A Visual Basic Software for Computing Fisher's Exact Probability

Haseeb Ahmad Khan

Research Center, Armed Forces Hospital, Riyadh, Saudi Arabia.

ABSTRACT

Fisher's exact test (FET) is an important statistical method for testing association between two groups. However, the computations involved in FET are extremely tedious and time consuming due to multi-step factorial calculations after the construction of numerous 2×2 tables depending on the smallest cell value. A Visual-Basic computer program, *CalcFisher*, has been developed to handle the complexities of FET resorting the techniques of looping subroutines and logarithmic conversions. The software automatically calculates the P-value after entering the respective cell values and has been validated for proper functioning with a wide range of frequencies (tens to several thousands). The important features of the program include, easy data entry, tail-selection, comprehensive report format and the facility of printing and saving of results.

Keywords: Fisher's exact test; Statistical analysis; P-value; Visual Basic software

INTRODUCTION

Fisher's exact test (FET) calculates the exact probability value for the relationship between two dichotomous variables (Campbell and Machin 1996; Siegel 1956; Armitage and Berry 1994; Kramer 1988). It is an extremely useful non-parametric method for analyzing statistical association between the two independent sample groups and is commonly used for analyzing clinical and

Address correspondence to: Haseeb Ahmad Khan PhD, MRSC (UK), Armed Forces Hospital, P.O. Box 7897 (T-835), Riyadh 11159, Kingdom of Saudi Arabia. E-mail: khan_haseeb@yahoo.com

experimental data in biomedical research. The results of FET are expressed in terms of exact probability (P-value), varying within 0 and 1. Two groups are considered statistically significant if the P-value is less than the chosen significance level, which is quite often 0.05. The data format for FET is conveniently represented by 2×2 table, made of 2 rows and 2 columns. The two rows are two independent groups and the two columns represent the two effects or conditions.

Although the calculations required for FET are fairly straightforward the construction of additional 2×2 tables and the computation of respective probabilities using factorial formula entail considerable time and effort, especially when the lowest cell value is high (Siegel 1956; Armitage and Berry 1994; Kramer 1988). The aim of this study was to develop a computer program to solve the complexities involved in factorial computations for data analysis using FET.

COMPUTATION METHODS

Standard Method

The design of 2×2 contingency tables provides a comprehensive view of data to be analyzed as shown in Table 1. There are four input parameters (frequencies) belonging to two different groups. The top row is one group and the bottom row is another group whereas '+' and '–' signs above the two columns indicate presence or absence of a certain condition respectively. The standard formula (Formula 1) for calculating P-value (Campbell and Machin 1996; Siegel 1956; Armitage and Berry 1994; Kramer 1988) is shown in Table 1. If the smallest cell value in the contingency table is 0 then only one exact probability has to be calculated which is the simplest form of FET. However, if none of the cell frequencies is 0, more extreme deviations from the distribution could occur with the same marginal totals; thus, all those possible deviations must be considered and respective probabilities summed for testing null hypothesis. For instance, if the smallest cell value is 2, then three exact probabilities (using smallest cell values 2, 1 and 0) must be determined and then summed to get the exact P-value.

Modified Procedure

Our preliminary efforts while developing this Visual Basic application showed that Formula 1 could only be used for up to a total of 113 subjects (X = 113), beyond that the output of factorial computations exceeds the range of Visual Basic. The use of Stirling approximation formula (Diem and Lentner 1975) was also avoided for the sake of exactness of resulting P-values. Consequently, a modified procedure (Formula 2, Table 1) based on logarithmic conversions was used to perform FET for a wider range of frequencies.

Table 1. Construction of 2×2 contingency table and formulae for P-value calculation.

2x2 Table Format

	Present (+)	Absent (–)	Total
Group 1	x_1	x_2	$t_1=x_1+x_2$
Group 2	x_3	x_4	$t_2=x_3+x_4$
	$t_3=x_1+x_3$	$t_4=x_2+x_4$	X

x_1, x_2, x_3 and x_4 are 4 frequencies, t_1 and t_2 are rows' totals, t_3 and t_4 are columns' totals and X is total number of subjects.

