ISSN: 1322-624X

RESEARCH PAPERS AND REPORTS IN ANIMAL HEALTH ECONOMICS

AN ACIAR THAI-AUSTRALIAN PROJECT

Working Paper No. 27

Sampling Considerations for Active Surveillance of Livestock Diseases in Developing Countries

by

Steve Harrison

November 1996

THE UNIVERSITY OF QUEENSLAND

ISSN 1322-624X

RESEARCH PAPERS AND REPORTS IN ANIMAL HEALTH ECONOMICS

Working Paper No. 27

Sampling Considerations for Active Surveillance of Livestock Diseases in Developing Countries[1]

by

Steve Harrison[2]

November 1996

[1] This series, Research Papers and Reports in Animal Health Economics, is supported by the Australian Centre for International Agricultural Research (ACIAR) Project No. 9204 and is published by the Department of Economics, University of Queensland, 4072, Brisbane, Australia. (See next page for more information).

[2] Contact Email: c.tisdell@economics.uq.edu.au School of Economics, The University of Queensland, St. Lucia Campus, Brisbane QLD 4072, Australia. Steve Harrison: s.harrison@uq.edu.au

RESEARCH PAPERS AND REPORTS IN ANIMAL HEALTH ECONOMICS is published by the Department of Economics, University of Queensland, Brisbane, 4072, Australia as a part of a research project sponsored by the Australian Centre for International Agricultural Research, viz., Project No. 9204, 'Animal Health in Thailand and Australia: Improved Methods in Diagnosis, Epidemiology, Economic and Information Management'.

The Commissioned Organization is the Queensland Department of Primary Industries. Collaborating institutions in Australia are CSIRO-ANHL, Geelong, Victoria and the University of Queensland (Department of Economics; Department of Geographical Sciences and Planning). In Thailand, the collaborating institutions are the Department of Livestock Development (National Institute of Animal Health; Disease Control Division), Chiang Mai University (Department of Agricultural Economics; Department of Animal Husbandry) and Thammasat University (Faculty of Economics). The collaborating institution in Laos is the Department of Livestock and Veterinary Services. Dr F.C. Baldock, Senior Principal Epidemiologist, Queensland Department of Primary Industries is the Project Leader in Australia and Dr P. Chamnanpood, Senior Epidemiologist, Thai Department of Livestock Development is the Project Leader in Thailand. Professor Clem Tisdell and Dr Steve Harrison, Department of Economics, University of Queensland are responsible mainly for the economic component of this project.

'The overall goal of this project is to develop and evaluate the .necessary tools to provide decision-makers with reliable animal health information which is placed in context and analysed appropriately in both Thailand and Australia. This goal will be achieved by improving laboratory diagnostic procedures; undertaking research to obtain cost-effective population referenced data; integrating data sets using modern information management technology, namely a Geographical Information System (GIS); and providing a framework for the economic evaluation of the impact of animal diseases and their control.

A number of important diseases will be targeted in the project to test the systems being developed. In Thailand, the focus will be on smallholder livestock systems. In Australia, research will be directed at the northern beef industry as animal health information for this sector of livestock production is presently scarce.'

For more information on *Research Papers and Reports Animal Health Economics* write to Professor Clem Tisdell (c.tisdell@economics.uq.edu.au) or Dr Steve Harrison,(s.harrison@uq.edu.au) Department of Economics, University of Queensland, Brisbane, Australia, 4072.

Sampling Considerations for Active Surveillance of Livestock Diseases in Developing Countries

ABSTRACT

This paper reviews statistical considerations in multistage sampling designs for active surveillance of livestock diseases, with particular emphasis on FMD in Thailand. Issues addressed include statistical efficiency concepts, estimator formulae for proportion of protected animals, and sampling design and cost considerations. A simulated sampling approach to comparing sampling designs with respect to their precision and cost is presented. Comments are also made about the application of active surveillance for livestock disease monitoring in Australia.

Keywords: Animal Health, Foot and Mouth disease, Thailand

JEL Codes:Q160

Sampling Considerations for Active Surveillance of Livestock Diseases in Developing Countries

1. Introduction

Livestock diseases are a major cause of lost production and lost trade in developing countries. Animal health improvement programs require sound information on levels of protection against particular diseases, and the movements over time in these levels of protection in response to vaccination and other measures. This information can be obtained from normal reporting of disease cases by livestock owners and district veterinary officers, or by surveys specifically designed to obtain information on the protection status of a random sample of animals. Such surveys are expensive, and designs are required which provide information at the required accuracy level at minimum cost.

