

A practical approach to calculate sample size for herd prevalence surveys

Roger W. Humphry^{a,*}, Angus Cameron^b, George J. Gunn^a

^a*Epidemiology Unit, Veterinary Science Division, Scottish Agricultural College,
Drummondhill, Stratherrick Road, Inverness IV2 4JZ, UK*

^b*AusVet Animal Health Services, 140 Falls Road, Wentworth Falls, NSW 2782, Australia*

Received 24 September 2002; received in revised form 13 July 2004; accepted 13 July 2004

Abstract

When designing a herd-level prevalence study that will use an imperfect diagnostic test, it is necessary to consider the test sensitivity and specificity. A new approach was developed to take into account the imperfections of the test. We present an adapted formula that, when combined with an existing piece of software, allows improved planning. Bovine paratuberculosis is included as an example infection because it originally stimulated the work. Examples are provided of the trade-off between the benefit (low number of herds) and the disadvantage (large number of animals per herd and exclusion of small herds) that are associated with achieving high herd-level sensitivity and specificity. We demonstrate the bias in the estimate of prevalence and the underestimation of the confidence range that would arise if we did not account for test sensitivity and specificity.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Prevalence; Survey design; Imperfect test; Sample size; Sensitivity; Specificity; Paratuberculosis; Johne's disease

1. Introduction

Valid estimates of herd-level prevalence can be obtained from population surveys using cluster sampling. Herds are selected at random and a diagnostic test is applied to randomly selected animals from these selected herds. Based on the results of individual-animal tests,

* Corresponding author. Tel.: +44 1463 243030; fax: +44 1463 711103.

E-mail address: roger.humphry@sac.ac.uk (R.W. Humphry).

each herd is assessed as either positive or not positive (thus providing a herd-level test so that the herd-level prevalence can be estimated).

The difficulty with this approach lies in that most tests have imperfect animal-level sensitivity (SENS) and specificity (SPEC), which means that the categorisation of the herd as either positive or negative (i.e. herd tests) is also imperfect.

In one of the most basic methods (which we refer to as Method 1 in this paper), calculation of sample size for both first and second stages in prevalence surveys is based on the assumption that the test used (both at the herd and animal level) is perfect. We describe a different process (Method 2) for calculation of both first-stage and second-stage sample sizes for a survey to estimate herd prevalence. Our process accounts for the imperfect test at both levels. This issue appears not to have been addressed elsewhere for survey design for herd prevalence. The process was designed to estimate the herd prevalence of bovine paratuberculosis (Johne's disease) in Great Britain (GB), using a test that is far from perfect but the process is equally applicable to other infections or diseases. This paper retains paratuberculosis as an example of the approach and, consequently, refers to "infection" throughout. We provide a derivation for a formula to estimate first-stage sample size, taking herd-test sensitivity (HSENS) and specificity (HSPEC) into account. The formula has been used in combination with available software to calculate a range of sample sizes based on varying assumptions about test performance and infection prevalence.

2. Methodology

2.1. Method 1: a basic approach

This approach is based on the assumptions of perfect classification of each herd tested and of each animal, as either positive or negative and that we test all the animals in a herd at the second stage.

First-stage sample size might be estimated using the simple formula based on the normal approximation to the binomial distribution (Snedecor and Cochran, 1989)

$$HN = \left(\frac{1.96}{L} \right)^2 \times HTP(1 - HTP) \quad (1)$$

where HN is the sample size (number of herds tested), HTP is the estimated true herd prevalence and L is the tolerance around the prevalence for the 95% confidence limits, i.e. the desired maximum size of confidence (Table 1).

The above is based on an assumption of an approximately infinite population size and perfect test. Cannon and Roe (1982) provide tables for sampling programmes to establish *freedom from infection* (as opposed to prevalence) that take into account a finite population, assuming a perfect test and use an approximation to the hypergeometric distribution.

In practice, however, tests are not perfect. Methods exist to:

- (a) establish the number of animals within a herd to be tested so as to achieve a set HSENS and HSPEC (Cameron and Baldock, 1998) taking into account the test SENS and SPEC;

Table 1
Symbolic notation used in formulae (using Jordan, 1998 as template)

Abbreviation	Meaning
HTP	True prevalence of infected herds (assumed)
HAP	Apparent herd prevalence (calculated)
TPWH	True prevalence within herds (assumed)
HSSENS	Herd-level sensitivity (chosen)
HSPEC	Herd-level specificity (chosen)
SENS	Sensitivity at the individual-animal level (assumed)
SPEC	Specificity at the individual-animal level (assumed)
D	Number of infected animals within the positive herd (assumed)
N	Number of animals within the herd (assumed)
L	The tolerance (chosen)
C	The level of confidence for the tolerance (chosen)
c	Cut-point number of test-positive animals denoting a test-positive herd (calculated)
HN	Number of herds sampled (calculated)
HT	Number of herds testing positive (calculated)
n	Number of animals sampled within the herd (calculated)
AACR	Approximate apparent confidence range
ATCR	Approximate true confidence range

- (b) estimate the true prevalence (HTP) based on the apparent prevalence (HAP) and on known test SENS and SPEC (Rogan and Gladen, 1978); and
- (c) calculate sampling programmes for substantiating freedom from infection (Cameron and Baldock, 1998; Cannon, 2001).

