EPISCOPE: Computer programs in veterinary epidemiology

K. Frankena, J. P. Noordhuizen, P. Willeberg, P. F. van Voorthuysen, J. O. Goelema

Veterinary Record (1990) **126**, 573-576

Veterinary epidemiology is a rapidly developing science. However, many veterinarians are unfamiliar with the relevant techniques because veterinary schools have not introduced biostatistics as a core element of their courses or adopted epidemiology as a specific discipline. EPISCOPE, the computer software presented in this paper, covers many epidemiological principles and calculations. It can assist both the teaching of epidemiology and the analysis of field data.

VETERINARY epidemiology is a rapidly changing science. In the past, the subject was mainly concerned with the causal agent of infectious diseases and the transmission of these agents. In the last decade, more attention has been paid to non-infectious diseases, and the emphasis has shifted from individual animals to herds or populations of animals. Modern veterinary epidemiology deals more and more with the identification and quantification of multifactorial diseases, the planning and evaluation of control programmes, and the assessment of the economic impact of disease (Thrusfield 1986). This tendency has resulted in the implementation of health and production control programmes for dairy and pig herds (Noordhuizen 1984, Buurman and others 1987).

The increasing importance of veterinary epidemiology is reflected in a series of developments. Societies for veterinary epidemiology have been established in the United Kingdom, Denmark and recently in the Netherlands, Scandinavia and France. New textbooks specifically relating to veterinary epidemiology are now available (Thrusfield 1986, Martin and others 1987). The need for systematic epidemiological and economic analysis of integrated veterinary, herd-management and economic records has increased (Dijkhuizen 1989) because advice to the farmer must be based on scientifically justified arguments for production policies, and the interpretation of herd data is unsound without the application of sound statistical techniques.

The application of quantitative epidemiology is germane to the collection and analysis of data (Rothman 1986). However, many veterinarians are unfamiliar with the relevant techniques because veterinary schools have not introduced biostatistics as a core element of their courses or adopted epidemiology as a specific discipline. As a result, computer programs which assist the teaching of epidemiology and the analysis of data have a valuable role in modern veterinary medicine. These programs should run on personal computers (PCs) because mainframe computers are not accessible to veterinarians outside universities or institutions.

Sophisticated statistical packages like the Statistical Analysis System (SAS 1985) or Generalised Linear Interactive Modelling (Baker and Nelder 1978) both for multivariate analysis, and MINITAB (Ryan and others 1985) perform few 'epidemiology-specific' analyses. For PCs, several 'homemade' programs deal with aspects of epidemiology. However, these are usually very limited, eg, the calculation of diagnostic sensitivity only, and documentation is often not available. In this paper, some epidemiological procedures are demonstrated with EPISCOPE, a series of programs for epidemiology-specific analyses which uses a spreadsheet program. Spreadsheet programs are commonly used on PCs for numerical modelling and data analysis (Carpenter 1984, Voorthuysen and others 1988, Kock and others 1989) and so are ideally suited to epidemiological calculations.

Materials and methods

Spreadsheet program

EPISCOPE is based on the spreadsheet program SUPERCALC version 4. This software can be run on IBM-compatible PCs (XT or AT) using MS-DOS 3.20 (or later versions) as the operating system.

The spreadsheet has a tabular structure which provides a minimum of 254 rows and 63 columns, resulting in 16,000 cells. The maximum size of the spreadsheet is 9999 rows and 2000 columns. Information (data, text, formulas) can be stored in each cell. One page of 160 cells (eight columns and 20 rows) can be displayed on the computer's video display unit (VDU). SUPERCALC can also present data graphically.

EPISCOPE

EPISCOPE consists of four modules, each module comprising a number of programs. The modules deal with the evaluation of diagnostic tests, sample size calculations, the analysis of cohort and case control studies, and models.

Each program is 'menu-driven' and has a page-like structure (Fig 1). The pages are ordered vertically and access to different pages is obtained by selecting the page on the menu line. Below the menu line, extra information (menuhelp) is given about the option highlighted. The RESULTS, INTRODUCTION and HELP sections may consist of more than one page.

