ABC of Computing

JOHN ASBURY

STATISTICS AND EPIDEMIOLOGY

Most investigational work requires some form of statistical manipulation, and the doctor contemplating a research project is often daunted by the arithmetic needed. The computer's ability to do repeated tasks, to store and retrieve information reliably, and to do all this fast, objectively, and efficiently makes it an ideal tool for statistical work.

Initial planning: assessing the amount of data

Consider a doctor starting a research project. He knows the medical problem but has little knowledge of statistics. The first step is to consult simple texts on statistics, or some of the papers that have appeared in the *BMJ*. The doctor must understand why statistics are used before he starts. The central concept to grasp is that of variability in results. Secondly, having further defined his problems the doctor should then consult a statistician, preferably one with experience of computers. The statistician will be able to advise on experimental design, the most suitable forms of statistical analysis, and, most important, the volume of data that may have to be collected to show any significant result. This is a crucial step and allows one to make decisions on how to handle the data.

If there is a small amount of data—for example, up to 30 records each containing five pieces of information—and a simple analysis is appropriate then this can reasonably be undertaken with a simple desk top calculator. The cost in money and temper in entering 150 numbers via a keyboard should not, however, be underestimated. Invariably there will be mistakes and the data will have to be corrected.

If, however, a more advanced form of analysis is needed then even this small amount of information may prove to be too much and the use of a computer package should be considered. One possibility is to use a commercial microcomputer statistical package. Many of the microcomputer packages are excellent but some may be unable to undertake complex calculations that need large memory capacity. Some seemingly small computer programs—for example, discriminant analysis—create large “tables of numbers” (arrays or matrices) which may exceed the capacity of a microcomputer’s memory.

If the volume of data is more than that described above the use of a statistical package with data entry facilities implemented on a mainframe computer should be considered from the outset. It is important to consider the implications of this decision because it might alter the duration, location, and costing of the project. The need to use a large computer package, such as the statistical package for the social sciences (SPSS) may mean that the data have to be transferred to a minicomputer or a mainframe machine, where the package can be implemented, and this might mean working on a local university computer. Additionally, the doctor may have to consider the cost of computer time and the cost of using data entry facilities in his budget.

Collecting the data and getting it into the computer

The management of the data collecting needs to be carefully planned with regard to the data source, data transfer, data vetting, and calculations. Particular attention should be paid to reducing the number of times that data are transferred from one medium to another because each time a transfer occurs the risk of introducing errors increases. No amount of elegant computer processing will produce good results from incorrect data.

Data source—At one extreme data may come from housewives giving answers to a doorstep pollster to, at the other extreme, rats jumping through hoops in the laboratory. The data management technique should be matched to the circumstances. In the first example data must obviously be recorded on paper initially before being transferred to the computer, but the measurement of murine acrobatics could be entered directly into the computer via a terminal in the laboratory. In each case data transfer and entry systems have to be designed for the project.

The commonest data transfer technique is to collect the data using a computer compatible paper record, which is then given to a data preparation clerk, who converts it into a computer compatible form such as punched cards, which can be read directly by the computer. A better method is for the clerk to type the coded data from the record directly into the computer via a terminal, thereby avoiding a data transfer step where errors could occur. This record needs careful design—more than just drawing boxes on paper. Most of the comments relating to paper records apply to visual displays for data entry.

Name				
Hospital number				
Date of birth				
Age in completed years				
Menstrual status	premenopause = 1	menopause = 2	postmenopause = 3	15
	not known = 9			
Side	left = 1	right = 2	bilateral = 3	16
Tumour size (cms)	not known = 99			
Fixation to skin	nil = 1	partial = 2	complete = 3	19
Ulceration				20
Peau d'orange - tumour area				21
breast				22
Satellite nodules skin				23
Invasive skin wide of breast				24
Deep fixation - fascia	no = 1	yes = 2		25
muscle				26
chest wall				27
Homolateral axillary nodes	not involved = 1	mobile = 2	matted = 3	28
Other nodes				29
Lymphoedema	no = 1	yes = 2		30
Metastasis at presentation - skin				31
distant				32
International stage	not known = 9			33
Primary treatment				
simple mastectomy	no = 1	yes = 2		34
radiotherapy	no = 1	yes = 2		35
radical mastectomy	no = 1	yes = 2		36
other	no = 1	yes = 2		37
Site of first metastasis				
visceral	no = 1	yes = 2		38
bone	no = 1	yes = 2		39
other breast	no = 1	yes = 2		40
Date of first treatment	not known = 0			
Date of first metastasis	not known = 0			
Date of death	not known = 0			
Cause of death	breast cancer = 1	other causes = 2	alive = 3	59
Status at 1980 anniversary	alive = 1	lost to follow up = 2	dead = 3	60
Survival in completed years				

