Epidémiol. santé anim., 1997, 31-32

PROVING FREEDOM FROM DISEASE USING IMPERFECT TESTS: THE FREECALC SAMPLE SIZE CALCULATOR AND SURVEY ANALYSIS PROGRAM

Cameron A.R.[1], Baldock F.C.[2]

Les ratifications de l'Organisation Mondiale du Commerce sont en cours de modification notamment concernant la gestion des risques de maladies dans le cadre des échanges internationaux des produits d'origine animale. Les partenaires seront amenés à exiger d'avantage de garanties vis-à-vis des maladies animales les plus importantes, les résultats et preuves devront se baser sur des principes épidémiologiques.
Au niveau national, de nombreux pays souhaitant contrôler et éradiquer les maladies enzootiques, ont retenu des programmes d'accréditation de troupeaux. Dans les deux situations, la surveillance soit pour détecter les maladies, soit pour garantir la bonne santé est nécessaire.
Dans le passé, de pareilles études se sont toujours basées sur les hypothèses soit d'un test parfait, soit d'une population infinie. Une formule de probabilité a été développée tenant compte de la sensibilité et de la spécificité du test ainsi que de la taille de la population. Cette formule permet de calculer la probabilité avec précision connaissant le nombre de réactions positives aussi bien quand l'échantillon provient d'un élevage infecté que non infecté.
Cette formule a été mise en œuvre à l'aide d'un programme informatique « FreeCalc », qui permet de calculer la taille de l'échantillon pour ce type d'enquête, puis d'analyser les résultats. La base théorique de cette formule et l'application pratique de ce programme sont discutés.

INTRODUCTION

Recent changes in regulations governing the international trade in animals and animal products have led to a greater need for epidemiologically sound surveys to prove the freedom of a country or zone from particular diseases. The World Trade Organisation, created by the General Agreement on Tariffs and Trade (GATT) (Anon., 1994) has adopted the codes of the Office International des Epizooties (OIE) to act as guidelines for international trade in animals. A critical component of these guidelines is the establishment of regional, national, or sub-national disease-free zones for the purpose of livestock exports. Many countries are endeavouring to eradicate high impact or trade limiting diseases such as Rinderpest, Tuberculosis, Foot and Mouth Disease, and Contagious Bovine Pleuropneumonia. Proof of the final success of these campaigns will depend on epidemiologically valid surveys.

Countries currently free from disease may also be forced to provide stronger evidence for their disease-free status than simply the absence of clinical reports. Countries such as Australia may well be required to prove the absence of such diseases as Bovine Spongiform Encephalopathy or Porcine Reproductive and Respiratory Syndrome. In the case of an exotic disease outbreak in a previously free country, the economic effects of trade limitations can be minimised by the rapid establishment of infected and disease free zones within the country, and the proof of this status by appropriately designed surveys.

PROBABILITY FORMULA

Two of the most basic questions asked in designing a survey are "How many animals do I need to test?" and "How do I interpret the results?". The answers to these questions depend on a knowledge of the probability of detecting a given number of test-positive animals (reactors) when sampling the target population using a test of known performance. In the past, one or two assumptions have been commonly made to simplify the calculation of these probabilities, namely 1) that the diagnostic test being used is perfect (sensitivity and specificity are both equal to one) and/or 2) that the population being studied is infinite (or that sampling is performed with replacement). However, virtually no diagnostic test is perfect, and populations under study are often small enough to be considered finite, especially with herd-level sampling. To overcome these limitations, a probability formula based on the hypergeometric distribution was developed. The formula calculates the exact probability of observing a given number of reactors (test positive animals) from a given population, while taking test sensitivity and specificity and finite population size into account.

If P(T+ = x) is the probability that the observed number of test positive animals will equal x, n is the sample size, N is the population size, d is the number of diseased animals in the population, Se is the test sensitivity and Sp is the test specificity, then

$$P(T_+ = x) = \binom{n}{x} \sum_{i=0}^{n} \sum_{j=0}^{i} \frac{\binom{d}{i}\binom{N-d}{n-i}\binom{x}{j}\binom{n-x}{i-j} Se^{i} (1-Sp)^{x-j} (1-Se)^{d-i} Sp^{n-x-i+j}}{\binom{n}{i}\binom{N}{n}}$$

[1] Lao-Australian Animal Health Project, PO Box 7042, Vientiane, Lao PDR
[2] AusVet Animal Health Services, 12 Thalia Crt, Corinda, Queensland 4075, Australia

FREECALC COMPUTER PROGRAM

This formula has the drawback that it is difficult to calculate, due to summations over a large number of terms. It is also impossible to equate it to n to calculate the sample size. To simplify this task, the formula has been incorporated into a computer program, FreeCalc. The program is written in the Borland Pascal programming language, and runs under the MS-DOS operating system. The program has two parts. The first calculates the sample size required for a survey, based on the test sensitivity and specificity, the population size, the minimum expected prevalence, and the desired type I and type II error rates. When a survey has been conducted, the second part of the program can be used to analyse the results.

