STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:2857–2866
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2146

Modified exact sample size for a binomial proportion with special emphasis on diagnostic test parameter estimation

Geoffrey T. Fosgate*,†

Department of Veterinary Anatomy and Public Health, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843-4458, U.S.A.

SUMMARY

The design of epidemiologic studies for the validation of diagnostic tests necessitates accurate sample size calculations to allow for the estimation of diagnostic sensitivity and specificity within a specified level of precision and with the desired level of confidence. Confidence intervals based on the normal approximation to the binomial do not achieve the specified coverage when the proportion is close to 1. A sample size algorithm based on the exact mid-P method of confidence interval estimation was developed to address the limitations of normal approximation methods. This algorithm resulted in sample sizes that achieved the appropriate confidence interval width even in situations when normal approximation methods performed poorly. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: sample size; binomial; exact; proportion; diagnostic test

1. INTRODUCTION

The calculation of the sample size is an important component of the epidemiologic study design process in general [1] and specifically for the validation of diagnostic tests [2]. The ability of the calculated sample size to yield statistically significant results upon completion of the study depends upon the choice of the assumptions and the statistical model used to make the calculations. The statistical method employed should parallel the method of data analysis to the extent possible. The use of overly conservative methods of sample size estimation would not be considered incorrect because findings should be 'statistically significant,' however, such methods may not be efficient due to the increased cost associated with the excessive number of study subjects.

The efficient use of resources is important in the field of diagnostic assay development because recent technological advances in the field of molecular biology has led to the rapid

*Correspondence to: Geoffrey T. Fosgate, Department of Veterinary Anatomy and Public Health, College of Veterinary Medicine, Texas A&M University, College Station, TX 77843-4458, U.S.A.
†E-mail: gfosgate@cvm.tamu.edu

Received 17 June 2004
Accepted 1 November 2004

development of new tests for the control of emerging, and re-emerging, infectious diseases of veterinary and public health importance. Newly developed assays need to be validated and subsequently compared to existing methods before they can be accepted for general use. Diagnostic assay validation can be considered to be comprised of five stages [3] that can be roughly grouped into two components: bench validation and field validation. Bench validation can be considered a laboratory process and field validation an epidemiologic consideration. Both aspects of the validation procedure should include sample size calculations to achieve desired precision in the results.

Newly developed assays, including polymerase chain reaction (PCR) based diagnostics, are often expected to be highly accurate as measured by diagnostic sensitivity (probability of a positive test in a truly affected individual) and specificity (probability of a negative test in a truly non-affected individual). The specificities of such assays are often reported to be greater than or equal to 0.98 [4–7]. Statistical procedures based on the uncorrected normal approximation to the binomial perform poorly for proportions close to 0 and 1 [8, 9] and therefore other methods would be recommended for the analysis of data in such instances. Sample size calculations for the evaluation and comparison of diagnostic assays has been the focus of previous reports [10–13], but the design of such studies based on sample sizes calculated using normal approximation methods are not likely to yield results at the desired level of precision. The design of evaluation studies for assays that are expected to be close to perfect would benefit from sample size estimates based on modified exact binomial theory.

Exact binomial methods have been previously developed for the estimation of the sample size necessary to compare an expected proportion to a fixed null value. These methods have been included in commercially available software packages including StatXact version 5.0 (Cytel Software Corporation, Cambridge, MA, U.S.A.) and SAS version 9.1 (SAS Institute Inc., Cary, NC, U.S.A.). However, the author is not aware of available statistical software or peer-reviewed published routines that will calculate the sample size necessary to estimate a proportion within specified limits at a predetermined level of confidence based on exact binomial methods. The objective of the paper reported here was to compare the performance of a modified exact binomial sample size computer algorithm to the usual normal approximation method based on inverting the Wald statistic.