The computer record form—It is most important to find out what sort of data format the package requires; much time will be saved if the record is in the same form. For example, numeric codes may be preferred over letters. The record should be simple and neat. It should have a natural order—for example, most doctors examine a patient in a certain sequence and the data collection should take this into account. The record should not ask for unnecessary information, because the more information that is asked for the less that will be willingly and accurately provided. The method of recording should be clear so that if, for example, the respondent does not know how to answer the question he should be able to say so by way of a "don't know" code, or if he is unsure by an "undecided" code. This method makes it easy to detect missing data as each box should contain something. A refinement is to allow the respondent to write his uncodable answer freehand, and to code it later. Ideally when information is entered on the record the response codes should be printed nearby, but this can lead to large and daunting forms. Once the questions have been written and the forms designed they should be tested to check that the correct data really are being collected. Changing one word in a question may alter the response enormously. Finally, it is important to check that the data preparation clerk can understand the form; it is pointless collecting data if the clerk cannot enter it into the computer correctly.

NOT Enter date: 1 0 M A R 8 3 **BUT**

Enter day No

1	0
---	---

Enter month No

0	3
---	---

Codes should be sensible—M for male and F for female, or coded 1 or 2. A mixture of numeric and alphabetic codes slows down the data entry clerk considerably. Most computer terminals have a separate set of keys specifically designed for entering numeric data (numeric keypad). Where times or ages are entered the units of measurement should be clearly stated.

Data vetting—Assuming that the data have been placed in a computer file the user must not then think that these are necessarily correct; errors can be introduced each time the data are transferred. The extent of the data may make it impossible to check them by hand (though a limited manual check is advised), so a data vetting program can be used. In centres where large statistical packages are implemented complementary data vetting packages are also available. A crude data vetting program could work by checking ranges of numbers—for example, ages, so that if there were some octogenarians in a paediatric trial the computer would indicate to the user that there might be something wrong. If more information is given to the data vetting program—on how to interpret dates, for example, a patient with a birth date after his death could be spotted.

Once errors have been located they can be corrected, and this can usually be done directly from a computer terminal.

Statistical packages

```

-----
ANALYSIS OF ITU DATA DR JOHN ASBURY START FIRST 1000
FILE NONAME (CREATION DATE = 25/01/83)
***** C R O S S T A B U L A T I O N
OUTCOME BY SEX
CONTROLLING FOR.. AGESQ VALUE = 7.00
*****

COUNT I SEX
ROW PCT I MALE FEMALE ROW
COL PCT I TOTAL
TOT PCT I 1.00I 2.00I
-----I-----I-----I
OUTCOME 1.00 I 61 I 22 I 83
DEP NOT IM I 73.5 I 26.5 I 65.4
I 64.9 I 66.7 I
I 48.0 I 17.3 I
-----I-----I-----I
9.00 I 33 I 11 I 44
I 75.0 I 25.0 I 34.6
I 35.1 I 33.3 I
I 26.0 I 8.7 I
-----I-----I-----I
COLUMN 94 33 127
TOTAL 74.0 26.0 100.0

CORRECTED CHI SQUARE = 0.00081 WITH 1 DEGREE OF FREEDOM
PHI = 0.00253
CONTINGENCY COEFFICIENT = 0.00253 SIGNIFICANCE = 0.9773
KENDALL'S TAU B = -0.01634 SIGNIFICANCE = 0.3926
KENDALL'S TAU C = -0.01364 SIGNIFICANCE = 0.4100
GAMMA = -0.03937
SOMER'S D = -0.01773