2.2. Method 2: survey design allowing for imperfections in the test

2.2.1. Sample size formula for prevalence estimation where the herd-level test is imperfect

We adapted the work of Rogan and Gladen (1978) by assuming a normal approximation to the binomial distribution (see Appendix A for derivation and Table 1 for parameter definitions).

$$HN = \left(\frac{Z(C)}{L} \right)^2 \times \frac{(\text{HSSENS}(\text{HTP}) + [1 - \text{HSPEC}][1 - \text{HTP}]) \times (1 - \text{HSSENS}(\text{HTP}) - [1 - \text{HSPEC}][1 - \text{HTP}])}{(\text{HSSENS} + \text{HSPEC} - 1)^2} \quad (2)$$

This allowed us to estimate the number of herds to be tested to establish a herd-level prevalence; we used an a priori estimate of 10% for paratuberculosis as our example with 95% confidence limits for a tolerance of 5%. Consequently, the estimate depends on the levels of confidence and tolerance required, the estimated herd prevalence (HTP) and also the selected HSSENS and HSPEC. The HSSENS and HSPEC are chosen (as opposed to assumed) and can be of any value such that HSSENS + HSPEC > 1.

2.2.2. Determination of herd status

For a range of herd sizes, SENSs and SPECs, we established the number of animals that need to be tested to achieve the HSENS and HSPEC used in stage one. The method of [Cameron and Baldock \(1998\)](#) and the latest version of FreeCalc 2.0 ([Cameron, 2001](#)) were used. This is based on sampling a finite population of animals (e.g. within a herd) and uses an approximation to the hypergeometric distribution. There were problems with these calculations for small herds and we will discuss them later.

2.3. Assumptions

Calculations were performed using a range of values for SENS, SPEC, the minimum within-herd prevalence (TPWH) for an infected herd and a single a priori estimate of herd prevalence ([Table 2](#)).

2.4. Comparison of proposed method with a survey design that assumes a perfect test used on all animals

For the two contrasting approaches we calculated: (a) the HAP; (b) the approximate apparent 95% confidence limits based on the HAP; and (c) the approximate true 95% confidence limits (formulae provided in [Appendix B](#)). This comparison was made because, whilst the methods already exist to account for SENS and SPEC for establishing the status of a herd ([Cameron and Baldock, 1998](#), [Cannon, 2001](#)) these methods have to consider HSENS and HSPEC as pre-set goals. Therefore, HSENS and HSPEC can be ignored only when *animal-level* SENS and SPEC also are considered to be 100% and all animals in the herd are tested.

Table 2
Range of values and references for assumptions associated with paratuberculosis

Parameter	Notation	Estimate (%)	Reference
Test sensitivity (animal-level)	SENS	35	Whitlock et al. (2000)
		50	Collins and Morgan (1991) ; Meylan et al. (1995) ; Reichel et al. (1999) ; Sweeney et al. (1995)
		65	See Section 4
Test specificity (animal-level)	SPEC	97	Collins and Morgan (1991) ; Meylan et al. (1995) ; Sweeney et al. (1995)
		99	Cox et al. (1991) ; Meylan et al. (1995) ; Reichel et al. (1999)
			Turnquist et al. (1991) ; Wells and Wagner (2000)
Minimum within-herd prevalence for a diseased herd	TPWH	3	
		5	Collins et al. (1994) ; Meylan et al. (1995)
		10	Collins et al. (1994) ; Obasanjo et al. (1997) ; Ott et al. (1999) ; Wells and Wagner (2000)
A priori estimate of herd prevalence	HTP	10	Kennedy (2000)

Selecting a herd prevalence of 10% (e.g. Johne's (Kennedy, 2000)), the number of herds tested (HN) based on Method 1 was 139 herds. Unsurprisingly, the number of herds to be tested was fewer than if the test imperfections were accounted for using Method 2. We calculated the HAP and approximate apparent confidence limits (AACR) if the test imperfections were ignored not only in survey design but also post sampling. In addition, we calculated the approximate true confidence range (ATCR) that such a sampling design would lead to if test imperfections then were accounted for, post sampling. The formulae used to calculate these numbers are presented in Appendix B. This comparison was carried out for a range of assumed parameter values with a common "standard" set of values based on Johne's as a good example of a disease with imperfect test sensitivity and specificity. It is common practice to describe such a study of the rate of change of an output parameter (e.g. apparent prevalence) with respect to an input parameter (e.g. test specificity) as "sensitivity analysis" (e.g. Thrusfield, 1995). For the purposes of this paper alone, in which test and herd sensitivity and specificity are the central themes, we hope to avoid confusion by describing rate of change of output with respect to input as "responsiveness".

3. Results

3.1. Number of herds to test

As HSENS or HSPEC increases, using Method 2, the number of herds (HN) to be tested is reduced (Table 3). To reduce the number of herds to be sampled further, the confidence can be relaxed (decreased) or the tolerance increased.

Table 3
Estimates of the number of herds to be tested based on alternative tolerance, confidence, herd-level sensitivity and specificity targets for a herd-level prevalence of 10%

Tolerance (%)	Confidence (%)	Herd-level specificity (%)	Herd-level sensitivity			
			55%	70%	85%	90%
5	95	55	38,169	6131	2400	1897
5	95	70	5394	2156	1164	984
5	95	85	1479	828	539	477
5	95	90	941	574	395	355
5	90	55	26,883	4319	1691	1336
5	90	70	3799	1518	820	693
5	90	85	1041	584	379	336
5	90	90	663	405	278	250
7.5	95	55	16,964	2725	1067	844
7.5	95	70	2398	958	517	438
7.5	95	85	657	368	240	212
7.5	95	90	419	255	176	158
7.5	90	55	11,948	1920	752	594
7.5	90	70	1689	675	365	308
7.5	90	85	463	260	169	150
7.5	90	90	295	180	124	111

Table 4

Estimates the number of animals to be tested within a herd to achieve a predetermined herd-level sensitivity and herd specificity (see Table 2)