2. EXACT BINOMIAL SAMPLE SIZE

The Clopper–Pearson method of exact confidence interval estimation [14] has been considered the ‘gold standard’ method of interval estimation because it guarantees that the coverage probability will be at or above the nominal level [8, 9, 15, 16]. The basis of the interval is the binomial probability function and the following equations can be used to find the typical exact confidence limits for a specified level of alpha.

Lower limit

$$\sum_{k=x}^{n} \binom{n}{k} \pi_{\mathrm{L}}^{k}(1-\pi_{\mathrm{L}})^{n-k} = \alpha/2$$

Upper limit

$$\sum_{k=0}^{x} \binom{n}{k} \pi_{\mathrm{U}}^{k}(1-\pi_{\mathrm{U}})^{n-k} = \alpha/2$$

The confidence limits are the solutions to these equations by adjusting π_U and π_L for the fixed (observed) values of x and n until the appropriate probability is obtained.

These same equations can be used to calculate the sample size necessary to achieve specified confidence limits and alpha error rate. In the sample size situation, π_U and π_L are fixed by the investigator as the desired limits of the confidence interval around the hypothesized proportion (p_0). The equations are then solved for the value of n yielding the appropriate sample size. The value of x is calculated as the integer that when divided by n yields the closest value to the proportion hypothesized by the investigator ($p_0 \approx x/n$).

This method results in the calculation of separate sample sizes for the lower and upper limits of the confidence interval. The conservative approach would be to report the maximum of these two sample sizes as the one necessary to obtain. The Clopper–Pearson interval estimate is considered to be overly conservative in certain situations [8, 15–17] due to the discreteness of the binomial distribution and the sample size calculated in this manner would therefore tend to be larger than necessary to achieve nominal precision.

3. MODIFICATION OF EXACT METHOD

The mid-P method [18, 19] of adjusting the traditional exact binomial confidence interval is less conservative and still achieves good coverage [8, 9, 16, 17, 19, 20]. Calculation of the sample size based on this modification is summarized in the two formulae below.

Lower limit

$$\frac{1}{2}\binom{n}{x}\pi_L^x(1-\pi_L)^{n-x}+\sum_{k=x+1}^{n}\binom{n}{k}\pi_L^k(1-\pi_L)^{n-k}=\alpha/2$$

Upper limit

$$\frac{1}{2}\binom{n}{x}\pi_U^x(1-\pi_U)^{n-x}+\sum_{k=0}^{x-1}\binom{n}{k}\pi_U^k(1-\pi_U)^{n-k}=\alpha/2$$

Solving the above formulae still results in the calculation of two different sample sizes, which may continue to be overly conservative. This can be controlled by combining the two above probabilities into the single formula below.

$$\frac{1}{2}\binom{n}{x}\pi_L^x(1-\pi_L)^{n-x}+\frac{1}{2}\binom{n}{x}\pi_U^x(1-\pi_U)^{n-x}+\sum_{k=x+1}^{n}\binom{n}{k}\pi_L^k(1-\pi_L)^{n-k}$$

$$+\sum_{k=0}^{x-1}\binom{n}{k}\pi_U^k(1-\pi_U)^{n-k}=\alpha \tag{1}$$

Using the above formula for the sample size situation necessitates that π_U and π_L be fixed by the investigator as the desired limits of the confidence interval around the hypothesized proportion (p_0). The equation is then solved for the value of n yielding the appropriate sample size. The value of x is calculated as the integer that when divided by n yields the closest value to the proportion hypothesized by the investigator ($p_0 \approx x/n$). Solving the above formula will result in a sample size that is less conservative than the previously mentioned procedures by allowing for asymmetry in the tail probabilities.

