Historical Software Issue 3: MINITAB

Thaller, Manfred

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

software

HISTORICAL SOFTWARE SECTION[+]

"Using EDP" means for most of the readers of this newsletter probably using SPSS. As long as one remains in the realm of methods and sources similiar to survey research, the possibilities presented by that package will usually suit the arising needs quite well. Additionally, SPSS has the great advantage that it has meanwhile developed into a kind of international standard. Therefore packages like CLUSTAN (for Clusteranalysis) or GRADAP (for the analysis of social networks) have either been designed from the very beginning to have a command language which is more or less an extension of the one employed by SPSS (as in the case of GRADAP) or allow at least to access SPSS system files.
Being that universal, the package has so many things, which recommend it, that they should not be repeated here. The valuable features of the system are probably obvious to most historians, which ever have used it - its shortcommings on the other hand may seem to be rather obscure for the casual user of EDP. So when in this issue we will seriously recommend the use of another statistical programming system to historians we have first of all to make clear, that this system, MINITAB, is and cannot be seen as a replacement to SPSS in the full width of applications the latter one is useful for. This is particularly true, if you are using more complex statistical methods, as e.g. path analysis, factor analysis or discriminant analysis. Routines for these purposes are included into SPSS, while they are missing in MINITAB. Just: how many quantitative studies are there which actually reach this level of statistical complexity, using SPSS beyond the CROSSTABS, BREAKDOWN and REGRESSION procedures?

Another of the attractions of SPSS is the high quality of output it produces. Version 9 - at the present accessible, though, only for a small minority of our members - will even contain possibilities for plots in four colors. While this makes computing comfortable, most historians will have to redraw their tables before they get them published. So the nice outlook of SPSS tables is useful, but not as indispensable as a real shortcut to printing would be. (As regards plotting: SPSS, not being famous for its cheap licences, reached this stage by a marriage with the software distributed by DISSPLA being almost anything but cheap. As a result the plotting options offered

---

[+]Address all communications to: Manfred Thaller, Max-Planck-Institut für Geschichte, Hermann-Föge-Weg 11, D-3400 Göttingen.

with release 9 have a price in allprobability prohibitive for many of at least the smaller computing centres.)
A third andpotentially the greatest attraction of SPSS is its universal availability. This is an undisputed fact - with an important restriction. As everyone knows hardware is becoming cheaper. Indeed mini computers at a price of about 30.000 DM which e.g. are based upon an LSI 11 and running under one of the PDP 11 operating systems can provide all the support for statistical computing nowadays a historian is used to get from his/her university computing centre. And, while at the present economic situation it would not make sense to talk about them being potentially within the financial limits of most institutes doing historical research, quite a few such systems are installed meanwhile by institutions like social science faculties. SPSS provides a version (SPSS-11(1)) for this family of computers which extends right thru to the VAX. Unfortunately this version, while very similiar in its control language to the familiar "large systems" SPSS has quite a few highly significant differences to the main dialect. And, whats even worse: this "small systems" version is in itself rooted very deeply in the structure of the 11 family, being therefore scarcely transportable to other small machines.
A feature of SPSS that ist usually considered very helpful by people wanting to make use of a computer, while not becoming EDP experts, is that the command language shields almost completely from you, what is going on in the machine. This very highlight of the system makes it somewhat dangerous though: you are seduced to arrange for results where you dont know, how they are reached. This is particularly important, when quantification shall not remain a very advanced knowledge of the fortunate few, but a tool that can be taught along with other ones in seminaries and classes within the normal curriculum of a university. Indeed, most people which have tried to teach historians statistics soundly, introducing them into things like the regression coefficient by teaching them how to compute one abhorr the seemingly unavoidable situation that their students do their first regression equation by a pocket calculator and than, when the data become more complex and therefore interesting, forget what they just have learned as it seems just to be necessarry to push a few SPSS buttons to get the "correct" results(2).
Just mentioning a final point: SPSS is excellent, if you have a large number of data, which you are going to analyze for a long time. It falls rather short in performance, though, if you just want to use the machine as a slightly more sophisticated pocket calculator, typing in various short datasets and analyzing them on the spot. (Either you have to punch your data out after every recoding operation, or you will have to keep track of an ever increasing number of system files - very many computing centres restricting severely the number of files one may simultaneously maintain, not a behaviour to be recommended.) SPSS offers a conversational system (called SCSS) which can overcome those difficulties but it is unfortunately priced as SPSS (to all appearance getting more and more used to "rich" commercial customers) is used to - and therefore much less commonly available as SPSS proper.

One might point out other deficiencies of SPSS. The ones we have

mentioned that far are overcome by the package that shall be presented here, i.e. by MINITAB(3).

