CONTENTS

1. Introduction
2. Available packages
3. Statistical Libraries
4. Old generation packages
5. New generation packages
6. Interfaces
7. Conclusion
1. Introduction

In this research report we will consider the problems of using statistical packages seen by the point of view of a social science researcher.

Use of statistical packages falls into two parts, namely first the administration of the data next to the statistical analysis. Very often the first part is omitted in the literature, but the fact is that the main part of the invested time, very well can be used concerning heavily data problems. One of the reasons can be to read the data in the right form. It is often necessary to transform data to a new filesystem before the statistical analysis. It can also be a specific subgroup of the data which is wanted for further analysis, or that two datafiles have to be merged before the statistical analysis.

It is obviously of great importance to have a good contact between data administration and the statistical procedures. We will consider these administration problems as well as the statistical procedures in the packages.

The diversity in the problems of statistical computing comes not only from the diversity of tasks addressed, but also from the diversity of persons. By "statistics" may be meant organization, presentation or evaluation of statistical data, or study of random processes, or the theory of decision-making under uncertainty, or many other kinds of mathematical theory considered to be statistical. The persons who may seek to use a computer for some statistical purpose vary greatly in their interests in, and knowledge of, both statistics and computing.
Two very different sorts of persons, in this respect, are:

1. The professional statistical theorist, interested in developing new ideas or methods, interested in particular statistical problems, mainly as examples of general ones. He demands if he comes to the computer, to understand fully what is going on, preferably his own programmer, so that he may do what he wishes rather than what someone else has wished.

2. The scientist in a field that yields statistical problems, himself not primarily a statistician, who wants to use good statistical tools to better his understanding of his field, but has no interest in statistical methods for their own sake, nor desires to experiment with them, and he prefers to be concerned as little as possible with the technicality of statistical calculations.

This work is concerned with the second group rather than the first, e.g. a researcher in the social science who wants to carry out a statistical analysis on his data easily.
2. Available packages

In this research report we will only consider packages available on the computer centres NEUCC (North European University Computing Center) and RECKU (Regnecentret ved Københavns Universitet). This is however not a real limitation, because the two centres have a wide range of programmes for statistical purpose.

A. NEUCC

- BMDP statistical package
- SPSS statistical package
- OSIRIS statistical package
- SAS statistical package
- GENSTAT statistical package
- GLIM statistical package
- HARWELL mathematical/statistical library
- IMSL mathematical/statistical library

B. RECKU

- BMDP statistical package
- GENSTAT statistical package
- GLIM statistical package
- SPSS statistical package
- NAG mathematical/statistical library

At NEUCC the installation is IBM 3033 and IBM 4341; at RECKU a UNIVAC system is used. Due to the UNIVAC installation, SAS is not available at RECKU because SAS demands an IBM compatible installation, which is not fulfilled by UNIVAC.

However, users at RECKU can start jobs on NEUCC and get prints due to a transmissionline between the two centres. This possibility is very often used by users who do not have direct access to NEUCC, and it has the advantage that SAS can be used more or less as a package under the RECKU system.
3. Statistical Libraries

NAG, HARWELL and IMSL are mathematical statistical libraries.

A library is a set of procedures or subroutines which user
programmes may access by means of the calling method of a high-
level language. A package is a single complete programme or
suite of programmes which is data driven, a logical distribution
being drawn between data records, which supply commands
and others which supply actual data values. A survey of all
packages can be found in Francis (1981). Packages are effective
because they provide easy access to a set of algorithms within
an appropriate data environment. A clear and easily
learned control language, good data management facilities,
sensible output formatting, and self explanatory error diag-
nostics are important to the package user. These items are
not fulfilled by libraries.

Libraries, by contrast are concerned primarily with algo-
rithms. The library user chooses a specific algorithm to
solve his problem. The library should enable the user to an
effective solution, by providing variants of the algorithm
in different circumstances.

The library is less suitable for performing the datamanage-
ment and output generation abilities of packages, and the
users should construct his own main programme.

IMSL and NAG have a significant statistical context;
HARWELL is rather sparse in this field, and is merely used
for mathematical and numeric calculations.

A. IMSL

IMSL (International Mathematical and Statistical Libraries)
contains about 400 Fortran subroutines addressing problems
in numerical analysis and statistics. The statistical contents
is about one half of the library and concentrates on simple
statistics, correlation, regression, nonparametric statistics,
time series, factor analysis and sampling. The programmes
cannot be used as subroutines in a public user specified pro-
cedure unless IMSL is implemented on the installation.
B. The NAG library

NAG (Numeric Algorithms Group Ltd.) contains 466 routines or more which perform tasks in numerical analysis and statistics. The library is available in Fortran and Algol and contains about 100 subroutines in statistics in the same field as IMSL.