Formula 1

$$\text{P-value} = \frac{t_1! \times t_2! \times t_3! \times t_4!}{X! \times x_1! \times x_2! \times x_3! \times x_4!}$$

Formula 2

$$\text{P-value} = \sum_{0}^{x_{min}} \exp\left[\Sigma \ln t_{(1\text{-}4)} - \Sigma \ln x_{(1\text{-}4)} - \ln X\right]$$

x_{min} in the smallest cell value in 2×2 table.

According to this procedure, the program finds out the P-value of the original frequencies using the antilogarithm of the value obtained by subtracting the logarithm of total subjects (X) and sum of logarithms of individual frequencies (x_1, x_2, x_3 and x_4) from the sum of logarithms of row and column totals (t_1, t_2, t_3 and t_4). Then the program identifies the minimum frequency in the 2×2 table, subtracts 1 from this frequency and adjusts the remaining frequencies in the table so that row (t_1 and t_2) and column totals (t_3 and t_4) remain constant. The resulting set of frequencies is also used to compute the respective P-value. The whole process of subtracting 1 from the current minimum frequency, adjusting remaining 3 frequencies and computing the P-value is repeated until the least frequency becomes 0. All the P-values (obtained by using the least frequency 0, 1, 2, x_{min}) are summed up to get the exact P-value (Formula 2, Table 1).

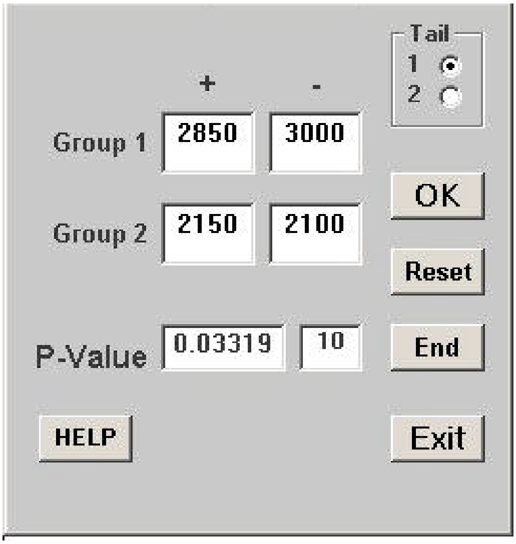
PROGRAM DESCRIPTION

Hardware and System Specification

The hardware used was a Pentium III computer with 20 GB hard disk capacity and 64 MB RAM. The program was developed in VISUAL BASIC 6.0 and will run as an executable file (48 KB) in any WINDOWS environment even without the presence of VB.

Design

The program is comprised of two windows (forms), which appear together on the screen when the program is run. The design and special features of both the forms are given below:



Tail
1 ●
2 ○

	+	-
Group 1	2850	3000
Group 2	2150	2100
P-Value	0.03319	10

Fig. 1. Form 1: the main window for data input.

Form 1 (data input)

The configuration of form 1 is shown in Fig. 1, which is the main window for data input and computations. Among the various controls, there are 4 text boxes for data entry, 2 option buttons for tail selection, 4 command buttons for different applications and 2 labels, one for the output of P-value and the other is a counter.

Form 2 (report)

The report window is comprised of one rich-text box and two command buttons namely ‘Print’ and ‘Save’ which are controlled by a common-dialog application. The report shows test number, P-value, input frequencies of the two groups and the tail type.

PROGRAM VALIDATION

The software was validated for its proper functioning and accuracy using representative frequencies and the results were compared with standard statistical programs, SPSS and EPI-INFO (Table 2). The results confirmed the efficiency of *CalcFisher* program for computing Fisher’s exact P-values for a wide range of frequencies.

