FREQUENTIST AND BAYESIAN APPROACHES TO PREVALENCE ESTIMATION USING EXAMPLES FROM JOHNE'S DISEASE

Locksley L. McV. Messam[1], Adam J. Branscum[2], Michael T. Collins[3] and Ian A. Gardner[1]*

[1]Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, Davis, California, USA [2]Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, Kentucky, USA [3]Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin-Madison, Madison, Wisconsin USA

***Corresponding Author**: Ian A. Gardner, Department of Medicine and Epidemiology, School of Veterinary Medicine, University of California, One Shields Ave, Davis, CA 95616, USA. Tel: 1-530-752-6992, Fax: 1-530-752-0414, e-mail: iagardner@ucdavis.

Abstract

Although frequentist approaches to prevalence estimation are simple to apply, there are circumstances where it is difficult to satisfy assumptions of asymptotic normality and nonsensical point estimates (greater than 1 or less than 0) may result. This is particularly true when samples sizes are small, test prevalences are low and imperfect sensitivity (Se) and specificity (Sp) of diagnostic tests need to be incorporated into calculations of true prevalence. Bayesian approaches offer several advantages including intervals between 0 and 1, direct probabilistic interpretation, and the flexibility to model the probability of zero prevalence. In this review, we present formulae for individual- and herd-level true prevalence estimation using both frequentist and Bayesian methods. This is done for both individual-level and pooled sampling schemes. We provide statistical methods for detecting differences between population prevalence, and frequentist methods for sample size and power calculations. All examples are motivated using *Mycobacterium avium* subspecies *paratuberculosis* (MAP) infection and we provide WinBUGS code for all examples of Bayesian estimation.

Keywords: Prevalence estimation, Bayesian methods, frequentist methods, pooled testing, *Mycobacterium avium* subspecies *paratuberculosis*

1. Introduction

Prevalence is a measure of disease frequency that focuses on existing status rather than new events (Rothman and Greenland, 1998). Though incidence is of primary importance in etiological research, prevalence is often used in the study of chronic diseases of insidious onset (e.g. Johne's disease) where incident cases are difficult to define even if herds are monitored longitudinally.

Diagnostic tests are commonly used for prevalence studies and ideally, true prevalence should be estimated from apparent (test) prevalence by adjusting for test sensitivity (Se) and specificity (Sp). Lack of knowledge of, or disregard for, test errors (false positives and negatives) can lead to inaccurate sample size calculations for surveys, misclassification of diseased and non-diseased states, and biased estimates of measures of effect in risk factor studies. All of these negatively impact disease surveillance, control and eradication programs, and consequently animal trade.

Frequentist methods have traditionally been used for animal-health prevalence surveys. In recent years, however, applications of Bayesian methods for the statistical analysis of veterinary epidemiologic data have increased (Cowling et al., 1999, Staubach et al., 2002, Clough et al., 2003, van Schaik et al., 2003, Borel et al., 2004, Branscum et al., 2004, Dorny et al., 2004, Geurden et al., 2004, Ngowi et al., 2004, Carabin et al., 2005, Durr et al., 2005, Benito et al., 2006, Rapsch et al., 2006, Wang et al., 2006). Bayesian statistical analysis of prevalence data is appealing because it formally incorporates previously-collected prevalence data and expert-elicited information into current calculations (Gardner, 2002). In addition, when sample sizes for

prevalence surveys are small, prevalence is low and results are based on inaccurate diagnostic tests, it may be difficult to satisfy large-sample normal approximations when frequentist confidence intervals are constructed. In these circumstances, use of Bayesian methods provides a practical alternative for data analysis.

In this non-exhaustive review, we present to the non-epidemiologist/statistician Bayesian and frequentist approaches to prevalence estimation with examples involving Johne's disease. We highlight advantages and limitations of both approaches while presenting methods for the statistical analysis of individual-level, herd-level and pooled-sample prevalence data as well review calculations involving existing prevalence estimates. We conclude with a brief summary of frequentist methods for sample size and power calculations motivated by practical examples. Although we focus on *Mycobacterium avium* subspecies *paratuberculosis* (MAP) infection, the issues and methods discussed are broadly applicable to chronic infectious diseases. We note that prevalence also can be estimated from latent class models, e.g. two-test in two-population model, that are designed primarily for evaluation of test accuracy. For brevity, we do not describe this approach and we refer interested readers to other papers (Hui and Walter, 1980, Johnson et al., 2001, Branscum et al., 2005).

2. Background

2.1 Prevalence definitions

Prevalence is a dimensionless, unit-free value ranging from zero to unity (zero to 100 if expressed as a percentage). Depending on the context, an investigator might be interested in

prevalence of infection, infectious animals or disease. We restrict our presentation to infection prevalence.

Two types of prevalence are usually estimated in epidemiologic studies; point and period prevalence. Point prevalence is the proportion of infected individuals in a defined population (or more generally that have a condition of interest) at a given point in time, hence:

Point Prevalence = Number of infected individuals/Total number of individuals at risk for infection in the population

Period prevalence is the proportion of infected individuals in a defined population found over a specified time period (e.g. 2 years). For non-resolvable infections, e.g. MAP infection, point prevalence provides valid estimates of the frequency of individuals that have ever been infected and thus we do not further consider period prevalence in this article.

2.2 Sampling

To justify descriptive inferences and probabilistic interpretations of prevalence estimates based on sample data, random selection of units from the population to be tested is necessary (Greenland, 1990, Greiner and Gardner, 2000a). Depending on the goals of the survey, availability of resources, and logistics, units may be enrolled using simple random, stratified random, systematic random, single or two-stage cluster sampling schemes (Scheaffer et al., 1995), among other random sampling schemes.

2.3 Statistical inference

A frequentist approach, assumes the prevalence of the infection (P) in the target population to be a fixed, unknown quantity. Frequentist inferences for P are summarized using a point

estimate $\left(\hat{P}\right)$, along with a $100(1-\alpha)\%$ confidence interval (CI) that quantifies inferential precision, where typically $\alpha = 0.05$. The interpretation of a $100(1-\alpha)\%$ CI is as follows: If sampling were repeated an infinite number of times using the same sample size and sampling procedure, 100(1-α) % of the CIs so generated would be expected to contain P.

A Bayesian approach assumes a probability distribution for P which characterizes the researcher's uncertainty about the population prevalence independently of the current survey data. Subjective knowledge about P and about test accuracy parameters (i.e. Se and Sp) is elicited from experts familiar with the subject matter or derived from historical or current data obtained from similar populations and is formally incorporated into the analysis in the form of probability distributions called prior distributions. Inferences for P are made by using a likelihood function and the prior distribution, respectively, to combine information from the data with information independent of the data to form the posterior distribution of P. The relationship between the likelihood, prior, and posterior is given by:

$$\underbrace{\Pr(\theta \mid Data)}_{Posterior} = K \times \underbrace{\Pr(\theta)}_{Prior} \times \underbrace{\Pr(Data \mid \theta)}_{Likelihood} \quad (1)$$

where $\Pr(\cdot)$ denotes a probability density or mass function, θ denotes parameters in the statistical model and K is a normalizing constant. Summary measures of posterior distributions, such as means or medians, and outer percentiles, form the basis for inference about P. Typically, results are presented in the form of the median and a 100(1-α)% probability interval (PI) (usually $\alpha = 0.05$) which provides a range of values that contains P with probability = 1-α, given the data and model. Hence, unlike CIs, PIs have probabilistic interpretations dependent only on currently observed data and a given set of priors. This renders interpretation and

application of results more straightforward for both scientists and policy-makers. For further explanation of fundamental concepts of Bayesian inference, we refer the reader to other introductory works (Berry, 1996, Bland and Altman, 1998, Dunson, 2001). All Bayesian computations presented in this paper were performed using the freely-available WinBUGS software (downloadable from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml).

3. Individual-level prevalence estimation from individual samples

3.1 Apparent prevalence

3.11 Frequentist approach

Apparent prevalence (P_A) is the probability that a randomly-chosen unit of observation will test positive (Greiner and Gardner, 2000b) and therefore is dependent on the true population within-herd prevalence (P_T) and *Se* and *Sp* of the diagnostic test used. A point estimate of P_A based on test outcomes of *n* randomly sampled individuals from the source population of size *N* is given by

$$\hat{P}_A = x/n \qquad (2)$$

where x denotes the number of sampled units that test positive. An approximate 100(1-α) % CI for P_A is given by the Wald CI:

$$\hat{P}_A \pm z_{\alpha/2}\sqrt{\hat{P}_A(1-\hat{P}_A)/n} \qquad (3)$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ percentile of the standard normal distribution. Equation (3) is derived assuming (i) $x \sim bin(n,\ P_A)$ and (ii) that this binomial distribution is well approximated

by $x \sim N\left(nP_A, nP_A\left(1-P_A\right)\right)$. It is usually recommended that both nP_A and $n(1-P_A)$ be sufficiently large, e.g. both greater than 5, (Agresti, 1996, Leemis and Trivedi, 1996) for the approximation to be valid. Others have suggested that the normal approximation is appropriate for 95% and 99% CIs if $0 < \hat{P}_A \pm k\sqrt{\hat{P}_A(1-\hat{P}_A)/n} < 1$ for k = 3 and 5, respectively (Enoe et al., 2000). Key limitations of Wald CIs for P_A are that they can provide coverage much less than the nominal level (Agresti and Coull, 1998), may have lower and upper bounds <0 and >1, respectively, and are inappropriate when no animals test positive. Alternative methods for CI calculation exist and here we highlight 3 common approaches. First, exact CIs can be calculated directly from the binomial distribution (see Appendix A.1 for details). Second, Agresti and Coull (1998) have shown that the score CI

$$\left(\hat{P}_A + \left(z^2_{\alpha/2}/2n\right) \pm z_{\alpha/2}\sqrt{\left[\hat{P}_A(1-\hat{P}_A) + \left(z^2_{\alpha/2}/4n\right)\right]/n}\right)\Big/\left(1+\left(z^2_{\alpha/2}/n\right)\right) \quad (4)$$

provides coverage closer to the nominal level than both the Wald and exact CIs for small sample sizes (i.e., 5 to 100). Third, to circumvent Wald and exact CI coverage limitations, as well as the cumbersomeness of (4), Agresti and Coull (1998) suggest a modified 95% Wald interval (the "add 2 successes and 2 failures" adjusted Wald interval):

$$\tilde{P}_A \pm 1.96\sqrt{\tilde{P}_A\left(1-\tilde{P}_A\right)/\left(n+4\right)} \quad (5)$$

where $\tilde{P}_A = \left(x+2\right)/\left(n+4\right)$. This interval is slightly conservative relative to the score interval but does not suffer from low coverage near 0 and 1 and is still closer to 95% coverage than exact intervals (for sample sizes ranging from 0 - 100) (Agresti and Coull, 1998).

Example

Assume that from a very large herd of dairy cows (N = 2000), $n = 120$ animals are randomly sampled and $x = 20$ animals test positive for MAP infection, yielding an apparent prevalence estimate of $\hat{P}_A = 0.167$. Using (3), a 95% CI for P_A is 0.100 to 0.233. Exact calculations yield a 95% CI from 0.105 to 0.246, while the 95% score CI is from 0.111 to 0.243. The adjusted Wald 95% CI is from 0.110 to 0.246.