The statistical programmes in IMSL and NAG are not used very often in the social science field, first because they demand that the user should write his own programmes to call the subroutines; second that the statistical packages are all superior to IMSL and NAG concerning statistical solutions. The two libraries are very useful in the mathematical field for numerical integration, eigen values, maximalisation of functions and more.

4. Description of old generation packages

If we take a closer look at the statistical packages, they can be divided into an old and a new generation.

To the older generation belongs:

Old

BMDP
SPSS
OSIRIS

and to the new belongs:

New

SAS
GENSTAT
GLIM

It is not unfair to say that the older generation is rather inflexible, compared to the newer. They provide a set of programmed procedures that can be used in relative isolation, rather than a set of tools, that can be combined in numerous ways to produce an appropriate analysis. It is possible to
transfer some results between procedures but often not particularly easy. The difference between the old and the new generations can also be stated in terms of the level of the language used. Generally, the higher the level, the easier it is to specify a small number of inflexible procedures. The introduction of flexibility, means lowering the level of language, which in turn seems to imply breaking the problem down into smaller components, thus giving the beginner a more difficult task. This can be omitted by describing the language to name a rich system of default settings. These defaults correspond to the most commonly needed requirements, so that the single common problem continues to have a simple specification. In other respects the language can and should contain the sort of facilities expected in a general purpose language. This is indeed not fulfilled by the old generation packages.

A. BMDP

BMDP is developed at the University of California. It can be used as a library of stand alone subroutines, and can by limitations be used as a package by a filesystem called BMDP SAVEFILES, which admits results from one subroutine to be used by another. The data to be read, have to be formed by FORTRAN statements. Many of the subroutines has a high international standard, especially for categorical data. The documentation is very careful, and the programmes has proved to be accurate.

The control language however gives some error messages which can be cryptic. The user is assumed to know statistics and the manual can be used as a reference, rather than an introductory manual.

The current version of BMDP includes a greater variety of state of the art methods for data analysis than the current version of SPSS, especially in such areas as data screening,
analysis of residuals, categorical data and graphical representation. The data management capabilities are rather primitive but this can be solved using SAS as a database manager as mentioned latter. The references in the BMDP manual indicate an awareness of numerical problems and recent work in numerical analysis. BMDP is essentially a series of excellent stand alone programmes with consistent overall control language and syntax. The weak point in BMDP is mainly the datamanagement.

B. SPSS

SPSS is developed on National Opinion Research Center of Chicago University. The package is still today probably the most widely used. The reasons are manyfold: The manual gives a beginner in statistical computing possibility in an easy way to run a programme on his data. SPSS requires generally less statistical background than BMDP, however some situations have shown that SPSS programmes are less accurate than BMDP, and also more computer time consuming due to creation of workfiles. SPSS has a much more powerful data documentation and management than BMDP. There is no doubt that SPSS is easy to learn and use as a system. The SPSS manual combines explanation of why a particular technique is useful with an introductory discussion of the statistical approach. There is however lack of documentation that combines programme information and statistical theory. Everyone can now run analyses of variance, regression, factor analyses and more on their data with very little knowledge of statistical theories and assumptions. At first, such unsophisticated users tend to accept any print that does not contain explicit error messages.

Statistical packages can give results, that can be disastrously misleading because of errors in data, outliers and the factors. Statistical models are used without really understanding of which assumptions are necessary, and SPSS is especially weak at these points. The social science researcher is often in a survey situation where cases easily can exceed thousands and variables in the hundreds, and SPSS claims to be an efficient package in this area, but most survey data are typically checked by other programmes before they reach SPSS,
e.g. SAS which have much better facilities for checking, editing cases, valid code ranges and logical consistency between questions.

Another point is that the social science researcher, caused by the survey data, is very interested in categorical data analysis. This means analysis of frequency tables by log linear models (and perhaps even latent structure analysis). SPSS is very weak here and is exceeded in every point by BMDP and SAS.

The concept of SPSS gives some weak facilities, in many situations the programmes can give erroneous results due to the numerical constructions, the statistical capability has now been surpassed by BMDP and the new generation packages.

C. OSIRIS

OSIRIS is developed by the University of Michigan, Institute for Social Research and the International University Consortium for Political and Social Research.

The package is not supported by the Danish computer centres, at NEUCC only an elder version OSIRIS III is implemented, and it is uncertain that the newest version will be installed. However there are some advantages worth mentioning.

