolation to handle incomplete data. References to literature which deals with missing values are not given. SAS is compareable with MINITAB. In most cases the complete case analysis is used, but in time series several methods are offered and the statistical background is explained very extensively in the manuals. S-PLUS offers mean imputation to fill up incomplete data what is necessary, because quite often actions will be refused because of missing values. This program is the only one which offers a method based on Kalman filter to deal with missing values in time series. Furthermore, it has many functions for missing values, but no examples are given and no information about the treatment of incomplete data sets is available in the online help. STATISTICA and Stata

[1] manuals / online help
[2] only in categorical variables in high-resoluted graphs
[3] χ^2-test for independence refuses incomplete variables.
[4] if the two-way-ANOVA design changes to unbalanced an error message arises.
[5] c if variables will be clustered. If observations will be clustered MV are not allowed.
[6] in multiple regression
[7] this action must be carried out with the data management before
[8] for metric variables
[9] χ^2-test for independence enables to include missing values in a new category (optional)
[10] possibility to write programs
[11] if missing values are in the center of the time series
[12] in ARIMA-models
[13] only in line plots
[14] user defined missings are listed separately
[15] with the MVA-module
[16] in a special book about numerical algorithms
[17] it can be decided whether missing values are included in calculations of statistics or not
[18] the change of the code is valid for all variables

are also compareable in its abilities to deal with missing values. Both have several imputation methods in addition to the complete case and the available case analysis. On the other hand it is impossible to define several values or an area of values as missing value code. Only STATISTICA allows to change the system missing code for each variable and the user may decide whether a new category for missing values in tables shall be created or not. Besides Stata enables imputation methods in calculations of tests and confidence intervals but offers no discriminant analysis. The statistical theory of the methods is explained in a rather short way or missing. SPSS without its additional MVA module has many procedures to treat incomplete data. Several imputation methods are offered which must in most cases be carried out by data transformation before applying statistical analysis. The user can also define codes for missing values in more than one way. With the MVA module, which is especially for incomplete data sets, SPSS allows to examine the missing value pattern and the system which possibly causes the nonresponding. Therefore it is the most comfortable program in this selection.

To sum up one must say that the problem of missing values is recognized, but considered in a different intensity within the examined software packages. The perfect program was not in this survey. Hence, it should be an incentive for all to enlarge the abilities for the missing value problem.

6 Bibliography

D. Altman (1997): *Practical Statistics for Medical Research*, Chapman & Hall, Weinheim.

K. Backhaus, R. Erichson, W. Plinke, R. Weiber (1996): *Multivariate Analysemethoden*, Springer Verlag, Berlin, Heidelberg.

G. Bamberg, F. Baur (1991): *Statistik*, Oldenburg Verlag 7.Auflage.

Bankhofer, Hilbert (1997): Statistical Software Packages for Windows: A Market Survey, *Statistical Papers* **38**: 393-407.

G. Brosius, F. Brosius (1998): *SPSS Base System und Professional Statistics, Fuer die Versionen 5.x und 6.x*, Bonn: International Thomson Publishing 2. Auflage.

A. Buehl, P. Zoepfel (1995): *SPSS fuer Windows 6.1, Praxisorientierte Einfuehrung in die moderne Datenanalyse*, Bonn: Addison-Wesley 2. Aufalge.

J.M. Chambers, T.J. Hastie (1992): *Statistic Models in S*, Wadsworth and Brooks Cole.

CYTEL Software Corporation (1996): *LogXact For Windows: User Manual.*

CYTEL Software Corporation (1996): *StatXact 4 For Windows: User Manual.*

T.R. Dawber (1980): *The Framingham Study: The Epidemiology of Atherosclerotic Disease*, Harvard University Press, Boston.

FDA, Food and Drug Administration, *2867fnl.pdf*, www.fda.gov/cder/guidance.

A. Fieger, H. Toutenburg (1994): *SPSS (fuer Windows) Tables: Arbeitsbuch fuer Praktiker*, Muenchen: Prentice Hall.

A. Fieger, H. Toutenburg (1995): *SPSS Trends fuer Windows: Arbeitsbuch fuer Praktiker*, Muenchen: Prentice Hall.

T. Hahl, R. Shelton: *Dropping Variables That Have Only Missing Values*, Observations Vol. 5, No. 4, The SAS Institute, v5n20pp1.html .

J. Hartung, B. Elpelt (1992): *Multivariate Statistik*, Oldenburg Verlag, Muenchen.

H. Kahn, C. Sempos (1989): *Statistical Methods in Epidemiology*, Oxford.

A. Krause (1997): *Einfuehrung in S und S-PLUS*, Springer Verlag.

R. Little, D. Rubin (1987): *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

S-PLUS 4, Guide to Statistics, Mathsoft (1997).

S-PLUS, User's Guide, Version 4.0, Mathsoft (1997).

S-PLUS, Programmer's Guide, Version 4.0, Mathsoft (1997).

MINITAB User's Guide, Release 11 for Windows (1996).

MINITAB Reference Manual, Release 11 for Windows (1996).

C. R. Rao, H. Toutenburg (1999): *Linear Models: Least Squares and Alternatives*, Springer Verlag, New York.

SAS Institute (1995): *JMP Version 3.1 Introductory Guide*, SAS Institute Inc.

SAS Institute (1995): *JMP Version 3.1 User's Guide*, SAS Institute Inc.

SAS Institute (1995): *JMP Version 3.1 Statistics and Graphics Guide*, SAS Institute Inc.

SAS Institute (1990): *SAS/STAT User's Guide, Version 6*, SAS Institute Inc.

SAS Institute (1990): *SAS/GRAPH Software: Reference, Version 6*, SAS Institute Inc.

SPSS Inc. (1993): *SPSS Base System User's Guide Release 6.0*, Mary Ann Hill: SPSS Inc.

SPSS Inc. (1997): *SPSS Missing Data Analysis 7.5*, Mary Ann Hill: SPSS Inc.

SPSS (1996): *SYSTAT 6.0 for Windows: Data.*

SPSS (1996): *SYSTAT 6.0 for Windows: Graphics.*

SPSS (1996): *SYSTAT for Windows: Statistics.*

SPSS (1997): *SYSTAT 7.0 for Windows: New Statistics.*

SPSS (1997): *SYSTAT 7.0 for Windows: Command Reference.*

Getting started with Stata for Windows, Stata Press (1999), College Station, Texas.

Stata Graphics Manual Release 6, Stata Press (1999), College Station, Texas.

Stata User's Guide Release 6, Stata Press (1999), College Station, Texas.

Stata Reference Manual Release 6, Volume 1-4, Stata Press (1999), College Station, Texas.

StatSoft (1997): *STATISTICA Benutzerhandbuch.*

H. Toutenburg (1992): *Lineare Modelle*, Physica-Verlag, Heidelberg.

H. Toutenburg, A. Fieger, Ch. Kastner (1998): *Deskriptive Statistik fuer Betriebs- und Volkswirte, Eine Einfuehrung in SPSS fuer Windows*, Muenchen: Prentice Hall.

H. Toutenburg, A. Fieger, Ch. Kastner (1995): *Induktive Statistik fuer Betriebs- und Volkswirte, Eine Einfuehrung in SPSS fuer Windows*, Muenchen: Prentice Hall.