If a 100(1-α) % CI is desired when there are no positive test results, an exact CI can readily be calculated using $\left(0,\ 1-\alpha^{1/n}\right)$ (Louis, 1981). When $n \geq 20$, and $\alpha = 0.05$, a good approximation to this interval is "the rule of three" 95% CI = $\left(0,\ 3/n\right)$ (Jovanovic and Levy, 1997). If in our last example, 0 animals test positive, then using $\left(0,\ 1-\alpha^{1/n}\right)$, a 95% CI for P_A is 0 to 0.0247 and using the "rule of three" is 0 to 0.0250. Hence for this scenario (n = 120), both these methods give practically the same results.

3.12 Bayesian approach

Here, the data are supplemented with prior information. Priors for P, require distributions ranging from 0 to 1. The beta family provides a class of distributions supported on the unit interval (0 - 1) with a range of flexible shapes, including U, J, and L-shaped, unimodal symmetric, right and left skewed, and uniform distributions. With a $\text{binomial}\left(n, P\right)$ likelihood, a $\text{beta}\left(a, b\right)$ prior for P results in a $\text{beta}\left(x+a,\ n+b-x\right)$ posterior for P. Hence, we can summarize posterior inference for P using either the mean or the median and 95% PI of this beta distribution.

To specify values for the hyperparameters a and b for a $beta(a,b)$ prior, we generally use historical data and/or expert opinion. For instance, to incorporate expert opinion for apparent prevalence, we elicit what the expert believes to be the most likely value of P_A in the herd. This serves as a prior point estimate, namely the prior mode of P_A. Then, with a specified high probability (usually 0.95 or 0.99), the expert provides a value that P_A is believed to be greater than (less than) if the prior mode is greater than (less than) 0.5. The prior mode and outer percentile (95th or 99th) are input into software programs e.g. BetaBuster (which is downloadable from http://www.epi.ucdavis.edu/diagnostictests/betabuster.html) and the values a and b are produced as output. When little is known of the prior testing history of the herd an uninformative prior may be used which prescribes equal likelihood to every possible value of the prevalence. A $beta(1,1)$ (equivalent to a $Uniform(1, 1)$) distribution is an example of an uninformative prior.

Example

Assume that we have no prior information on disease prevalence in the herd sampled in 3.11. Since $sample\ size\ (n)/herd\ size\ (N) \le 0.1$, we assume a binomial distribution $(x \sim \text{bin}(n, P_A))$ for the number of animals testing positive (Berry and Lindgren, 1996). For simplicity, we also assume that $Se = Sp = 1$ (i.e. $P_A = P_T$). The apparent prevalence can easily be estimated in WinBUGS (see appendix B.1a). Using a beta(1, 1) prior, the posterior median and 95% PI for P_A were 0.170 and 0.111 to 0.243, respectively. Thus, we are 95% certain that the apparent prevalence is between 11.1 and 24.3 %. If prior information were available on the prevalence of infection of MAP in the herd, this could easily be incorporated into the Bayesian analysis. If

prior to data collection, the herd veterinarian indicates that based on a previous survey, he expects that 1 in 10 cows will be seropositive for MAP (mode = 0.1) and he is 95% sure that $P_A <$ 0.2. Inputting this information into BetaBuster, we get a beta(5.62, 42.57) prior. This beta prior has a median of 0.11 and 95% PI from 0.043 to 0.220. From Winbugs (see appendix B.1b) the posterior median and 95% PI for P_A are 0.151 and 0.102 to 0.210, respectively. Thus, we are 95% certain that the apparent prevalence is between 10.2 and 21.0 %. Note that the posterior point estimate (0.151) is intermediate between the point estimate obtained from the data (0.167) and the median of the prior distribution (0.11). In addition, we note that the posterior distribution (95% PI: 0.102 to 0.210) is narrower than the prior distribution (95% PI: 0.043 to 0.220).

3.2 Diagnostic tests

Diagnostic tests commonly used in prevalence studies of Johne's disease in live animals are based on either the identification of the etiologic agent (MAP), or detection of a serum or milk antibody response to its presence (Manning and Collins, 2001, Collins et al., 2006). Researchers typically use fecal culture as a reference MAP-detection test against which to evaluate the performance of serological tests such as the enzyme-linked immunosorbent assay (ELISA) (Manning and Collins, 2001). However, the intermittent fecal shedding of MAP particularly during sub-clinical infection, coupled with late-stage antibody production makes accurate and precise estimation of test characteristics challenging. Consequently, prevalence estimation methods that rely on these tests should incorporate uncertainty in values of test accuracy parameters like *Se* and *Sp*. Priors for diagnostic tests can be derived based on the results of recent validation studies involving animals with similar characteristics to the ones about to be

tested. If during a validation study for test Se, x out of n infected animals test positive for MAP then a beta$(x+1,\ n-x+1)$ is an appropriate choice of prior for Se (Vose, 2000). This beta distribution is obtained as the posterior of Se when the validation study data are modeled as $\text{bin}(n,\ Se)$ with a uniform$(\text{U}(0,\ 1))$ (the same as a beta (1, 1) prior on Se. For example, if in a recent validation study, 30 of 100 infected cows tested positive for MAP infection, then a $\text{beta}(31, 71)$ prior for Se can be used in analysis of the future prevalence survey data. The use of a uniform prior here is equivalent to incorporating the opinion of an expert who believes that all values of sensitivity are equally likely in the new study. To the extent that the new study's sample differs from that used for diagnostic test validation, the prior should be more informative. For our examples, we use 2 commonly-used diagnostic tests, serum (IDEXX) ELISA and fecal culture on Herrold's egg yolk (HEY) medium. Priors for these tests (see table 1) are based on both the results of validation tests (Collins et al., 2005) (Sockett et al., 1992) and informed expert opinion .

3.3 True prevalence

3.31 Frequentist approach

True prevalence (P_T) of infection is the probability that a randomly-chosen unit from the source population is infected. Apparent prevalence is related to P_T, Se, and Sp through the equation $P_A = P_T Se + (1-P_T)(1-Sp)$. Among the first estimators of P_T was the Rogan-Gladen-estimator (Rogan and Gladen, 1978), which is obtained by substitution of estimates of P_A, Se and Sp into this equation to yield

$$\hat{P}_T = \left(\hat{P}_A + Sp - 1\right) \Big/ \left(Se + Sp - 1\right). \qquad (7)$$

If Se and Sp are known with certainty, then the standard error of $\hat{P}_T$ is given by

$$s.e.\left(\hat{P}_T\right) = \sqrt{\hat{P}_A\left(1-\hat{P}_A\right)/nJ^2} \quad (8)$$

where $J = Se + Sp - 1$ denotes the Youden index (Youden, 1950), a combined measure of test accuracy. In practice, Se and Sp are not known with certainty (Greiner and Gardner, 2000a) in which case

$$s.e.\left(\hat{P}_T\right) = \sqrt{\left(A + B\hat{P}_T^{\,2} + C(1-\hat{P}_T)^2\right)/J^2} \quad (9)$$

provides a more valid quantification of uncertainty. Here,

$A = \hat{P}_A\left(1-\hat{P}_A\right)/n$, $B = Se(1-Se)/n_1$, and $C = Sp(1-Sp)/n_2$,

where n_1 and n_2 are the numbers of infected and non-infected animals, respectively, in the original test validation study.

Example

Assume that an IDEXX ELISA $(Se = 0.30, Sp = 0.96)$ was used to test the herd described in section 3.11. Then $\hat{P}_T$ = 0.346. If we assume that Se and Sp are known with certainty, then using (8), a 95% CI is 0.167 to 0.525. Using (9) with n_1 and n_2 = 415 and 359, respectively (Collins et al., 2005), a more realistic 95% CI that captures the uncertainty in Se and Sp is 0.160 to 0.533. As is evident from this example, the discrepancy between $\hat{P}_T$ (0.346) and $\hat{P}_A$ (0.13) can be substantial. In general, the more accurate the test, i.e. the closer the Youden index is to 1, the smaller the discrepancy. Though most often $P_T > P_A$ for a given herd and diagnostic test, the converse is also possible. Using a test with $Se = 0.30$ and $Sp = 0.96$, we note that $\hat{P}_T < \hat{P}_A$ when

$\hat{P}_T < 0.057$. Comparing the CIs obtained using formulae (5), (8) and (9), we also note that precision is greater when P_A is estimated or when P_T is estimated assuming that the diagnostic test's characteristics are known perfectly. In general, while differences between the precision of estimates obtained assuming known and unknown test characteristics decrease with size (n_1 and n_2) of the studies originally used to estimate Se and Sp, the difference in bias increases since the bias in $\hat{P}_T$ decreases but not in $\hat{P}_A$ (Rogan and Gladen, 1978).

The application of the Rogan-Gladen-estimator is sometimes problematic because $\hat{P}_T$ can be < 0 or > 1. In the context of MAP infection, only the former is of practical importance and can occur in a small sample of low infection prevalence when the proportion of false-positive results is greater than the apparent prevalence. The limitations of (7) are further exacerbated by the fact that when sample sizes are small and true prevalence low, the normal approximation used for calculation of CIs may not be justifiable. In particular, the condition $0 < \hat{P}_T \pm 3s.e.(\hat{P}_T) < 1$ becomes difficult to satisfy when (9) is used. For these reasons, we advise against using (7) when samples are small and P is believed to be close to 0.

3.32 Bayesian approach

A Bayesian analysis naturally accounts for uncertainty in the values of P_T, Se and Sp. Under binomial sampling, the distribution of the number of animals testing positive is given by $x \mid (P_T, Se, Sp) \sim \text{bin}(n, P_A)$, where $P_A = P_T Se + (1 - P_T)(1 - Sp)$. Implicit in the use of a beta prior for the prevalence is the certainty that $P_T \neq 0$. If we wish to allow for the possibility of zero infection prevalence, then we can use the mixture prior distribution:

$P_T \sim \text{beta}(a,\ b)$ with probability = λ and $P_T = 0$ with probability =1- λ. The parameter λ denotes the probability that the herd is infected, and in the single herd setting is typically set equal to an expert-specified value, but it can alternatively be modeled using a beta distribution (Branscum et al., 2004). Use of a mixture distribution in these circumstances is a more accurate way of incorporating our prior beliefs. Omission of the mixture distribution, is likely to result in point estimates biased towards unity and unrealistically narrow probability intervals. As previously mentioned, in addition to modeling P_T with a beta prior, beta distributions are commonly used as priors for *Se* and *Sp* (see table 1) (Hanson et al., 2003b, Branscum et al., 2004). For most of the Bayesian examples in this article, we use data collected from randomly-selected cows in 29 California dairy herds that were tested for Johne's disease using serum (IDEXX) ELISA and/or HEY fecal culture.