In Thailand, foot-and-mouth disease (FMD) is a virus disease of economic significance in cattle, buffalo and pigs, and a national vaccination program has been introduced with the ultimate aim of eradication of this disease. FMD reduces incomes of beef and dairy producers, is an impediment to exports of pigmeats. "Passive surveillance" has been unsuccessful in that while FMD is a notifiable disease, livestock owners often fail to report cases. A project supported by the Australian Centre for International Agricultural Research (ACIAR)[1] is providing assistance in developing a system of "active surveillance" to monitor levels protection of cattle and buffalo, involving taking periodic blood specimens of a sample of livestock and testing for FMD immunity.

The current project is concerned with development of an information system for animal health in Thailand. A trial in active surveillance is being conducted over three provinces in north-west Thailand, viz. Lampang, Lamphun and Chiang Mai. It is envisaged that sampling will be extended to other provinces, and to neighbouring countries. A two-stage sampling procedure is being adopted, with villages being selected with probability proportional to size in the first stage and a constant number of livestock in selected villages being selected randomly in the second stage. Blood titres for FMD above a critical level (positive diagnoses)

[1] ACIAR Project No. 9204, titled "Animal Health in Thailand and Australia: Improved Methods of Diagnosis, Epidemiology, Economics and Information Management". This and previous ACIAR projects have assisted in developing diagnosis facilities for FMD and other diseases at the Northern Veterinary Diagnostic and Research Centre at Hang Chat, Lampang Province.

indicate animals are protected from the disease, due to one or both of exposure to the disease and vaccination.

National FMD eradication is expected to be a long-term program, with strategies varying over time and differing between regions. Eradication in southern provinces is likely to be achieved earlier than in north-west (the project area) and the north-east, due to long land borders with neighbouring countries where FMD is endemic. Vaccination is central to early stages of the eradication program, with "stamping out" only adopted once a low incidence of outbreaks has been achieved.

Progress towards eradication is only possible if a sufficiently high level of protection of animals is achieved, thought to be of the order of 70% to 80%. Continuous monitoring is needed to ensure the effectiveness of vaccination. If disease incidence on a national or regional basis is to be monitored over a number of years, repeated large-scale surveys involving substantial field work will be required. Deploying veterinary teams to villages, mustering and bleeding stock, and testing blood specimens in the laboratory involve considerable expense. Hence a sampling design is needed which is both cost effective and within budget, and estimates of the proportion of animals protected to within accuracy targets.

This paper reviews statistical considerations in multistage sampling designs for active surveillance of livestock diseases, with particular emphasis on FMD in Thailand. Issues addressed include statistical efficiency concepts, estimator formulae for proportion of protected animals, and sampling design and cost considerations. A simulated sampling approach to comparing sampling designs with respect to their precision and cost is presented. Comments are also made about the application of active surveillance for livestock disease monitoring in Australia.

2. Sampling Designs for Active Surveillance

While simple random sampling (SRS) is the easiest form of sampling design to comprehend and to implement, it is in general not suitable for monitoring human or animal diseases on a regional or national basis. Two important limitations are that a sampling frame of the whole population generally is not available, and that this design would impose an inordinate cost of travel.

Stratified and multistage sampling designs offer potential for more cost-effective monitoring of animal health. Stratification by region, province or district could be important in assessing the spatial nature of protection levels. Multistage sampling involves sampling primary sampling units (PSU) or clusters, then drawing samples from within these primary sampling units[2]. Relative to simple random sampling, this design leads to some loss in statistical efficiency for a given sample size, but can greatly reduce cost where sampling frames are limited and travel costs are high.

A two-stage design, involving approximately 30 villages and 15 livestock from each, and six-monthly visits, has been trialed for FMD monitoring in the three Thai provinces. Villages have been selected with probability proportional to village cattle numbers. The fixed number of livestock per village is efficient in terms of the time field team members need to spend in a village to collect blood specimens. That is, PPS sampling has advantages with respect to convenience (hence lower cost per sample member) and hence to precision for a given survey budget. Also, in general it leads to greater statistical efficiency than placing an equal probability of selection on each village.