4. FORTRAN ALGORITHM

An iterative routine was written in FORTRAN [21] to solve the sample size equation discussed in the previous section. The input for the algorithm is the hypothesized proportion (p_0), the desired error limit (forming the limits π_L and π_U), and the desired level of confidence ($1-\alpha$). The binomial probability function is difficult to use when the sample size is large due to the factorial component of the formula. The functional limit would be a sample size of 170 as the largest number to calculate the factorial component using double precision (8 bytes) real numbers ($170! = 7.3 \times 10^{306}$). The following relationship was used to remove this limitation to the sample size calculation and included in the FORTRAN algorithm.

$$0 < x < n: \quad \binom{n}{x} = \frac{n!}{x!(n-x)!} = \prod_{k=x+1}^{n} \frac{k}{n-(k-1)}$$

$$x = n, 0: \quad \binom{n}{x} = 1$$

The sample size algorithm starts the procedure at the minimum n necessary to observe the entered proportion exactly. For example, a proportion of 0.5 would start at $n = 2$ and for a proportion of 0.99 the starting point would be 100. The value of x yielding the entered proportion is always 1 at the first iteration of the sample size procedure because of this starting point. The algorithm simply adds 1 to the sample size at each iteration and sums the probabilities in each tail for these values of x and n (such that $p_0 \approx x/n$). The appropriate sample size has been reached when the sum of the tail probabilities is less than the specified alpha level (1 – confidence). To calculate the tail probabilities equation (1) was modified for the computer algorithm to reduce the computational complexity and improve efficiency. The modified equation transforms both summations to be over the same range and is included below (2) for the interested reader.

$$\frac{1}{2}\binom{n}{x}\pi_L^x(1-\pi_L)^{n-x} + \sum_{k=x+1}^{n}\binom{n}{k}\pi_L^k(1-\pi_L)^{n-k} + \left[1 - \left(\sum_{k=x+1}^{n}\binom{n}{k}\pi_U^k(1-\pi_U)^{n-k} + \frac{1}{2}\binom{n}{x}\pi_U^x(1-\pi_U)^{n-x}\right)\right] = \alpha \tag{2}$$

The FORTRAN program was compiled as a DOS-based application and can be obtained by contacting the author.

5. SAMPLE SIZE EVALUATION

The modified exact algorithm was used to calculate the sample size necessary to estimate a range of proportions (0.5–0.99), error limits (0.01, 0.05, and 0.1), and confidence levels (0.90, 0.95, and 0.99). Statistical software [22] was used to perform the corresponding sample size calculations using the large sample normal approximation methods based on Wald confidence limits ($\hat{\pi} \pm z_{(1-\alpha/2)}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$) [15]. The sample size formula for this method is $n = \hat{\pi}(1-\hat{\pi})(z_{1-\alpha/2})^2/e^2$ [23], where e is the desired error to form the limits of the confidence

interval. For each calculated sample size, the value of x (number of binomial successes) was determined as the integer that yielded a proportion closest to the hypothesized proportion. Mid-P adjusted exact confidence intervals were calculated for these proportions (x/n) using standard statistical software [22]. The width of the resulting confidence interval was determined and compared to the nominal (desired) width. A percent deviation from the nominal width was calculated as

$$\text{per cent width deviation} = [(\text{observed width} - \text{expected width})/\text{expected width}] * 100$$

6. RESULTS

The sample sizes estimated using the modified exact algorithm were very similar to the usual normal approximation methods for evaluated proportions between 0.5 and 0.8 (Table I). The width of the mid-P confidence interval formed using the calculated sample sizes from both methods were less than or equal to the nominal width for evaluated proportions between 0.5 and 0.85 (Figure 1). The sample sizes calculated for proportions greater than 0.8 were larger using the modified exact method and resulted in confidence interval widths being noticeably different for the two methods (Figure 2). The sample sizes based on the normal approximation resulted in confidence intervals that were often too wide (larger than nominal width). The corresponding intervals for the exact method tended to be narrower than the specified length. Figure 2 also suggests a cyclical pattern to the confidence interval width. These fluctuations are most likely the result of the data being discrete and the point estimate (x/n) not being exactly the same for the different sample sizes. The difference in confidence interval width between the modified exact and normal approximation methods becomes more dramatic as the proportion approaches 1 (Figure 3).

Table I. Comparison of sample sizes for the estimation of binomial proportions at two specified levels of precision and three levels of confidence.