Test sensitivity (%)	Test specificity (%)	Herd size	Herd sensitivity								
			≥55%			≥70%			≥85%		
			Herd specificity			Herd specificity			Herd specificity		
			≥55%	≥70%	≥85%	≥55%	≥70%	≥85%	≥55%	≥70%	≥85%
35	97	20	17	*	*	*	*	*	*	*	*
35	97	50	18	*	*	*	*	*	*	*	*
35	97	100	46	*	*	*	*	*	*	*	*
35	97	200	46	150	*	175	*	*	*	*	*
35	99	20	*	*	*	*	*	*	*	*	*
35	99	50	32	32	*	46	*	*	*	*	*
35	99	100	38	88	*	56	*	*	*	*	*
35	99	200	39	90	191	58	173	*	*	*	*
50	97	20	14	*	*	*	*	*	*	*	*
50	97	50	16	*	*	47	*	*	*	*	*
50	97	100	18	88	*	80	*	*	*	*	*
50	97	200	18	88	*	81	180	*	*	*	*
50	99	20	19	19	*	*	*	*	*	*	*
50	99	50	25	25	*	35	35	*	*	*	*
50	99	100	31	31	*	45	90	*	*	*	*
50	99	200	32	32	114	47	94	183	126	175	*
65	97	20	12	*	*	17	*	*	*	*	*
65	97	50	14	33	*	41	*	*	*	*	*
65	97	100	16	58	*	48	*	*	*	*	*
65	97	200	16	59	143	49	118	*	142	*	*
65	99	20	15	15	15	20	20	*	*	*	*
65	99	50	20	20	47	29	29	*	41	*	*
65	99	100	26	26	60	38	75	*	56	95	*
65	99	200	27	27	61	39	79	117	105	105	183

This is based on alternative test sensitivity, test specificity and herd size assumptions. The minimum within-herd prevalence is fixed at 3%. *No estimate possible; insufficient animals to achieve the required herd-level sensitivity or herd-level specificity.

3.2. Number of animals to test

Table 4 provides estimates of the number of animals to be tested within each herd to achieve the HSENS and HSPEC presented in Table 3 (assuming within-herd prevalence of 3%). It is possible to provide estimates (full cells in table) only for the larger herds and where desired HSENS and HSPEC are low. The number of animals to be tested increases as the herd size increases. For a given herd size, SENS and SPEC, the greatest number of animals required is for the highest values of HSENS and HSPEC sought. In general, one would expect that as test performance improves the number of animals to be tested falls. The results show that this is not always true (e.g. in Table 4 compare the estimate for a herd size of 100, herd sensitivity of 55%, herd specificity of 55%, test sensitivity of 50% and two values of test specificity: 97% and 99%). This is because, all else being equal, HSENS decreases as SPEC increases. Table 5 presents similar information to Table 4 but provides a

Table 5

Estimates of the number of animals to be tested within a herd to achieve a predetermined herd-level sensitivity and herd specificity (see Table 3)

Herd size	Minimum within-herd prevalence (%)	Herd sensitivity								
		$\geq 55\%$			$\geq 70\%$			$\geq 85\%$		
		Herd specificity			Herd specificity			Herd specificity		
		$\geq 55\%$	$\geq 70\%$	$\geq 85\%$	$\geq 55\%$	$\geq 70\%$	$\geq 85\%$	$\geq 55\%$	$\geq 70\%$	$\geq 85\%$
20	5	19	19	*	*	*	*	*	*	*
50	5	25	25	44	27	27	*	40	*	*
100	5	22	22	51	33	33	66	50	86	*
200	5	23	23	52	34	34	68	52	91	127
20	10	12	12	12	17	17	*	*	*	*
50	10	13	13	13	19	19	38	28	28	48
100	10	13	13	13	20	20	39	30	30	52
200	10	13	13	13	20	20	40	31	31	54

This is based on a test sensitivity of 50% and test specificity of 99% with alternative herd size and minimum within-herd prevalence assumptions. *No estimate possible; insufficient animals to achieve the required herd-level sensitivity or herd-level specificity.

comparison between 5% and 10% minimum within-herd prevalence, demonstrating that desired levels of HSENS and HSPEC are more easily achieved for higher values of the minimum within-herd prevalence.

Table 6 provides more detail of the number of animals to be sampled for a given herd size and includes the cut off number of reactors that define a positive herd (cut point). For the purposes of this illustration, we have considered the minimum number of infected animals within a positive herd to be fixed at five. This avoids the complexities (discussed by Cameron and Baldock, 1998) that arise from rounding the minimum number of infected animals when a fixed prevalence is used.

3.3. Comparison with the Method 1

Fig. 1 demonstrates that as the SPEC increases, the HAP decreases from close to 100% down to less than the true prevalence of 10% (HTP, Tables 1 and 2). The HAP is most responsive to SPEC as the latter approaches 100%. We see that the 95% approximate apparent confidence range (AACR) is greatest when the SPEC is around 99% (unsurprisingly where the HAP is 50%). The 95% approximate true confidence range (ATCR) based on the HAP with adjustments to take into account the HSENS and HSPEC is, like the HAP, highly responsive to the changes in SPEC over the range 97%–100%, and gets smaller as the SPEC approaches 100%.

Fig. 2 shows that, for the standard settings chosen in reference to Johne's, the HAP and AACR are not responsive to the test SENS. However, the ATCR increases rapidly as the sensitivity approaches 0. This is as expected because as SENS approaches zero, $\text{SENS} + \text{SPEC}$ approaches its lower bound of 1 (at which point Eq. (2) becomes unstable due to a denominator of zero).

Fig. 3 shows that, for our “standard” settings, the HAP, 95% AACR and 95% ATCR, adjusted for the HSENS and HSPEC, show most (albeit low) responsiveness to

Table 6

Estimates, using FreeCalc V2, of the within-herd sample size to achieve 70% herd sensitivity and specificity for a range of herd sizes where the number of infected animals in the herd is assumed to be 5 (for all herd sizes), test sensitivity is 50% and test specificity 99%

Herd size	Sample size	Cut point number of reactors
20	8	0
30	12	0
40	16	0
50	19	0
60	22	0
70	25	0
80	28	0
90	30	0
100	33	0
110	35	0
120	75	1
130	79	1
140	84	1
150	87	1
200	105	1
250	175	2
300	253	3
350	272	3
400	355	4

The cut point is the number of reactors above which, a herd should be defined as testing positive.

true within-herd prevalence (TPWH) at low values for TPWH. The HAP is not very sensitive but is substantially higher than the TPWH. The AACR and ATCR are both higher than the desired tolerance but are scarcely responsive to changes in TPWH at a prevalence > 0.05 .

