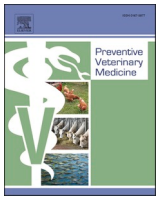
Preventive Veterinary Medicine 235 (2025) 106397

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed



A new web application for determining sample size in freedom-from-disease testing with imperfect tests

Darren Michael Green

Institute of Aquaculture, University of Stirling, Stirling, Stirlingshire FK9 4LA, UK

ARTICLE INFO

Keywords:
Diagnostic test
Epizootiology
Epidemiology
Sampling
Sample size

ABSTRACT

Veterinary surveillance frequently requires study design for freedom-from-disease testing, specifying a sample size to balance higher statistical power with larger sample sizes against increased research and ethics costs, with the recognition that tests can generate false positive and negative results: i.e., tests exhibit imperfect sensitivity and specificity. In this paper, we revisit the mathematics behind exact calculations of sample size in terms of the binomial and hypergeometric distributions, and present a new algorithm – implemented and available to use in *R* as a *Shiny* application with a graphical user interface – to determine sample size for practical situations. Often, sample size calculations are based upon simulations or approximations, but we show here that exact calculations are feasible. In addition, we relax the liberal assumption – which provides conservative sample-size estimates – that sensitivity and specificity are known exactly, and instead assume both are Beta distributed with known hyperparameters. The application presented here was originally designed as a learning tool for students and is now made available for wider use.

1. Introduction

A common task in veterinary epidemiology is surveillance for freedom from disease using fixed sample sizes and a diagnostic test of known sensitivity and specificity. Recent examples of designed sampling studies include surveillance of brucellosis in small ruminants in Algeria (Ramdani et al., 2022) and for zoonotic *Coxiella burnetii* in European bison in Poland (Krzysiak et al., 2021). Maximum sample size is limited only by the population size, but it is important to appreciate the diminishing returns of increased sampling – notably the standard error of a mean shrinks only according to the square root of sample size. Furthermore, there is the potential for diagnostic tests to give incorrect results at individual or herd level, and the costs of increasing testing in terms of finance and welfare (Bacchetti et al., 2011; Krzywinski and Altman, 2013). When designing sampling schemes for this purpose, it is necessary to estimate sample sizes based on prevalence estimates and a known population size. There is therefore a circular argument problem – intrinsic to power calculations: In order to plan to determine prevalence, you need to know something about the prevalence! There is also often an assumption that sensitivity and specificity are known exactly, and the more general case is considered below, where both quantities have some uncertainty in their estimation.

Sample size calculations for freedom from disease were previously devised by Cameron and Baldock (1998), and their approach is implemented by *Ausvet* (Sergeant, 2018), and (Paterson et al., 2020) . This method, amongst other aspects of epidemiological sampling, is recently reviewed by Stevenson (2021) and Meletis et al. (2024), who consider surveillance aspects such as time-series use in testing, structured populations, and the costs of potentially getting the answer wrong. Where a test has specificity below 100 %, some accommodation needs to be made for occasional false positives in the sample: The number of required positive tests is known in this case as the cutpoint. Cameron and Baldock (1998) devise an algorithm to solve the two-dimensional problem of determining both sample size (below, n) and the required cutpoint number of reactors (below, c). Their approach is developed further by Johnson et al. (2003), who relax the assumption that the population prevalence, test sensitivity, and specificity are perfectly known, and use a Bayesian approach applying Beta-distribution priors to both quantities, an approach also taken by Booth et al. (2023) while considering disease prevelance and sample size in wild-animal surveillance.

Frequently, sample size estimates are based upon approximations or simulation, as the exact calculation is thought to be complex or resource intensive. These approximations include substituting the binomial distribution for the hypergeometric distribution or rules of thumb based on simple approximations to the more complex mathematics. However, this assumption of complexity is worth revisiting with increased computer

E-mail address: darren.green@stir.ac.uk.

https://doi.org/10.1016/j.prevetmed.2024.106397
Received 17 May 2024; Received in revised form 28 November 2024; Accepted 29 November 2024
Available online 2 December 2024

processing power year on year, and increased ease of vectorising the equations in statistical programming environments such as *R* (R Core Team, 2024). As Gautam et al. (2019) argue, approximations – particularly where inherited from earlier work and poorly attributed – can lead to spurious results and are usually no longer necessary.

One such approximate method is to deal with confidence intervals of proportions by using approximations such as the normal distribution. Vallejo et al. (2013) note such approximations tend to be inaccurate close to the ends of the distribution at probabilities of zero or one, which is exactly the part of the distribution we are trying to work with when considering freedom from disease as a potentially rare event. Other concerns, such as the difference between sampling with and without replacement, and correction factors which can be applied to account for this, are reviewed by Fosgate (2009).

As part of development of teaching materials in epidemiology for MSc programmes here at the university, a *Shiny* application (Chang et al., 2024) has been developed that provides a web-based interface to an updated calculator for sample-size for freedom-from-disease testing, incorporating uncertainty for sensitivity and specificity. It is effectively exact, subject to some shortcuts in dealing with a large matrix outer product. This is combined with a new simple *R* algorithm to solve for required sample size. Shiny is an *R* package which provides a graphical user interface (GUI) which can be server deployed, run directly locally, or run locally from remotely hosted source files. This is not the first use of Shiny in developing epidemiological tools: Alba et al. (2017) have developed a Shiny application – *Optisample* – to optimise herd sampling from repeated samples with imperfect tests, with probability of freedom from disease as the key output. Thus, *Optisample* is the reverse calculation from what we are building here. SRUC (2024) also maintain a selection of Shiny applications online, including one for sample size for freedom from disease, however only for perfect sampling.

