Available online at www.sciencedirect.com
ScienceDirect
Preventive Veterinary Medicine 82 (2007) 151–158
PREVENTIVE VETERINARY MEDICINE
www.elsevier.com/locate/prevetmed

Short communication

Monte Carlo approaches for determining power and sample size in low-prevalence applications

Michael S. Williams *, Eric D. Ebel, Bruce A. Wagner

Animal and Plant Health Inspection Service, USDA 2150 B Centre Avenue, Mail Stop 2E6, Fort Collins, CO 80526, USA

Received 6 November 2006; received in revised form 27 April 2007; accepted 18 May 2007

Abstract

The prevalence of disease in many populations is often low. For example, the prevalence of tuberculosis, brucellosis, and bovine spongiform encephalopathy range from 1 per 100,000 to less than 1 per 1,000,000 in many countries. When an outbreak occurs, epidemiological investigations often require comparing the prevalence in an exposed population with that of an unexposed population. To determine if the level of disease in the two populations is significantly different, the epidemiologist must consider the test to be used, desired power of the test, and determine the appropriate sample size for both the exposed and unexposed populations. Commonly available software packages provide estimates of the required sample sizes for this application. This study shows that these estimated sample sizes can exceed the necessary number of samples by more than 35% when the prevalence is low. We provide a Monte Carlo-based solution and show that in low-prevalence applications this approach can lead to reductions in the total samples size of more than 10,000 samples.
Published by Elsevier B.V.

Keywords: Two population; Proportion; Eradication; Animal surveillance

1. Introduction

Consider the problem of determining an adequate sample size to detect a specified difference in the prevalence of disease in two populations (e.g., Adcock, 1997). In such a study, the null hypothesis is that no difference in prevalence exists in the populations (i.e., H_0: $\pi_1 = \pi_2$). However, in many epidemiological investigations, the goal is to determine whether the difference in the prevalence exceeds a pre-determined threshold. For example, suppose one wishes to determine if the prevalence in an exposed population is at least five times higher than the prevalence in the

* Corresponding author. Tel.: +1 970 494 7306; fax: +1 970 494 7174.
E-mail address: michael.s.williams@aphis.usda.gov (M.S. Williams).

0167-5877/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.prevetmed.2007.05.015

unexposed population (i.e., H_1: $\pi_1 \geq 5\pi_2$). If α is the significance of the test and if π_1 and π_2 are the true prevalences, then the power of the test, denoted by $1 - \beta$, is the probability that the null hypothesis is rejected whenever $\pi_1 \geq 5\pi_2$ (i.e., $Pr[\text{reject } H_0|H_1] = 1 - \beta$). A common value for β is 0.20, but lower values should be considered when the cost of not detecting the difference is high in comparison to the cost of collecting the data.

Casagrande et al. (1978) and Fleiss et al. (1980) provide estimators to determine the sample size needed to detect a difference between two populations with a specified significance level and power. Due to the discrete nature of the data and the reliance of these estimators on the asymptotic behavior of the test statistic, a number of different continuity corrections have been suggested to the original estimator given by Fleiss et al. (1980). Gordon and Watson (1996) summarize the results of numerous authors and conclude that continuity correction is rarely beneficial.

Commonly available software packages have implemented many a number of different sample size estimators, with examples being EpiInfo (CDC, 2006), the *Hmisc* library for *R* and *S+* (Alzola and Harrell, 2006), the *sampsi* function in Stata (StataCorp, 2003), and the *Power* procedure in SAS (O'Brien, 1998).

These sample size estimators have been shown to work well in many applications. However, the range of prevalences considered in these studies is often orders of magnitude larger than the prevalence levels encountered in many animal surveillance applications, particularly when the disease has been nearly eradicated from the populations in question. In this study, we consider the performance of two of these estimators and show that the suggested sample sizes are often very inaccurate when the prevalence of the disease is low. We propose a Monte Carlo simulator, combined with a binary search algorithm, to determine the appropriate sample size to achieve a test with a given power. A simulation study shows that while the Monte Carlo-based solution performs well, the estimated sample sizes provided by the other two methods can exceed the necessary number of samples by more than 35% when the prevalence is low. Computer code has been made available to implement the Monte Carlo-based solution in either *R* or *S+*.