The WELCOME page contains information about: the name (function) of the program; the required input parameters; the output parameters; the menu on the status line which shows the next possible action, and the menuhelp on the bottom line

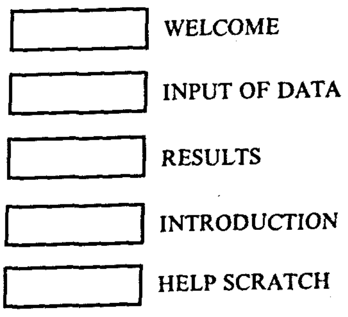


FIG 1: Page-like structure of EPISCOPE

K. Frankena, MSc, PhD, **J. P. Noordhuizen,** DVM, PhD **J. O. Goelema,** BSc, Agricultural University, Department of Animal Husbandry, PO Box 338, 6700 AH Wageningen, The Netherlands
P. Willeberg, DVM, PhD, DVSc, Royal Veterinary and Agricultural University, Department of Forensic and State Veterinary Medicine, Copenhagen, Denmark
P. F. van Voorthuysen, DVM, MUMPS Business and Veterinary Systems, Houten and State University, College of Veterinary Medicine, Department of the Science of Food of Animal Origin, Utrecht, The Netherlands

which gives extra information about the next possible action.

Page 2 is the INPUT OF DATA page. Some text or lines, eg, a 2 × 2 contingency table, will be present on this page. Values for parameters can be entered by choosing the 'Edit values' option. The spreadsheet will be recalculated automatically. The results of this calculation can now be monitored by choosing the 'Results' option. In general, there is no input capacity on the RESULT pages, but in some programs a confidence level might be changed, eg, from 95 per cent to 90 per cent.

The INTRODUCTION and HELP pages are below the results. These give information about the subject with which the program is dealing. In some programs, these pages are followed by SCRATCH pages, for example for intermediate results, or graph titles, which are not informative to the user.

EPISCOPE is started via a batch file that loads SUPERCALC and a 'macro'. A macro is an additional program that minimises the number of keystrokes and commands which need to be performed by the user. Within EPISCOPE, macros control the menu, making it very user-friendly. The first macro calls for a starting file that contains a list of the four modules within EPISCOPE. Next, the programs within the chosen module are displayed and the program of choice is loaded by typing its number. The program is entered via the WELCOME page.

Data

After loading the program initial data are available. These data are derived from examples in the accompanying workbook, which contains both fictitious data and data from published papers. The workbook has been developed for teaching purposes but is also useful for a first acquaintance with EPISCOPE. The initial data can be overwritten by other data.

In some EPISCOPE programs the input consists of frequencies, eg, the number of diseased and exposed animals in case-control studies. EPISCOPE is not meant for the calculation of these frequencies because SUPERCALC was not developed as a database-oriented program. It is often easier to use a real database program for that purpose.

Statistics

The statistics relating to case-control and cohort studies are based on the formulae of Rothman (1986).

Examples

Two examples of EPISCOPE programs are presented. Program 1 concerns the calculation of sample sizes for surveys. Program 2 determines the odds ratios and confidence intervals in case-control studies.

Results

Example 1: Sample sizes for surveys

When designing a survey for estimating the prevalence (P) of a disease, the sample size needs to be determined (number of herds, number of animals per herd). The formula to calculate the minimal sample size depends on the allowable error (L), the standard deviation (sd) of the expected prevalence and the desired level of confidence (Snedecor and Cochran 1980). The sd equals P(1-P), because the prevalence P is a binomial proportion. Thus, to assess the sd, an advance estimate of the prevalence, is needed. The following formula is used to determine the sample size, *n*:

$$n = t^2P(1-P)/L^2$$

where t = the value of Student's *t* for normally distributed data at a specified confidence level