This paper takes the approach that exact statistical approaches are sometimes simpler to explain, and in the possession of sufficient computing power, easier to explore. Also, stemming from the fact that this paper emerged from a teaching project, we take pains here to explain the logic as clearly as possible given the fairly lengthy equations.

2. Methods

2.1. Distribution of positive test results

In this paper, we start with the approach taken by Cameron and Baldock (1998), and present the equations in three forms: first, in simplified form as probability distributions to demonstrate ease of vectorisation and implementation; second, in full with the probabilities explicitly calculated; and third, the corresponding *R* code. The code for the Shiny application is made available on *GitHub* at https://github.com/pinkmongoose/ShinySampleSize. With the *Shiny* library loaded, this application can be run by the single line of *R* code:

```
runGitHub[redacted for blind peer review]
  "ShinySampleSize","Pinkmongoose")
```

We start with a sample of individuals of size n drawn from a wider population of size N where the number of true positive and negative individuals meeting our case definition (as determined by a gold standard) are fixed at n^+ and $n^- = n - n^+$. (Therefore, at this stage we consider a particular sample from the population, rather than the potential range of sample prevalences found due to the sample itself being randomly drawn from this wider population.) A complete list of the mathematical symbols used for the models is given in Table 1. In this case, on application of an imperfect test, the distributions of the numbers of positive test results in the two subsets of gold-standard true-positive and gold-standard true-negative individuals (respectively t^+ and t^-, superscripts representing the nature of the sample, not our current test result), are both binomial and determined by the *sensitivity* $1-\beta$ and *specificity* $1-\alpha$ of the test used, where β is the *false negative rate* and α is the *false positive rate*.

Table 1
Model parameters, outcomes, and internal variables.

Symbol	Description	Domain
Parameters		
N	Population (herd) size	$\{1, 2\ldots\infty\}$
p	Fixed prevalence: proportion infected	$0 \leq x \leq 1$
$1-\beta$	Test sensitivity	$0 \leq x \leq 1$
$1-\alpha$	Test specificity	$0 \leq x \leq 1$
$1-B^{req}$	Desired herd sensitivity	$0 \leq x \leq 1$
$1-A^{req}$	Desired herd specificity	$0 \leq x \leq 1$
Hyperparameters		
η^+	Beta sensitivity prior: test positives on true positives	$0 < x$
θ^+	— test negatives on true positives	$0 < x$
η^-	Beta specificity prior: test positives on true negatives	$0 < x$
θ^-	— test negatives on true negatives	$0 < x$
Internal variables		
n^+	Sampled individuals which are true positive	$\{0, 1\ldots\infty\}$
n^-	— true negative	$\{0, 1\ldots\infty\}$
t^+	Positive test results within sampled positives	$\Pr(x), x \in \{0, 1\ldots n^+\}$
t^-	— within sampled negatives	$\Pr(x), x \in \{0, 1\ldots n^-\}$
t	— within whole sample $t^+ * t^-$	$\Pr(x), x \in \{0, 1\ldots n\}$
T	— across average sample of infected herd	$\Pr(x), x \in \{0, 1\ldots N\}$
T^0	— across average sample of uninfected herd	$\Pr(x), x \in \{0, 1\ldots N\}$
i, j, k, m	Index variables	$\{0, 1\ldots\infty\}$
Outcomes		
n	Sample size	$\{1, 2\ldots\infty\}$
c	Cutpoint number of positive tests	$\{0, 1\ldots n\}$
$1-B$	Achieved herd sensitivity	$0 \leq x \leq 1$
$1-A$	Achieved herd specificity	$0 \leq x \leq 1$

$$\begin{aligned} t^+(n^+) &\sim \text{Binom}(n^+, 1-\beta) \\ t^+(n^+)_i &= \binom{n^+}{i}(1-\beta)^i\beta^{n^+-i} \end{aligned} \tag{1}$$

where i indexes across the distribution.

$$\begin{aligned} t^-(n^+) &\sim \text{Binom}(n-n^+, \alpha) \\ t^-(n^+)_i &= \binom{n-n^+}{i}\alpha^i(1-\alpha)^{n-n^+-i} \end{aligned} \tag{2}$$

In R, this is naturally vectorised, resulting in a single expression to calculate the whole of each distribution vector. We assume individual test results are independent and thus these two distributions are independent. In this case, we can multiply the two vectors to produce a matrix outer product, and sum the reverse diagonals to obtain the distribution of total positive test results – both true and false – in the whole sample, t. This is the convolution of the two distributions (1) and (2). In standard equation format, this operation looks a little different.

$$\begin{aligned} t(n^+)_k &= (t^+ * t^-)_k = \sum_{i=0}^{k} t_i^+ t_{k-i}^- \\ &= \sum_{i=0}^{k}\binom{n^+}{i}(1-\beta)^i\beta^{n^+-i}\binom{n-n^+}{k-i}\alpha^{k-i}(1-\alpha)^{n-n^+-k+i} \end{aligned} \tag{3}$$

The range for i only needs to cover $i \in \{0\ldots k\}$ because either i or $k-i$ are out of range otherwise. In terms of computer code, this stage of identifying the diagonal is one that is not easily performed without resorting to iterating over loops, which loses some of the efficiency of working with vectorised operations. Where sample size n is large, and

neither n^+ or $n - n^+$ are small, the size of $n^+(n - n^+)$ can in turn be large. To simplify this multiplication, we observe that the probability density of t^+ and t^- is in most cases near zero in the case of very large n, and for the elements of the outer product, these near-zero regions are even more noticeable. As an approximation, we therefore restrict this multiplication to the centre of mass of the probability distribution for both distributions, excluding the extremes of the tails by iterating through the centre 99.8 % for each. The tails of the binomial distribution are easily calculated in *R* by using the *qbinom* function, and they do not need to be searched for programmatically. This slightly goes against the 'exact' philosophy of this paper, however the exact approach can be specified in the application options, at the expense of speed.