SAMPLE SIZE CALCULATION

When conducting a survey to prove freedom from disease, the null hypothesis H_0 is that disease is present at a specified prevalence and the alternative hypothesis, H_A, is that the disease is not present. This prevalence for the null hypothesis may be chosen in two ways. For herd level surveys, the prevalence represents the minimum prevalence expected for a disease should it be present. For national level surveys, the prevalence of disease-positive herds may be very low, while the within-herd prevalence amongst positive herds might be quite high. Choosing a minimum expected prevalence of one percent of herds affected is the same as saying that the survey will not be able to detect a prevalence of lower than one percent. The results of the survey, if they indicate freedom from disease, are in fact saying that, if disease is present, it is present at a prevalence of below one percent.

While the choice of the minimum expected within-herd prevalence is usually based on the epidemiology of the disease, the choice of between-herd prevalence is based on a combination of different factors. Proof of the absence of disease at very low levels requires a very large survey, so economic factors become important. Equally relevant are the requirements of trading partners and regulatory guidelines, which may demand proof at a specified prevalence.

Testing the hypothesis involves conducting a survey of n animals (or herds), and recording the number of positive reactors. Positive reactors are those animals with a positive test result, which may be either diseased (true positives) or non-diseased (false positives). The probability of observing this number of reactors under the null hypothesis is calculated. If this probability is small enough, the null hypothesis can be rejected. Using this traditional approach to statistical hypothesis testing, rejection of the null hypothesis usually implies acceptance of the alternative, H_A. In this case, if the probability of detecting the observed number of reactors from a herd initially assumed to be positive (with prevalence p) is small enough, we conclude that the herd is free from disease.

When testing multiple animals with an imperfect test, the number of reactors is related to the properties of the test (sensitivity and specificity), the number of animals tested, and the prevalence in the population. At a given prevalence, the probability of detecting a certain number of reactors can be calculated. A probability distribution can be created by calculating the probability over all possible numbers of reactors (zero to the sample size n).

Let α, the type I error, be the probability of rejecting the null hypothesis when the null is true (disease is in fact present), and β, the type II error, be the probability of accepting it in the absence of disease. If we assume that there is no disease present (the alternative hypothesis), a probability distribution can be created based on a given sample size, showing the probability of observing 0,1,2...n reactors. This distribution (or its cumulative equivalent) can be used to calculate the maximum number of reactors that would be observed with a probability of 1- β, if the population was free from disease. The value β marks the upper tail of the distribution, and is used as the cut-point for the maximum number of reactors if the disease is not present.

If the probability distribution based on the null hypothesis (disease present at prevalence = p) is drawn, the cumulative probability of observing a number of reactors less than or equal to this same cut-point can be calculated. If this probability is high, then the maximum number of reactors likely to be drawn from a population *without* disease is also quite likely to be drawn from a population *with* disease, making it difficult to distinguish between the two populations. If the probability is low, then it is unlikely that a number of reactors less than or equal to the cut-point would be drawn from a diseased population.

At low sample sizes, the distributions (for the null and alternative hypotheses) have a wide overlap. As the sample size increases, the distributions become more separated, with a smaller overlap. The required sample size is the value at which the number of reactors at the cut-point with a probability of 1-β from the distribution with zero prevalence is equal to the number of reactors occurring with probability at the left tail of the distribution with prevalence = p.

A trial and error procedure is used to calculate the required sample size. An arbitrary starting sample size is chosen, and the probability of falsely concluding that disease is not present is calculated. If this probability is higher than the required level of the sample size is increased, and the calculation starts again. If it is lower than the required level of α the sample size is decreased. The process is continued until a value equal to or just smaller than α is reached.

The cut-point for a sample size represents the maximum number of reactors that could be observed with 1-β probability, if the population is disease free. This is calculated by first determining the probability of observing no reactors. If this probability is less than 1-β, the cumulative probability of observing 1 or fewer reactors is calculated. This is repeated until the probability is equal to or slightly greater than 1-β. The calculation of the cut-point number of reactors based on a disease free population represents the role of the alternative hypotheses, that disease is not present.