P = the estimated prevalence

and L = the accepted error

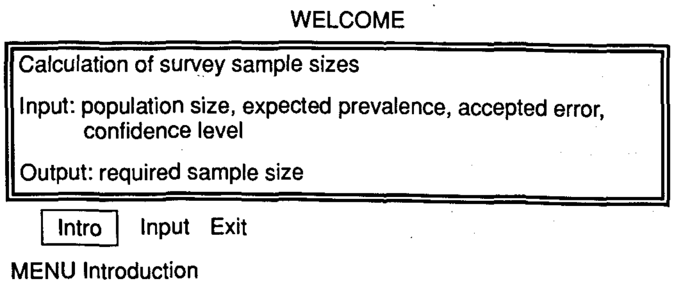
This formula shows that the minimal sample size, n, is larger when the desired confidence level is higher (*t* increases), when the allowable error is smaller, or when the expected prevalence is closer to 0·50 (or 50 per cent). However, the calculated value of n is only approximate, and needs to be adjusted when the sample size exceeds 5 per cent (rule of thumb) of the total population (N):

$$n\ (\text{adjusted}) = n/(1+f)$$

where f = the sampling fraction n/N

This adjustment is often necessary when small populations are sampled.

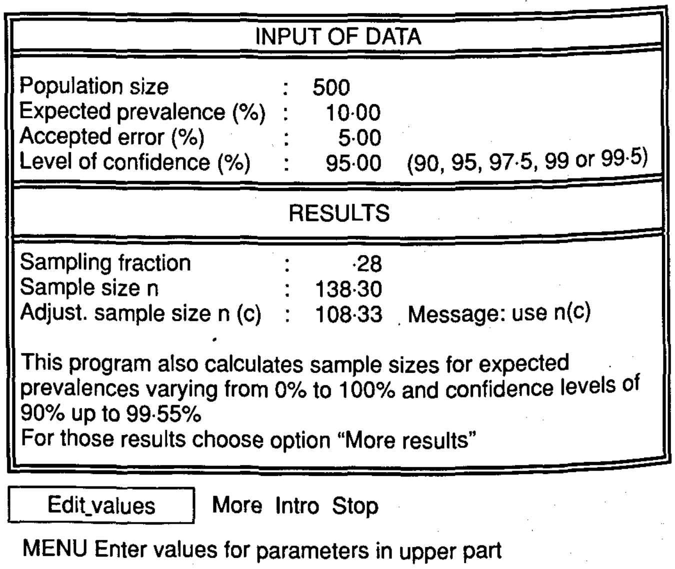
This calculation can be done by hand or with a pocket calculator but will be time-consuming, especially when it is necessary to evaluate the effect of changes in L, *t* or the expected prevalence. The EPISCOPE program SSPROPOR (sample size to estimate a proportion) has been developed to make quick calculations. After SSPROPOR is loaded the WELCOME page will appear (Fig 2).



```
WELCOME

Calculation of survey sample sizes

Input: population size, expected prevalence, accepted error,
       confidence level

Output: required sample size

 Intro   Input  Exit
MENU Introduction
```

FIG 2: WELCOME page of program SSPROPOR

The two INTRODUCTION pages of this program explain the formulae that are used. The page for the INPUT OF DATA and RESULTS (a combined page), including the fictitious data, is shown in Fig 3.

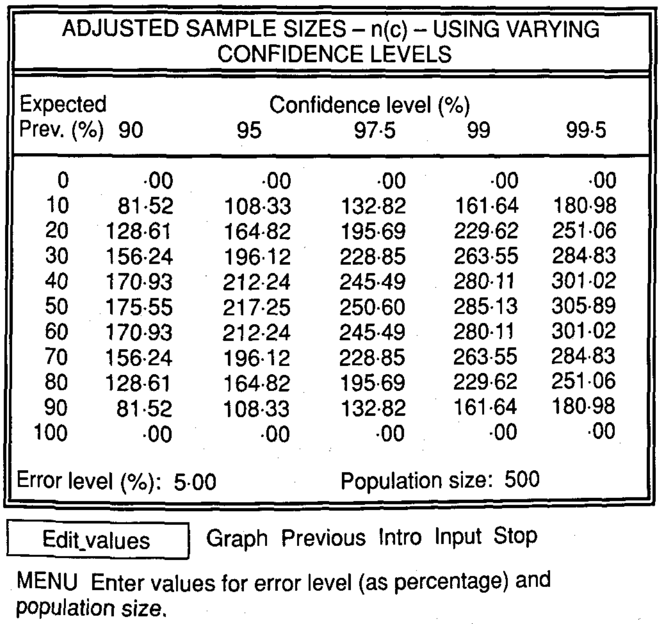


```
                         INPUT OF DATA

Population size           :  500
Expected prevalence (%)   :   10·00
Accepted error (%)        :    5·00
Level of confidence (%)   :   95·00  (90, 95, 97·5, 99 or 99·5)

                            RESULTS

Sampling fraction         :     ·28
Sample size n             :  138·30
Adjust. sample size n (c) :  108·33  Message: use n(c)

This program also calculates sample sizes for expected
prevalences varying from 0% to 100% and confidence levels of
90% up to 99·55%
For those results choose option "More results"

 Edit_values   More  Intro  Stop
MENU Enter values for parameters in upper part
```