Need for information on a regional basis may in the future dictate approximately equal representation from each region, and hence stratification by region prior to selection of villages and livestock, i.e. use of stratified two-stage sampling.

3. Efficiency of PPS Sampling

Developing multistage sampling designs requires an appreciation of the concepts of statistical *efficiency* and *cost of sampling*.

Definition of statistical efficiency

Sampling designs can be compared on the basis of their expected precision of estimators for a given sample size. Combined with economic efficiency (or cost-effectiveness) of the sample design, this provides the basis for choosing between alternative sampling designs. An estimator is efficient if it has a low variance, i.e. if the *sampling distribution* of that estimator is relatively narrow and peaked. Efficiency is a relative concept, by which alternative sampling designs can be compared. Two estimators frequently used are the mean and the

[2] The terms “multistage sampling” and “cluster sampling” are sometimes used interchangeably, while sometimes the latter term is used only when all members of selected PSU are included in the sample

proportion. The sample mean $\bar{x}$ is an estimator of the population mean μ and the sample proportion p is an estimator of the population proportion π of a random variable. An example of the latter is the proportion of the overall cattle population in a particular region which have high blood titres with respect to FMD.

The measures of efficiency for estimators can be expressed in terms of variances of sampling distributions, e.g. $V(\bar{X})$ and *V(P)*, where $\bar{X}$ and *P* are random variables representing the sampling distributions of the mean and proportion respectively, from repeated samples of fixed size from a given population. These measures may be defined in terms of the population parameters, given the sampling design. They are theoretical measures, in the sense that population parameters (means, variances, number of members) normally are not known, hence estimates must be made on the basis of sample statistics.

If variances of the sampling distributions are small, reflected in small sample standard errors, confidence intervals (e.g. the 95% confidence interval for the population proportion) will be narrow, so the estimator will be precise or have high information value. If the confidence interval is wide, the survey results may be of little assistance for decision support.

Formulae for variance of the sampling distribution of means for multistage sampling are quite complex. Fortunately, they are considerably simplified for probability proportional to size (PPS) sampling.

Definitions of symbols

Consider a livestock population which can be divided into *M* primary sampling units (PSU). The subscript *i* will be used to represent an individual PSUs. Each PSU has N_i members, for a total population of *N* members ($N = \Sigma N_i$). Further, *m* of the PSUs are to be selected with probability proportional to size, i.e. the probability of selecting each is $p_i = N_{ii}/N$. A constant number ($\bar{n}$) of secondary or elementary sampling units (SSU) will be drawn from each selected PSU. The clusters or primary sampling units will correspond to villages, and the SSUs to individual animals. The objective will be to estimate π, the *proportion* in the population having the particular characteristic, e.g. high blood titre against FMD.

Sampling theory is normally developed with respect to the mean. Hence efficiency formulae will be considered first for the sample mean, and then for the sample proportion. Greek symbols will be used for population *parameters* (mean μ, proportion π and variance σ^2).

Upper case letters will be used for *population numbers* and lower case for *sample numbers* (e.g. N_i and n_i). Following Yamane (1967), a single subscript will be used for cluster *totals* (x_i).

Efficiency formulae (based on population parameters)

Suppose two-stage sampling is carried out, with PSU *selected with replacement* and SSU selected *without replacement.*[3] The variance of the sampling distribution of sample means (Yamane, 1967, p. 254, Levy and Lemeshow, 1991, p. 267) is

$$V(\hat{\mu}) = \frac{1}{mN}\sum_i^M N_i\,(\mu_i - \mu)^2 + \frac{1}{mN}\sum_i^M \frac{N_i - \bar{n}}{N_i - 1}\ \sum_j^{N_i} \frac{(x_{ij} - \mu_1)^2}{\bar{n}}$$

The first of these terms represents variance of the sampling distribution of means between PSUs or clusters and the second represents variance within PSUs.