2.2. Hypotheses for prevalence

Up to this point, we have considered a fixed number of sample positives n^+. In this section, we consider two alternate scenarios: disease present in the population, and disease absent. As with Johnson et al. (2003) we reverse the framing of the hypotheses compared with Cameron and Baldock (1998) such that our null hypothesis H_0 is disease absence, and the alternative hypothesis H_1 is disease presence, so strictly this is formulated as testing for the *presence* of disease, not absence. Our sample sensitivity then becomes $1 - B$ where B is the likelihood of failing to reject an incorrect null hypothesis (type-II error). And our sample specificity then becomes $1 - A$ where A is the likelihood of incorrectly rejecting a correct null hypothesis (type-I error).

We follow the Cameron and Baldock (1998) approach of devising our testing schedules such that sensitivity judged against a fixed worst-case *minimum expected prevalence* level, p across a finite population size of size N, while meanwhile also judging specificity against a population entirely free from disease. If prevalence is fixed and population size fixed, then the number of sample positives in a random sample taken without replacement follows the hypergeometric distribution.

$$\begin{aligned} n^+ &\sim \mathrm{Hyp}(pN, (1-p)N, n) \\ n_j^+ &= \frac{\binom{pN}{j}\binom{(1-p)N}{n-j}}{\binom{N}{n}} \end{aligned} \tag{4}$$

where j indexes across the distribution. For large populations, this is asymptotically approached by the binomial distribution, a frequently used approximation.

$$\begin{aligned} n^+ &\sim \mathrm{Binom}(n, p) \\ n_j^+ &= \binom{n}{j} p^j (1-p)^{n-j} \end{aligned}$$

We can then calculate (from (3) and (4)) the distribution of positive test results in all samples from the wider population, T, as the expectation of t across n^+.

$$\begin{aligned} T_k &= \sum_{j=0}^{n} n_j^+ t(j)_k \\ &= \sum_{j=0}^{n} \frac{\binom{pN}{j}\binom{(1-p)N}{n-j}}{\binom{N}{n}} \\ &\cdot \sum_{i=0}^{k} \binom{j}{i} (1-\beta)^i \beta^{j-i} \binom{n-j}{k-i} \alpha^{k-i} (1-\alpha)^{n-j-k+i} \end{aligned} \tag{5}$$

The above allows us to calculate the herd sensitivity of the sampling effort in the case of presence of disease at a prevalence level of p, but we also need to consider herd specificity, and the situation of a healthy population. Where disease is absent, $p = 0$, there can only be false-positive test results (denoted T^0), and the above equation simplifies to just the binomial distribution.

$$\begin{aligned} T^0 &\sim t^-(0) \sim \mathrm{Binom}(n, \alpha) \\ T_k^0 &= \binom{n}{k} \alpha^k (1-\alpha)^{n-k} \end{aligned} \tag{6}$$

2.3. Beyond known test parameters

A good reason for relaxing some of the assumptions made above and allowing for more variation in model parameters is that such variation is most likely to widen probability distributions based on them, with heavier tails, and a likely increase in our sample size calculation estimates. Or to put it the other way, our sample size calculations may not be conservative enough.

We assume above our testing is performed with known sensitivity and specificity according to binomial distributions. A more conservative approach is to assume sensitivity and specificity are not known with certainty and have a likely distribution, for which a flexible approach is to describe them with a Beta distribution (Johnson et al., 2003; Booth et al., 2023). Summing binomial distributions where the mean is beta distributed produces the Beta-binomial distribution, written below in the form of gamma functions, as these can handle non-integer parameters, unlike factorials:

$$\begin{aligned} x &\sim \mathrm{BetaBinom}(N, \eta, \theta) \\ x_i &= \frac{\Gamma(N+1)\Gamma(i+\eta)\Gamma(N-i+\theta)\Gamma(\eta+\theta)}{\Gamma(i+1)\Gamma(N-i+1)\Gamma(N+\eta+\theta)\Gamma(\eta)\Gamma(\theta)}. \end{aligned}$$

Hyperparameters η and θ (we have already used the conventional symbols α and β or A and B) can be thought of as representing in a Bayesian context a prior body of evidence of positive (η) and negative (θ) results in earlier studies of the performance of the diagnostic test. Where η and θ are arbitrarily large but still have a defined ratio $0 \leq \frac{\eta}{\eta+\theta} \leq 1$, our Beta-binomial distribution reduces to a binomial distribution, recovering our earlier model.