FIG 3: INPUT OF DATA page of the program SSPROPOR

For example, on a pig fattening unit of 500 pigs it is expected that 10 per cent of the animals carry the microbes implicated in atrophic rhinitis. To confirm this, a number of pigs will be investigated to identify these microbes. The accepted absolute error is set at 5 per cent and the confidence level at 95 per cent.

Use of these data results in a sample size of 139 animals but, since the sampling fraction is larger than 5 per cent, the sample size can be reduced to 109. This means that by sampling 109 animals it can be demonstrated at a defined level of significance (P <0·05) that the prevalence of the disease is between 5 per cent and 15 per cent (the advance estimate of 10 per cent ± 5 per cent). When the measured prevalence is higher than the advance estimate of the prevalence the sample size of 109 animals is not sufficient for the 5 per cent level of significance. When the real prevalence is 50 per cent, 218 animals should be sampled. The next page of results (option 'More') shows the minimal sample size using several expected prevalences and confidence levels (Fig 4).



```
        ADJUSTED SAMPLE SIZES – n(c) – USING VARYING
                    CONFIDENCE LEVELS

Expected              Confidence level (%)
Prev. (%)  90        95        97·5       99        99·5

   0        ·00       ·00       ·00        ·00       ·00
  10      81·52    108·33    132·82     161·64    180·98
  20     128·61    164·82    195·69     229·62    251·06
  30     156·24    196·12    228·85     263·55    284·83
  40     170·93    212·24    245·49     280·11    301·02
  50     175·55    217·25    250·60     285·13    305·89
  60     170·93    212·24    245·49     280·11    301·02
  70     156·24    196·12    228·85     263·55    284·83
  80     128·61    164·82    195·69     229·62    251·06
  90      81·52    108·33    132·82     161·64    180·98
 100        ·00       ·00       ·00        ·00       ·00

Error level (%):  5·00          Population size:  500

[ Edit_values ]  Graph  Previous  Intro  Input  Stop

MENU  Enter values for error level (as percentage) and
population size.
```

FIG 4: Option 'More' of SSPROPOR showing sample sizes at several prevalences and confidence levels

Example 2: Case-control studies

In case-control studies, the relationship between exposure to a factor and the occurrence of disease is analysed. Thus, there are four categories of animals (or herds) in a 2 × 2 contingency table (Table 1).

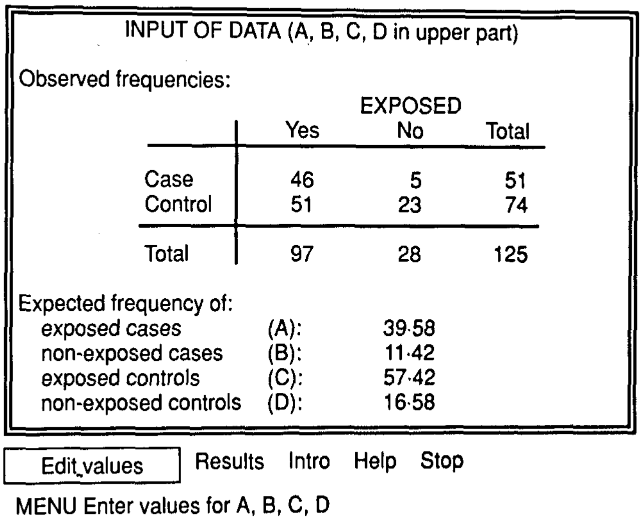
TABLE 1: General structure of a 2 × 2 contingency table for a case-control study