Example

Assume that $n = 60$ cows are randomly selected out of a herd of size 675 and suppose $x = 2$ cows test positive for MAP infection using a serum ELISA. Note that here the Rogan-Gladen-estimator yields a negative prevalence estimate ($\hat{P}_T = -0.142$). Since $60/675 = 0.089 \leq 0.10$, the assumption of binomial sampling is reasonable. The probability that the herd is infected was set at λ = 0.9 since 90% of herds in California were presumed infected with MAP at the time of sampling. Since no records were available from the herd regarding previous monitoring or diagnostic testing we assume that, conditional on the herd being Johne's positive, all prevalences are *a priori* equally likely. This translates to $P_T \sim \text{beta}(1, 1)$ prior distribution. Using beta priors for ELISA *Se* and *Sp* (see table 1), the model was fit in WinBUGS (code in appendix B.2). The

estimated (posterior median) P_T of MAP infection in the herd was 0.02 with 95% PI from 0 to 0.453. Thus, given that 2 animals out of 60 tested positive using the IDEXX ELISA, we are 97.5% certain that $P_T < 0.453$. The analysis also permits inferences regarding the posterior probability that the herd is infected. Given 2 of 60 animals tested positive, we are 57% sure that the herd is infected.

A Bayesian approach is readily implemented for all realizations of x, including $x = 0$ and $x = n$. For instance, if no animals had tested positive out of 60 the posterior median and 95% PI would be $P_T = 0$ (95% PI; 0 to 0.237), and the posterior probability of infection would decrease to 0.37.

When sample size (n)/herd size $(N) > 0.1$, the binomial sampling approximation is no longer appropriate and the hypergeometric distribution, which is based on finite population sampling, should be used for analysis, i.e. $x \mid (P_T, Se, Sp) \sim \text{hypergeometric}(N, n, P_A)$. The Bayesian Disease Freedom (BDFree) software, which can be downloaded from www.epi.ucdavis.edu/diagnostictests, contains a module for estimating P_T under finite population sampling. Assume that in the previous example, all 675 cows were tested, 23 tested positive, and that the herd was known to be infected. Using BDFree (see appendix B.4), $P_T =$ 0.025 (95% PI; 0.001 to 0.152). The posterior probability that the herd is infected given that 23 cows test positive is 0.96.

Computational methods of prevalence estimation are much more complex for finite population sampling than for binomial sampling and we refer readers elsewhere for details (Cameron and Baldock, 1998, Su et al., 2004).

3.4 True vs. apparent prevalence

In the context of prevalence estimation, inferences for P_T are preferred to inferences for P_A. Nevertheless, if test *Se* and *Sp* is assumed to be constant and we want to ascertain only if P_T is increasing or decreasing over time in a herd, we can make use of the linear relationship resulting from (7); $P_T = mP_A + C$ where $m = 1/J$ and $C = (Sp - 1)/J$ are both constants. Consequently, if P_A increases (decreases) we infer that P_T increases (decreases). For these conclusions to be valid, the same diagnostic test with constant Se and Sp over time at the same cut-off must be used on each occasion and the sampling scheme must be unchanged.

4. Individual-level prevalence estimation from pooled samples

4.1 Uses of pooled testing: Advantages vs. disadvantages

Though pooling has been used for detection of fish pathogens in aquaculture for at least 20 years (Worlund and Taylor, 1983), the use of artificially-created pools for diagnostic testing in livestock species is a recent occurrence (Nielsen et al., 2000, Borel et al., 2004, Tavornpanich et al., 2004, Letellier et al., 2005, Brinkhof et al., 2006). One pooled sample can be used to determine the pathogen status of a herd (infected or non-infected) while individual-level infection prevalence can be estimated if multiple pools are tested (Cowling et al., 1999,

Christensen and Gardner, 2000). In this section, we focus mainly on issues related to individual-level prevalence estimation using data from pooled samples.

When sample collection costs are low relative to test cost and $P < 0.1$(i.e. the condition is rare) (Sacks et al., 1989, Tu et al., 1994), pooled testing can be more cost effective than individual testing and can provide more precise prevalence estimates per number of tests used (Kline et al., 1989). These advantages are based on the assumption that there is no loss of test accuracy resulting from aggregation of samples and that bias is negligible (implies moderate to small pool sizes and large sample sizes). Disadvantages of pooled testing include the need to store and handle larger volumes and the probable loss of Se and possibly Sp compared with use of individual samples (Christensen and Gardner, 2000).

4.2 Sensitivity and specificity of pooled tests

The pooled sensitivity (PSe) is the probability that a pool will test positive given that it contains at least one sample from an infected animal, while pooled specificity (PSp) is the probability that a pool will test negative given that it does not contain a sample from any infected animal (Munoz-Zanzi et al., 2006). Although several authors mention the necessity for the "no loss of test accuracy assumption" $(PSe = Se$ and $PSp = Sp)$, especially as it applies to PSe (Kline et al., 1989, Tu et al., 1995, Cowling et al., 1999, Christensen and Gardner, 2000), there is a dearth of information in the animal health literature on this topic. It is assumed that PSe depends on the prevalence of the agent, its concentration in infected samples, the number of samples per pool, the mechanism by which the test identifies positivity, and the analytic and diagnostic sensitivity of the assay to be used (Munoz-Zanzi et al., 2006). For instance, an increase in pool size when

testing for bovine viral diarrhea using real-time PCR (Munoz-Zanzi et al., 2006) and when testing for MAP using fecal culture (Wells et al., 2002) has been found to decrease *PSe.* Others have found an increase in number of infected animals per pool when testing for MAP using fecal culture resulted in increased *PSe* (Tavornpanich et al., 2004). Pooled specificity (PSp) is a function of the degree of cross contamination of the pool by other related agents. Some authors have suggested that larger pool sizes may increase *PSp* because of dilution of false-positive-causing analytes (Christensen and Gardner, 2000). Others (Munoz-Zanzi et al., 2006) have found opposite results and have speculated that larger pool sizes increase the probability of more contaminants being included in the pool. The effect of different factors on *PSp* is likely to be disease and test dependent, and would not apply to bacterial or virus isolation which should be perfectly specific.

Culture of fecal pools has been the testing method of choice for pooled-sample estimation of MAP infection prevalence at both the individual (Kalis et al., 2000, Wells et al., 2002, Tavornpanich et al., 2004) and herd level (Wells et al., 2003, Kalis et al., 2004). Compared to individual animal testing for MAP infection, pools of feces from 5 to 10 animals lead to substantial economic savings with commensurately minor losses in test accuracy $(PSe \approx Se$ and $PSp \approx Sp)$. In particular, $PSp = Sp$ when testing for MAP using fecal culture.

4.3 Frequentist estimation

4.31 Fixed pool size and no available test information.

If test *Se* and *Sp* information is not available, a simple approach uses the following formula:

$\hat{P}_A = 1-\left(1-\left(x_p/n_p\right)\right)^{1/s_p}$ (Kline et al., 1989, Sacks et al., 1989) (10)

where s_p = pool size, x_p = number of pools testing positive and n_p = total number of pools. An approximate 100 $(1-\alpha)$ % CI for P_A is:

$$\hat{P}_A \pm z_{\alpha/2}\sqrt{p_p\ \left(1-p_p\right)^{(2/s_p)-1}\Big/ s_p^2 n_p} \quad (11)$$

where $p_p = x_p/n_p$ (prevalence of positive pools).

Example

Assume we randomly select 120 cows from a large herd. Fecal samples are obtained from all 120 and pooled into groups of 10. Culture is performed on all pools using fecal culture and 5 pools test positive for MAP. Using (10 and (11) yields $\hat{P}_A = 0.053(95\%\ \text{CI};\ 0.007 \text{ to } 0.098)$. Apart from the implicit assumption of perfect *Se* and *Sp*, one disadvantage of this method is that for low numbers of positive pools, negative lower confidence limits may result (Cowling et al., 1999). If in our example, x_p = 2, using (10) and (11), $\hat{P}_A = 0.018(95\%\ \text{CI};\ -0.007 \text{ to } 0.043)$.

When asymptotic methods fail, one can first calculate an approximate CI for p_p using

$$\left[p_p\right]_{Lower} = \left(\left(2n_p p_p + z_{\alpha/2}^2 - 1\right) - z_{\alpha/2}\sqrt{z_{\alpha/2}^2 - \left(2+\left(1/n_p\right)\right) + 4p_p\left(n_p q_p + 1\right)}\right)\Big/ 2\left(n_p + z_{\alpha/2}^2\right) \quad (12)$$

and

$$\left[p_p\right]_{Upper} = \left(\left(2n_p p_p + z_{\alpha/2}^2 + 1\right) + z_{\alpha/2}\sqrt{z_{\alpha/2}^2 - \left(2+\left(1/n_p\right)\right) + 4p_p\left(n_p q_p + 1\right)}\right)\Big/ 2\left(n_p + z_{\alpha/2}^2\right) \quad (13)$$

(where $q_p = 1-p_p$) for the lower and upper bounds, respectively (Tu et al., 1994). The lower and upper 100(1-α) % confidence limits for P_A are then obtained by substituting $\left[p_p\right]_{Lower}$ and

$\left[p_p\right]_{Upper}$ for $p_p\left(=x_p/n_p\right)$ in (10). Using (12) and (13) for $x_p = 2$, an approximate 95% CI would range from 0.003 to 0.05. Equations (12) and (13) make use of a continuity correction, bringing normally-distributed variables into closer agreement with the binomial distribution (Fleiss, 1981). Another alternative is to calculate exact confidence intervals (see Appendix A.1) using x_p as the binomial variate and then as above, substituting substituting $\left[p_p\right]_{Lower}$ and $\left[p_p\right]_{Upper}$ for p_p in (10).

4.32 Fixed pool size and imperfect test sensitivity and specificity

If information on Se and Sp is available, the conditions $1-Sp \le p_p \le Se$ and $Se+Sp-1>0$ (preferably both Se and $Sp>0.5$) must hold in order to estimate P_T (Tu et al., 1994). We assume $PSe = Se$ and $PSp = Sp$ and if Se and Sp are known, then the following formula can be used to estimate P_T:

$$\hat{P}_T = 1-\left[\left(PSe-p_p\right)/J_p\right]^{1/s_p} \qquad (14)$$

where, $J_p = PSe+PSp-1$. An approximate 100(1-α) % CI for P_T is given by:

$$\hat{P}_T \pm z_{\alpha/2}\sqrt{p_p\left(1-p_p\right)\left(1-\hat{P}_T\right)^{\left(2/s_p\right)-2}\Big/ s_p^2 n_p J_p{}^2} \text{ (Cowling et al., 1999).} \qquad (15)$$

If in our previous example, we assume that HEY fecal culture ($Se = 0.60$, Sp =0.999: Table 1) was used to detect MAP, $n_p = 12$, $s_p = 10$ and $x_p = 5$, then $\hat{P}_T = 0.112\left(95\%\text{ CI; }0.060\text{ to }0.163\right)$. When asymptotic assumptions fail, (12) and (13) may be used to obtain lower and upper confidence limits for p_p, which can then be substituted into (14) for p_p to obtain a CI for P_T (Tu et al., 1994, Cowling et al., 1999). Alternatively exact confidence limits may be calculated

for p_p and similarly substituted into (14) (Cowling et al., 1999). The Pooled Prevalence Calculator (PPC) (http://www.ausvet.com.au/pprev/) facilitates easy calculations of both exact and asymptotic CIs for pooled prevalence (see Appendix A.2).