Let us assume that these hyperparameters are described by prior information where our test produced η^- positive results from negative samples, θ^- negative results from negative samples, η^+ positive results from positive samples, and θ^+ negative results from positive samples. In these cases, from a given number of positive or negative samples, we can revise our formulation for t^+ and t^- as follows:

$$\begin{aligned} t^+(n^+) &\sim \mathrm{BetaBinom}(n^+, \eta^+, \theta^+) \\ t^+(n^+)_i &= \frac{\Gamma(n^+ + 1)\Gamma(i+\eta^+)\Gamma(n^+ - i + \theta^+)\Gamma(\eta^+ + \theta^+)}{\Gamma(i+1)\Gamma(n^+ - i + 1)\Gamma(n^+ + \eta^+ + \theta^+)\Gamma(\eta^+)\Gamma(\theta^+)} \end{aligned} \tag{7}$$

$$\begin{aligned} t^-(n^+) &\sim \mathrm{BetaBinom}(n - n^+, \eta^-, \theta^-) \\ t^-(n^+)_i &= \frac{\Gamma(n - n^+ + 1)\Gamma(i+\eta^-)\Gamma(n - n^+ - i + \theta^-)\Gamma(\eta^- + \theta^-)}{\Gamma(i+1)\Gamma(n - n^+ - i + 1)\Gamma(n - n^+ + \eta^- + \theta^-)\Gamma(\eta^-)\Gamma(\theta^-)}. \end{aligned} \tag{8}$$

If we insert these two terms (7) and (8) into our previous equation for T_k we obtain the following monster:

$$T_k = \sum_{j=0}^{n} \frac{\binom{pN}{j}\binom{(1-p)N}{n-j}}{\binom{N}{n}} \cdot \sum_{i=0}^{k} \frac{\Gamma(j+1)\Gamma(i+\eta^+)\Gamma(j-i+\theta^+)\Gamma(\eta^+ + \theta^+)}{\Gamma(i+1)\Gamma(j-i+1)\Gamma(j+\eta^+ + \theta^+)\Gamma(\eta^+)\Gamma(\theta^+)} \cdot \frac{\Gamma(n-j+1)\Gamma(k-i+\eta^-)\Gamma(n-j-k+i+\theta^-)\Gamma(\eta^- + \theta^-)}{\Gamma(k-i+1)\Gamma(n-j-k+i+1)\Gamma(n-j+\eta^- + \theta^-)\Gamma(\eta^-)\Gamma(\theta^-)}. \quad (9)$$

We also need to update our expression for T_k^0 as a Beta-binomial rather than a binomial distribution:

$$T^0 \sim \text{BetaBinom}(n, \eta^-, \theta^-)$$
$$T_k^0 = \frac{\Gamma(n+1)\Gamma(k+\eta^-)\Gamma(n-k+\theta^-)\Gamma(\eta^- + \theta^-)}{\Gamma(n+1)\Gamma(n-k+1)\Gamma(n+\eta^- + \theta^-)\Gamma(\eta^-)\Gamma(\theta^-)}. \quad (10)$$

Under *Implementation* below, we will see how to work through all these gamma distributions more easily. To facilitate the same simplification of the convolution of t^+ and t^- as before for the binomial distribution, we perform a search to identify the tails of the distributions before calculating the outer product, a quantile function lacking in *R*.

2.4. Algorithm for determining sample size

Two user-preference parameters must be specified in order to determine required sample size, which are the population target sensitivity $1 - B^{\text{req}}$ and specificity $1 - A^{\text{req}}$ required by the user, which can be considered as forms of type I and type II error for the hypothesis test of disease presence. The values specified for both will depend on the cost–benefit trade-off of testing that generates either false negatives (low sensitivity) or false positives (low specificity). Where test specificity is perfect, we have only the sample size to consider. Where false test positives may occur, we also need to consider the case where a low rate of positive test results may need to be overlooked – the cutpoint. In this case, we have a two-dimensional optimisation problem: to identify the minimum sample size and number of positive test results which satisfies our requirements both in terms of herd sensitivity and specificity. As before, we attempt to use built-in language features of *R* where possible in favour of coded loops to maximise computational efficiency and minimise complexity of the code. A combination of exhaustive search and bracketing is used, to avoid the potential for non-optimal solutions to be found.

1. The algorithm starts with an initial trial sample size of $n = 1$ and cutpoint number of positive tests $c = 0$, meaning that for a sample size of n, only c positive test results are accepted before the hypothesis of freedom from disease is rejected.
2. Achieved herd sensitivity $1 - B$ and specificity $1 - A$ at herd level are determined as follows. For herd sensitivity, T_k (using (5) or (9)) is summed across $k = c + 1 \ldots n$, which are all cases where the number of positive test results exceeds the cutpoint.

$$1 - B = \sum_{k=c+1}^{n} T_k$$

For herd specificity, the simpler case where prevalence is zero need only be considered, in which case the equation for T collapses to a similar binomial or Beta-binomial distribution (using (6) or (10)). We sum over only those cases where the number of positive test results does not exceed the cutpoint:

$$1 - A = \sum_{k=0}^{c} T_k^0$$

3. Sample size is bracketed by an initial $n_{lo} = 1$ and $n_{hi} = N$, to provide an algorithm which scales $O(\log N)$ in complexity with increasing population size.
4. Looping from Step 7 returns to here.
5. If at the current parameters, target sensitivity is not achieved, n_{lo} is set to n, otherwise n_{hi} is set to n. The new value of n is then set at the geometric mean of n_{lo} and n_{hi}, rounding up.
6. The sample is then re-evaluated, looping until the bracketing converges on the lowest n which satisfies target sensitivity. If target sensitivity is not achieved, the algorithm terminates with an error message, indicating that even testing the entire population is not sufficient.
7. At this point, if target specificity is achieved, the algorithm is finished.
8. If not, c is incremented and n_{hi} is reset to N, and the algorithm continues from Step 4. Increasing c always reduces the sensitivity of the test, and so a larger sample size than currently under consideration will be needed, and there is no need to reset n_{lo} to 1.
9. If c reaches the total population size N, target specificity is not achieved, and the algorithm terminates with an error.