```

```

-----
ANALYSIS OF ITU DATA DR JOHN ASBURY START FIRST 1000
FILE NONAME (CREATION DATE = 25/01/83)
***** C R O S S T A B U L A T I O N
OUTCOME BY SEX
CONTROLLING FOR.. AGESQ VALUE = 8.00
*****

COUNT I SEX
ROW PCT I MALE FEMALE ROW
COL PCT I TOTAL
TOT PCT I 1.00I 2.00I
-----I-----I-----I
OUTCOME 1.00 I 9 I 9 I 18
DEP NOT IM I 50.0 I 50.0 I 78.3
I 75.0 I 81.8 I
I 39.1 I 39.1 I
-----I-----I-----I
9.00 I 3 I 2 I 5
I 60.0 I 40.0 I 21.7
I 25.0 I 19.2 I
I 13.0 I 8.7 I
-----I-----I-----I

```

A statistical package is a collection of pretested computer programs designed to undertake statistical and epidemiological operations quickly, accurately, and efficiently. In epidemiological research operations such as counting and plotting scattergrams are often made by computer and most of the packages have appropriate facilities. Many of the packages have been available for some time and have been objectively evaluated and most of the problems are known. The package allows the user, who may not be mathematically minded, to use powerful statistical techniques with confidence, knowing that the computation will be correct. Typically a package requires the user to know a little of a simple "job control and specification language." Usually this language is simpler to use than a conventional programming language such as BASIC and it should not be beyond the wit of any doctor. If problems occur the packages are equipped with comprehensive error message systems. All the major packages have user manuals usually written for the non-computer expert, and even if the user does not understand them he can ask his tame statistician to explain. These manuals are usually available in libraries.

Most of the packages have data modification facilities so that the user can, for example, change the arrangement of the data. If information on mortality has been recorded by health districts these facilities would allow it to be reclassified by region. Most also have calculation facilities, so that given the birth date and the current date the program could calculate the patient's age in years, months, and days and enter it into the analysis. One particularly useful facility is the built in method for handling missing data—a common problem in medical research.

Conclusions

Statistical packages usually produce a tidy, well arranged print out and this can often be annotated to the extent that anybody unfamiliar with computing can understand it. This is a great advantage when regular analyses are used as the printout can be circulated unaltered.

Unfortunately the ease of use of statistical packages has encouraged users to use them mindlessly for "fishing," on the off chance that a significant relationship may appear. This is an inappropriate use of a statistical package and it may even produce invalid results.

It is important to check that the package really is doing the right sort of analysis, and again the advice of a statistician is important. Furthermore, though a package may produce an answer to the statistical question it remains the user's task to interpret the answer: it may be meaningless. There is something about computer print out that causes the naive user to become uncritical; and just because the computer produces the answer that the user wants it does not mean that it is the correct one.

The following are some of the commonly available packages.

Statistical package for social sciences—This American package is one of the easiest to use and is available for batch or interactive processing. SPSS has a wide range of statistical programs and also has useful tabulation facilities which can be specified in a few words of the job control language. It is usually implemented on mainframes.

The BioMedical computer package (BMD) is one of the earlier packages and contains a wide selection of statistical programs. It is usually implemented on mainframes.

Genstat offers facilities similar to the first two but is very general in its application and hence more difficult to use. An understanding of the Genstat language is necessary to exploit fully the considerable power of this package. It is best used for non-standard analyses.

General Linear Modelling (GLIM) is designed to allow the user to fit generalised linear models to data, which is a very powerful epidemiological technique. Many forms of data transformation are allowable—for example, if the relation between age and height proved unhelpful the program could be instructed to examine the relation between age and the logarithm of height.

Clustan is designed to define and quantify natural groupings (clusters) in a set of data. It has, for example, been used to classify psychiatric patients on the basis of their responses to a standard questionnaire. Mathematically cluster analysis is satisfactory but it is often hard to define a meaningful cluster. Different mathematical techniques often give different answers in all but the obvious cases.

The main message is that the doctor does not need to be daunted by the prospect of a large survey. Statistical packages are there to help, though they must be used with care.

Dr A J Asbury, FFARCS, PHD, is lecturer in anaesthetics, University of Sheffield.