Notes on computer programmes for statistical analysis

James Stanley
Biostatistical Consulting Service
University of Otago, Wellington
james.stanley@otago.ac.nz

Original file 28/04/2009
Last updated 09/02/2012

Contents

1 General notes

This document is intended to give a brief summary of the pros and cons of several commonly used data analysis programmes. All of the programmes listed below are used by one or more biostatisticians at the University of Otago, Wellington, and so at least some support can be provided for these programmes (be aware that different biostatisticians have different preferences, and different levels of expertise in any given package.)

All of the programmes described beyond the "basic" level here are what might be called "multipurpose" statistical software[1]: they can handle a variety of statistical procedures. Many more programmes exist that either perform just a single type of analysis (e.g. sample size estimation, or a particular type of test such as analysis of variance) or are developed for a more specialised purpose (e.g. detection of a change in the shape of a trend over time), and are beyond the (current) scope of this document.

If you'd like to see a more comprehensive list of statistical packages, or to get more details on a particular package discussed here, it might be worth visiting Wikipedia at

http://en.wikipedia.org/wiki/List_of_statistical_packages

Feedback on the contents or usefulness of this document is appreciated.

2 Summary of available statistical software

Table 1 is a brief rundown of some common statistical software programmes, classified into "basic", "intermediate", and "advanced" programmes. This designation describes the complexity of the statistical analysis available in the programme rather than the skill level required to use the programme: for example, most statistical tests are harder to calculate in Excel than in EpiInfo or SPSS. An appendix summarises several details on each programme, such as cost, supported operating systems (OS), and other details[2].

2.1 General features and points to note

Some of the following details are also noted in the appendix.

Both Excel and Access are often pre-installed on Windows computers in the University.

EpiInfo is freeware software originally developed by the Center for Disease Control in Atlanta, Georgia[3]. OpenEpi is opensource software based on the EpiInfo programme: it is implemented and run through a web-browser interface[4], and therefore does not require any installation (although this also limits the functionality.)

SPSS (formerly known as SPSS for many years, and for a brief period as PASW before the company was purchased by IBM, who probably wondered why not to use the well-known brand-name), SAS, and Stata are commercial software programmes. S-Plus is also commercial software – it is listed here next to R, which is freeware software that is an implementation of the S programming language, and was originally developed at Auckland University. The differences between R and S-Plus are essentially to do with the "front end" or engines – which for most users boils down to the graphical interfaces one uses to perform tasks. A more detailed discussion of the differences between these engines is available at

http://cran.r-project.org/doc/FAQ/R-FAQ.html

under section 3.1, with the question "What-is-S?". Of course this account is written by people who use/develop R.

[1] with the possible exception of OpenEpi.
[2] Note that all of these programmes work under Windows XP and Windows 7 at the time of writing.
[3] http://www.cdc.gov/EpiInfo/
[4] http://www.openepi.com

As a general note regarding the user complexity of the programmes – the hard part of learning a new statistical analysis programme is usually working out the interface and how to interact with the software. While it is definitely more work to learn a more complex programme, when you progress to more complex analysis methods you will be able to make this methodological step without having to learn a new programme from the ground up. In other words – facing the steeper learning curve now may well save you time in the future. Thus your future plans should be a consideration in choosing a package.

A final note here regarding reproducibility in analysis – the more advanced packages listed here (SAS, Stata, R) have the added advantage that it is in most cases necessary to keep a file of the commands you have run to produce an analysis. This is **vital** for anyone undertaking a complex analysis or studying towards a research degree – nothing is more awkward than needing to revise an analysis and being unable to reproduce your initial analysis. SPSS and Epi Info both have facilities by which you can save those commands that have been run so far, but these need to be explictly undertaken (which increases the risk of not bothering.)

3 Details on specific programmes

3.1 The basic options

3.1.1 Excel and Access

Advantages:

Usually installed on most (work) computers.

Most computer users have at least some experience with Excel.

Easy to use for data entry and data storage.

(Relatively) easy to use for basic descriptive statistics.

Excel has some basic analysis tools built in (you will need to activate analysis toolpack first[5]) e.g. t-tests, correlation, chi-squared test.

Access database features can prevent some data storage and access issues (as noted below).

Access has Forms that can be set up for easy and consistent data entry.

Disadvantages:

Excel has NO restriction on data type storage (e.g. can enter dates inconsistently across entire dataset).

There is no simple way (in Excel) to include or exclude particular cases from an analysis without either deleting the data or making a copy of the numbers (both bad practices, making it hard to review your work later on).

In a nutshell, Excel allows multiple user-errors to slip through the gaps.

Access can be difficult to learn, particularly the way in which queries work.

Access also lacks facilities for calculating some simple descriptive statistics (e.g. median, quartiles).

[5]Statistical analysis in Excel is usually accessed through the Tools menu option, under Data Analysis. If this last option is not visible, you can add this option by going to Tools, selecting the Add-Ins option, and then ticking the boxes that have Analysis Toolpak in the title[a].