2.5. Implementation

The algorithm was developed in *R* using *Shiny*, hosted on *GitHub* and the user interface kept simple to avoid confusing language. The app is available on shinyapps.io at https://pinkmongoose.shinyapps.io/ShinySampleSize/. For the fixed sensitivity/specificity model, the user needs to specify the population size $N > 0$, the proportion prevalence $0 \leq p \leq 1$, test sensitivity and specificity $0 \leq \beta \leq 1$ and $0 \leq \alpha \leq 1$, and the target herd sensitivity and specificity $0 \leq B^{\text{req}} \leq 1$ and $0 \leq A^{\text{req}} \leq 1$. A push button then runs the algorithm and provides the user with certain output. Before running the algorithm, the input is sanity-checked for the ranges above and some situations where input parameters guarantee a solution cannot be found. The completed application was tested using cohorts of an MSc class as victims, revising the UI after the first cohort. For the Beta-distributed sensitivity/specificity model, to simplify the input, the parameter sets η and θ are reparameterised as estimates and sample sizes n_α and n_β, such that $\eta^+ = \beta n_\beta$, $\theta^+ = (1-\beta)n_\beta$, $\eta^- = (1-\alpha)n_\alpha$ and $\theta^- = \alpha n_\alpha$. These are not required to be integers. It may be helpful to know that the standard deviation of the Beta distribution can be given by:

$$\sigma = \frac{\sqrt{\eta\theta}}{(\eta+\theta)\sqrt{\eta+\theta+1}}$$

(Weisstein, 2021), which simplifies using our reparameterisation (where p stands for η or θ) to:

$$\sigma = \sqrt{\frac{p(1-p)}{n_p + 1}}.$$

These 'advanced' tools – less easily interpreted by the casual user – are restricted to an 'advanced' input panel within the application. Options in the application include choice for sensitivity and specificity to be modelled as binomial *versus* Beta-binomial, and for coverage of the convolution operation for the distribution t, defaulting to the high value of 0.999 (0.001 exclusion of each tail).

The long but elegant formulation of the Beta-binomial distribution is used here as it can be evaluated using the sum of a number of log-gamma functions, provided in *R* by *lgamma*. This therefore requires no additional package dependencies, and avoids the problem of multiplying large numbers of gamma functions, which can quickly lead to numerical errors, as the largest gamma function which can be calculated with 64-bit floating-point arithmetic is $\Gamma(171)$. There are two things to be wary of with using the log-gamma function. First, numerical errors (infinities)

can occur for unusual cases $\log\Gamma(x)$ where $x \leq 0$ but these mostly correspond to the case where $\text{BetaBinom}(\cdot) = 0$. This also occurs where $\eta = 0$ or $\theta = 0$, which can be avoided by adding a very small positive value to the operand, set arbitrary at 10^{-10}.

3. Results

3.1. Application output

Achieved herd sensitivity and specificity $1 - B$ and $1 - A$ are provided to the user, which in the case of a successful calculation are always in excess of $1 - B^{req}$ and $1 - A^{req}$, as well as the required sample size n and cutpoint c. A prose interpretation of these values is given. As well as textual output, the *Shiny* application also provides further diagnostics in the form of a receiver operator characteristic (ROC) analysis, plotting herd sensitivity versus herd specificity for a variety of cutpoints $c \in \{0, 1 \ldots n\}$ for the determined sample size n. This plot includes the values of $1 - B^{req}$ and $1 - A^{req}$ as well as the line $B = 1 - A$ (corresponding to an uninformative test) for comparison. This model output is shown in Fig. 1. The area under the curve is informative, where the higher the area, the more informative the test, with coverage of half the plot being no better than guesswork.

3.2. Case studies

To demonstrate agreement with other approaches to sample size calculation, and explore sensitivity to model assumptions, results are shown for three baseline scenarios.

'MSc' scenario This scenario is used as a demonstration for our students. Here, 10 % minimum infected prevalence exists amongst a population of 50, test sensitivity and test specificity are set to 98 %, and desired herd sensitivity and herd specificity are set to 95 %. In class, we consider the simpler case of where test sensitivity and test specificity are known with high accuracy, and test positives and test negatives use the binomial distribution, with sampling without replacement (hypergeometric). This is the model output shown in Fig. 1. We obtain a required sample size of 39 and a cutpoint number of reactors of 2. Following the approach of Sergeant (2018) the prose description of the results is described as *If a sample size of 39 is taken and 2 or fewer reactors are found, then the probability that the population is free from disease at a prevalence of 5 / 50 (0.1) is 0.9586*. This is written below in short as 2/39.

'Cameron & Baldock' This scenario considers a survey for foot and mouth disease (FMD) considered by Cameron and Baldock (1998). Here, a herd of 265 animals is considered, where in the event of infection a minimum prevalence of 30 %. Target herd and test sensitivity and specificity parameters are as for the 'MSc' scenario. Testing is performed as above. Our application generates for this scenario a sample size of 1/14, which agrees with that of Cameron and Baldock. Further comparisons are made with this *Ausvet* model output in Table 2 showing close agreement between the two approaches, with some deviation where the sample size is so large it approaches the whole population. Table 2 contrasts two testing situations: the second with a test with lower sensitivity and specificity (90 %) but with a condition that is easier to find (25 % expected minimum prevalence).

'Johnson et al.' Derived from the above, this scenario also considers a herd of 265 animals and minimum infected prevalence of 30 %. However, here we use the Beta-distribution parameters specified in Johnson et al. (2003) corresponding to $\eta^{+} = 68.74$, $\theta^{+} = 4.57$, $\eta^{-} = 3.17$ and $\theta^{-} = 107.2$ and test sensitivity and specificity of 0.938 and 0.971, reflecting that these test statistics are not known exactly. Again, we fix target herd sensitivity and specificity both to 95 %. Our application generates a sample size here of 2/19, which is marginally more conservative than the Johnson et al. sample size of 2/18.