If Se and Sp are not known with certainty, an approximate 100(1-α) % CI for P_T is given by:

$$\hat{P}_T \pm z_{\alpha/2}\sqrt{\left(A_p^{(2/s_p)-2} \Big/ s_p^2 J_P{}^2\right)\left[\left(p_p\left(1-p_p\right)/n_p\right)+B_p\left(1-A_p\right)^2+C_p A_p^2\right]} \qquad (16)$$

where $A_p=\left(PSe-p_p\right)/J_p=\left(1-\hat{P}_T\right)^{S_p}$, $B_p=PSe\left(1-PSe\right)/n_1$, $C_p=PSp\left(1-PSp\right)/n_2$ and n_1 and n_2 (as before) refer to the number of infected and non-infected animals, respectively, used in the original test validation study. Using n_1 =182 and n_2 =100 (Sockett et al., 1992), $\hat{P}_T=0.112\left(95\%\text{ CI; -0.026 to }0.249\right)$. The (unrealistic) negative lower confidence limit reflects the increased difficulty in satisfying normality assumptions when test characteristic uncertainty is considered. Currently, no frequentist exact methods exist that incorporate uncertainty in Se and Sp. Thus for reasons previously mentioned (see section 3.31), we advise against using (16) when prevalence is low.

4.4 Bayesian estimation

As pooled testing is most cost effective when prevalence is low, approaches which both incorporate test Se and Sp and provide realistic results (i.e. estimates and limits within the range 0 to 1) are particularly desirable for those situations. A Bayesian approach ensures intervals are bounded below by 0. We assume a binomial distribution (Branscum et al., 2004) for the number of pools that test positive and thus $x_p \mid P_T, PSe, PSp \sim bin(n, p_p)$, where

$p_p = PSe - (PSe + PSp - 1)(1 - P_T)^{1/s_p}$. As in the case of prevalence estimation based on individual samples, beta distributions are chosen to model uncertainty in P_T, PSe and PSp, and mixture models can be used to allow for zero prevalence (analogous to the development in section 3.32).

Example

Fecal samples from $n = 60$ cows of a herd of $N = 2223$ are obtained, randomly aggregated into 6 pools of 10 each and each pool is tested for MAP using fecal culture $(Se = 0.60$ and $Sp = 0.999)$ (see table 1). Two pools test positive. In a previous testing within other herds in the region, the most frequently occurring prevalence was 13% and the herd veterinarian is of the opinion that if MAP is present in this herd, no more than 1 in 4 animals would be infected. The herd is located in a region which has a herd prevalence of Johne's disease of 95%. Using BetaBuster, we obtain a $P_T \sim \text{beta}(5.87,\ 33.61)$ given that the herd is infected. The prior probability that the herd was infected was set equal to 0.95. The results of the Bayesian analysis (see appendix B.5 for code) were $\hat{P}_T = 0.18(95\%\ \text{PI};\ 0.08\ \text{to}\ 0.32)$. Given that 2 of 6 pools test positive, the posterior probability that the herd is infected is 0.997.

5. Herd prevalence estimation

5.1 General concepts

5.11 Definition of a herd

A herd is defined as any cluster or grouping of animals (Christensen and Gardner, 2000), often defined by location (Boelaert et al., 2000, Tavornpanich et al., 2004) but may also be defined by size (Baggesen et al., 1996), intended purpose (Christensen et al., 2002), clinical signs (Wells et al., 2003), age (Boelaert et al., 2000), or combinations of the above. The herd (or herd-level) prevalence (HP) of infection in a population of herds is defined as the proportion of infected herds in the population:

HP = number of infected herds/total number of herds in the population (of herds).

In contrast to individual-level prevalence surveys, the unit of analysis in herd-level surveys is the herd or whole group.

5.12 Herd testing

Herd testing is typically undertaken to classify the status of a herd (or herds) with respect to disease or infection. If the goal is to make statistical inferences regarding HP, then herds to be tested should be chosen randomly in order to ensure disperse coverage of the targeted region. Typically, a two-stage sampling design is used where M herds are randomly sampled in the first stage, and then n_i animals are sampled from each herd i, $i = 1,\ldots,M$. Selection of animals within herds may be done randomly or not depending on the survey goals (e.g. estimation of within-herd prevalence versus classification of herds by infection status). In the veterinary literature, herd tests and herd-prevalence estimation have been conducted for health certification (Cameron and Baldock, 1998, Kalis et al., 2004) and diagnostic test evaluation (Sergeant et al., 2002, Wells et al., 2003, Tavornpanich et al., 2004), to identify herd-level risk factors for individual and herd-level infection (Delafosse et al., 2006), and for assessments of exposure to etiologic agents (Turnquist et al., 1991).

Unlike individual-level testing, herd-level testing is a function of two cut-off points; the positivity threshold of the diagnostic test(s) used to classify individual animals as infected or not, and the herd cut-off value for classifying herds as infected or not (Christensen and Gardner, 2000). The herd cut-off value may be an absolute number (e.g. ≥ 2 infected animals classifies the herd as infected), or a percent (e.g. $\geq 1\%$ of animals infected classifies the herd as infected).

5.12 Herd sensitivity and specificity

Herd-level sensitivity (HSe) is the proportion of infected herds in which the number of animals with positive test results equals or exceeds the herd-level cut-off (Martin et al., 1992). With the decision rule that uses a cutoff of ≥ 1 positive test result to classify the herd as positive, $HSe = 1-(1-P_A)^n$ (Christensen and Gardner, 2000), where n denotes the number of animals tested from the herd. Herd-level sensitivity increases with increased n, P_A (and hence P_T), and Se, and decreases with increased within-herd cutoff for a positive test result (Martin et al., 1992, Donald et al., 1994).

Herd-level specificity (HSp) is the proportion of non infected herds that test negative in the population of herds, i.e. the proportion of non infected herds in which the number of animals with positive test results is less than the within-herd cutoff (Martin et al., 1992). With a cutoff of 1 for herd infection, $HSp = Sp^n$ (Christensen and Gardner, 2000). Herd-level specificity increases with increased cutoff and Sp, and decreases as n increases (Martin et al., 1992, Donald et al., 1994).

Herd-level test characteristics are also affected by within-herd correlation in disease or infection states (Donald et al., 1994). For a detailed account of the effect of correlation on herd-level test characteristics, the reader is referred to Donald et al., (1994).

5.2 Frequentist approach

Analogous to individual-level prevalence, true herd-level prevalence is estimated by:

$$HP_T = \left(HP_A + HSp - 1 \right) \Big/ J_H \quad \text{(Wagner and Salman, 2004)} \tag{17}$$

where HP_A is the estimated apparent herd prevalence

$\left(\text{i.e number of test-positive herds} \Big/ \text{total number of herd tested}\right)$ and $J_H = HSe + HSp - 1$. An approximate $100(1-\alpha)\%$ CI for HP_T is given by

$$HP_T \pm z_{\alpha/2} \sqrt{HP_A(1-HP_A) \Big/ n_H J_H^2} \tag{18}$$

where n_H is the number of herds tested. The application of (17) and (18) to herd prevalence estimation is problematic as it is necessary to assume common values for HSe and HSp. This also implies a common value for P_A which would unrealistically imply constant disease prevalence from herd to herd. As a result, we discourage routine use of these formulae in favour of the Bayesian approach to herd prevalence estimation presented in the next section.

5.3 Bayesian approach

In the Bayesian setting with binomial sampling, we model within-herd prevalences as independent and identically distributed according to a common prevalence distribution, and we

assume that Se and Sp are invariant from herd to herd. The number of animals that test positive in each herd is modeled as $x_i \sim bin(n_i, P_{Ti}Se + (1-P_{Ti})(1-Sp))$ with independent beta priors for Se and Sp. As in the single herd setting, within-herd prevalences are modeled according to a mixture distribution that models zero infection prevalence with probability $1-HP_T$ where HP_T denotes the herd-level prevalence in the region. For infected herds, the within-herd prevalence distribution is given by $P_{Ti} \sim beta(\mu\psi, \psi(1-\mu))$ where μ is the mean true prevalence in the population and ψ is a parameter related to the variance of the prevalence distribution that flexibly allows for disperse (e.g. uniformly distributed) or very similar infection prevalences among herds (Branscum et al., 2004). Large values of ψ indicate less variability so that within-herd prevalences concentrate around μ, while small values of ψ correspond to more heterogeneity of prevalence values across herds (Donald et al., 1994, Branscum et al., 2004).

The mean prevalence (μ) and ψ are typically modeled with beta and gamma distributions, respectively (Hanson et al., 2003a, Branscum et al., 2004), but other distributions can be used. A Bayesian approach allows not only for herd-level prevalence estimation but also provides a straightforward means to make predictive inferences about herds not tested during the survey.

Example

Using a two-stage sampling scheme, 29 dairy herds were randomly selected from a dairy practices in central California. Random samples of 60 cows were selected from each herd and

tested for Johne's disease using serum ELISA and if 1 or more cows had a positive test result the herd was considered infected. Data from the 29 herds are summarized in table 2 and we assume a binomial distribution for the numbers of animals testing positive. Based on expert opinion, 60% of herds in this region of California were expected to be infected at the time of sampling with a one-sided 95% limit of 30% (i.e. the expert was 95% sure that the proportion of MAP infected herds was greater than 0.30). BetaBuster yields a $\text{beta}(4.8, 3.6)$ prior for HP_T. The ELISA test Se and Sp are modeled as in previous sections (see Table 1). The mean prevalence (μ) and ψ were modeled using beta(3.283, 17.744) and gamma(4.524, 0.387) priors, respectively (see appendix B and Hanson et al. 2003a for rationale and method of derivation). Finally, a mixture model was used because some herds in the region might be free of MAP infected animals. The model was fit using WinBUGS (see appendix B.6 for code).

The estimated posterior median of HP_T was 0.51 with 95% PI from 0.27 to 0.80, while the mean prevalence among infected herds was $\mu = 0.18(95\%\ \text{PI};\ 0.08 \text{ to } 0.33)$. The estimated posterior prevalence distribution characterizes the within-herd prevalences in the region. At the time of sampling, 49% percent of the herds in the region were estimated to be free of MAP infection, 59% of the herds in the region had infection prevalence < 0.05 and 82% had infection prevalence < 0.2.

6. Comparison of prevalence

6.1 Sequential testing in herds

Investigators often sequentially test herds as a means of evaluating the effectiveness of management practices or interventions aimed at controlling or eradicating MAP infection, or to longitudinally monitor changes in infection prevalence. On other occasions, prevalence might be compared between 2 herds or 2 groups of animals within herds to determine whether or not they differ. Depending on the goal of the survey, the samples may or may not be regarded as statistically independent.