[a]The missing *c* makes the toolpa(c)k more edgy.

Apparently has some poorly-coded statistical distributions that may make some calculations inaccurate.

3.1.2 OpenEpi

Advantages:

OpenEpi has an online, web-browser based software interface. Therefore no installation is required on the user's computer, and likewise it will work under any operating system.

It is open-source software... so if you are really keen, you can see the calculation methods (although understanding the programming code is a bit complex). This also means that if you wanted, you could download the OpenEpi code and build your own additions.

Has several quick and easy to use functions for some basic and intermediate tasks – particular strengths include confidence intervals for proportions, 2x2 tables (including chi-squared tests, and odds ratios/rate ratios), and sample size calculations.

Disdvantages:

The flip-side of the web-based format means that OpenEpi is restricted to relatively simple functions, usually based on pre-calculated summary statistics. Therefore it is not a suitable tool for analysing an entire dataset.

3.2 The intermediate options

3.2.1 EpiInfo

Advantages:

Freeware! See appendix for link to website.

Easy to select subsets of data for analysis, without having to delete records or make multiple copies of datasets.

Keeps a saveable record of the analysis steps you have performed.

Automatically includes all VALID data points in calculations/tests.

Performs both descriptive statistics and a lot of basic to intermediate analysis (e.g., comparison of means, proportions; many regression methods)

Can be used for data entry – uses an Access-like data storage system. However, might be simpler to enter data in another programme (e.g. Excel), and then import this file into EpiInfo.

Update (09/02/2012) – The CDC has recently released Epi Info v7., which is a nicer looking implementation of the old Epi Info engine, and includes support for newer Excel file formats (Office 2007 onwards.) At present the older version of Epi Info seems to have a bit more functionality, so we are recommending people stick with that version for the moment.

Disdvantages:

Runs under Windows only.

Can be difficult to pick up (true for any statistics programme, though).

Limited analysis options beyond the basic methods. (UPDATE – version 7 is missing some analysis features from earlier version.)

Graphics can look quite sloppy – good for interpretation, or sharing with colleagues, but not so good for publication/presentation. (UPDATE – version 7 has much nicer graphical outputs, although at last check it did not include the capacity to draw a histogram.)

(UPDATE – with the release of v. 7, Epi Info is now being developed again – although it remains to be seen whether this is going to be just a re-skinned version of the old Epi Info or whether new analysis tools will be added to the mix.) (The following note is now obsolete:) No longer being developed by the CDC, and so what you can do now might be the most you will ever be able to do.

3.2.2 SPSS/PASW

Advantages:

Widely used in psychological and social science research, so some people will already have experience in using it.

Good range of statistics from descriptive methods (means, medians, frequencies etc.) through to common tests (t-tests, regression, ANOVA) and some more advanced statistical measures (e.g. Factor analysis)

Most likely covers everything you might want to do yourself – for many users, you might be consulting a statistician for more complex analysis anyway.

Can produce some nice looking graphs.

Disdvantages:

A few quirks here and there, but otherwise very user-friendly, and non-intimidating to people who are familiar with Excel.

Can be a bit rigid with regards to advanced options for tests sometimes; Lacking some of the more complex statistical procedures and data handling capabilities of the more advanced programmes.

3.3 The advanced options

3.3.1 SAS

Advantages:

Pretty much industry standard.

Widely used in medical research and pharmaceutical industry.

Well supported by the consulting biostatisticians!

Is very adept at data manipulation (e.g. counting elapsed days between two dates) as well as analysis.

Range of procedures from descriptive statistics through to simple analysis and on to complex analysis.

Usually good help files, with many worked examples available (although sometimes obscure topics!)

Can write programmes to automate some time-consuming processes.

You save your analysis script files to keep track of what analysis you actually did... so that you can amend it or re-run it later on.

For those who are "allergic" to text-based programming, there are some graphical user interfaces (GUIs) available that can help with analysis[6], although more complex tasks may only be possible through scripting.

Disadvantages:

SAS is a script-based programming system (some GUIs available, as noted above)

There is a steep learning curve at the start, even for very simple analyses;

Essentially a programming language: can be tricky to get your head around, and seems intimidating when you start using it.

3.3.2 Stata

Advantages:

Performs a large number of statistical analyses

Easier to get to grips with than SAS in first instance

Has some quick to use commands that give results for simple questions quickly (might take 1 minute to run, rather than 5 minutes to set up in SAS).

Has a large amount of example data available within the package, as well as online.

Large number of downloadable extensions that can be used to do more complex analysis/data presentation – note that these are unofficial packages and so caveat (non)emptor applies.

Main Stata files frequently updated to fix bugs and errors in the programme.

Disadvantages:

Help files are frequently more than a little obscure.

Less flexible than SAS in terms of data manipulation.

Main Stata files frequently updated to fix bugs and errors in the programme[7].

3.3.3 R/S-Plus

All of the following notes are written with reference to R: most points should also apply to S-Plus, with the (already noted) proviso that R is free, while S-Plus is not so free. As this is currently the programme which I have the least experience with, some of the following information might well be inaccurate.