This last scenario provides an opportunity to demonstrate sensitivity of sample size calculation to the amount of prior information n_β and n_α upon which sensitivity and specificity estimates are based. Both values are varied in the range 3–1000 in Fig. 2, on a log scale, showing the outputs of required sample size n and cutpoint c. In this scenario, a smaller amount of prior information results in more uncertainty and an increase in both required sample size and in the cutpoint, and required sample size and cutpoint are themselves seen to be correlated. However, the approach is considerably more tolerant of uncertainty in sensitivity than in uncertainty in specificity as can be seen in the graph, and the effect is equivalent to a modest decrease in sensitivity, which sample-

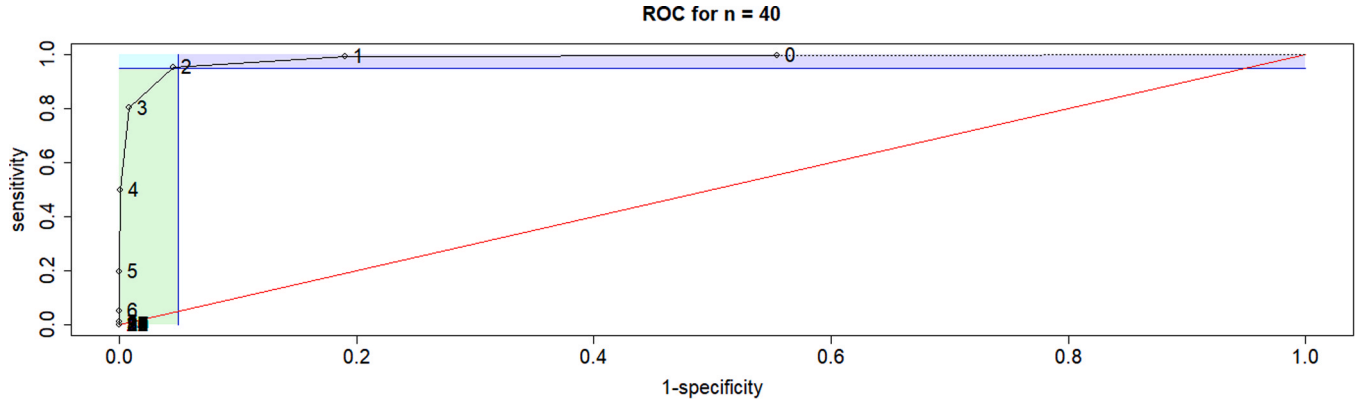


Fig. 1. ROC plot for 'MSc' sample calculation scenario. Herd sensitivity and specificity are shown for various cutoff numbers of reactors for a population size of $N = 50$ and $p = 10$ % prevalance, using test sensitivity and test specificity both of $1 - \beta = 1 - \alpha = 98$ %. Test positives and negatives use the binomial distribution, sampling without replacement. Shaded areas show acceptable range above target herd sensitivity and specificity both of $1 - B^{\text{req}} = 1 - A^{\text{req}} = 95$ %. The diagonal shows the result for an uninformative test.

Table 2
Comparison and diagnostics of sample size estimates.

				Results from this paper				*Ausvet*		
N	p	$1-\beta$	$1-\alpha$	$1-B$	$1-A$	c/n	$\sum t$	$1-B$	$1-A$	c/n
100	0.1	0.95	0.95	0.951	0.956	7/79	0.998	0.953	0.956	7/79
1000				0.952	0.956	9/108	0.999	0.953	0.956	9/108
10,000				0.951	0.953	9/109	0.999	0.952	0.953	9/109
40	0.25	0.90	0.90	0.951	0.964	6/32	0.999	0.964	0.977	7/36
100				0.955	0.963	7/39	0.999	0.956	0.963	7/39
1000				0.955	0.952	7/41	0.999	0.956	0.952	7/41

$1-B^{\text{req}}$ and $1-A^{\text{req}}$ are not listed, but the achieved $1-B$ and $1-A$ are instead presented to allow better comparison between models. The proportion coverage by the algorithm of the mass of distribution t is shown as $\sum t$

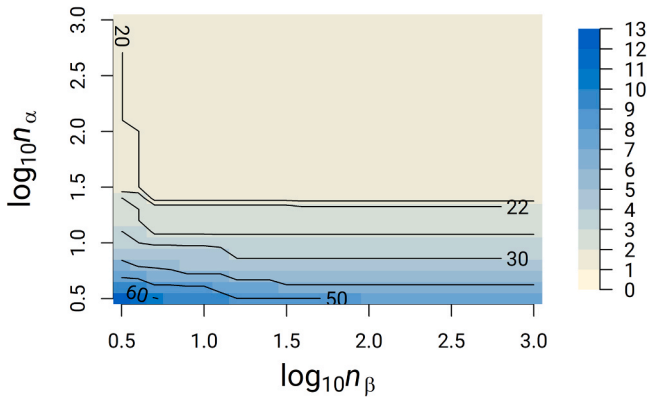


Fig. 2. Sensitivity analysis of Beta hyperparameters for 'Johnson' sample calculation scenario. The required sample size n is shown by the contour lines and the cutpoint number of positive tests c by coloured regions for different sizes of prior information for determining test sensitivity n_β and test specificity n_α. Prior information sizes were gridded on a $\log_{10}$ scale across several orders of magnitude with step size $\sqrt[10]{10}$. Other parameters are test sensitivity $1-\beta = 93.8$ %, test specificity $1-\alpha = 97.1$ %, population size $N = 265$, number infected $pN = 80$, and required sensitivity and specificity both $1-B^{\text{req}} = 1-A^{\text{req}} = 95$ %.

size calculations are generally more tolerant of than decreases in specificity.