Fig. 4 demonstrates that over the range of herd sizes tested, the HAP is an over-estimate of the HTP and increases as the number of animals in the herd (N) increases. The HAP increases asymptotically towards 100% as the herd size increases thus demonstrating greatest responsiveness for lower herd numbers. The 95% AACR shows a maximum at around 60 animals when the HAP is close to 50%. The 95% ATCR increases rapidly and with increasing gradient (responsiveness) as the number of animals in the herd increases.

4. Discussion

A new and practical process is presented for calculating sample sizes for estimating herd prevalence for a disease whose tests are not perfect. Our method (Method 2) combines the work of (Cameron and Baldock, 1998) with a separate equation (Eq. (2)).

The process suggested here is practical, but assumes a large number of herds from which to sample. In cases where the number of herds in the population is small, HSENS and HSPEC can be taken into account using Freecalc 2 (Cameron, 2001) so that imperfections

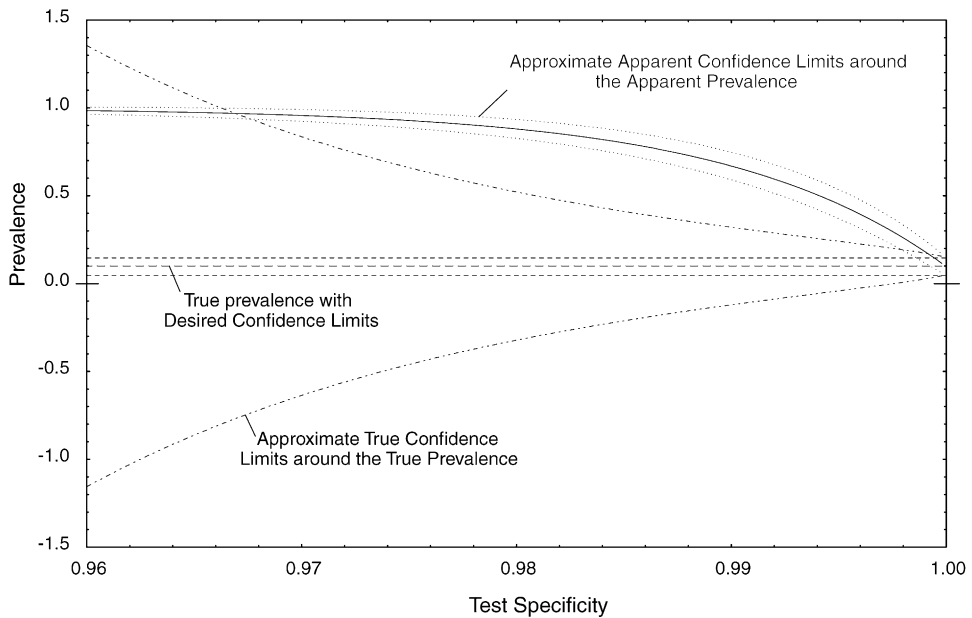


Fig. 1. Comparison of Methods 1 (naive) and 2 (corrected for test imperfection) regarding herd-level prevalences as the specificity of the test changes. The test sensitivity is 50%, the true within-herd prevalence is 5%, the number of animals in the herd is 100, the true herd prevalence (assumed) is 10%, a 95% confidence range is used and the desired tolerance is 5%. One hundred thirty-nine herds are tested based on Method 1.

in classification and finite populations are being accounted for, at both the within-herd and herd levels. We have not presented this method because Eq. (2) is more easily calculated, is more practical and is acceptable at the herd level if the number of herds is large. Furthermore, for the example infection illustrated here, one can see that the high number of herds required for testing would dictate that the total number of herds in the population be high. Therefore, in cases where the number of herds in the population is low and the test is highly imperfect, all herds would have to be tested to achieve meaningful confidence limits (and, therefore, estimates of the number of herds requiring to be tested become inappropriate).

Table 3 illustrates how the number of herds that need to be tested vary given four different HSENS values and four different HSPEC values. As HSENS or HSPEC improves (increases), the number of herds to be tested is reduced. However, if we wish to include smaller herds in the survey, then there is a problem: there are insufficient animals in these herds to achieve the desired HSENS and HSPEC. Therefore, any estimate of HTP would only include larger herds. If smaller herds were to be included, then the required low HSENS and HSPEC would dictate that a larger number of herds be tested.

The researcher also has the option to reduce the number of herds to be tested by adjusting the tolerance and confidence of the estimate (Table 3). To achieve a small reduction, the confidence in the estimate can be relaxed (decreased) to 90%. Alternatively, a larger reduction in herd numbers can be achieved when the tolerance is increased to