	Exposed Yes	No	Totals
Case	A	B	A+B
Control	C	D	C+D
Totals	A+C	B+D	A+B+C+D

A, B, C and D represent the numbers of animals in each category. The strength of association between exposure and disease is expressed by the odds ratio (OR) which is calculated as:

$$\text{OR} = (A/C)/(B/D),$$
$$= AD/BC$$

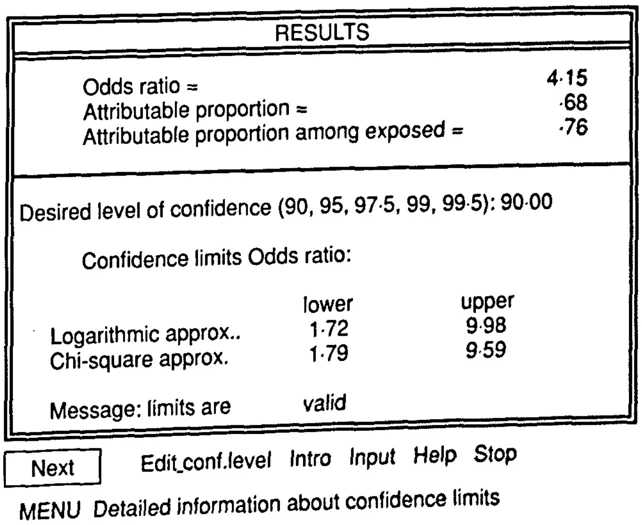
When the odds ratio is greater than 1·0 there is a positive relationship between exposure to the factor and the disease, and the factor is thus a risk factor. When the odds ratio is less than 1·0 the factor is associated with a reduced risk of disease. Once the frequencies in each cell are known the calculation of the point estimate of the odds ratio is simple. However, what really is of interest is whether the odds ratio is significantly different from 1·0. A confidence interval can be constructed for the odds ratio, and when this interval does not include 1·0 exposure to the factor is significantly associated with the disease (either positively or negatively, depending upon the values of the lower and upper confidence limits). The interval can be calculated in several ways. Two approximate methods (logit and test-based) are used in this program.



```
           INPUT OF DATA (A, B, C, D in upper part)

Observed frequencies:
                              EXPOSED
                        Yes     No      Total
                     ____________________________
           Case      |  46       5        51
           Control   |  51      23        74
                     ____________________________
           Total     |  97      28       125

Expected frequency of:
  exposed cases          (A):     39·58
  non-exposed cases      (B):     11·42
  exposed controls       (C):     57·42
  non-exposed controls   (D):     16·58

[ Edit_values ]  Results  Intro  Help  Stop

MENU Enter values for A, B, C, D
```

FIG 5: INPUT OF DATA page of the program OACASECO, using data from Schukken (1988)

The data in the following example (Fig 5) were collected by Schukken (1988) on dairy farms having low somatic cell counts in bulk milk. If coliform mastitis was recorded on a farm, the farm was designated as a case, and if not, it was designated as a control. Exposure to the risk factor concerned the use of a teat dip. Teat dip was applied on 46 or 51 case farms and on 51 of the 74 control farms. When the values of A, B, C and D are entered the relationship between the incidence of coliform mastitis and the use of teat dip is evaluated (Fig 6).



```
                          RESULTS

      Odds ratio =                                     4·15
      Attributable proportion =                         ·68
      Attributable proportion among exposed =           ·76

Desired level of confidence (90, 95, 97·5, 99, 99·5): 90·00

      Confidence limits Odds ratio:

                              lower           upper
    Logarithmic approx..      1·72            9·98
    Chi-square approx.        1·79            9·59

    Message: limits are       valid

[ Next ]    Edit_conf.level  Intro  Input  Help  Stop

MENU  Detailed information about confidence limits
```