	90 per cent confidence level				95 per cent confidence level				99 per cent confidence level			
	±0.10		±0.05		±0.10		±0.05		±0.10		±0.05	
Proportion	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx	Exact	Approx
0.50	68	68	270	271	96	97	384	385	164	166	662	664
0.55	68	67	268	268	95	96	380	381	162	165	655	657
0.60	65	65	260	260	92	93	369	369	160	160	635	637
0.65	62	62	248	247	88	88	351	350	151	151	605	604
0.70	59	57	229	228	80	81	323	323	140	140	560	558
0.75	52	51	204	203	72	73	288	289	128	125	500	498
0.80	45	44	175	174	64	62	249	246	115	107	430	425
0.85	39	35	140	138	53	49	200	196	99	85	353	339
0.90	29	25	100	98	47	35	148	139	80	60	260	239

Exact = sample size calculated using the modified exact computer algorithm.
Approx = sample size calculated using the usual large sample normal approximation method.

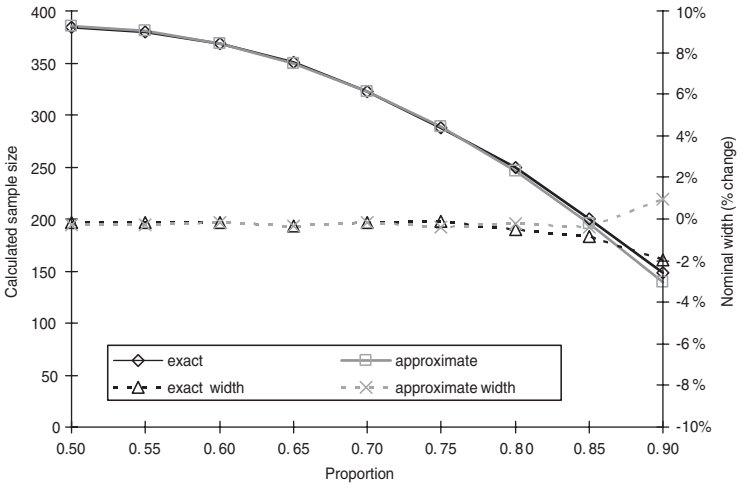

Figure 1. Comparison of sample size estimates for modified exact and normal approximation methods and their ability to achieve nominal width of mid-P confidence intervals. Sample sizes estimated for specified proportion with 0.05 precision and 95 per cent confidence level. Lines (dashed and solid) are only included to help visualize trends among data points.

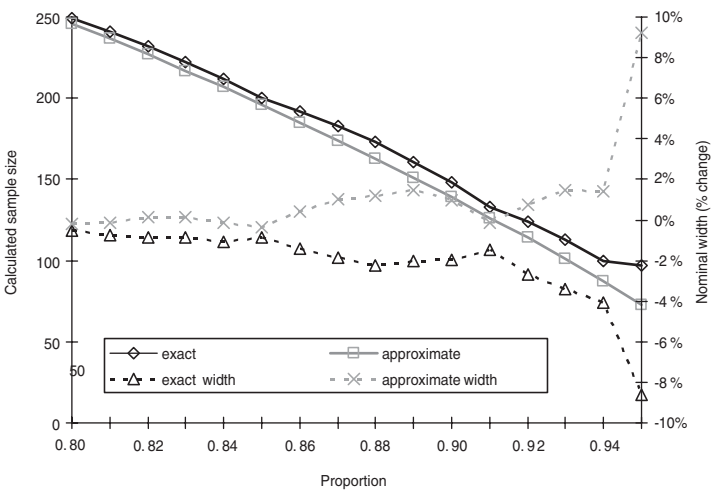

Figure 2. Comparison of sample size estimates for modified exact and normal approximation methods and their ability to achieve nominal width of mid-P confidence intervals. Sample sizes estimated for specified proportion with 0.05 precision and 95 per cent confidence level. Lines (dashed and solid) are only included to help visualize trends among data points.