MINITAB:

- has a command language that is certainly not more complicated than the one employed by SPSS,
- is a genuine conversational system,
- is with many respects more flexible in producing "simple" statistics,
- has been prepared to run, without any change apparent to the user and with minimal effort in implementation on practically all machines from very small minis right through to the mainframes of computing centres,
- can be used to administrate several small datasets concurrently,
- is pretty transparent. You can use it as a straight forward expansion to the pocket calculator of your statistics class as well as a rather sophisticated programm package,
- has a number of procedures for time series analysis and Exploratory Data Analysis (EDA) which are missing in SPSS or only available from version 9 onward. (Telefunken being out of business one probably can say quite bluntly that those procedures will very late, if ever, become available to those of you who have to use TR-440ies.)

There are shortcomings:

- the range of analytical procedures is much more restricted than with SPSS,
- the size of the system files one may sensibly work with is severely limited. (on most installations you will be restricted to less than a hundred variables and a couple of hundred cases. This, as should be emphasized, is an installation decision, not a limitation necessarily inherent in the package. If your relations to the staff of your computing centre are sufficiently good, they may be ready to generate a special "large-scale" version for a research project demonstrating reasonable interest in such a version. Still, even if you can not arrange for that, it can pay off to use MINITAB for preparatory analysis within large scale projects, using SPSS only in the later stages. This is particularly true, if the turnaround time of your computing centre is longer than about two hours during day time for batch jobs.),
- while you can "name" your variables, those naming comes much less natural than with SPSS,
- there are no variable and value labels. If you have to interrupt an ongoing project for a couple of months and want to take up things where you left them some months ago, you will be very badly surprised how unreadable your output has become meanwhile.

That much should be said to explain, why you should be interested

in MINITAB. The following presentation tries to summarize its capabilities, based on and following closely the reference manual as of October 1st,1980. That manual, while not as broad and no introduction into statistics as the one available for SPSS, is in a style understandable by everybody who has understood that counterpart.

1. MINITAB commands consist generally of a word starting in column 1 of a line ("card") that can be abbreviated to 4 characters. The specification consists of a number of arguments following, which can be explained by mnemonic text interspersed freely between them. The command lines

   DIVIDE the data in C3 by 3, store the result in C3

   and

   DIVI C3 3, C3

   are equivalent therefore.

2. The basic entity of MINITAB is a "column of data" (representing a variable). Additionally the user can specify stored constants and matrices. The later feature ist particularly useful for somebody trying to work with a large number of very small tables, as they may be administrated within the same system file, though having different numbers of rows and/or columns.

3. "Columns" are denoted by C1, C2, C3..., constants by K1, K2, K3.... and matrices by M1, M2, M3 ..... With all three types, the maximal number of items allowed is installation dependent. Columns (and only columns) can be assigned mnemonic names with the help of a command like:

   NAME for C6 is 'STATUS', for C8 is 'INCOME'.

   After this or a similiar command is given, column numbers and symbolic names can be used interchangeably. So

   PLOT variable in C6 against variable in C8

   and

   PLOT 'STATUS' against 'INCOME'

   would be equivalent.

4. Missing values have - unfortunately - to be recoded into one internal value, accesible for the user as '+'. This is one of the definite shortcomings of MINITAB. It is not possible to access the value of a variable it had before it has been declared as missing!

5. The user can enter as many columns, matrices and stored constants

concurrently as the limit of the respective installation allows. If that limit is reached - or some columns are too long, or some matrices too big - in MINITAB terminology "the worksheet overflows" - he can delete any combination of those items he does not need any more to provide more space. This is the most serious shortcoming of the system: it requires all data to reside in core simultaneously.

6. The current state of the users "worksheets", i.e., the combination of columns, matrices and constants in use at a given time can be stored for further use (as a kind of systemfile).

7. MINITAB was designed as a conversational system. It differentiates between the severity of errors encountered. When you use it interactively, it will present you with an analysis of any errors encountered in your command for immediate correction. If you use it as batch system, it will terminate execution only when really severe errors are encountered.

8. The system has a rather good help facility. Theoretically, somebody with any previous exposure to statistical computing should be able to learn the use of the system by activating it and typing HELP as his first command, as by that he will prompt very detailed explanatory modules.

9. Input can be a mixture of fixed field and free field. The system defaults to free field. The command

   READ THE FOLLOWING DATA INTO COLUMNS C2, AND C3

   or

   READ C2, C3

   would read

```
   2 4
   3.5   27.0    9
 1 2
```

   into the columns as follows:
   2, 3.5 and 1 are put into C2, 4, 27 and 2 are put into C3. 9 is ignored as not provided for in the read command. For formatted input the FORTRAN E, F, A, X and T formats can be used together with the slash. Data can be entered either from the main input medium (i.e., be typed in via video terminal) or from any file of the computing centre.
   Any subset of the worksheet (i.e., any combination of rows and columns, or of "cases" and "variables" can be entered and/or overwritten selectively. Parts of the input data can be "read" repeatedly in free field input. (1972, 1973, 1974)2 is e.g. equivalent to 1972, 1972, 1973, 1973, 1974, 1974.