4. Discussion

4.1. Implementation and assumptions

There are some implementation points of our approach here to note. The algorithm here fits one parameter, sample size n by bracketing, and cutoff c stepwise. This works well for most scenarios since in general $c \ll n$ and is close to zero where specificity is high. This means there are far fewer options for c to search through, while benefitting from the fast search for n. For $N = 5000$, $p = 0.1$, test specificity $1-\alpha = 1$, and sensitivity $1-\beta = 1$ it is nearly instantaneous on the author's laptop (Dell Inspiron 14 7000 [7490]; Intel Core i5-10210U processor). With higher population sizes and lower specificity, the application takes notably longer to process, up to a few seconds for $1-\beta = 0.95$.

A potential speed saving can be made by tolerating exclusion of more of the distribution tails while performing the convolution of the probability distributions, however this is limited as an option in the Shiny application settings as it does affect sample sizes if taken too far. For example, for the example of $N = 40$, $1-\alpha = 0.9$, $1-\beta = 0.9$, $p = 0.25$ from Table 2 and default coverage of tails (0.999) we obtain $c/n = 6/32$. Reducing coverage to 0.99 affects the result producing 6/33. Higher values for the coverage slow the program at higher population sizes.

Where multiple replicate measures of sensitivity or specificity exist, suitable hyperparameters could be chosen on the basis of knowing the mean and variance of a Beta distribution are $\frac{\eta}{\eta+\theta}$ and $\frac{\eta\theta}{(\eta+\theta)^2(\eta+\theta+1)}$. Another approach to choosing hyperparameters on the basis of known bounds for sensitivity and specificity was made by Johnson et al. (2003). This model produces an expectation for sample size based on specificity and sensitivity uncertainty, however there may be circumstances where, e.g. for precautionary purposes, a worst-case scenario is considered. In this case, the user may wish to consider the simpler binomial assumption for sensitivity and specificity, but where these are drawn as a particular percentile, e.g. 5 %, from the Beta-binomial distribution, e.g. using the *R* statement

```
qbeta(0.05, eta, theta).
```

An assumption made in the model above is that sensitivity and specificity are uncorrelated. There may be correlations amongst parameter estimates, e.g. if for whatever reason, overall positive test rates vary amongst trials, a positive correlation between estimates for both true positives and false positives would be expected, which implies negative correlation between sensitivity and specificity. Potential reasons for such correlations are discussed by Li and Fine (2011), particularly relating to prevalence itself, e.g. where higher prevalence also means more severe disease, or where pathologists are more willing to accept positive test results amongst high-prevalence populations. Similar complexities exist where tests are combined, pooled, or duplicated, where sensitivity and specificity must be considered as compound values with underlying reasons for test failure due to multiple factors, including the subject itself, sample collection, operator, interpretation, and chance (Greiner and Gardner, 2000).

Our algorithm for determining n and c is based on these variables only taking discrete, integer values. This makes sense in a practical capacity, but relaxing this assumption would allow greater precision where numbers such as cutpoints are small. Would not an output showing a cutpoint of $c = 0.5$ not be quite different from $c = 1.5$? Mathematically, this might be achieved by altering the expression for T replacing all occurrences of factorials and binomial coefficients with their equivalent formulation in terms of the gamma function, and replacing sums with integrals. This could potentially allow a whole new range of approaches to optimising n and c based on, for example, gradient-following algorithms, where points close together in parameter space are evaluated.

4.2. Comparison with simpler approaches

Des Clers (1994) discusses two approximations for sample sizes in the context of aquaculture prevalence studies, for small and large populations. It is useful to compare these approaches – easily calculated – with our more computationally intensive approach. For large populations (and, implicitly, high specificity), this is given as (refactored)

$$n = \frac{-\log B^{\text{req}}}{(1-\beta)p}.$$

For $1 - B^{\text{req}} = 0.95$, the numerator becomes 2.996, giving rise to the 'rule of three' sample size approximation name.

For small populations where the differences between the binomial, Poisson, and hypergeometric distributions become important, Des Clers gives this as (again, refactored)

$$n = \left(1 - B^{\text{req}\frac{1}{(1-\beta)pN}}\right)\left(N - \frac{(1-\beta)pN - 1}{2}\right)$$

Comparison with the algorithm described above shows this remains a useful approximation when specificity is high. For example, for test sensitivity $1 - \beta = 0.95$, $N = 50$, $p = 0.1$, and herd sensitivity $1 - B^{\text{req}} = 0.95$, Des Clers suggests a sample size of 22.5 (limit of large population, 31.5) which compares well with our result of 23. For $N = 1000$, we get 29.6 and 30 respectively. For a rarer disease, e.g. $N = 1000$ and $p = 0.01$, we get 269.3 versus 272 (limit of large population, 315.3). This approximation will become less appropriate, and very liberal, where test specificity is substantially less than perfect, a scenario not considered in the Des Clers model.

4.3. Conclusion

In short, this new application and associated algorithm provides another tool in the epidemiologists toolbox, provided here open source and with a front end which is easy to use. This application may be used by researchers as part of experimental design, or for planning routine surveillance, and it is also scalable as a teaching aid for large classes given its ease of deployment directly from *Github*. The subtle differences in sample size generated by different models may seem small, but it is exactly in the tails of such distributions we tend to be operating.