6.2 Frequentist comparison of population prevalence

6.21 Independent samples

To determine if 2 apparent prevalences $\left(P_{A_1} \text{and } P_{A_2}\right)$ are statistically different based on test results obtained from random samples of animals from distinct herds, a chi-square $\left(\chi^2\right)$ test of homogeneity of proportions may be performed;

$$\chi^2 = \sum\left(O_{ij} - \hat{E}_{ij}\right)^2 \Big/ \hat{E}_{ij} \text{ (Agresti, 1996)} \qquad (19)$$

where both i and j = 1, 2, O_{ij} and $\hat{E}_{ij}$ are the observed and (estimated) expected values, respectively, in the ith row and jth column of a 2×2 contingency table (see table 3) and $\hat{P}_{A_1} = n_{11}/n_{+1}$ and $\hat{P}_{A_2} = n_{12}/n_{+2}$. The estimated expected value, $\hat{E}_{ij} = r_i c_j / n$, where r_i and c_j are the ith row and jth column totals, respectively. The statistical hypothesis to be tested is: $H_0 : P_{A_1} = P_{A_2}$ *vs.* $H_1 : P_{A_1} \neq P_{A_2}$ with degrees of freedom $\left(df\right) = \left(R-1\right)\left(C-1\right) = 1$ where R and C are the number of rows and columns, respectively $\left(R = C = 2 \text{ for } 2\times2 \text{ tables}\right)$. The null hypothesis $\left(H_0\right)$ is rejected if $\chi^2 > \chi^2_{df=1,1-\alpha}$ where $\chi^2_{df=1,1-\alpha}$ is the value corresponding to the $1-\alpha$ percentile of the χ^2_1 distribution and $\alpha =$ type I error probability.

Example

Assume that one year after an initial testing, the herd described in section 3.11 is again tested for MAP, and 30/200 cows have positive test results, yielding an apparent prevalence $\left(\hat{P}_{A_2}\right) = 0.15$ (table 3). To test whether P_{A_2} and P_{A_1} are statistically different, we use (19) and $\alpha = 0.05$. This yields $\chi^2 = 0.33$ (p = 0.56) $< \chi^2_{1,\,0.95} = 3.84$ hence we do not reject H_0 (i.e., there is no statistical evidence suggesting a change in apparent prevalence). A $100(1-\alpha)\%$ CI for $P_{A_2} - P_{A_1}$ is

$$(\hat{P}_{A_2} - \hat{P}_{A_1}) \pm z_{\alpha/2}\sqrt{\left(\hat{P}_{A_2}\left(1-\hat{P}_{A_2}\right)/n_{+2}\right)+\left(\hat{P}_{A_1}\left(1-\hat{P}_{A_1}\right)/n_{+1}\right)}. \qquad (20)$$

Hence, $\hat{P}_{A_2} - \hat{P}_{A_1} = 0.02$ with 95% CI -0.05 to 0.09. For statistically independent samples, if any $\hat{E}_{ij} < 5$, the Fisher exact test is recommended to compare proportions (Rosner, 2000).

6.22 Dependent samples

In the previous example, if on both occasions the same cows were sampled (matched pair or repeated measures design), the data may be presented as in table 4 and a McNemar's chi-square test of homogeneity of proportions that accounts for correlated data should be performed. Here,

$$\chi^2 = \left(n_{12} - n_{21}\right)^2 / \left(n_{12} + n_{21}\right) \text{ (Agresti, 1996)} \qquad (21)$$

where n_{12} and n_{21} are the numbers of cows with discordant test results $\left(T+T- \text{ and } T\text{-}T+, \text{ respectively}\right)$ at the initial and subsequent sampling. If $n_{12} \neq n_{21}$ then $\hat{P}_{A_1} \neq \hat{P}_{A_2}$ and (21) can be used to test $H_0 : P_{A_1} = P_{A_2}$ *vs.* $H_1 : P_{A_1} \neq P_{A_2}$. Again H_0 is rejected if $\chi^2 > \chi^2_{df=1,1-\alpha}$. Using $\alpha = 0.05$, $\chi^2 = 0.333$ (p-value = 0.56) < 3.84 hence we do not reject H_0.

For paired binary data, a 100(1-α) % CI for $P_{A_2} - P_{A_1}$ is

$$(\hat{P}_{A_2} - \hat{P}_{A_1}) \pm z_{\alpha/2} \left(\sqrt{(n_{11} + n_{22})(n_{12} + n_{21}) + 4n_{12}n_{21}} \right) \Big/ n_p \sqrt{n_p} \quad \text{(Fleiss, 1981)} \qquad (22)$$

where n_{11} and n_{22} are the number of pairs with both test results positive and negative respectively. Hence, if the same cows were tested on both occasions, the estimated difference between the apparent prevalences is 0.02, with 95% CI of -0.05 to 0.09. Both this and the CI calculated in the previous section indicate that the data are consistent with P_{A_2} being as large as 9% more, and as small as 5% less than P_{A_1}. If $n_{12} + n_{21} < 20$, an exact version of McNemar's test is recommended (Rosner, 2000) (see Appendix A.3).

With data subject to misclassification (i.e. *Se* and *Sp* both <1), a primary limitation of these frequentist methods is that differences between true prevalences cannot be formally tested. An informal test of the difference between 2 true prevalences may be performed by checking for overlap of their confidence intervals. However, caution should be exercised in using this procedure as, when applied to Wald CIs, this method has less power than standard hypothesis testing (Schenker and Gentleman, 2001).

6.3 Bayesian comparison of population prevalence

Using a Bayesian approach, inferences can be made for differences in true prevalences between 2 (or more) population prevalences. Assuming independent binomial sampling and that both herds (section 6.21) were tested for MAP infection using the same serum ELISA (table 1), a Bayesian model for the data is $x_j \mid (P_j, Se, Sp) \sim Bin(n_j, P_{A_j})$ where $P_{A_j} = P_j Se + (1 - P_j)(1 - Sp)$

for $j = 1, 2$ and beta priors are used for P_j, Se and Sp. We require informative priors for Se and Sp to mitigate lack of identifiability (i.e. ensure that there are enough degrees of freedom for parameter estimation). Here, based only on the data, x_j / n_j estimates P_{A_j} and we require additional input about test accuracy to estimate true prevalence. With the data and expert-elicited information about Se and Sp, inferences for $P_{T_2} - P_{T_1}$ can be made.

Example

Assume that based on the presence of culture-confirmed clinical cases of Johne's disease, both herds are known to be infected with MAP. Suppose, however, that no records of previous test results are available and hence, the magnitude of within-herd prevalence of MAP infection is unknown. We thus use uniform priors for P_{T_2} and P_{T_1} $\left(P_{T_2}, P_{T_1} \sim \text{beta}(1,1)\right)$. The ELISA Se and Sp are modeled as in previous sections (table 1), and the model is fit in WinBUGS (see appendix B.7 for code). The posterior median of $P_{T_2} - P_{T_1}$ is 0.08 with 95% PI from -0.24 to 0.44. On the basis of the posterior probability, $\Pr\left(P_{T_2} > P_{T_1} \mid \text{data}\right) = 0.71$, we conclude that we are 71% certain that herd 2 has a higher true prevalence than herd 1.

A Bayesian approach also facilitates comparisons of prevalence when different diagnostic tests are used in the 2 herds. For example, if herd 2 was tested with fecal culture and Se and Sp are modeled as in previous sections (see appendix B.8 for code), the posterior median of $P_{T_2} - P_{T_1}$ is -0.09 with 95% PI from -0.62 to 0.29. Given that the posterior probability

$\Pr\left(P_{T_2} > P_{T_1} \mid \text{data}\right) = 0.31$, we are 31% certain that herd 2 has a higher true prevalence than herd 1 ($\hat{P}_{T_1} = 0.38\left(95\%\text{ PI; }0.14\text{ to }0.90\right)$ and $\hat{P}_{T_2} = 0.28\left(95\%\text{ PI; }0.16\text{ to }0.60\right)$).

6.4 Reporting of hypothesis tests' results

For presentation of results based on frequentist or Bayesian analysis of prevalence data, we encourage the use of interval estimates (e.g. 95% CIs or PIs) rather than, or in conjunction with, p-values or posterior probabilities. Confidence intervals and probability intervals help to reveal whether different statistically significant (or insignificant) results have different implications for policy or intervention decisions. The use of p-values, which reduce results into a significant-non-significant dichotomy, obscures quantitative differences between results.

7. Sample size and power

7.1 Relevance of sample size estimation

A critical aspect of the planning phase of a prevalence survey is an estimation of the number of animals needed. The sample size determines the precision of our final estimates and ultimately influences whether the survey (or surveys) will be able to address the research questions of interest. Limited time and resources dictate that prior to performing a study, an investigator should be reasonably assured that enough individuals are available so that estimates can be generated with the requisite precision. Consequently, sample size calculations are standard requirements for research grant proposals. Because the components of these calculations must be hypothesized, only estimates can be obtained for any proposed study. This uncertainty becomes more pronounced in prevalence surveys when accuracy of case identification and precision of

estimates is also dependent on diagnostic test characteristics. Though Bayesian approaches provide a wide variety of methods for dealing with uncertainty in prior information they are complex and have not yet been regularly employed in WinBUGs. For this reason we present only methods for frequentist sample size calculations in this section and point the interested reader to the following bibliographic sources (Adcock, 1997, Rahme et al., 2000).

7.2 Sample size required to estimate true and apparent prevalence with a given precision

Precision of an estimate may be defined as the width of its $(100-\alpha)$% CI. The sample size necessary for estimating P_A and P_T, respectively, within a given error margin (Δ) is given by:

$n = \dot{P}_A\left(1-\dot{P}_A\right)z^2_{\alpha/2}/\Delta^2$ (Greiner and Gardner, 2000b) (23)

and

$n = \dot{P}_A\left(1-\dot{P}_A\right)z^2_{\alpha/2}/J^2\Delta^2$ (Greiner and Gardner, 2000a) (24)

respectively, where $\dot{P}_A$ is an *a priori* estimate of P_A (for example, obtained via pilot surveys or expert opinion), $J = Se + Sp - 1$ and Δ = half the CI width which may be either an absolute number or a percentage of $\dot{P}_A$ (or $\dot{P}_T$). Table 5 provides estimates of the number of cows necessary for testing to estimate P_T for MAP infection when Δ ranges from 0.10 to 0.20, given *a priori* estimates of P_A for testing with serum ELISA $(Se = 0.30,\ Sp = 0.96)$ and fecal culture $(Se = 0.60,\ Sp = 0.999)$ assuming that the test characteristics are known.. For fixed $\dot{P}_T$ $(\dot{P}_A)$, as Δ increases (decreases), the required sample size decreases (increases), while for fixed Δ, as

Se and/or Sp increase (decrease), the required sample size decreases (increases). Thus, the sample size required to estimate $P_T\left(P_A\right)$ will be inversely proportional to the desired precision. If Se and Sp are not assumed to be known, then (24) will underestimate the sample size as it does not incorporate the uncertainty in Se and Sp . A more accurate formula is

$$n = \dot{P}_A\left(1-\dot{P}_A\right)z^2_{\alpha/2}\Big/\left(\Delta^2 J^2 - z^2_{\alpha/2}\left(B\dot{P}_T^2 - C(1-\dot{P}_T)^2\right)\right) \text{ (Greiner and Gardner, 2000a)} \qquad (25)$$

where $B = Se\left(1-Se\right)/n_1$, $C = Sp\left(1-Sp\right)/n_2$, and n_1 and n_2 are as previously defined for (9).

7.3 Sample size to detect a difference in prevalence

Sample size calculations to detect prevalence differences have 4 fundamental components; the significance level or the probability of falsely concluding that there is a difference when none exists $\left(\alpha = \text{probability of type I error}\right)$, the probability of falsely concluding that there is no difference when there is one $\left(\beta = \text{probability of type II error}\right)$, the variance of the parameter to be estimated and most critically, the size of the parameter itself i.e., the difference to be detected (Wittes, 2002).