Advantages:

[6] One option, supported by the UOW, is called SAS Enterprise Guide, which basically sticks a GUI on top of SAS to make the functions more accessible.

[7] This was listed as an advantage too, but if you are an infrequent user (like me) then it can be frustrating if Stata wants to go through the process of updating itself every other time you launch it.

Freeware; plus, R was developed in New Zealand[8].

Highly flexible scripting based language.

Downloadable packages freely available online[9] that have been developed to perform specialised statistical analyses.

Excellent graphing capabilities – highly flexible output, easy to overlay multiple graphs in the same figure, and you can customise almost every aspect of a figure (with a bit of work).

Can be installed on any operating system (installation may be more complex for Mac or Linux users)

Can be installed on removable media (e.g. USB flashdrive), so can be run on any computer.

Again, for those allergic to scripting-based languages, there are packages available that implement a graphical user interface (GUI) over the top of the R interface: such packages (e.g. RCommander) remove some of the stress from learning R, while letting the user explore the more options R has to offer available[10].

Disadvantages:

As already noted, R is a scripting based language, and therefore has a steep learning curve. Probably more complex to learn than SAS or Stata.

Help files are variable in usefulness: can often be complicated to understand the command structure.

Getting results quickly out of R (e.g. copying tables to paste into Word/Excel for a paper or report) can be convoluted (exporting figures is very easy though).

4 Concluding remarks

My advice would be to take several things into account when selecting a software package, most of which are common sense:

1. What statistical functionality do you require now? (and what might you require in the near future?)

   If you are likely to continue developing your statistical skills over several more projects, it may well be worth your time and effort to learn a more complex package. If however, you just need something for *right now* then choose the package that meets your immediate needs.

2. What can you afford to install?

   Keep in mind here the duration of your project and likely future use: it might be wasteful to install expensive commercial software for a one-off analysis.

3. What will your computer be able to run?

   The appendix has some information on compatibility between the different software options and common operating systems. In most instances, older computers should be able to run many programmes, but will work at a slower speed for complex analysis or large datasets. Check the system requirements before installing (and definitely before purchasing!).

[8] Although as it is freeware, downloading R does not count as buying NZ-made

[9] The installation of these packages is done through R, making adding a new package very simple.

[10] And as RCommander also shows you the script that is being run based on the GUI interaction, it serves to show you how functions work.

4. What computing experience do you have, and how much are you prepared to learn about a new programme[11]?

   If you are comfortable with using Excel, then a programme like PASW/SPSS will be a good fit. For the more complex programmes, Stata is reasonably easy to interact with, while SAS is more complex, and R is more complex again. Keep in mind point number 1 above.

5. What support will you be able to get for the programme you plan to use?

   Learning a more complex package can be tricky in terms of getting even basic procedures correct. Advice can come in two areas: (1) learning how to use the programme itself, and (2) learning which statistical procedures to use. The programme specific information can often be found online with minimal work, but having access local experts (or even fellow novices) can be a real boon.

6. Will I get accurate answers? Will I be able to work out what I did six months ago?

   Just as a final note, I would seriously recommend never using Excel for anything other than descriptive statistics (even then, under duress) and making graphs/arranging tables. Having a programme that you can include text annotations in is very useful for leaving yourself notes as to **why** you did a particular step.

At any rate, whatever option you choose, good luck!

[11] As an aside, I decided in early 2009 that I wasn't going to learn any new programmes, but didn't even last the calendar year before I started using R.

Table 1: Statistical software, ordered by methodological capabilities

Statistical capabilities		
Basic	Intermediate	Advanced
Excel	EpiInfo	SAS
Access	SPSS/PASW	Stata
OpenEpi		R/S-Plus

Appendix:
Summary of programme licensing and compatibility

Programme	License type	Cost*	Platform compatibility			Website
			Windows	Mac	Linux	
Excel	Perpetual	No cost to staff	✔	✔	✗	http://www.otago.ac.nz/its/software/msopen.html
Access	Perpetual	No cost to staff	✔	✔	✗	http://www.otago.ac.nz/its/software/msopen.html
OpenEpi	Freeware	—	✔	✔	✔	http://www.openepi.com (Under information and help)
EpiInfo	Freeware	—	✔	✗	✗	http://www.cdc.gov/epiinfo/
SPSS/PASW	Yearly license	$40	✔	✔	✗	http://www.otago.ac.nz/its/software/spss.html
SAS	Yearly license	$150	✔	✗	✗	http://www.otago.ac.nz/its/software/sas.html
Stata	Perpetual	$233	✔	✔	✗	http://www.otago.ac.nz/its/software/stata.html
R	Freeware	—	✔	✔	✔	http://www.r-project.org/
S-Plus	*Not known*†	*Not known*‡	✔	✗	✔	http://spotfire.tibco.com/Products/S-Plus-Overview.aspx

*Costs checked 9/2/2012
†Appears to be periodic licensing
‡But apparently quite expensive