The sample proportion may be viewed as a special case of the mean, i.e. the mean of a binary or 0-1 variable. Hence, efficiency formulae for a proportion may be developed from those for the mean. The population variance for the proportion under two-stage sampling (from Yamane, p. 278) is

$$V(\hat{P}) = \frac{M - m}{M(M-1)\bar{N}^2}\ \frac{\sum_i^M (N_i\pi_i - \overline{N\pi})^2}{m} + \frac{M}{mN^2}\sum_i^M \frac{N_i^2(N_i - n_i)}{N_i - 1}\ \frac{\pi_i(1-\pi_i)}{n_i}$$

where $\bar{N}$ is the average cluster size, n_i is the sample size from the *ith* PSU, π_i and $(1 - \pi_i)$ are the proportions of protected and non-protected members in the *ith* PSU and *P* is the overall proportion protected.

Efficiency comparisons

The efficiency of two-stage sampling and simple random sampling may be compared using the expression (Yamane, 1967, p. 227; Murthy, 1967, p. 327)

[3] At both stages of sampling, selected members could be excluded from re-selection or alternatively allowed to be selected two or more times. Development of variance formulae is simpler if sampling is conducted with replacement, because probability of selection does not vary as sampling proceeds. However, sampling without replacement tends to be more practical in that it is usually undesirable to have individual population members repeated in a sample. In two-stage sampling, the favoured practice would appear to be to select the PSUs with replacement (WR) and the SSUs without replacement (WOR), and this is the basis on which formulae are developed here.

$$V(\bar{X}) = V(\overline{X_{srs}})[1 + (\bar{n} - 1)\rho] = \frac{\sigma^2}{m\bar{n}}\,[1 + (\bar{n} - 1)\rho]$$

where $V(\bar{X}_{srs})$ is the efficiency measure for simple random sample, and *mn* is the total sample size. Here ρ is the *intracluster correlation coefficient*, or measure of homogeneity within clusters relative to the population mean. Members within a PSU tend to be more uniform or homogeneous than members of the overall population. That is, diseases tend to cluster or form pockets, and as well vaccination coverage tends to be more complete in some areas than in others. Where such homogeneity within PSUs exists, ρ will tend to be positive and the sampling error can be expected to be greater than that which would be obtained under simple random sampling. The design effect approach uses the term in square brackets as an inflation factor to increase the sample size relative to that which would be needed with a simple random sample.

Statistical estimator formulae (and precision estimates based on sample data)

The sample mean

While efficiency formulae are needed in considerations of sample design and size, once a survey has been conducted the task becomes one of estimating population parameters based on sample data. Both point estimates (especially the mean and proportion) and confidence intervals may be derived. The unbiased estimator of the population mean under two-stage sampling (Yamane, 1967, p. 218) is

$$\hat{\mu} = \frac{1}{N}\,\frac{M}{m}\sum_i^m \frac{N_i}{n_i}\,\sum_j^{n_i} x_{ij}$$

Provided each population member has an equal probability of selection in the sample, the sample mean is an unbiased estimate of the population mean. This is the case under PPS sampling, which is a *self-weighting* design in that the fixed number of SSU compensates for the unequal probabilities attached to the PSU, so:

$\hat{\mu}_{pps} = \frac{1}{mn}\sum^m\sum^n \bar{x}_{ij}$ (Yamane, 1969, p. 255)

The estimated variance of the mean, and the standard error

The estimate of the variance of the sampling distribution of means may be obtained by substituting sample estimates into the above efficiency formula, and summing over the

sample numbers. However, a simpler formula (Yamane, 1967, p. 255; Cochran, 1977, p. 309; Levy and Lemeshow, 1991, p. 268), is

$$\hat{V}(\bar{X}_{pps}) = \frac{1}{m}\frac{1}{m-1}\sum^{m}(\bar{x_i} - \bar{x}_{pps})^2 \ , \qquad \bar{x_i} = x_i/\bar{n}$$

The standard error is the square root of this value. An interesting feature of this formula is that "the estimator depends only on the variation between ultimate clusters and not on the variation within an ultimate cluster" (Yamane, 1967, p. 255). The size of the samples from within PSUs does not enter directly into the formula, although the distribution of PSU means about the overall population mean will depend on *n*.