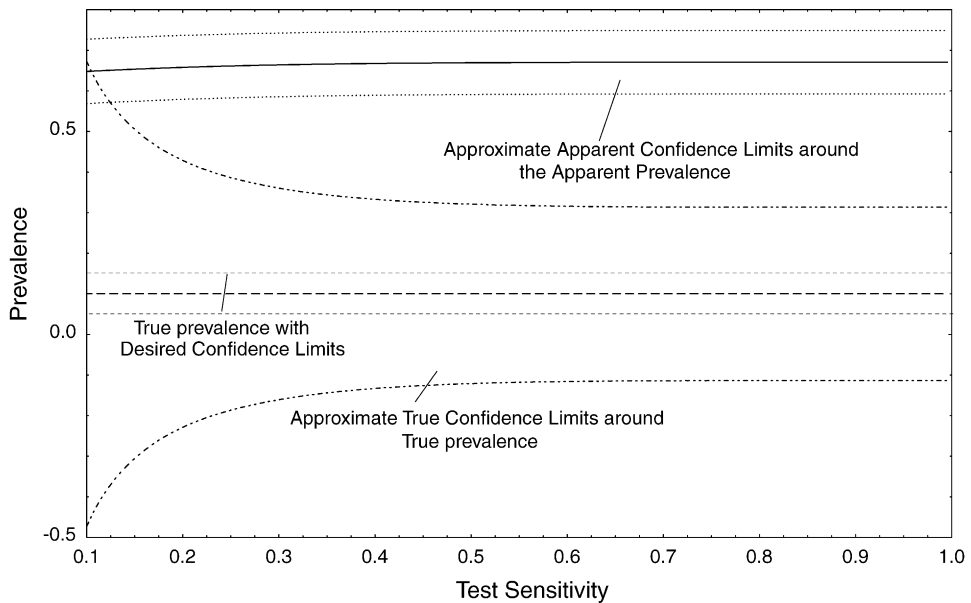


Fig. 2. Comparison of Methods 1 (naive) and 2 (corrected for test imperfection) regarding herd-level prevalences as the sensitivity of the test changes. The test specificity is 99%, the true within-herd prevalence is 5%, the number of animals in the herd is 100, the true herd prevalence (assumed) is 10%, a 95% confidence range is used and the desired tolerance is 5%. One hundred thirty-nine herds are tested based on Method 1.

$\pm 7.5\%$. Obviously, in our example, the smallest number of herds needs to be tested where confidence is relaxed to 90% and tolerance is increased to 7.5%.

The second stage using this method is to consider the number of animals to sample within each herd (n) for a given SENS and SPEC. However with Method 2, attention has to be paid to the predetermined (target) levels of HSENS and HSPEC (Tables 3 and 4). Each sample size was calculated using the updated, FreeCalc 2.0 (Cameron and Baldock, 1998) which is based upon an approximation to the hypergeometric distribution. Previous work has used an extension of the binomial distribution (Jordan, 1996) at the within-herd level. Use of the binomial distribution is not always reasonable because it assumes that the number of animals in the herd is infinite. The binomial distribution is inaccurate for finite populations and that inaccuracy increases as population size decreases. Thus, the hypergeometric approach is commonly more appropriate (Cameron and Baldock, 1998).

Table 4 illustrates that as SENS improves, fewer animals (n) require to be sampled from each herd. Furthermore, as SENS improves, higher levels of HSPEC and HSENS can be achieved. It is clear that higher HSENS can be attained only in large herds (and in these cases, most animals must be sampled).

If SPEC is over-estimated, then the effects are important. As SPEC falls, it is increasingly difficult to achieve high HSENS and HSPEC.

Now we can clearly illustrate, by working between Tables 3 and 4, the problems encountered when HSENS and HSPEC are included (as they should be) in sample size

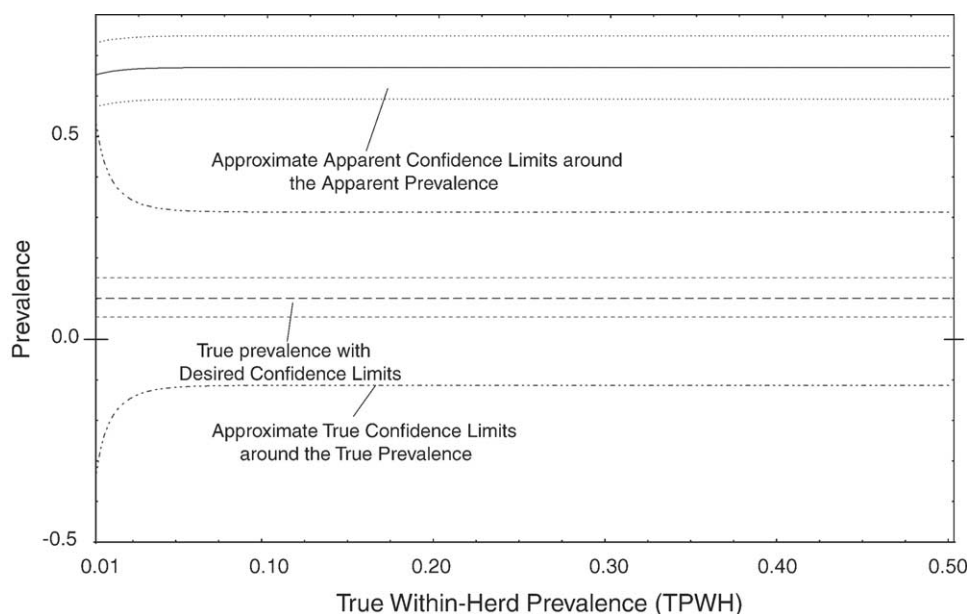


Fig. 3. Comparison of Methods 1 (naive) and 2 (corrected for test imperfection) regarding herd-level prevalences as the true within-herd prevalence changes. The test sensitivity is 50%, test specificity is 99%, the number of animals in the herd is 100, the true herd prevalence (assumed) is 10%, a 95% confidence range is used and the desired tolerance is 5%. One hundred thirty-nine herds are tested based on Method 1.

considerations. Table 4 uses a minimum within-herd prevalence of 3% and the same values for HSENS and HSPEC as in Table 3 (discussed above). Table 3 demonstrates that, the higher the selected HSENS and HSPEC, the fewer herds need to be screened. However, there are insufficient animals within the small herds to provide the number of tests required to achieve high HSENS and HSPEC. The result is that many of the cells of Table 4 remain unfilled. The best compromise we can find in Table 3 is for HSENS of 70% and HSPEC of 70%. By screening approximately 2200 herds, we retain relatively small tolerance with high confidence in our estimate (at the expense of not being able to include small herds). Alternatively, we could settle for 55% HSENS and HSPEC, and include small herds but have to test 38,000 herds (an impractical survey).