2. Review

Consider two large populations where the true proportion of diseased animals is given by π_1 and π_2, respectively. From each of the populations, a random sample of size n_1 and n_2 is drawn and x_1 and x_2 diseased animals are found. Thus, x_1 and x_2 are such that $X_1|n_1,\pi_1 \sim Binomial(n_1,\pi_1)$, $X_2|n_2,\pi_2 \sim Binomial(n_2,\pi_2)$, and $p_1 = x_1/n_1$ and $p_2 = x_2/n_2$ are the estimators of π_1 and π_2.

The statistic used in the test is

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\text{var}(p_1 - p_2)^{1/2}},$$

where $\text{var}(p_1 - p_2) = \text{var}(p_1) + \text{var}(p_2) - 2\text{cov}(p_1, p_2) = \text{var}(p_1) + \text{var}(p_2)$ because the samples in each population are assumed to be independent.

Under H_0, the prevalence of the disease is the same in each population, so $\pi_1 = \pi_2$ and

$$z = \frac{(p_1 - p_2)}{\text{var}(p_1 - p_2)^{1/2}}.$$

The sampling distribution of this statistic is approximately Normal and the distribution of the test statistic agrees well with a standard Normal distribution when both the sample size and proportion of diseased animals are high.

For the typical significance level of $\alpha = 0.05$, the test statistic will fall in the interval $(-1.96, 1.96)$ for roughly 95% of all samples. In other words, if the null hypothesis is true and the sample size is sufficient for the distribution of z to be approximately Normal, then

$$P[-z_{\alpha/2} < z < z_{\alpha/2}] \approx 0.95.$$

The interpretation of failing to reject H_0 can be misleading because even though a test fails to reject the hypothesis that $\pi_1 = \pi_2$, it does not imply that no difference exists. Rather the result can imply that, for the given sample size, the difference in the two populations was too small to be detected. It can be misleading to use statistical tests without considering their power, where the power of a statistical test is the probability that H_0 will be rejected when the difference between the two population parameters is $\pi_1 - \pi_2$.

Assume that the specified significance level is $\alpha = 0.05$ and *a priori* it is known that the appropriate alternative hypothesis is H_1: $\pi_1 > \pi_2$. Then the power of the α-level test for the null hypothesis H_0: $\pi_1 = \pi_2$ is given by

$$P[\text{reject } H_0 | H_0 \text{ false}] = P[z > z_{\alpha/2}].$$

The power of the test can be determined for a given $\pi_1 - \pi_2$ as follows;

$$P[z > z_{\alpha/2}] = P\left[\frac{(p_1 - p_2)}{\text{var}(p_1 - p_2)^{1/2}} > z_{\alpha/2}\right] = P\left[\frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\text{var}(p_1 - p_2)^{1/2}} > z_{\alpha/2} - \frac{(\pi_1 - \pi_2)}{\text{var}(p_1 - p_2)^{1/2}}\right]$$

$$= P\left[z > z_{\alpha/2} - \frac{(\pi_1 - \pi_2)}{\text{var}(p_1 - p_2)^{1/2}}\right] \simeq 1 - \Phi\left(z_{\alpha/2} - \frac{(\pi_1 - \pi_2)}{\text{var}(p_1 - p_2)^{1/2}}\right).$$

This result forms the basis for the derivation of the sample size calculation provided by Fleiss et al. (1980). Casagrande et al. (1978) derive the appropriate sample size for the case where an equal number of samples is taken from each population. However, in many cases an unequal sample size is desirable because of the factors such as the difference in cost to collect samples from each population. Let r define the relationship between the sample sizes drawn from each population. If n_1 is the sample size in the first population and $n_2 = rn_1$, with r specified in advance, then Fleiss et al. (1980) give

$$n_1 = \frac{[z_{\alpha/2}\sqrt{(r+1)\bar{\pi}(1-\bar{\pi})} + z_{\beta}\sqrt{r\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}]^2}{r(\pi_1 - \pi_2)^2}, \tag{1}$$

with $\bar{\pi} = (\pi_1 + r\pi_2)/(r+1)$. This formula is used to determine the approximate sample sizes in the *Hmisc* for *R* and the Power function in SAS using the Pearson's Chi-squared test option.

Fleiss et al. (1980) and Ury and Fleiss (1980) add the following continuity correction factor

$$n_1^{cc} = \frac{n_1}{4}\left[1 + \sqrt{1 + \frac{2(r+1)}{rn_1(\pi_1 - \pi_2)}}\right]^2. \tag{2}$$

This formula is used to determine the approximate sample sizes in the EpiInfo package. The utility of this correction factor has been questioned by Gordon and Watson (1996).