DDA (Danish Data Archive) has build up a data library for the social science. The data sets are implemented in the OSIRIS dictionary file system, which is especially designed for numerous data. Recently OSIRIS has been extended by Zumapack which is not available in Denmark. SPSS and SAS have features to read OSIRIS dataset, and very often these two packages are used to analyze the data, because in the statistical sense they are superior to OSIRIS. The OSIRIS data format with a card deck containing the data and a dictionary giving all relevant information about the datastructure is however a nice way to archive data, and more it is a portable way to store data.
5. New generation packages

A. SAS

SAS (Statistical Analysis System) is a complete system for data analysis, which in a natural way combines information storage and retrieval, data modifications and programming, report writing, file handling and statistical analysis. The control language is rather advanced and it is close to PL/1 in the syntax. The system falls in two parts:

1. Data step

   This step performs all data operations. Reading of other datasets, cardfiles, a.o. computing new variables, subsetting, merging and updates new datasets.

2. The PROC step

   Once the SAS data set is created, any SAS procedure is available, and as described in the next section, any programme in BMDP, SPSS and OSIRIS can be used.

The SAS procedures are runned automatically and different options can be chosen corresponding to the procedure in question.

   The system is very easy to learn and SAS has a wide potential for statistical analysis. The data part is the most advanced found in statistical packages. As we shall see the other statistical packages can take advantage by this fact.

B. GENSTAT

GENSTAT (A General Statistics Programme) is an integrated statistical system and is evaluated at Rothamsted Experimental Station, England. It is based on a matrix-orientated language which means that matrices are natural datastructures which can be used for all computations. The database facilities are not easily accessed as apparent in SAS. GENSTAT generates a workspace area, where the users can copy files to implement the filesystem on the specific installation.
The statistical programmes in GENSTAT include generalized linear models and a wide range of other programmes. GENSTAT allows branching, loops and macros.

However, GENSTAT is rather difficult to learn due to a complex manual and the general language defined. It is obviously not dedicated to a social science researcher, but merely to a professional statistician, who can take advantage of the wide range of possible modelspecifications.

C. GLIM

GLIM (Generalised Linear Interactive Modelling) is also evaluated at Rothamsted and is dedicated for interactive use. GLIM is mainly used for constructions of complicated statistical models and have the same syntax as the GENSTAT package.

GLIM is a very valuable tool for statisticans, but it has the same complexity as GENSTAT, not suited for the social science.

6. Interfaces between packages

Most serious data analysers realize that no single system is apparent to be the best for all their needs. In fact, a single problem may sometimes call for more than one system, such as combining the data-management capabilities of one package with the statistical power of another. The most flexible thing to do would be a standard interchange file format for data definition and documentation, however no such system is apparent, and no package can claim to be the best in all senses.

Three criteria should be mentioned:

A. The algorithms should be reliable and welldocumented.

B. The systems shall allow the user to read any data-structure defined by the user, this means that the system should be open, thus allowing simple and free of maintenance linkage systems as well for the integration of user programmes.
C. There shall be homogeneity in the language and options between the programmes.

The next table shows how the packages fulfils these criterias:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPSS</td>
<td>-</td>
<td>-</td>
<td>YES</td>
</tr>
<tr>
<td>BMDP</td>
<td>YES</td>
<td>-</td>
<td>YES</td>
</tr>
<tr>
<td>SAS</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>GENSTAT/GLIM</td>
<td>YES</td>
<td>-</td>
<td>YES</td>
</tr>
</tbody>
</table>

A very impressive analysis of the facilities in statistical packages has been carried out by Francis (1981). One of the main results is that users of SAS, BMDP and GENSTAT were knowledgeable and critical, while users of SPSS do not seem to be so familiar with the package. SPSS users are under the impression that the package has been extensively tested, but the developer states that this is not the case and claims only that internal, unpublished tests have been performed.

A conclusion must be that neither OSIRIS, SPSS or BMDP can be used as central statistical systems. However each of the packages has advantages in specific fields as mentioned in the previous chapters. A common data interface would be a powerful tool, but the producers of statistical systems seem apparently to avoid this.

What is really important is to reduce the technical problems of users, which frequently have to work simultaneously with different statistical systems, in order to fulfil the needs of proper data analysis. The basic idea of the concept is to call from a central analysis system all other systems and stand alone programmes. Furthermore, this system shall allow easy access to the data base system and high level data management, even with interactive full screen editing, merging among others. The full screen editing facility now supported by SAS/FSP.