As the SENS improves, it becomes possible to attain higher HSENS and HSPEC or to include smaller herds in the survey (or a combination of all three). If it were believed that the SPEC is 97%, then, generally, we would have to test a higher proportion of each herd. In this case, we can achieve only a HSENS of 55% and HSPEC of 70%. If we believe that the true SENS is only 35% then we can achieve only a HSENS of 55% and HSPEC of 55%. The results of combining both these assumptions (i.e. use a test with SPEC of 97% and SENS of 35%) are also illustrated in Table 4. If scarce resources were focussed upon developing an improved test with a sensitivity of 65%, then it might be possible to design a survey where a HSPEC of 70% and HSENS of 70% can be attained and also include small herds. Equally, a survey with HSENS (=85%) and HSPEC (=55%) is possible (but would

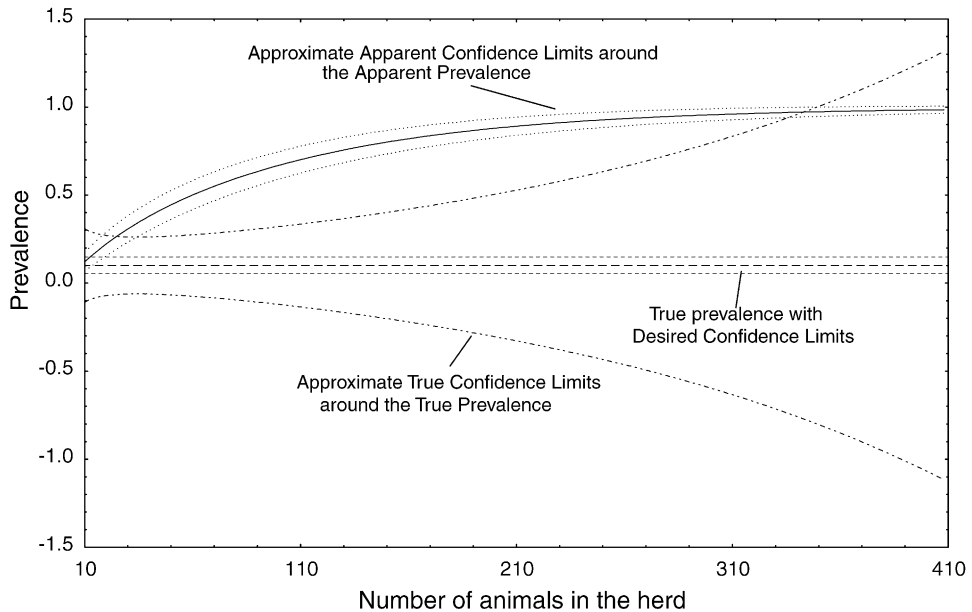


Fig. 4. Comparison of Methods 1 (naïve) and 2 (corrected for test imperfection) regarding herd-level prevalences as the herd size changes. The test sensitivity is 50%, test specificity is 99%, the true within-herd prevalence is 5%, the true herd prevalence (assumed) is 10%, a 95% confidence range is used and the desired tolerance is 5%. One hundred thirty-nine herds are tested based on Method 1.

require 2400 herds to be tested). Attaining higher HSENS and HSPEC reduces the risk of misclassification of herds which in turn reduces the number of herds to sample.

In these examples, it must be apparent that a sampling strategy including small herds and based upon a minimum within-herd prevalence estimate of just 3% will result in an unfeasibly large sampling programme. Our options are either to restrict the survey to large herds or to select one of the higher alternative estimates for minimum within-herd prevalence. In the case of Johne's, several authors suggest higher prevalences for both beef and dairy, and most authors from other countries agree that within-herd prevalence is higher in dairy herds than beef herds. If we decide to survey to detect herds where the prevalence is 5% (Collins et al., 1994; Meylan et al., 1995) or even 10% (Collins et al., 1994; Obasanjo et al., 1997; Ott et al., 1999; Wells and Wagner, 2000), then not only do sample sizes fall, but smaller herds can be tested with increased HSENS and HSPEC.

With the cut point held constant, the probability of misclassifying a negative herd as positive goes up as the sample size increases; conversely, the probability of misclassifying a positive herd as negative falls as the sample size increases. Achieving the balance between HSENS and HSPEC is the objective of the FreeCalc software (Cameron and Baldock, 1998). If the number of true positives is constant and the number tested (n) is constant then as herd size (N) increases HSENS falls and the HSPEC improves. To counteract this decrease in HSENS it becomes necessary to increase the sample size (n). At a certain stage, the risk of false-positive animals becomes so high that a cut point of one

positive animal might need to be introduced. Then, a herd containing zero or even one test-positive animal is still categorised as a negative herd. The program FreeCalc Version 2 calculates the minimum sample size in which a cut point can be found such that the desired HSENS and HSPEC are achieved. Table 6 illustrates the results of these calculations based on a minimum within-herd level of infection of five infected animals.

It is now possible to determine the sample frame. If the minimum within-herd prevalence is accepted to be 5%, the survey only includes herds with at least 30 cattle aged over 2 years of age.

As described above, we concluded that, to survey either beef cattle or all cattle together for Johnes, we should use a minimum TPWH estimate of 5% (SENS 50%; SPEC 99%). An alternative strategy for dairy herds alone would be with minimum TPWH of 10% (SENS 50%; SPEC 99%). In this latter case, it would be possible to achieve HSENS of 85% and HSPEC of 90% for herd sizes of 45 and above.

The comparison of the proposed method (Method 2) with the very basic method (Method 1) (Figs. 1–4) highlights that, if SENS and SPEC are ignored and all animals in the herd are tested (i.e. HSENS and HSPEC are assumed to be 100%), then there can be large biases in the HAP, AACR and ATCR. The bias in the HAP is particularly large for low SPEC and large herd sizes. This is because, for the basic method, the HSPEC is equal to the SPEC raised to the power of the number of animals and thus is extremely responsive to both the SPEC and N . The large difference between the desired tolerance, the AACR and the ATCR demonstrates that by ignoring animal-level and herd-level test accuracy, confidence ranges can be underestimated greatly.