7. DIAGNOSTIC TEST APPLICATIONS

The development of PCR-based diagnostics has resulted in assays with high degrees of accuracy, especially diagnostic specificity. It is important to choose the appropriate

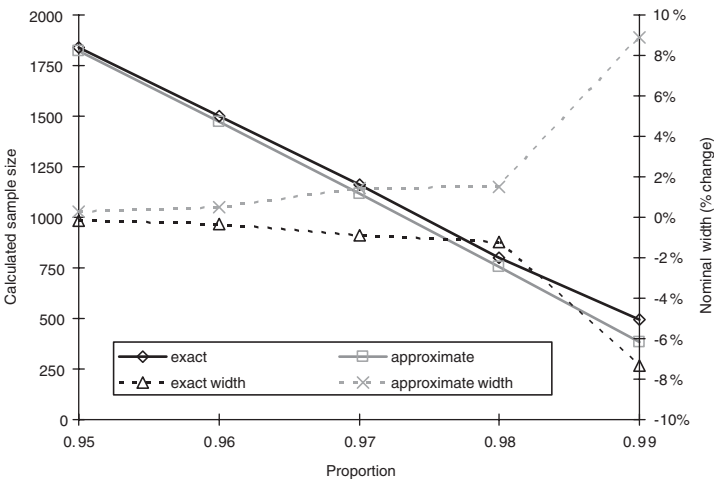

Figure 3. Comparison of sample size estimates for modified exact and normal approximation methods and their ability to achieve nominal width of mid-P confidence intervals. Sample sizes estimated for specified proportion with 0.01 precision and 95 per cent confidence level. Lines (dashed and solid) are only included to help visualize trends among data points.

confidence interval estimation method when reporting the diagnostic accuracy of tests that have near perfect sensitivity or specificity. To evaluate the effect of near perfect diagnostic parameters (sensitivity and specificity), coverage probabilities were determined for several situations in which the observed proportion was greater than or equal to 0.9. Three common methods of confidence interval formation—Wilson's score method [24] with continuity correction [25], mid-P exact, and Wald were determined using a binomial distribution with $n = 200$ and success probabilities ranging from 0.90 to 0.99 by 0.01. The score method incorporating continuity correction is referred to as the Fleiss quadratic method by the employed software [22] and is derived from the standardization of a proportion using the following formula [25]:

$$z = \frac{|p - \pi_0| - 1/(2n)}{\sqrt{\pi_0(1 - \pi_0)/n}}$$

where π_0 is the null value for the proportion and $1/(2n)$ is the continuity correction factor. Confidence limits are then determined by solving this equation for p (upper and lower) using the appropriate $Z_{1-\alpha/2}$ value for the desired level of confidence.

A sample size of 200 was chosen as one that would be considered adequate for most diagnostic test evaluations, numerators would yield the desired proportion exactly, and $n\pi > 5$ and $n(1 - \pi) > 5$ would be satisfied for most evaluated proportions (0.90–0.97). The $n\pi > 5$ and $n(1 - \pi) > 5$ rule is often cited in introductory statistics books as situations when the usual Wald intervals are expected to perform adequately [26]. Confidence intervals for the number of binomial successes were performed using available software [22] and coverage probabilities were determined using the following formula [15]:

$$C_n(\pi) = \sum_{k=0}^{n} I(k, \pi) \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