The sample sizes (n_1, n_2) and (n_1^{cc}, n_2^{cc}) will be referred to as the uncorrected and continuity corrected sample sizes, respectively.

3. Performance for low-prevalence populations

In the articles relating to the derivation of Eqs. (1) and (2) the prevalence levels considered typically ranged from $\pi_1 = 0.05$–0.80 and the differences between the π values in the two populations are relatively small. However, in many animal surveillance applications the proportion of diseased animals in the two populations can differ by orders of magnitude and at these low-prevalence levels a small number of infected animals can drastically change the estimated prevalence. The example used in this section relates to the prevalence of tuberculosis in an exposed and an unexposed population of wild deer. An initial small sample from the exposed population suggested an apparent prevalence of tuberculosis of four animals per 1000 ($\pi_1 = 0.004$). It was determined that samples from both populations should be collected so that at least a 10-fold difference in the prevalence between the two populations could be detected ($\pi_2 = 0.004$). The relatively small size of the geographical area that was thought to be exposed limited the total number of samples that could be collected, so the relationship chosen for the sample sizes was $r = 4$. Using Eqs. (1) and (2), the estimated sample sizes to achieve a power of 0.80 were ($n_1 = 1246$, $n_2 = 4984$) and ($n_1^{cc} = 1574$, $n_2^{cc} = 6296$), respectively.

Given the large discrepancy between the two sample sizes, a Monte Carlo simulation was performed to estimate the true power of the test for the different sample sizes. The simulator draws samples of size (n_1, n_2) and (n_1^{cc}, n_2^{cc}) from the appropriate binomial distributions and calculates the z statistic for each sample. This process is repeated 500,000 times to form a Monte Carlo approximation of the sampling distribution. Using this process, the achieved power for the two different sample sizes was 0.858 and 0.911, when using the uncorrected and continuity corrected sample sizes, respectively. The simulator was then used to determine that a sample size of only ($n_1^{mc} = 974$, $n_2^{mc} = 3896$) was sufficient to achieve a power of 0.80. This constitutes a reduction of 1360 and 3000 samples when compared to the sample sizes derived from Eqs. (1) and (2).

Fig. 1 illustrates the large discrepancy between the nominal and achieved power levels. Clearly, at these low-prevalence levels, the assumption that the distribution of the z statistics approaches that of a unit Normal is not appropriate. Extensive simulation suggests that the sample size estimates derived from Eqs. (1) and (2) consistently overestimate the required sample size.

4. A Monte Carlo approach to sample size determination

At the low-prevalence levels encountered in some surveillance applications, the assumption that the statistic z follows an underlying Normal distribution in repeated samples of equal size is not tenable. One option for determining the appropriate sample size is to use a Monte Carlo approach to "search" for a sample size that achieves the desired level of power. While an exhaustive search for the appropriate sample size is possible, a more efficient approach takes advantage of the fact that the power of the test increases monotonically with increasing sample size (Fig. 1). So rather than perform an extensive search for possible sample sizes, a search can be employed to efficiently find the appropriate sample size to within a user specified tolerance. A binary search, which is a technique for finding a particular value in an order list by ruling out half of the data at each step, is an efficient method. The algorithm for finding the appropriate sample size is as follows:

(1) Choose a tolerance value that describes the acceptable discrepancy between the nominal and actual power of the test.