Whilst this methodology was devised in response to paratuberculosis, it is applicable to all diseases or infections with imperfect tests. Indeed it is recommended that, wherever herd prevalence estimates are sought using an imperfect test, the consequences of test sensitivity and specificity be taken into account. Our methodology described here is a practical means by which this can be achieved.

Acknowledgements

We thank Graham Horgan and Claus Mayer of BioSS for statistical advice. We thank M. Thrusfield and K. Frankena for their thorough scrutiny of the manuscript. SAC received financial support for this project from the Scottish Executive Environment and Rural Affairs Department (SEERAD). This project was jointly funded by SEERAD and the UK Department of Environment, Fisheries and Rural Affairs (DEFRA).

Appendix A. Derivation of sample-size formula

The number of test positives (=HT) in a sample of HN:

$$HT \approx \text{Binom}(HN, PH_{\text{pos}})$$

where PH_{pos} is the probability of a single herd testing positive (= apparent herd prevalence

(HAP)). Thus,

$$\text{PHpos} = \text{HSENS}(\text{HTP}) + (1 - \text{HSPEC})(1 - \text{HTP})$$

where HTP is the true herd prevalence and HSENS is the herd-level sensitivity and HSPEC the herd-level specificity.

$$\text{HT} \approx \text{Binom}(\text{HN}, \text{PHpos}) \Rightarrow \text{var}(\text{HT}) = \text{HN}(\text{PHpos})(1 - \text{PHpos})$$

Now the estimator, e , for the true prevalence is given by the following formula (Rogan and Gladen, 1978):

$$e = \frac{(\text{HT}/\text{HN}) + \text{HSPEC} - 1}{\text{HSENS} + \text{HSPEC} - 1}$$

where HT/HN is the apparent prevalence (=HAP).

Therefore:

$$\begin{aligned} E(e) &= E\left\{\frac{(\text{HT}/\text{HN}) + \text{HSPEC} - 1}{\text{HSENS} + \text{HSPEC} - 1}\right\} = \frac{E(\text{HT})/\text{HN} + \text{HSPEC} - 1}{\text{HSENS} + \text{HSPEC} - 1} \\ &= \frac{(\text{HN}(\text{PHpos})/\text{HN}) + \text{HSPEC} - 1}{\text{HSENS} + \text{HSPEC} - 1} = \frac{\text{PHpos} + \text{HSPEC} - 1}{\text{HSENS} + \text{HSPEC} - 1} \\ &= \frac{\text{HSENS}(\text{HTP}) + (1 - \text{HSPEC})(1 - \text{HTP}) + E - 1}{\text{HSENS} + \text{HSPEC} - 1} \\ &= \frac{\text{HSENS}(\text{HTP}) + \text{HSPEC}(\text{HTP}) - \text{HTP} - \text{HSPEC} + 1 + \text{HSPEC} - 1}{\text{HSENS} + \text{HSPEC} - 1} \\ &= \frac{(\text{HSENS} + \text{HSPEC} - 1)\text{HTP}}{\text{HSENS} + \text{HSPEC} - 1} = \text{HTP} \end{aligned}$$

(unsurprisingly)

and,

$$\begin{aligned} \text{var}(e) &= \text{var}\left\{\frac{(\text{HT}/\text{HN}) + \text{HSPEC} - 1}{\text{HSENS} + \text{HSPEC} - 1}\right\} = \frac{\text{var}(\text{HT})}{(\text{HN}^2)(\text{HSENS} + \text{HSPEC} - 1)^2} \\ &= \frac{\text{HN} \times \text{PHpos}(1 - \text{PHpos})}{(\text{HN}^2)(\text{HSENS} + \text{HSPEC} - 1)^2} = \frac{\text{PHpos}(1 - \text{PHpos})}{\text{HN}(\text{HSENS} + \text{HSPEC} - 1)^2} \end{aligned}$$

(assuming, that HSENS and HSPEC are constants).

Since the number of herds testing positive can be considered as the sum of several bernoulli trials we can assume, as in for a classic binomial sampling, through the central limit theorem that, above a certain number of herds (we do not provide any rule of thumb for this number in this paper), e can reasonably be approximated to a normal curve of mean $E(e)$ and variance $\text{var}(e)$ then we want to choose n such that

$$\text{var}(e) = \left(\frac{L}{Z(C)}\right)^2,$$

where L is the tolerance, C the confidence and $Z(C)$ the inverse distribution function of a standard normal curve based on the confidence C .

Thus:

$$\begin{aligned}\text{var}(e) &= \frac{\text{PHpos}(1 - \text{PHpos})}{\text{HN}(\text{HSENS} + \text{HSPEC} - 1)^2} = \left(\frac{L}{Z(C)}\right)^2 \Leftrightarrow \text{HN} \\ &= \left(\frac{Z(C)}{L}\right)^2 \left(\frac{\text{PHpos}(1 - \text{PHpos})}{(\text{HSENS} + \text{HSPEC} - 1)^2}\right) \\ &= \left(\frac{Z(C)}{L}\right)^2 \left[\frac{[\text{HSENS}(\text{HTP}) + (1 - \text{HSPEC})(1 - \text{HTP})] \times [1 - \text{HSENS}(\text{HTP}) - (1 - \text{HSPEC})(1 - \text{HTP})]}{(\text{HSENS} + \text{HSPEC} - 1)^2}\right]\end{aligned}$$

This is the equation we used and formulated to estimate the number of herds that need to be tested given a set herd-level sensitivity and specificity.