7.31 Independent random samples

To calculate the sample size $\left(n_S = n_1 + n_2\right)$ necessary to detect a difference $\dot{P}_{A_1} - \dot{P}_{A_2}$ from 2 independent random samples we use:

$$n_1 = \left[\sqrt{\bar{\dot{P}}_A\bar{\dot{Q}}_A\left(1+\left(1/k\right)\right)}z_{\alpha/2} + \sqrt{\dot{P}_{A_1}\dot{Q}_{A_1} + \left(\dot{P}_{A_2}\dot{Q}_{A_2}/k\right)}z_\beta\right]^2 \Big/ \left(\dot{P}_{A_1} - \dot{P}_{A_2}\right)^2 \text{ (Rosner, 2000)} \qquad (26)$$

and n_2 is computed based on $k = n_2/n_1$ where k is usually determined by the investigator. Here, $\dot{Q}_{A_j} = 1 - \dot{P}_{A_j}$, $\bar{\dot{P}}_A = \left(\dot{P}_{A_1} + k\dot{P}_{A_2}\right)\Big/(1+k)$, $\bar{\dot{Q}}_A = 1 - \bar{\dot{P}}_A$. A two-tailed hypothesis test is implied by (26) i.e., $H_0 : P_{A_1} = P_{A_2}$ *vs.* $H_1 : P_{A_1} \neq P_{A_2}$. Common choices of α and 1-β are 0.05 and 0.8, respectively, and total sample size (n_S) is minimized for $k = 1$ $(n_1 = n_2)$. To estimate sample sizes based on one-tailed hypotheses (to identify either an increase or decrease in P_A), $z_{\alpha/2}$ is substituted with z_α in (26). However, if the inherently two-tailed chi-square (χ^2) statistic is to be used to test differences in prevalences, then the sample size estimation must be two-tailed to avoid sample size underestimation. One-tailed sample size calculations should be performed when a one-tailed test is planned to be used in data analysis to determine a difference in prevalence (Lachin, 1981).

Example

Assume that 2 large herds have prevalences of MAP infection of $\dot{P}_{A_1} = 0.10$ and $\dot{P}_{A_2} = 0.15$. We wish to determine if $\left|P_{A_2} - P_{A_1}\right| \geq 0.05$. We therefore calculate the required number of animals to sample in order to have a probability of 0.8 of detecting $\left|P_{A_2} - P_{A_1}\right| = 0.05$ if it exists. Using $\alpha = 0.05$ and $1\text{-}\beta = 0.8$ and assuming that equal numbers of animals $(k = 1,$ i.e. $n_2 = n_1)$ are available to randomly sample, (26) yields $n_1 = 685$ and $n_s = 1370$.

Sometimes sample sizes need to be calculated to detect a change in test prevalence of Johne's disease between 2 time points in a single herd. Sampling at the first time point may already have

occurred when the need for the sample size calculation arises and economic considerations may influence whether the same number of cows can be tested at the second sampling.

Example

Assume that we have a very large herd. In an earlier random sample of 600 cows, 60 tested positive i.e., $\hat{P}_{A_1} = 0.10$. New management practices were introduced to reduce the prevalence of MAP infection and one year later, we wanted to evaluate if the apparent prevalence has decreased by 5% $\left(P_{A_2} = 0.05\right)$. Costs of testing have increased and total costs would be reduced if testing less than 600 animals were possible. First, we need to calculate the ratio of numbers of animals necessary to sample in order to have a probability of 0.8 of detecting this difference (0.05) in prevalence if it exists. Using $\alpha = 0.05$ and 1-$\beta = 0.8$ and assuming that equal numbers of animals are available to sample (randomly), the implied hypotheses are $H_0 : P_{A_1} = P_{A_2}$ *vs.* $H_1 : P_{A_1} > P_{A_2}$. We assume for simplicity that the data are biologically independent because of the time transpired between the 2 sampling dates.

We first calculate the total sample size as if equal sized samples $\left(\text{i.e. } n_2 = n_1\right)$ were to be used, denoted $n_{s(k=1)}$, using $z_{\alpha=0.05}$ (one-sided hypothesis) in (26). This yields $n_{s(k=1)} = 684$. We then make use of the following relationships (Altman, 1991):

$$n_s = n_1 + n_2 = n_{s(k=1)}\left(1+k\right)^2 \big/ 4k \tag{27}$$

$$n_2 = kn_s \big/ \left(1+k\right) \tag{28}$$

$$n_1 = n_s \big/ \left(1+k\right) \tag{29}$$

.

From (27) and (28), $n_1 = n_{s(k=1)}(1+k)/4k$. Since 600 animals have already been sampled, we let $n_1 = 600$ and $n_{s(k=1)} = 684$ (from above). This yields $k \approx 0.398$ and $n_2 = 239$. Hence, we sample 239 cows at the second sampling and 839 cows in total.

7.32 Dependent samples

The number of paired observations (n_p) necessary to detect a difference in P_A estimated from testing each cow at 2 different time points is given by:

$$n_p = \left[\left(z_{\alpha/2}\sqrt{\dot{p}_{12}+\dot{p}_{21}} + z_{\beta}\sqrt{(\dot{p}_{12}+\dot{p}_{21})-(\dot{p}_{12}-\dot{p}_{21})^2}\right)\Big/(\dot{p}_{12}-\dot{p}_{21})\right]^2 \text{ (Lachin, 1992)} \qquad (30)$$

where $\dot{p}_{12}$ and $\dot{p}_{21}$ are the *a priori* probabilities that a cow will have test results T+T- and T-T+, respectively, at the 2 testing times. In the context of a group of animals tested at 2 time points, n_p refers to the number of pairs of tests to be done (i.e. the number of animals to be tested) (see appendix A.4).

Example

Assume that we have a large herd known to be infected with MAP. We wish to determine if the prevalence of infection is increasing in the herd and thus plan to select a group of cows and test them longitudinally at two time points. We assume based on previous experience with other herds and with the diagnostic test to be used that $\approx 87\%$ of cows tested will have the same test

result at both testings. In addition, we assume that $P_{A_1}\left(p_{1+}\right)=0.10$ and $P_{A_2}\left(p_{2+}\right)=0.05,$ respectively (see appendix A.4) . Hence $\dot{p}_{12}+\dot{p}_{21}=0.13$ and $\dot{p}_{21}-\dot{p}_{12}=0.05$. Using $\alpha=0.05,$ z_α (one-sided hypothesis) and $1\text{-}\beta=0.8,$ from (30), $n_p\approx 318$ pairs of tests (318 cows).

7.4 Sample size based on true prevalence

Sample sizes to detect differences in P_T are done by first determining the difference in P_T that is important to identify.. The Rogan-Gladen estimator (7) is used to calculate the equivalent difference on the P_A scale and then equation (26) or (30) is used to find the sample size necessary to detect the difference on the P_A scale. We recommend that sample size estimation be done based on true rather than apparent prevalence estimates (see table 6), wherever possible.

7.5 Sample size required to estimate individual-level prevalence from pooled samples

Formula for numbers of pools required to estimate individual level true and apparent prevalence can easily be derived given the corresponding variance formulas (Cowling et al., 1999).

For fixed pool sizes, the formula for number of pools necessary to estimate P_A within a given error margin (Δ) is

$$n_p=z^2_{\alpha/2}\left(1-\left(1-\dot{P}_A\right)^{s_p}\right)\left(1-\dot{P}_A\right)^{2-s_p}\Big/\Delta^2 s_p^2. \qquad (31)$$

If an *a priori* estimate is given of the prevalence of positive pools $\left(\dot{p}_p\right)$, then:

$$n_p = z_{\alpha/2}^2 \dot{p}_p \left(1-\dot{p}_p\right)^{\left(2/s_p\right)-1} \Big/ \Delta^2 s_p^2 . \quad (32)$$

Given an *a priori* estimate of true prevalence, when Se and Sp are assumed known

$$n_p = z_{\alpha/2}^2 \dot{p}_p \left(1-\dot{p}_p\right)\left(1-\dot{P}_T\right)^{\left(2/s_p\right)-2} \Big/ J_p^2 \Delta^2 s_p^2 \quad (33)$$

If Se and Sp are not known,

$$n_p = z_{\alpha/2}^2 \dot{p}_p \left(1-\dot{p}_p\right) A^{\left(2/s_p\right)-2} \Big/ \left(J^2 \Delta^2 s_p^2 - z_{\alpha/2}^2 A^{\left(2/s_p\right)-2} \left(B_p \left(1-A_p\right)^2 + C_p A_p{}^2 \right) \right) \quad (34)$$

where $A_p = \left(PSe - p_p\right) \big/ J_p = \left(1-\hat{P}_T\right)^{s_p}$, $B_p = PSe\left(1-PSe\right)/n_1$, $C_p = PSp\left(1-PSp\right)/n_2$ and n_1 and n_2 (as before) refer to the number of infected and non-infected animals, respectively, used in the original test validation study. As in the case of individual samples, Δ may either be a constant or a percentage of the prevalence to be estimated. We note that for a given *a priori* P_A $\left(\text{or } P_T \text{ or } p_p\right)$ and Δ, larger pool sizes result in smaller total numbers of pools (see table 7).

7.6 Costs of pooling vs. costs of individual sampling

From a financial perspective, pooled estimation of individual prevalence is more feasible than using individual samples when for a given half-width $\left(\Delta\right)$

$$c_p < \left(\left(n/n_p\right)-1\right)c_T + \left(\left(n/n_p\right)-s_p\right)c_S \quad (35)$$

(see appendix A.5 for derivation) where c_T = cost of performing one diagnostic test, c_S = cost of obtaining an individual sample, s_p = pool size, c_P = cost associated with combining s_p samples into a single pool (mixing and storage of individual samples), and n and n_p are the numbers of individuals and pooled samples required, respectively, to estimate P_T within a given error of its true value.

Example

Assume that for a given herd, c_S = US \$1, c_T = US \$19 (for fecal culture). Given that we wish to estimate $\dot{P}_T = 0.10$ within an error margin $\Delta = 0.05$, using the information in table 9 for a pool size s_p =10 , it is economically feasible to use pooled rather than individual testing as long as the cost per pool (i.e., creating and storing) c_P < US \$10.52 . From table 8, if c_P < US \$1.93, then it is more feasible to use pool sizes of 5 than individual samples.

7.7 Sample size required to estimate herd prevalence

The number of herds required to estimate true herd prevalence $\left(HP_T\right)$, within a given error margin $\left(\Delta\right)$ is given by:

$$n_h = z^2_{\alpha/2} H\dot{P}_A(1-H\dot{P}_A)\big/\Delta^2 J_h^2 = z^2_{\alpha/2}\left(\Omega+\Pi\right)\left(1-\Omega+\Pi\right)\big/\Delta^2 J_h^2 \qquad (36)$$

(Humphry et al., 2004)

where n_h = the estimated number of herds, $\Omega = HSe\left(H\dot{P}_T\right)$, $\Pi = \left(1-HSp\right)\left(1-H\dot{P}_T\right)$, Δ = prescribed error margin, $J_H = HSe + HSp - 1$ and $z_{\alpha/2}$ as before, refers to the $1-\alpha/2$ percentile of the standard normal distribution. The test's herd-level characteristics $\left(HSe \text{ and } HSp\right)$ are specified by the investigator and therefore assumed known without error. Tables 10 and 11 provide estimates of the number of herds required to estimate $HP_T = 0.7$ and 0.9, respectively, within an error margin (Δ) = 0.1 with 95% confidence.