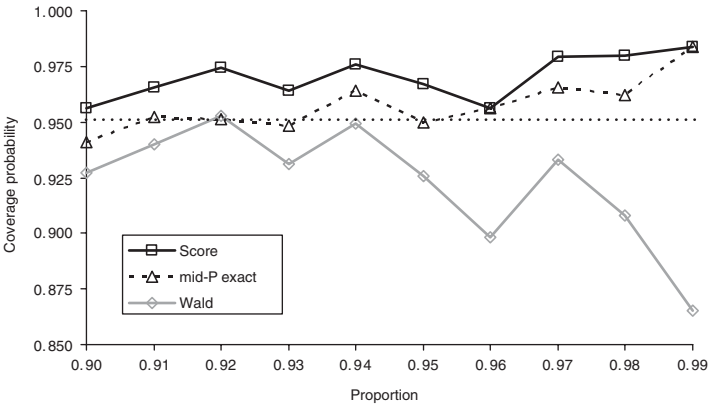

Figure 4. Comparison of coverage probability for three methods of 95 per cent interval estimation at multiple proportions relatively close to 1 and a sample size of 200. Lines (dashed and solid) are only included to help visualize trends among data points.

where $I(k, \pi)$ is equal to 1 if the interval contains π for the particular value of k and 0 otherwise.

The continuity-correct score method consistently resulted in confidence intervals with greater than nominal coverage for the evaluated proportions (Figure 4). Coverage probability for the Wald method was below nominal level for many of the evaluated proportions and was worse for values closest to 1. The mid-P adjusted exact method seemed to perform the best of the three methods and yielded coverage probabilities closest to the nominal level.

Confidence intervals based on inverting the Wald test tend to have erratic coverage probabilities that do not seem to be resolved simply by increasing the sample size [8, 15, 20]. The mid-P adjusted exact method, however, appears to give more consistent results over ranges of sample sizes and point estimates [9, 19]. These observations were supported by the results reported here and the coverage probability of the Wald intervals appear to worsen as the proportion approaches 1, which exacerbates the problem for the evaluation of some diagnostic tests.

8. DISCUSSION

The method of sample size estimation is important for the proper design and planning of an epidemiologic study concerning the validation of a diagnostic test, or the estimation of another population proportion. The author is not aware of another software routine that will calculate the sample size necessary to estimate a proportion within specified limits at a predetermined level of confidence that is not based on usual large sample normal approximation methods. The noted deficiencies in intervals estimated using such methods should caution investigators not to rely on similar formulae when determining the necessary size of a study. The sample size method presented here results in intervals that achieve nominal width (or narrower) even in situations when the hypothesized proportion is relatively close to 1.

The developed computer algorithm is based on the mid-P adjusted exact method and can therefore be computationally intense for certain combinations of proportion, interval width, and

confidence level. The algorithm was designed to be most efficient for proportions greater than 0.5 due to the employed probability formula (2). The program converts proportions <0.5 to values >0.5 ($1-$ entered proportion) before starting the algorithm. Results concerning proportions greater than or equal to 0.5 were only presented for this reason.

The computational intensity of the binomial probability function prevents the sample size algorithm from being able to solve all possible combinations of proportion, interval width, and confidence level. The program will fail to find the correct sample size when individual binomial probabilities approach 1×10^{-310}. When individual probabilities become functionally zero the overall sum of probabilities will start to decrease and the algorithm will continue to cycle without ever finding a suitable sample size. The program was therefore designed to end and print an error message upon reaching an individual probability of zero. This limitation could be improved by incorporation of variables with higher precision (e.g. 16 byte variables), but these were not available in the computing environment used by the author. This limitation is most severe for proportions close to 0.5 and is not much of an issue as the hypothesized proportion approaches 1. Therefore, the algorithm still appears to function well for most situations in which the normal approximation sample size methods perform poorly ($\pi > 0.85$).

Finally, the design of diagnostic test evaluation studies where the sensitivity or specificity is expected to be close to perfect would benefit from a new sample size method that is not based on the usual normal approximation methods. The modified exact method for sample size estimation is an improvement that would facilitate the evaluation of diagnostic tests. The development of other new sample size routines should be encouraged that would also aid in the design of future studies. It is important to provide these newly developed tools to the practicing epidemiologist to allow for their mainstream use.

ACKNOWLEDGEMENT

I would like to thank the anonymous referee for helpful suggestions, which resulted in a better overall manuscript.