Appendix B

Formulae used to calculate the number of herds sampled using the most basic method and then the apparent prevalence that would arise from such a sampling programme together with the apparent 95% confidence limits and the 95% confidence limits based on the apparent prevalence but adjusted (upwards) to take into account the herd level sensitivity and specificity. RoundUp() indicates a rounding up to the next whole number. An explanation of the symbols is provided in Table 1.

$$\text{HN} = \text{RoundUp} \left[\left(\frac{1.960}{L} \right)^2 \times \{ \text{HTP}(1 - \text{HTP}) \} \right] = 139$$

HN = 139 for the values of $L = 0.05$ and $\text{HTP} = 0.1$ as used in Figs. 1–4.

$$\text{HSENS} = 1 - [(1 - \text{SENS})^{\text{TPWH} \times N}] [\text{SPEC}^{(1 - \text{TPWH})N}] \text{HSPEC} = \text{SPEC}^N$$

$$\text{HAP} = \text{HSENS}(\text{HTP}) + [1 - \text{HSPEC}][1 - \text{HTP}]$$

$$\text{AACR} = 1.960 \left[\frac{\text{HAP}(1 - \text{HAP})}{\text{HN}} \right]^{0.5}$$

$$\text{ATCR} = 1.960 \left[\frac{\text{HAP}(1 - \text{HAP})}{\text{HN}(\text{HSENS} + \text{HSPEC} - 1)^2} \right]^{0.5}$$

References

- Cameron, A.R., 2001. FreeCalc software Version 2.
- Cameron, A.R., Baldock, F.C., 1998. A new probability formula for surveys to substantiate freedom from disease. *Prev. Vet. Med.* 34, 1–17.

- Cannon, R.M., 2001. Sense and sensitivity – designing surveys based on an imperfect test. *Prev. Vet. Med.* 49, 141–163.
- Cannon, R.M., Roe, R.T., 1982. *Livestock Disease Surveys – A Field Manual for Veterinarians*. Canberra.
- Collins, M.T., Morgan, I.R., 1991. Economic decision-analysis model of a paratuberculosis test and cull program. *J. Am. Vet. Med. Assoc.* 199, 1724–1729.
- Collins, M.T., Sockett, D.C., Goodger, W.J., Conrad, T.A., Thomas, C.B., Carr, D.J., 1994. Herd prevalence and geographic-distribution of, and risk- factors for, bovine paratuberculosis in Wisconsin. *J. Am. Vet. Med. Assoc.* 204, 636–641.
- Cox, J.C., Drane, D.P., Jones, Ridge, S., Milner, A.R., 1991. Development and evaluation of a rapid absorbed enzyme immunoassay test for the diagnosis of Johne's disease in cattle. *Aust. Vet. J.* 68, 157–160.
- Jordan, D., 1996. Aggregate testing for the evaluation of Johne's disease herd status. *Aust. Vet. J.* 73, 16–19.
- Jordan, D., Macewan, S.A., 1998. Herd-level test performance based on uncertain estimates of individual test performance, individual true prevalence and herd true prevalence. *Prev. Vet. Med.* 36, 187–209.
- Kennedy, D., 2000. Surveillance strategies in Australia. Assessment of surveillance and control of paratuberculosis in farm animals in GB.
- Meylan, M., Nicolet, J., Busato, A., Burnens, A., Martig, J., 1995. Paratuberculosis – a prevalence study in the Plateau-De-Diesse. *Schweizer Archiv Fur Tierheilkunde* 137, 22–25.
- Obasanjo, I.O., Grohn, Y.T., Mohammed, H.O., 1997. Farm factors associated with the presence of *Mycobacterium paratuberculosis* infection in dairy herds on the New York State Paratuberculosis Control Program. *Prev. Vet. Med.* 32, 243–251.
- Ott, S.L., Wells, S.J., Wagner, B.A., 1999. Herd-level economic losses associated with Johne's disease on US dairy operations. *Prev. Vet. Med.* 40, 179–192.
- Reichel, M.P., Kittelberger, R., Penrose, M.E., Meynell, R.M., Cousins, D., Ellis, T., Mutharia, L.M., Sugden, E.A., Johns, A.H., de Lisle, G.W., 1999. Comparison of serological tests and faecal culture for the detection of *Mycobacterium avium* subsp. *paratuberculosis* infection in cattle and analysis of the antigens involved. *Vet. Microbiol.* 66, 135–150.
- Rogan, W.J., Gladen, B., 1978. Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.* 107, 71–76.
- Snedecor, G.W., Cochran, W.G., 1989. *Statistical Methods*, 6th ed. Iowa State University Press, Ames, Iowa, p. 121.
- Sweeney, R.W., Whitlock, R.H., Buckley, C.L., Spencer, P.A., 1995. Evaluation of a commercial enzyme linked immunosorbent assay for the diagnosis of paratuberculosis in dairy cattle. *J. Vet. Diag. Investig.* 7, 488–493.
- Thrusfield, M., 1995. *Veterinary Epidemiology*, 2nd ed. Blackwell Science Ltd, Oxford, pp. 296–311.
- Turnquist, S.E., Snider, T.G., Kreeger, J.M., Miller, J.E., Hagstad, H.V., Olcott, B.M., 1991. Serologic evidence of paratuberculosis in Louisiana beef-cattle herds as detected by Elisa. *Prev. Vet. Med.* 11, 125–130.
- Wells, S.J., Wagner, B.A., 2000. Herd-level risk factors for infection with *Mycobacterium paratuberculosis* in US dairies and association between familiarity of the herd manager with the disease or prior diagnosis of the disease in that herd and use of preventive measures. *J. Am. Vet. Med. Assoc.* 216, 1450–1457.
- Whitlock, R.H., Wells, S.J., Sweeney, R.W., Van Tiem, J., 2000. ELISA and fecal culture for paratuberculosis (Johne's disease): sensitivity and specificity of each method. *Vet. Microbiol.* 77, 387–398.