7.8 Sample size required to substantiate freedom from infection.

Strictly speaking, to unequivocally demonstrate freedom from MAP infection in a herd or in a group of herds from a well-defined geographical region, all animals should be tested using a perfect test. However, a sample may be used to establish with prescribed probability, that infection, if present, is below a predetermined level, i.e., a certain minimum expected prevalence (also known as design prevalence). For a herd, the minimum expected prevalence of infection would be the minimum prevalence expected if infection were actually present (Cameron and Baldock, 1998). This prevalence estimate should be based on knowledge of the disease and its risk factors, knowledge of the herd itself (including previous results of surveillance and disease-related interventions) and knowledge of the *Se* and *Sp* of the diagnostic test to be used. At the regional level, the minimum accepted herd prevalence would be the lowest (non-zero) unacceptable herd-prevalence, based on economic or political factors (Cameron and Baldock, 1998). For substantiation of freedom from infection, two-stage sampling is usually employed. Sample sizes and cut-offs for both herd-level and within-herd testing to establish that infection is below a specified prevalence in a given region may be estimated using the "Freedom from Disease" drop-down menu within the "Survey Toolbox" or "FreeCalc" which are both available at www.epiweb.massey.ac.nz (see appendix A.6).

Example

Assume that in a region with 1000 dairy herds (herd sizes of 300 to 1500 cows) where the herd prevalence of MAP infection was initially 90%, substantial eradication efforts are undertaken and it is now claimed that the herd prevalence is below 50%. Veterinary authorities expect that within any given herd, 2% or more of the animals will be infected if MAP is present. Fecal culture $\left(Se = 0.60, Sp = 0.999\right)$ is to be used for testing. Using the Survey Tool Box with

$HSe = 0.95$ and $HSp = 0.98$, and 95% confidence and power, we calculate that if 8 herds are chosen randomly and ≤ 1 herd tests positive, we can be $\approx 95\%$ confident that $HP_T < 0.5$. The Survey Tool Box also calculates the numbers of animals necessary to sample (based on herd sizes) in order to be $\approx 95\%$ certain that infection if present is $\leq 2\%$, given type I and type II errors of 0.05 and 0.02, respectively (see table 12). The survey would then be conducted by randomly sampling 8 herds and then choosing random samples of between 296 to 453 cows from each herd (see table 12), depending on herd size. If > 2 animals test positive for MAP in at least 2 of the 8 herds, then the survey has not provided evidence that $HP_T < 0.5$. However, if either 0 or 1 herd has > 2 animals testing positive, then we can conclude with more than 95% confidence that the true herd prevalence (HP_T) of MAP infection in the region is < 0.5 and where present, the true individual-level prevalence (P_T) of MAP infection is < 0.02.

For a given herd, the lower the minimum expected prevalence and the more inaccurate the diagnostic test, the greater the sample size necessary to substantiate freedom from infection (Cameron and Baldock, 1998). Given a herd of 1000 cows and that diagnostic testing is to be done using fecal culture $(Se = 0.60$ and $Sp = 0.999)$ or ELISA $(Se = 0.30$ and $Sp = 0.96)$, table 13 provides sample size estimates and maximal numbers of animals testing positive in order to establish with 95% confidence that MAP infection, if present, is below a given true prevalence.

Bayesian approaches to the substantiation of disease freedom are also possible. However due to their complexity we don't present them here but refer the interested reader to (Johnson et al., 2004, Branscum et al., 2006)

7.9 Power calculations

7.91 Relevance of power calculations

In the context of detecting differences in prevalences, power $(1-\beta)$ is the probability that the results of the study will be statistically significant given a specified difference between groups (Goodman and Berlin, 1994). Consequently, power is of direct relevance in the design phase of a study, as a means of determining the likelihood of ultimately detecting a true difference between prevalences given constraints on sample size (due to cost, availability of subjects or other factors). Post-hoc power calculations, though often performed by researchers, are of limited value in both study interpretation and evidence-based decision making. Instead, confidence intervals (CIs) should be used to evaluate the extent to which the differences (in prevalence) are or are not compatible with the data (Goodman and Berlin, 1994).

7.92 Power estimation (statistically independent samples)

From (26), the power to detect a difference $\left|P_{A_2} - P_{A_1}\right|$ is given by:

$$\Phi\left(Z_\beta\right) = \Phi\left[\left(\sqrt{n_1}\left|\dot{P}_{A_1} - \dot{P}_{A_2}\right| - z_{\alpha/2}\sqrt{\bar{\dot{P}}_A\bar{\dot{Q}}_A\left(1+\left(1/k\right)\right)}\right)\Big/\sqrt{\dot{P}_{A_1}\dot{Q}_{A_1} + \left(\dot{P}_{A_2}\dot{Q}_{A_2}/k\right)}\right]$$ (Rosner, 2000) (37)

where n_1, k, $\dot{P}_{A_1}$, $\dot{P}_{A_2}$, $\dot{Q}_{A_1}$, $\dot{Q}_{A_2}$ $\bar{\dot{P}}_A$, $\bar{\dot{Q}}_A$ and $z_{\alpha/2}$ are as defined in 7.31 and $\Phi\left(x\right)$ denotes the standard normal cumulative distribution function evaluated at x.

Example

Assume as in the second example of *7.31,* that having previously sampled 600 cows $\left(\hat{P}_{A_1} = 0.10\right)$, we wish to detect a possible decrease in prevalence of Johne's disease of 5% at a

subsequent sampling i.e., $P_{A_2} = 0.05$. However, funding dictates that only 160 cows can be tested. Thus, we estimate the probability of detecting this difference in prevalence, if it exists. Using $\alpha = 0.05$ and k = 0.267, power $= \Phi(0.395) = 0.65$. If 160 cows are sampled, there is a probability of 0.63 of detecting a change of 0.05 in apparent prevalence. If we were to sample 240 animals (k = 0.398), power = 0.80.

7.93 Power calculations (statistically dependent samples)

From (30), with matched pairs or repeated measures data, the power to detect a difference $\left|P_{A_2} - P_{A_1}\right|$ is given by:

$$\Phi(Z_\beta) = \Phi\left[\left(\sqrt{n_p}\,|p_{12} - p_{21}| - z_{\alpha/2}\sqrt{p_{12} + p_{21}}\right) \Big/ \sqrt{(\dot{p}_{12} + \dot{p}_{21}) - (\dot{p}_{12} - \dot{p}_{21})^2}\right] \qquad (38)$$

where n_p, p_{12}, p_{21} and $z_{\alpha/2}$ are as defined in section *7.32*.

Example

Assume that for the herd described in section *7.32*, testing costs dictate that only 540 tests can be performed (270 animals). Using the information in section *7.32* with $n_p \approx 270$, from (38) $\Phi(Z_\beta) = 0.64$ and power = 0.74. Hence, there is a 74% chance of detecting an increase in apparent prevalence of 0.05, if it does occur.

8. Conclusion

Frequentist approaches to true prevalence estimation are easy to use but suffer from serious statistical limitations when prevalence is low, sample sizes are small and uncertainty in *Se* and *Sp* are to be incorporated into the precision of estimates. These limitations are easily overcome in a Bayesian context. It is hoped that this article will provide impetus for increased application of Bayesian methods in animal health research.

Acknowledgements

The study was funded through the Johne's Disease Integrated Project and was supported by USDA-CSREES-NRI grant number 2004-35605-14243

References

Adcock CJ (1997). Sample size determination: A review. Statistician 46: 261-283.

Agresti A (1996). An introduction to categorical data analysis. New York: Wiley-Interscience.

Agresti A and Coull BA (1998). Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician 52: 119-126.

Altman DG (1991). Practical statistics for medical research. London; New York: Chapman and Hall.

Baggesen DL, Wegener HC, Bager F, Stege H and Christensen J (1996). Herd prevalence of Salmonella enterica infections in Danish slaughter pigs determined by microbiological testing. Preventive Veterinary Medicine 26: 201-213.

Benito A, Carmena D, Joseph L, Martinez J and Guisantes JA (2006). Dog echinococcosis in Northern Spain: Comparison of coproantigen and serum antibody assays with coprological exam. Veterinary Parasitology 142: 102-111.

Berry DA (1996). Statistics: A Bayesian Perspective. Belmont, Calif.: Duxbury Press.

Berry DA and Lindgren BW (1996). Statistics: Theory and Methods. Belmont, Calif.: Duxbury Press.

Bland JM and Altman DG (1998). Bayesians and frequentists. British Medical Journal 317: 1151-60.

Boelaert F, Walravens K, Biront P, Vermeersch JP, Berkvens D and Godfroid J (2000). Prevalence of paratuberculosis (Johne's disease) in the Belgian cattle population. Veterinary Microbiology 77: 269-281.

Borel N, Doherr MG, Vretou E, Psarrou E, Thoma R and Pospischil A (2004). Seroprevalences for ovine enzootic abortion in Switzerland. Preventive Veterinary Medicine 65: 205-216.

Branscum AJ, Gardner IA and Johnson WO (2004). Bayesian modeling of animal- and herd-level prevalences. Preventive Veterinary Medicine 66: 101-112.

Branscum AJ, Gardner IA and Johnson WO (2005). Estimation of diagnostic-test sensitivity and specificity through Bayesian modeling. Preventive Veterinary Medicine 68: 145-163.

Branscum AJ, Johnson WO and Gardner IA (2006). Sample size calculations for disease freedom and prevalence estimation surveys. Statistics in Medicine 25: 2658-2674.

Brinkhof JMA, Houwers DJ and van Maanen C (2006). Development of a sample pooling strategy for the serodiagnosis of small ruminant lentiviral infections using the ELITEST-MVV ELISA. Small Ruminant Research doi:10.1016/j.smallrumres.2006.03.003

Cameron AR and Baldock FC (1998). Two-stage sampling in surveys to substantiate freedom from disease. Preventive Veterinary Medicine 34: 19-30.

Carabin H, Balolong E, Joseph L, McGarvey ST, Johansen MV, Fernandez T, Willingham AL and Olveda R (2005). Estimating sensitivity and specificity of a faecal examination method for Schistosoma japonicum infection in cats, dogs, water buffaloes, pigs, and rats in Western Samar and Sorsogon Provinces, The Philippines. International Journal for Parasitology 35: 1517-24.

Christensen J, Baggesen DL, Nielsen B and Stryhn H (2002). Herd prevalence of Salmonella spp. in Danish pig herds after implementation of the Danish Salmonella Control Program with reference to a pre-implementation study. Veterinary Microbiology 88: 175-188.

Christensen J and Gardner IA (2000). Herd-level interpretation of test results for epidemiologic studies of animal diseases. Preventive Veterinary Medicine 45: 83-106.

Clough HE, Clancy D, O'Neill PD and French NP (2003). Bayesian methods for estimating pathogen prevalence within groups of animals from faecal-pat sampling. Preventive Veterinary Medicine 58: 145-169.

Collins MT, Gardner IA, Garry FB, Roussel AJ and Wells SJ (2006). Consensus recommendations on diagnostic testing for the detection of paratuberculosis in cattle in the United States. Journal of the American Veterinary Medical Association 229: 1912-1919.