The cumulative probability of observing a number of reactors equal to or less than the cut-point is calculated by summing the probabilities of observing 0, 1, 2... reactors up to the cut-point, based on a population with disease

prevalence as set by the null hypothesis. This cumulative probability is compared with α to assess whether the sample size is correct, too large or too small.

ANALYSIS OF SURVEY RESULTS

Analysis of survey results is a much simpler probability calculation. The same parameters as used for sample size calculation are required, in addition to the actual sample size used and the number of reactors observed. The program then reports the probability of observing this number of reactors under the null hypothesis, and provides an interpretation as to whether the null can be rejected or not.

TWO-STAGE SAMPLING

Two-stage sampling for surveys of large populations is necessary, both because of the clustered nature of disease, and from a practical point of view (eg. the problem of generating a comprehensive sampling frame). The first stage involves sampling herds (or any other convenient grouping of animals), and the second, individual animals from the herd. A two stage analysis approach must be used, first to classify each herd sampled as infected or uninfected based on the results of the individual animal tests, and then to classify the population of herds as infected or uninfected, based on the herd or aggregate tests.

A screening test for an individual animal is characterised by its sensitivity and specificity. When a screening test is applied to a sample of animals from a herd for the purpose of classifying the herd as diseased or non-diseased, the combined procedure may be thought of as a single, herd-level screening test, with its own sensitivity and specificity. The sensitivity and specificity of herd tests are influenced by the sensitivity and specificity of the individual animal test used, the number of animals tested, as well as the way in which individual animal results are interpreted. Various authors have discussed the interpretation of herd tests (Martin et al., 1992; Donald et al., 1994; Jordan, 1995). The most common approach is to use a cut-point number of reactors to classify the herd as either diseased or non-diseased. If the cut-point chosen is zero, herd-level specificity will be low, but the sensitivity will be high. As the cut-point number of reactors is increased the sensitivity decreases and the specificity increases. Designing an optimal herd-level survey is therefore a question of determining the best sample size and the cut-point level, that will give the optimal combination of sensitivity and specificity for the purposes of the survey. Ideally, the researcher should be able to determine the herd test sensitivity and specificity required and then calculate the corresponding sample size and cut-point number of reactors.

To calculate the sample size for single-stage surveys, it is necessary to first define the levels of Type I and II errors. In the context of herd testing, the herd-level sensitivity is the probability that a diseased herd will be classified as diseased. This is equal to one minus the probability that a diseased herd will be classified as disease-free, or 1- α. Similarly, the herd-level specificity is equal to 1-β. These simple relationships allow us to specify the required herd-level sensitivity and specificity by specifying the animal-level power and confidence levels, and calculate the animal-level sample size and cut-point number of reactors needed to achieve them.

DISCUSSION

The FreeCalc program is able to carry out all the calculations necessary for the determination of sample size and the analysis of surveys to prove freedom from disease. The program allows field officers with little statistical training to conduct precisely designed surveys easily and with confidence. The flexibility of the program also offers the opportunity to calculate least-cost sample sizes for two-stage sample surveys. The program is available free of charge over the Internet on the World Wide Web at the EpiVetNet Web site (http://epiweb.massey.ac.nz). No restriction is placed on the distribution of the program, and users are encouraged to pass it on to colleagues. Use of the program should always be acknowledged in reports, scientific papers and presentations.

ACKNOWLEDGEMENTS

The formula and program described in this paper were developed under a project funded by the Australian Centre for Agricultural Research and working in cooperation with the Thai Department of Livestock Development. Angus Cameron is supported by a Junior Research Fellowship from the Australian Meat Research Corporation.

BIBLIOGRAPHY

Anon. 1994. General agreement on tariffs and trade (GATT), Sanitary & Phytosanitary Measures (MTN/FA II-A1A-4). Agreement on the application of sanitary and phytosanitary measures

Donald A.W., Gardner I.A., Wiggins A.D., 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

Jordan D. 1995, Aggregate testing for the evaluation of Johne's disease herd status. In: Morton J. (ed) Epidemiology Chapter, Australian College of Veterinary Scientists Proceedings. Australian Veterinary Association Annual Conference, Melbourne, May 21-26, 1995. Australian College of Veterinary Scientists, Indooroopilly, 60-67

Martin S.W., Shoukri M., Thorburn M.A., 1992. Evaluation the health status of herds based on tests applied to individuals. Preventive Veterinary Medicine 14, 33-43