FIG 6: RESULT page of program OACASECO, using data from Schukken (1988)

On the lower part of the INPUT OF DATA page the expected frequencies are calculated. If any of these frequencies is less than five more complex methods need to be used to evaluate the relationship between disease and exposure (Rothman 1986). For the data in this example the odds ratio is 4·15 (Fig 6) and the lower limit of the 90 per cent confidence interval is greater than 1, indicating that there is strong evidence of a positive relationship between the use of teat dip and coliform mastitis. According to these data, teat dip cannot be recommended to prevent coliform mastitis on farms with low cell counts.

The program also calculates the attributable proportion (AP) (Fig 6), that is the proportion of cases due to exposure. The

TABLE 2: Program available in EPISCOPE

Module 1:		Diagnostic tests
	1	Test agreement
	2	Test evaluation
Module 2:		Sample size calculations
	3	Sample size for detection of disease
	4	Sample size to estimate a proportion
	5	Sample size to estimate a difference between 2 proportions
	6	Sample size to estimate a difference between 2 means
Module 3:		Observational-analytical studies
	7	Case-control studies
	8	Stratified case-control studies
	9	Matched case-control studies
	10	Cohort studies (cumulative incidence data)
	11	Cohort studies (incidence rate data)
	12	Stratified cohort studies (cumulative incidence data)
	13	Stratified cohort studies (incidence rate data)
Module 4:		Models
	14	Classical Reed-Frost model

AP indicates the relative importance of exposure to the occurrence of disease. It is the proportion of cases that would not have occurred if the exposure factor had been absent from the population. This parameter reflects causality. It is calculated as (OR-1)/OR. In the coliform mastitis example, the AP = 0·68 which means that 68 per cent of all cases were related to the use of teat dip. Moreover, farms using teat dip have a probability of 0·76 of the disease. The message 'limits are valid' indicates that none of the expected frequencies is less than five.

EPISCOPE programs available

At the moment EPISCOPE consists of the modules and programs listed in Table 2. For the evaluation of diagnostic tests, two programs are available. The first calculates, by means of the KAPPA value (Martin and others 1987), the agreement between the results of two tests (or clinicians). The second evaluates diagnostic tests by calculating the sensitivity, specificity, predictive values and apparent prevalence. Programs 3 to 6 are for determining the minimal sample sizes in prevalence studies (see example 1) or in studies for the detection of disease in a population (or herd). Program 4 also calculates the maximum number of positive animals in a population, given that all samples taken show negative results. Programs 7 to 13 evaluate several types of case-control and cohort studies and are meant for the quantification of risk factors. The procedures for these programs are similar to the procedure in example 2. The last program, 14, deals with the classical Reed-Frost model, which stimulates an epidemic curve when an infection occurs in a population with susceptible (and immune) animals.

Discussion

EPISCOPE facilitates epidemiological calculations. These calculations are often necessary for a correct and easy interpretation of data, but in practice are often not performed because veterinarians are unfamiliar with the underlying epidemiological principles and formulae. For example, data are often presented as percentages, regardless of absolute numbers, and the 'analysis' may merely involve comparing these percentages. As a result false interpretations may easily occur, with detrimental consequences. EPISCOPE has been developed because of this lack of knowledge. The program can be used by people who are unfamiliar with PCs. Furthermore, a knowledge of SUPERCALC is not needed because of the macrocontrolled menus. The EPISCOPE 'manual' consists of only one page.

EPISCOPE has been developed from the prototype EPIDEMO which was the result of a joint European Community project between the State University of Utrecht and the Royal Veterinary and Agricultural University, Copenhagen (Voorthuysen and others 1988). EPISCOPE is meant as a dual purpose package, for teaching and practical 'field' use.

For educational purposes, it assists animal health workers in understanding some elementary epidemiological procedures. INTRODUCTION and HELP pages are added to each program. A workbook with examples, suitable datasets, questions and answers is also available. EPISCOPE can also be used in lectures, when the VDU is replaced by a liquid crystal display in combination with an overhead projector.