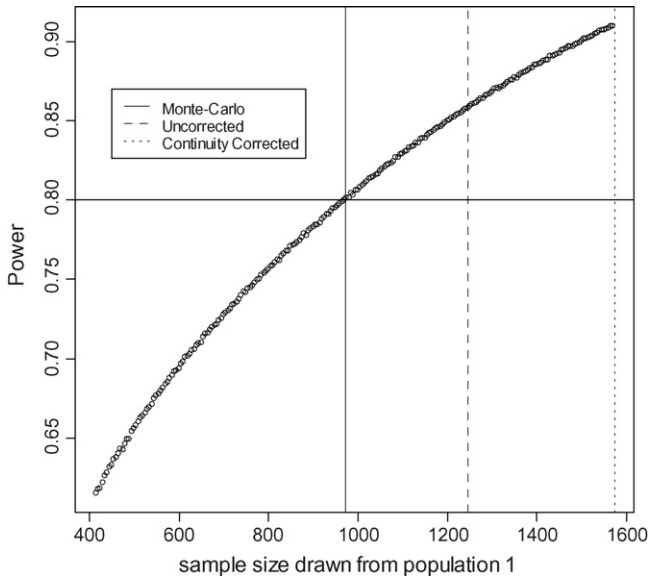

Fig. 1. The achieved power of the test as a function of the sample size in the exposed population (n_1). The design prevalences in the exposed and unexposed populations were $\pi_1 = 0.004$, $\pi_2 = 0.0004$, respectively. The vertical lines show the power achieved by sample sizes in the exposed population dictated by the Monte Carlo, uncorrected, and continuity correction-based approaches.

(2) Select an upper and lower bound for the sample size. In low-prevalence applications the uncorrected sample size (i.e., n_1) serves as a reasonable upper bound and $n_1/3$ is an acceptable choice for the lower bound.
(3) Assess the power at the upper and lower bounds of this interval with the Monte Carlo simulator.
(4) Use the binary search algorithm to choose new upper and lower bounds of an interval that contains the desired sample size.
(5) Repeat steps 3 and 4 until the desired tolerance level is obtained.

The sample sizes derived from the search algorithm above will be denoted by (n_1^{mc}, n_2^{mc}). *R* and *S+* code to implement the Monte Carlo sample size calculations is available at http://www.aphis.usda.gov/vs/nahss/resources.htm#software.

5. Simulations

A series of examples illustrate the potential reduction in sampling effort associated with using the Monte Carlo approach to sample size determination. The goal is to illustrate the factors and situations where the use of the Monte Carlo approach is most beneficial. A tolerance of 0.00025, for the discrepancy between the nominal and achieved power of the test, was chosen for this study. The three factors were:

- The size of the effect that was to be detected. The values chosen were a 2-, 4-, and 10-fold difference in the prevalence. These will be referred to as the *effect size*.

Table 1
Summary statistics for the simulation study

Prevalence, π_1	Effect size	Allocation to each population, r	Achieved power for each mc, uc, cc	Percent increase in sample size, Per(uc, mc)	Percent increase in sample size, Per(cc, mc)	Monte Carlo sample size, n_1^{mc}	Sample size difference, Δ(uc, mc)	Sample size difference, Δ(cc, mc)
0.1	2	1	79.9, 81.1, 84.1	2.3	10.4	423	20	98
0.01	2	1	80.2, 80.9, 83.9	2.1	9.6	4,578	190	974
0.001	2	1	80.1, 81.0, 84.1	2.5	10.0	45,816	2482	10,326
0.1	4	1	80.2, 82.5, 88.0	5.5	18.6	306	18	70
0.01	4	1	80.1, 83.7, 88.7	8.8	20.6	1,584	296	812
0.001	4	1	80.0, 83.9, 88.7	9.0	20.8	15,847	3166	8,322
0.1	10	1	80.3, 85.4, 92.1	11.0	26.4	89	22	64
0.01	10	1	80.1, 86.1, 92.0	14.1	28.5	910	298	724
0.001	10	1	80.1, 86.1, 92.0	14.6	28.8	9,100	3104	7,356
0.1	2	4	79.8, 80.3, 83.6	0.8	9.5	247	10	130
0.01	2	4	80.0, 80.4, 83.4	0.8	9.1	2,648	80	1,305
0.001	2	4	80.0, 80.3, 83.5	0.6	8.9	26,697	590	12,825
0.1	4	4	80.2, 82.3, 87.7	5.9	20.8	80	25	105
0.01	4	4	80.0, 81.5, 86.8	4.1	18.6	863	170	970
0.001	4	4	80.0, 81.6, 86.8	4.2	18.6	8,633	1945	9,950
0.1	10	4	79.9, 85.4, 91.0	21.3	38.3	37	50	115
0.01	10	4	80.0, 85.8, 91.0	22.1	38.2	386	555	1,210
0.001	10	4	80.0, 85.8, 91.1	22.3	38.5	3,889	5520	12,080

The differences in the estimated samples sizes necessary to achieve a test with power of 0.8 in low-prevalence applications are summarized. The achieved power and metrics describing the difference in the estimated number of samples using a Monte Carlo approach and two alternatives.