Table 2. Comparison of P-values obtained from *CalcFisher* program with those resulted from other statistical packages

	Group 1		Group 2		P-Value (1-tailed)		
No.	+	-	+	-	SPSS	EPI-INFO	*CalcFisher*
1	1	9	8	2	0.003	0.0027	0.0027
2	0	7	3	6	0.150	0.1500	0.1500
3	3	7	8	1	0.015	0.0149	0.0149
4	5	4	2	7	0.167	0.1674	0.1674
5	2	6	5	5	0.278	0.2783	0.2783
6	4	6	7	3	0.185	0.1849	0.1849
7	8	1	4	5	0.066	0.0656	0.0656
8	500	450	350	400	NP	NP	0.0083
9	1000	900	850	900	NP	NP	0.0078
10	2850	3000	2150	2100	NP	NP	0.0332

NP, not possible

DISCUSSION

Earlier studies from our center (Khan 2003) and other investigators (Todd and Wang 1996; Martin et al 1997; Runciman et al 1998) have demonstrated potential utility of Visual Basic software for various biomedical applications. Visual Basic programs are user friendly because of their object-oriented feature and the familiarity of most of the users with Microsoft windows environment. Unfortunately the use of Visual Basic for large factorial computations is greatly hampered by its upper limit (1.79E308) that is roughly closer to factorial of 170 (7.25E308); the factorial of 171 crosses the range of Visual Basic. This problem was successfully resolved by using logarithmic approach that drops out the exponential power of numeric values by converting them into respective log values. Moreover, the involvement of lesser number of operators in log-based factorial formula simplifies the Visual Basic code with a direct impact on software's efficiency.

The results of software validation clearly demonstrated the ability of *CalcFisher* program to accurately compute Fisher's exact P-values for a wider range of frequencies (Table 2). The commonly used statistical packages including SPSS and EPI-INFO can also be used for computing FET, but in a condition-bound strategy. The former program calculates Fisher's exact P-value only when the total subjects are twenty or less (SPSS for Windows 1999) whereas the later performs FET when any of the expected values is less than five (EPI-INFO). On the other hand, *CalcFisher* computes P-values irrespective of cell frequencies and therefore can be utilized for universal application of FET for any data sets. Both SPSS and EPI-INFO basically compute $\chi 2$ statistics in the 2×2 table format when the cell values are high. In fact, the $\chi 2$ test is an approximation to FET and when applied with appropriate continuity correction leads to a fair approximation to exact probability (Campbell and Machin 1996). However, the estimate of probability in the $\chi 2$ test may not be very accurate if the marginal is very uneven or if one of the values is very small (Cochran 1954). Whereas, FET is a valid procedure for any number of frequencies and can easily be performed using *CalcFisher*.

In conclusion, the complexity of factorial computations can be greatly simplified by using logarithmic methodology. Log-based computations are highly suitable for developing Visual Basic applications as they involve lesser number of operations and also keep the output of intermediate steps within the permissible range of Visual Basic. The operational simplicity and integrated report format of *CalcFisher* render a handy tool for performing Fisher's exact test.

REFERENCES

Armitage, P. and Berry, G. (ed.) (1994). *Statistical Methods in Medical Research*, Blackwell Scientific Publication, London, England.

Campbell, M.J. and Machin, D. (ed.) (1996), *Medical statistics: A Comprehensive Approach*, John Willey and Sons, West Sussex, England.

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics*, 10, 417-451.

Diem, K. and Lentner, C. (1975). *Scientific Tables*, Ciba-Geigy Ltd, Basle, Switzerland.

EPI-INFO, *Public Domain Software for Epidemiology and Disease Surveillance*, Atlanta, Georgia, U.S.A.

Khan, H.A. (2003). Calcdose: a software for drug dosage conversion using metabolically active mass of animals, *Drug and Chemical Toxicology*, 26, 53-60.

Kramer, M.S. (ed.) (1988). *Clinical Epidemiology and Biostatistics*, Springer-Verlag, New York, USA.

Martin, J.M., Dartois, E., Bontoux-Frank, G., Ricour, B., Pariset, L., Robaux, P. and Giampietro, V. (1997). A software for the description of workplaces in the PRS system, *Computer Methods and Programs in Biomedicine*, 52, 53-66.

Runciman, W.B., Helps, S.C., Sexton, E.J. and Malpass, A. (1998). A classification for incidents and accidents in the health care system, *Journal of Quality in Clinical Practice*, 18, 199-211.

Siegel, S. (ed.) (1956). *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, Inc, Tokyo, Japan.

SPSS for Windows (1999). *Release 9.0.0, Standard Version*, SPSS Inc., USA.

Todd, B.A. and Wang, H. (1996). A Visual Basic program to pre-process MRI data for finite element modeling, *Computers in Biology and Medicine*, 26, 489-495.