Ethical approval

This study was approved under the University of Stirling General University Ethical Panel (GUEP) review reference EC 2024 18301 13155.

Financial support

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Darren Michael Green: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of Competing Interest

The author declares no conflict of interest.

Data availability

The models used here are available at https://github.com/pinkmongoose/ShinySampleSize and at https://pinkmongoose.shinyapps.io/ShinySampleSize/

Acknowledgements

Thanks are due to the MSc students experimented on while developing this app, and particularly to Jimmy Turnbull, who first sparked my attention on the concepts here and helped with running the classes.

References

Alba, A., Morrison, R.E., Cheeran, A., Rivira, A., Alvarez, J., Perez, A.M., 2017. Optisample: open web-based application to optimize sampling strategies for active surveillance activities at the herd level illustrated using Porcine Respiratory Reproductive Syndrome (PRRS). PLoS One 12, e0176863.

Bacchetti, P., Deeks, S.G., McCune, J.M., 2011. Breaking free of sample size dogma to perform innovative translational research. Sci. Transl. Med. 3 (87), 87ps24. https://doi.org/10.1126/scitranslmed.3001628.

Booth, J.G., Hanley, B.J., Hodel, F.H., Jennelle, C.S., Guinness, J., Them, C.E., Mitchell, C.I., Ahmed, M.S., Schuler, K.L., 2023. Sample size for estimating disease prevalence in free-ranging wildlife populations: a Bayesian modeling approach. J. Agric., Biol. Environ. 29, 438–454. https://doi.org/10.1007/s13253-023-00578-7.

Cameron, A.R., Baldock, F.C., 1998. A new probability formula for surveys to substantiate freedom from disease. Prev. Vet. Med. 34, 1–17.

Chang W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., Borges, B., 2024.Shiny: Web Application Framework for R. R Package Version 1.8.0.9000. ⟨https://github.com/rstudio/shiny⟩, https://shiny.posit.co (Accessed 15 May 2024).

Des Clers, S., 1994. Sampling to Detect Infections and Estimate Prevalence in Aquaculture. Pisces Press.

Fosgate, G.T., 2009. Practical sample size calculations for surveillance and diagnostic investigations. J. Vet. Diagn. Investig. 21, 3–14.

Gautam, R., Wagener, A., Nerette, P., Bruneau, N., 2019. The inappropriate use of formulae and references and the possible domino effect of spurious results. Prev. Vet. Med. 170, 104728.

Greiner, M., Gardner, I.A., 2000. Epidemiologic issues in the validation of veterinary diagnostic tests. Prev. Vet. Med. 45, 3–22. https://doi.org/10.1016/S0167-5877(00)00114-8.

Johnson, W.O., Su, C.-L., Gardner, I.A., Christensen, R., 2003. Sample size calculations for surveys to substantiate freedom of populations from infectious agents. Biometrics 60, 165–171.

Krzysiak, M.K., Puchalska, M., Olech, W., Anusz, K., 2021. A freedom of *Coxiella burnetii* infection survey in European bison *Bison bonasus* in Poland. Animals 11, 651. https://doi.org/10.3390/ani11030651.

Krzywinski, M., Altman, N., 2013. Power and sample size. Nat. Methods 10, 1139–1140. https://doi.org/10.1038/nmeth.2738.

Li, J., Fine, J.P., 2011. Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. Biostatistics 12, 710–722.

Meletis, E., Conrady, B., Hopp, P., Lurier, T., Frössling, J., Rosendal, T., Faverjon, C., Carmo, L.P., Hodnik, J.J., Ózsvári, L., Kostoulas, P., van Schaik, G., Comin, A., Nielen, M., Knific, T., Schulz, J., Šerić-Haračić, S., Fourichon, C., Santman-Berends, I., Madouasse, A., 2024. Review state-of-the-art of output-based methodological approaches for substantiating freedom from infection. Front. Vet. Sci. 11. https://doi.org/10.3389/fvets.2024.1337661.

R Core Team, 2024. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing. ⟨https://www.R-project.org⟩ (Accessed 27 November 2024).

Paterson, J.T., Butler, C., Garrott, R., Proffitt, K., 2020. How sure are you? A web-based application to confront imperfect detection of respiratory pathogens in bighorn sheep. PLoS One 15, e0237309.

Ramdani, N., Boussena, S., Bouaziz, O., Moula, N., 2022. Brucellosis in small ruminant: seroprevalence, risk factors, and distribution in the southeast of Algeria. Trop. Anim. Health Prod. 54, 245. https://doi.org/10.1007/s11250-022-03236-1.

Sergeant, E.S.G., 2018.Epitools Epidemiological Calculators. Ausvet. ⟨http://epitools.ausvet.com.au⟩ (Accessed 15 May 2024).

SRUC, 2024.SRUC Epidemiology Resources: Sample size calculation.⟨https://epidemiology.sruc.ac.uk/shiny/apps/samplesize/⟩ (Accessed 15 May 2024).

Stevenson, M.A., 2021. Sample size estimation in veterinary epidemiologic research. Front. Vet. Sci. 7, 1115.

Vallejo, A., Muniesa, A., Ferreira, C., de Blas, I., 2013. New method to estimate the sample size for calculation of a proportion assuming binomial distribution. Res. Vet. Sci. 95, 405–409.

Weisstein, E.W., 2021.Beta Distribution. From MathWorld–A Wolfram Web Resource.⟨https://mathworld.wolfram.com/BetaDistribution.html⟩ (Accessed 15 May 2024).