For practical purposes, in addition to its value in designing field studies and analysing data, EPISCOPE could be used in simulation. It is especially useful to those who have no access to mainframe computers running suitable statistical packages, eg, practitioners. For example, EPISCOPE gives the same output for stratified case-control studies as a SAS procedure.

The data-analysis does not include multivariate techniques, such as analysis of variance or logistic regression. These techniques should not be used by the novice because the interpretation and the validity of the results requires deep insight into statistics (Rothman 1986).

EPISCOPE performs epidemiological calculations at great speed and it can evaluate alternatives by changing the input parameters. These characteristics might facilitate the widespread adoption of epidemiological procedures and stimulate motivation (Clarkson 1987).

A disadvantage of EPISCOPE is that the input in some programs consists of summary data, eg, frequencies in case-control and cohort studies. EPISCOPE does not calculate these summary data from the raw data. Summary data should be derived in another way, for example by hand or from a database program. This disadvantage is related to the fact that SUPERCALC is designed as a 'super calculator' and not as an efficient and easy-to-learn data handling program. Production of summary data using SUPERCALC involves a considerable knowledge of the SUPERCALC program.

Acknowledgements.— The authors express their gratitude to Dr M. V. Thrusfield of the University of Edinburgh, Department of Veterinary Clinical Studies, for his valuable assistance and critical comments.

References

BAKER, R. J. & NELDER, J. A. (1978) GLIM, release 3. Oxford, Numerical Algorithms Group

BUURMAN, J., LEENGOED, L. A. M. G. van, VERNOOY, J. C. M., WIEDA, A. & VALK, P. C. van der (1987) *Veterinary Quarterly* **9**, 15

CARPENTER, T. E. (1984) *American Journal of Epidemiology* **120**, 943

CLARKSON, M. J. (1987) Proceedings of the Society of Veterinary Epidemiology and Preventive Medicine, Solihull. Ed M. V. Thrusfield. p57

DIJKHUIZEN, A. A. (1989) *Veterinary Quarterly* **11**, 116

KOCK, M. D., CLARK, R. K. & JESSUP, D. A. (1989) *Preventive Veterinary Medicine* **7**, 137

MARTIN, S. W., MEEK, A. H. & WILLEBERG, P. (1987) Veterinary Epidemiology – principles and methods. Ames, Iowa, Iowa State University Press. p343

RYAN, B. F., JOINER, B. L. & RYAN, T. A. (1985) Minitab Handbook. 2nd edn. Boston, Prindle, Weber and Schmidt. p379

NOORDHUIZEN, J. P. T. M., WILBRINK, H. J. & BUURMAN, J. (1985) *Veterinary Quarterly*, **7**, 3

ROTHMAN, K. J. (1986) Modern Epidemiology. Boston/Toronto, Little Brown

SAS (1985) User's Guide: Basics, Version 5 edn. Cary, NC, USA, SAS Institute Inc.

SCHUKKEN, Y. H. (1988) Proceedings of the Dutch Society for Veterinary Epidemiology and Economics, Ed K. Frankena. Wageningen, The Netherlands. p17.

SNEDECOR, G. W. & COCHRAN, W. G. (1980) Statistical Methods. 7th edn. Ames, Iowa, Iowa State University Press, p 507

THRUSFIELD, M. (1986) Veterinary Epidemiology. London, Butterworth. p 280

VOORTHUYSEN, P. F. van, NOORDHUIZEN, J. P. T. M., AGGER, J. F. & WILLEBERG, P. (1988) Proceedings of the 5th International Symposium on Veterinary Epidemiology and Economics, Copenhagen, Denmark. Eds P. Willeberg, J. Agger, H. Riemann. p539

BVA guides binder

AN attractive green binder is available for all BVA guides. Price £3.00 including postage from TGS Subscriber Services, 6 Bourne Enterprise Centre, Wrotham Road, Borough Green, Kent TN15 8DG, telephone 0732 884023, fax 0732 884034. Cash with order, please