Collins MT, Wells SJ, Petrini KR, Collins JE, Schultz RD and Whitlock RH (2005). Evaluation of five antibody detection tests for diagnosis of bovine paratuberculosis. Clinical and Diagnostic Laboratory Immunology 12: 685-692.

Cowling DW, Gardner IA and Johnson WO (1999). Comparison of methods for estimation of individual-level prevalence based on pooled samples. Preventive Veterinary Medicine 39: 211-225.

Delafosse A, Castro-Hermida JA, Baudry C, Ares-Mazas E and Chartier C (2006). Herd-level risk factors for Cryptosporidium infection in dairy-goat kids in western France. Preventive Veterinary Medicine 77: 109-121.

Donald AW, Gardner IA and Wiggins AD (1994). Cut-off points for aggregate herd testing in the presence of disease clustering and correlation of test errors. Preventive Veterinary Medicine 19: 167-187.

Dorny P, Phiri IK, Vercruysse J, Gabriel S, Willingham IAL, Brandt J, Victor B, Speybroeck N and Berkvens D (2004). A Bayesian approach for estimating values for prevalence and diagnostic test characteristics of porcine cysticercosis. International Journal for Parasitology 34: 569-576.

Dunson DB (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. American Journal of Epidemiology 153: 1222-1226.

Durr PA, Tait N and Lawson AB (2005). Bayesian hierarchical modelling to enhance the epidemiological value of abattoir surveys for bovine fasciolosis. Preventive Veterinary Medicine 71: 157-172.

Enoe C, Georgiadis MP and Johnson WO (2000). Estimation of sensitivity and specificity of diagnostic tests and disease prevalence when the true disease state is unknown. Preventive Veterinary Medicine 45: 61-81.

Fleiss JL (1981). Statistical Methods for Rates and Proportions. New York: Wiley.

Gardner IA (2002). The utility of Bayes' theorem and Bayesian inference in veterinary clinical practice and research. Australian Veterinary Journal 80: 758-61.

Geurden T, Claerebout E, Vercruysse J and Berkvens D (2004). Estimation of diagnostic test characteristics and prevalence of *Giardia duodenalis* in dairy calves in Belgium using a Bayesian approach. International Journal for Parasitology 34: 1121-1127.

Goodman SN and Berlin JA (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Annals of Internal Medicine 121: 200-6.

Greenland S (1990). Randomization, statistics, and causal inference. Epidemiology 1: 421-9.

Greiner M and Gardner IA (2000a). Application of diagnostic tests in veterinary epidemiologic studies. Preventive Veterinary Medicine 45: 43-59.

Greiner M and Gardner IA (2000b). Epidemiologic issues in the validation of veterinary diagnostic tests. Preventive Veterinary Medicine 45: 3-22.

Hanson T, Johnson WO and Gardner IA (2003a). Hierarchical models for estimating herd prevalence and test accuracy in the absence of a gold standard. Journal of Agricultural Biological and Environmental Statistics 8: 223-239.

Hanson TE, Johnson WO, Gardner IA and Georgiadis MP (2003b). Determining the infection status of a herd. Journal of Agricultural Biological and Environmental Statistics 8: 469-485.

Hui SL and Walter SD (1980). Estimating the error rates of diagnostic tests. Biometrics 36: 167-171.

Humphry RW, Cameron A and Gunn GJ (2004). A practical approach to calculate sample size for herd prevalence surveys. Preventive Veterinary Medicine 65: 173-188.

Johnson WO, Gastwirth JL and Pearson LM (2001). Screening without a "gold standard": The Hui-Walter paradigm revisited. American Journal of Epidemiology 153: 921-924.

Johnson WO, Su CL, Gardner IA and Christensen R (2004). Sample size calculations for surveys to substantiate freedom of populations from infectious agents. Biometrics 60: 165-171.

Jovanovic BD and Levy PS (1997). A look at the Rule of Three. The American Statistician 51: 137-139.

Kalis CHJ, Collins MT, Barkema HW and Hesselink JW (2004). Certification of herds as free of Mycobacterium paratuberculosis infection: Actual pooled faecal results versus certification model predictions. Preventive Veterinary Medicine 65: 189-204.

Kalis CHJ, Hesselink JW, Barkema HW and Collins MT (2000). Culture of strategically pooled bovine fecal samples as a method to screen herds for paratuberculosis. Journal of Veterinary Diagnostic Investigation 12: 547-551.

Kline RL, Brothers TA, Brookmeyer R, Zeger S and Quinn TC (1989). Evaluation of Human Immunodeficiency Virus seroprevalence in population surveys using pooled sera. Journal of Clinical Microbiology 27: 1449-1452.

Lachin JM (1981). Introduction to sample size determination and power analysis for clinical trials. Controlled Clinical Trials 2: 93-113.

Lachin JM (1992). Power and sample size evaluation for the McNemar test with application to matched case-control studies. Statistics in Medicine 11: 1239-51.

Leemis LM and Trivedi KS (1996). A comparison of approximate interval estimators for the Bernoulli parameter. The American Statistician 50: 63-68.

Letellier C, De Meulemeester L, Lomba M, Mijten E and Kerkhofs P (2005). Detection of BVDV persistently infected animals in Belgium: Evaluation of the strategy implemented. Preventive Veterinary Medicine 72: 121-125.

Louis TA (1981). Confidence intervals for a binomial parameter after observing no successes. The American Statistician 35: 154-154.

Manning EJB and Collins MT (2001). Mycobacterium avium subsp paratuberculosis: pathogen, pathogenesis and diagnosis. Revue scientifique et technique de l'Office international des epizooties 20: 133-150.

Martin SW, Shoukri M and Thorburn MA (1992). Evaluating the health-status of herds based on tests applied to individuals. Preventive Veterinary Medicine 14: 33-43.

Munoz-Zanzi C, Thurmond M, Hietala S and Johnson W (2006). Factors affecting sensitivity and specificity of pooled-sample testing for diagnosis of low prevalence infections. Preventive Veterinary Medicine 74: 309-322.

Ngowi HA, Kassuku AA, Maeda GEM, Boa ME, Carabin H and Willingham IAL (2004). Risk factors for the prevalence of porcine cysticercosis in Mbulu District, Tanzania. Veterinary Parasitology 120: 275-283.

Nielsen SS, Thamsborg SM, Houe H and Bitsch V (2000). Bulk-tank milk ELISA antibodies for estimating the prevalence of paratuberculosis in Danish dairy herds. Preventive Veterinary Medicine 44: 1-7.

Rahme E, Joseph L and Gyorkos TW (2000). Bayesian sample size determination for estimating binomial parameters from data subject to misclassification. Journal of the Royal Statistical Society Series C-Applied Statistics 49: 119-128.

Rapsch C, Schweizer G, Grimm F, Kohler L, Bauer C, Deplazes P, Braun U and Torgerson PR (2006). Estimating the true prevalence of Fasciola hepatica in cattle slaughtered in Switzerland in the absence of an absolute diagnostic test. International Journal for Parasitology 36: 1153-1158.

Rogan WJ and Gladen B (1978). Estimating prevalence from the results of a screening test. American Journal of Epidemiology 107: 71-6.

Rosner B (2000). Fundamentals of Biostatistics. Pacific Grove, CA: Duxbury.

Rothman KJ and Greenland S (1998) Measures of disease frequency. IN: Rothman KJ and Greenland S (Eds.) Modern Epidemiology. Philadelphia, PA: Lippincott-Raven, 29-46.

Sacks JM, Bolin SR and Crowder SV (1989). Prevalence estimation from pooled samples. American Journal of Veterinary Research 50: 205-206.

Scheaffer RL, Mendenhall W and Ott L (1995). Elementary Survey Sampling. Belmont, Calif.: Duxbury Press.

Schenker N and Gentleman JF (2001). On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician 55: 182-186.

Sergeant ESG, Whittington RJ and More SJ (2002). Sensitivity and specificity of pooled faecal culture and serology as flock-screening tests for detection of ovine paratuberculosis in Australia. Preventive Veterinary Medicine 52: 199-211.

Sockett DC, Carr DJ and Collins MT (1992). Evaluation of conventional and radiometric fecal culture and a commercial DNA probe for diagnosis of Mycobacterium paratuberculosis infections in cattle. Canadian Journal of Veterinary Research-Revue Canadienne De Recherche Veterinaire 56: 148-153.

Staubach C, Schmid V, Knorr-Held L and Ziller M (2002). A Bayesian model for spatial wildlife disease prevalence data. Preventive Veterinary Medicine 56: 75-87.

Su CL, Gardner IA and Johnson WO (2004). Diagnostic test accuracy and prevalence inferences based on joint and sequential testing with finite population sampling. Statistics in Medicine 23: 2237-2255.

Tavornpanich S, Gardner IA, Anderson RJ, Shin S, Whitlock RH, Fyock T, Adaska JM, Walker RL and Hietala SK (2004). Evaluation of microbial culture of pooled fecal samples for detection of Mycobacterium avium subsp paratuberculosis in large dairy herds. American Journal of Veterinary Research 65: 1061-1070.

Tu XM, Litvak E and Pagano M (1994). Screening tests - Can we get more by doing less? Statistics in Medicine 13: 1905-1919.

Tu XM, Litvak E and Pagano M (1995). On the informativeness and accuracy of pooled pesting in estimating prevalence of a rare disease - Application to HIV screening. Biometrika 82: 287-297.

Turnquist SE, Snider TG, Kreeger JM, Miller JE, Hagstad HV and Olcott BM (1991). Serologic evidence of Paratuberculosis in Louisiana beef-cattle herds as detected by ELISA. Preventive Veterinary Medicine 11: 125-130.

van Schaik G, Schukken YH, Crainiceanu C, Muskens J and VanLeeuwen JA (2003). Prevalence estimates for paratuberculosis adjusted for test variability using Bayesian analysis. Preventive Veterinary Medicine 60: 281-295.

Vose D (2000). Risk analysis: A quantitative guide. Chichester ; New York: Wiley.

Wagner B and Salman MD (2004). Strategies for two-stage sampling designs for estimating herd-level prevalence. Preventive Veterinary Medicine 66: 1-17.

Wang X-H, Wu X-H and Zhou X-N (2006). Bayesian estimation of community prevalences of Schistosoma japonicum infection in China. International Journal for Parasitology 36: 895-902.

Wells SJ, Godden SM, Lindeman CJ and Collins JE (2003). Evaluation of bacteriologic culture of individual and pooled fecal samples for detection of Mycobacterium paratuberculosis in dairy cattle herds. Journal of the American Veterinary Medical Association 223: 1022-1025.

Wells SJ, Whitlock RH and Lindeman CJ (2002). Evaluation of bacteriologic culture of pooled fecal samples for detection of Mycobacterium paratuberculosis. American Journal of Veterinary Research 63: 1207-1211.

Wittes J (2002). Sample size calculations for randomized controlled trials. Epidemiologic Reviews 24: 39-53.

Worlund DD and Taylor G (1983). Estimation of disease incidence in fish populations. Canadian Journal of Fisheries and Aquatic Sciences 40: 2194-2197.

Youden WJ (1950). Index for rating diagnostic tests. Cancer 3: 32-35.