SAS is obvious the one and only system which has the linking possibilities to the other packages. Beutel (1980) has evaluated some important linking programmes, which I will repeat here. They extends the already built in functions in SAS. The following figure will show how SAS fulfils the linking concept.
Figure 1 shows the linking capabilities of SAS
1. Link from SPSS to SAS

The procedure CONVERT converts SPSS system files to a SAS dataset: This example shows how easy it can be done:

```sas
// EXEC SAS
// FILEIN DD DSN = SPSS. SAVE. FILE. DISP=SHR
PROC CONVERT SPSS = FILEIN.
```

2. Link from SAS to SPSS

The procedure SPSS generates SPSS system files and SPSS can be used directly under SAS. The procedures SPSS and CONVERT gives completely two-way communication between SPSS and SAS.

Example:

```sas
// EXEC SAS

SAS control cards

PROC SPSS;
PARMCARDS;

SPSS control cards

Back to SAS if requested.
```

3. Link from BMDP to SAS

The procedure CONVERT converts BMDP save files to SAS data sets:

```sas
// EXEC SAS
// SAVEFILE DD DSN = BMDP, DISP = SMR
PROC CONVERT BMDP = SAVEFILE;
```
4. Link from SAS to BMDP

The procedure BMDP can be used under SAS to analyse SAS data sets with BMDP programmes:

```sas
// EXEC SAS BMDP
PROC BMDP PROG = BMDP3F DATA = BASECOUNT:

PARMCARDS;
.
.
BMDP control cards
.
```

3 and 4 gives completely communications between BMDP and SAS.

5. Link from OSIRIS to SAS

The procedure CONVERT converts OSIRIS dictionary files to SAS data sets:

```sas
// EXEC SAS
// DICTIN DD DSN = OSIRIS. DICT, DISP = SHR
// DATAIN DD DSN = OSIRIS. DATA, DISP = SHR

DATASET, DISP = SHR

PROC CONVERT OSIRIS = DATAIN DICT = DICTIN;
```

Karsten Boye Rasmussen, Danish Data Archive (DDA) has evaluated a SAS programme which generates SAS statement to read the information from the codebook. This gives the advantage that not only the data but also labels, categories for the Proc Format statements and missing values are transported to SAS. Proc Convert combined with these lines gives a nice way to convert all OSIRIS information into SAS.

6. Link from SAS to OSIRIS

The procedure calls the OSIRIS monitor directly under SAS. 5 and 6 gives complete communications between OSIRIS and SAS.
7. Link from OSIRIS to SPSS

The OSIRIS-SPSS interface enables SPSS to read a type 1 OSIRIS dictionary file and fixed-format BCD OSIRIS data file as input for an SPSS run:

```
RUN
FILE
OSIRIS
INPUT
N OF CASES

NAME EXAMPLE
NAME OSIRIS
VARS V1 TO V20
MEDIUM TAPE
1000
```

8. Link from user programmes to SAS

This link gives the potential to implement user defined programmes as procedures in SAS. This gives the advance that the SAS datamanagement can be used directly, e.g. Weinreich (1981) has implemented a categorial data analy­sis programme which estimates the Rasch model for latent structures.

**Categorial data analysis**

In the social science the computer has until now mainly been used for collecting questionnaire data in survey ana­lysis.

The statistical analysis of these data has mainly been restricted to examination of all the two way marginal tables. These tables have of course some information, but the real important object namely the associations between the variables cannot be described correctly. For example association be­tween two variables may very well be due to a third variable, in the sense that this variable explains this association.
In the contingency table analysis, the log linear model now establish a very interpretative tool to describe these associations. An examination shows that BMDP, SAS and GENSTAT/GLIM have implemented programmes to solve these problems.

7. Conclusion

There is no doubt that all the packages mentioned will continue to expand and improve. There is however some huge structural advantages in the new generation packages not found in BMDP, SPSS and OSIRIS.

It is my point of view that BMDP mainly is trying to improve with better statistical procedures, remaining the data part unsolved. This problem is solved with the SAS interface.

SPSS now releases their version X, but they still have the handicap of repairing a system, which more and more turns out to be a former system.

OSIRIS seems now to run out of progress, so the solution is undoubtly the SAS system eventually with SAS/FSP and GRAPH if wanted. This gives a very easy way to maintain data and access statistical procedures in one operation. The result is that the social science researcher or other using statistics, have more time to investigate concerning the variables, which he wants to describe.
REFERENCES


Francis, Ivor (1981): ....... Statistical Software:
   A comparative Review. North Holland


Rasmussen, Karsten Boye: ... Archival Storage of SAS files, DDA in press.


Weinreich, Mogens (1981): .. A SAS implementation of the Rasch Item
   Analysis Model.
   University of Copenhagen. USI research report, nr. 82.