REFERENCES

1. Kelsey JL, Whittemore AS, Evans AS, Thompson WD. *Methods in Observational Epidemiology* (2nd edn). Oxford University Press: New York, 1996; 327–339.
2. Greiner M, Gardner IA. Epidemiologic issues in the validation of veterinary diagnostic tests. *Preventive Veterinary Medicine* 2000; **45**(1–2):3–22.
3. Office International des Epizooties. Manual of standards: Diagnostic Tests and Vaccines 2000, http://www.oie.int/eng/normes/mmanual/a_00013.htm, accessed 10 June 2004.
4. Checkley SL, Waldner CL, Appleyard GD, Campbell JR, Forsythe L, Janzen ED. The comparison of 5 diagnostic tests for diagnosis of Johne's disease in western Canada. *The Journal of Applied Research in Veterinary Medicine* 2003; **1**(3):196–205.
5. Kurabachew M, Enger Ø, Sandaa RA, Skuce R, Bjorvatn B. A multiplex polymerase chain reaction for genus-, group and species-specific detection of mycobacteria. *Diagnostic Microbiology and Infectious Disease* 2004; **49**(2):99–104.
6. Kelley GO, Zagmutt-Vergara FJ, Leutenegger CM, Myklebust KA, Adkison MA, McDowell TS, Marty GD, Kahler AL, Bush AL, Gardner IA, Hedrick RP. Evaluation of five diagnostic methods for the detection and quantification of *Myxobolus cerebralis*. *Journal of Veterinary Diagnostic Investigation* 2004; **16**(3):202–211.
7. Farcas GA, Zhong KJY, Mazzulli T, Kain KC. Evaluation of the RealArt Malaria LC real–time PCR assay for malaria diagnosis. *Journal of Clinical Microbiology* 2004; **42**(2):636–638.
8. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; **16**(2):101–133.

9. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**(8):857–872.
10. White DB, James L. Standard error and sample size determination for estimation of probabilities based on a test variable. *Journal of Clinical Epidemiology* 1996; **49**(4):419–429.
11. Obuchowski N. Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research* 1998; **7**(4):371–392.
12. Alonzo TA, Pepe MS, Moskowitz CS. Sample size calculations for comparative studies of medical tests for detecting presence of disease. *Statistics in Medicine* 2002; **21**(6):835–852.
13. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. *Journal of Clinical Epidemiology* 2003; **56**(11):1118–1128.
14. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934; **26**(4):404–413.
15. Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician* 1998; **52**(2):119–126.
16. Agresti A. Exact inference for categorical data: recent advances and continuing controversies. *Statistics in Medicine* 2001; **20**(17–18):2709–2722.
17. Agresti A. Dealing with discreteness: making 'exact' confidence intervals for proportions, differences for proportions, and odds ratios more exact. *Statistical Methods in Medical Research* 2003; **12**(1):3–21.
18. Lancaster HO. Significance test in discrete distributions [corrections **57**(300):919]. *Journal of the American Statistical Association* 1961; **56**(274):223–234.
19. Berry G, Armitage P. Mid-p confidence intervals: a brief review. *Statistician* 1995; **44**(4):417–423.
20. Vollset SE. Confidence intervals for a binomial proportion. *Statistics in Medicine* 1993; **12**(9):809–824.
21. Compaq Computer Corporation. *Compaq Visual Fortran*: *Professional Edition* (*version 6.6*). Hewlett-Packard Company: Palo Alto, CA, 2002.
22. Centers for Disease Control and Prevention. *Epi Info* (*version 6.04*). Epidemiology Program Office, Centers for Disease Control and Prevention (CDC): Atlanta, GA, 1996.
23. Daniel WW. *Biostatistics*: *A Foundation for Analysis in the Health Science* (7th edn). Wiley: New York, 1999; 183–184.
24. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**(x):209–212.
25. Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions* (3rd edn). Wiley: New York, 2003; 26–36.
26. Daniel WW. *Biostatistics*: *A Foundation for Analysis in the Health Science* (7th edn). Wiley: New York, 1999; 176–177.