10. Facilities for printing and punching data exist as well as for

tapehandling from within the system. (Necessarily heavily installation dependent.)

11. Histograms can be drawn and variables can be plotted against each other. You have more control over your histograms and plots than with SPSS FREQUENCIES and SCATTERGRAM, though the output produced by SPSS will, while generally less informative, look much better due to the labeling features.

12. Besides providing a SAMPLE command, MINITAB lets you choose between 7 random number generators - what makes it very handy if you get somewhat more intimately involved with any probabilistic estimates.

13. With the exception of MOD10, RND and TRUNC (which I personally miss considerably) you can use the same functions for the construction of arithmetic expressions as with SPSS. (I.e. for the logical equivalences of COMPUTE, COUNT and IF.) What is a real highlight of MINITAB, though, are functions called "column operations" which usually return some basic distributional prameter of a variable.

    LET 'ST.SCORE'=('SCORE'-AVER('SCORE'))/STAN('SCORE')

    is a command which by means of the column operations AVERage() and STANdardeviation() assigns a standardized score to a newly created variable. If you ever have produced an SPSS output just to feed the mean back in, you know what that feature is good for.
    Whats even more attractive from the point of view someone using advanced methods of contingency analysis will take: the various arithmetic commands are able to perform matrix algebra.

14. Fully developed recoding facilities exist.

15. The most general descriptive statistics (mean, standard deviation etc.) can be called for by the same mnemonic you use within the arithmetic expressions.

16. Procedures for the T-test and the other conventional tests are provided, along with several commands which can prepare other tests which are not yet implemented, but can be computed with the help of the arithmetic commands included in the package.

18. Analysis of Variance is supported.

19. Correlation coefficients can be computed.

20. While for relationships between non-intervall scaled variables few things are provided (mainly the Chi Square Test), the possibilities for tabulation are really excellent and surpass SPSS considerably. Unlike in SPSS you can fill the cells of the tables you create not only with the absolute frequencies and the main percentages, but with all the basic parameters of the distributions of all variables in the dataset, together with the possibility to list the values of a set of variables for all the cases, defining a given cell, within that cell. There is a limitation: the variables defining the table, called "classi-

fication variables" within MINITAB, have to be between -1000 and +1000, which at least in some cases may lead to rather cumbersome recoding operations.

21. While not very ample in this respect, MINITAB provides some basic statistics - and plotting possibilities - for time series analysis. Moving Averages, autocorrelation and crosscorrelations are provided, together with possibilities to fit non-seasonal and seasonal models. Very interesting is the lagging facility introduced together with those analytic procedures, being more comfortable than the one SPSS offers.

22. Two of the Exploratory Data Analysis methods for the display of data are included the stem-and-leaf display and the box-and-whisker plot. While I admit, that I see both of them as somewhat artificial and personally prefer using more conventional means of descriptive statistics for exploratory purposes, personal conservatism should not prevent one to point to the potential of MINITAB in a certainly expanding field of statistics(4).

23. The control language of MINITAB includes a looping facility. Similar to the DO REPEAT / END REPEAT of SPSS (as far as conversational systems are similar to batch ones), it is definitely less powerful.

FOOTNOTES

1 Norman H. NIE et al.: SPSS-11. The SPSS Batch System for the DEC PDP-11, New York etc., 1980.

2 B. PAGE and H. SCHWITTERS: Erfahrungen mit dem "didaktischen" Programmpaket MINITAB II in der Statistikausbildung, in: EDV in Medizin und Biologie, 11 (1980), pp. 44-52; Michael PROCTER in: SSRC Survey Archive Bulletin, 20(Sptember 1981) p. S1.

3 The package is described in T.A. Ryan, Jr. et al.: MINITAB Student Handbook, North Scituate, Mass.: Duxbury Press, 1976. The development is rather active, so one should try to get the latest available version of the more general reference manual, which is distributed together with the release tape in machine readable form. If the system is not yet at your site, manual, information, and the system itself are available from: MINITAB Project, Statistics Department, 215 Pond Laboratory, Penn. State University, University Park, Pa. 16802, USA. (Information about the fee are somewhat ambiguous 500 and 750 US $ are quoted for the release tape.)

4 EDA not yet being particularly well known among German historians, a few references may be given:
J. W. TUKEY: Exploratory Data Analysis, Reading, 1977 might become the classic of the field;
Frederick HARTWIG, Brian E. DEARING: Exploratory Data Analysis, Beverly Hills etc., 1979 (= Sage University Papers, Quantitative Applications in the Social Sciences 07-016) is an introduction, cheap and easy to read. The manual itself hints to Paul VELLEMAN and David HOAGLIN as publishing the algorithms used by MINITAB. So their "ABCs of EDA" (North Scituate, Mass: Duxbury Press, 1981) should be recommended for prospective users of MINITABS EDA facilities. I have not seen it that far, though, and cannot comment upon it.