- The proportion of affected animals in each population. Three different prevalence levels were considered for the exposed population. These were $\pi_1 = 0.1, 0.01, 0.001$. The prevalence levels in the unexposed population were determined by the effect size.
- The ratio, r, determines the allocation of the sample size to each subpopulation. The values $r = 1$ and 4 were used.

For each combination of these three factors, the study determined the sample size using the Monte Carlo, uncorrected and continuity corrected approaches and compared the results using a series of metrics. The first metric for comparison is the percent reduction in total sample size resulting from the use of the Monte Carlo sample size, which is

$$\text{Per}(\text{uc}, \text{mc}) = 100\frac{(n_1 + n_2) - (n_1^{\text{mc}} + n_2^{\text{mc}})}{(n_1 + n_2)}$$

and

$$\text{Per}(\text{cc}, \text{mc}) = 100\frac{(n_1^{\text{cc}} + n_2^{\text{cc}}) - (n_1^{\text{mc}} + n_2^{\text{mc}})}{(n_1^{\text{cc}} + n_2^{\text{cc}})}$$

for the uncorrected and Monte Carlo-based techniques, respectively. The achieved power for each of the methods is also given. The final metric is the total reduction in sample size in comparison to the uncorrected and continuity correct methods is also provided (i.e., $\Delta(\text{uc}, \text{mc}) = (n_1 + n_2) - (n_1^{\text{mc}} + n_2^{\text{mc}})$ and (i.e., $\Delta(\text{cc}, \text{mc}) = (n_1^{\text{cc}} + n_2^{\text{cc}}) - (n_1^{\text{mc}} + n_2^{\text{mc}})$). If the cost of collecting and testing each sample is known, this metric represents the total potential reduction associated with using the Monte Carlo-based sample sizes.

6. Results

The results are given in Table 1 where there are a number of clear patterns. The first is that the difference between the achieved power for the non-Monte Carlo methods is always greater than the nominal value of 80%, with the continuity corrected sample sizes overestimating the required sample size by a substantial amount. The level of the bias is determined by the effect size, with bias in the power increasing in accordance with the effect size. The allocation (r) and the prevalence level had little or no affect on the achieved power for the various sample sizes.

In contrast, both the allocation of the sample (r) and prevalence levels significantly influenced the difference in the estimated sample size provided by the non-Monte Carlo methods. As the prevalence decreased, the percentage of excess samples increased from 0.7% to as much as 38.5%. The number of excess samples that these methods estimate ranges from as little as 10 to nearly 13,000 samples.

7. Conclusions

The results of this study suggest that sample size calculations that rely on the assumption of a Normal distribution often overestimate the required number of samples to achieve a specified power. This poor performance is due to the failure of the distributional assumptions when the prevalence of the disease is low. In contrast, the proposed Monte Carlo approach returns sample sizes such that the achieved power of the test closely matches the nominal value. The examples also illustrate that the use of Monte Carlo methods can reduce the overall sample size by hundreds to thousands of samples while still meeting the study objectives.

References

Adcock, C.J., 1997. Sample size determination: a review. Statistician 46, 261–283.

Alzola, C.F., Harrell, F.E., 2006. An introduction to S and the Hmisc and design libraries. http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/Hmisc.

Casagrande, J.T., Pike, M.C., Smith, P.G., 1978. An improved simple approximate formula for calculating sample sizes for comparing binomial distributions. Biometrics 34, 483–486.

Centers for Disease Control, 2006. EpiInfo Center for Disease Control and Prevention (CDC). http://www.cdc.gov/epiinfo.

Fleiss, J.L., Tytun, A., Ury, H.K., 1980. A simple approximation for calculating sample sizes for comparing independent proportions. Biometrics 36, 343–349.

Gordon, I., Watson, R., 1996. The myth of continuity corrected sample size formulae. Biometrics 52, 71–76.

O'Brien, R.G., 1998. Tour of UnifyPow: a SAS module/macro for sample-size analysis. Proceedings of the Twenty-Third Annual SAS Users Group International Conference, SAS Institute Inc., Cary, NC, pp. 1346–1355. Software and updates to this article can be found at http://www.bio.ri.ccf.org/UnifyPow.

StataCorp, 2003. Stata Statistical Software: Release 8.0. Stata Corporation, College Station, TX.

Ury, H.K., Fleiss, J.L., 1980. On approximate sample sizes for comparing two independent proportions with the use of Yates' correction. Biometrics 36, 347–351.