The estimator of the proportion

Under two-stage sampling, in which $\bar{n}$ items are selected at random from each of *m* PSU, the unbiased estimator of the population proportion π is

$$\hat{\pi}\ \frac{M}{Nm}\ \sum^{m} N_i p_i \text{ , where } p_i = \sum^{n_i} x_{ij}/n_i \text{ (Yamane, p. 274)}$$

Here p_i is the sample proportion with the characteristic in the *ith* PSU, $N_i p_i$ is an estimate of the population number in the *ith* PSU, and these are summed to obtain an estimate of the number in the *m* selected PSUs. The factor *M/m* inflates this to an estimate of the total number in the *M* population PSUs, and dividing by *N* converts this to an estimate of the average (i.e. the proportion with the characteristic). Under PPS sampling, this formula simplifies to

$$\hat{\pi} = \frac{1}{m\bar{n}}\sum^{m}\sum^{n}\bar{x}_{ij} = \frac{1}{m}\sum p_i$$

where x_{if} (the value of the *jth* observation of the *ith* PSU) takes a value of 1 or 0, depending on whether a condition or attribute is present or absent, respectively (e.g. x_{ij} = 1 for a blood titre indicating protection against FMD).

The estimated variance for the population proportion under two-stage sampling is

$$V(\hat{\pi}) = \frac{1}{N^2}\left[M^2\,\frac{M-m}{M}\,\frac{1}{m}\,\frac{1}{m-1}\sum^{m}\left(N_i p_i - \frac{1}{m}\sum_i^m N_i\,p_i\right)^2 + \frac{M}{m}\sum_i^m N_i^2\,\frac{N_i - n_i}{N_i}\,\frac{1}{n_i}\,\frac{1}{n_i-1}\,n_i p_i q_i\right]$$ (Yanane, p. 281)

where p_i and q_i are sample proportions.

Under two-stage PPS sampling, where a constant number of elementary sampling units is selected from each PSU, the n_i become

$$n_i = n_2 = \cdots = \bar{n}$$

Under certain assumptions, the variance may be estimated by the simplified formula:

$$\hat{V}(P_{pps}) = \frac{1}{m}\,\frac{1}{m-1}\sum^{m}(p_i - \hat{\pi})^2$$ (Yanane, p. 288)

An interesting feature of this formula is that variance depends only on the variation between clusters, and not on the variation within clusters.

The classic $\{\boldsymbol{m}, \bar{\boldsymbol{n}}\} = \{\mathbf{30}, \mathbf{7}\}$ design

A standard design employed for a variety of World Health Organisation (WHO) surveys of immunisation coverage involves selection of 30 villages and 7 people from within each village (Henderson and Sundaresan, 1982; Lemeshow and Robinson, 1985; Fredrichs, 1989). The logic underlying this design is that a target precision in estimation of a population proportion of plus or minus 10% is sought. Under simple random sampling, and assuming the worst case in terms of sample size (where π = 0.5), the half-width of the 95% confidence interval would be

$$z_{0.025}\,S_P \leq 0.1$$

$$\text{i.e. } 1.95\sqrt{p(1-p)/n} \leq 0.1$$

$$\text{or } n \geq (1.06/0.1)(0.5) = 98$$

Using a *design factor* of 2.0 (i.e. assuming that two-stage sampling is half the efficient as simple random sampling), a sample size of at least about 200 is indicated. The standard

design of $m\bar{n} = (30)(7) = 210$ approximately meets this requirement, and hence provides a "ballpark" figure for sample size.

Although this design has been widely used, Lemeshow and Robinson (1985) note that the "30 clusters is based more on tradition and intuition than on statistical theory" and caution against its mechanical application without regard to any specific population. Lemeshow and Robinson (1985) note that much larger samples are needed to obtain reliable estimates with respect to rare diseases (where $\pi << 0.5$). In practice, there is often some loss of randomness in the second stage of sampling, due to difficulties in selecting households and individuals in developing countries, and this too could dictate need of a larger sample.

Where information is needed on protection levels over a number of regions, samples within each of those regions would need to be sufficiently large to achieve a target precision level, e.g. not less than about 200 beasts per region. Also, where the objective is to detect changes over time in protection levels, the relevant sampling distribution is that of differences between sample proportions, and the improvement between years may be modest, again suggesting a larger sample could be needed to achieve accuracy requirements.

4. Optimal Sample Design and Size

A number of approaches may be employed to determine a recommended sampling design. These will be discussed in terms of two-stage sampling of livestock for protection against disease, where first stage is PPS sampling with replacement, and second stage sampling is without replacement. These approaches include:

- determining the sampling design which will provide the most precise information for a given budget.
- obtaining information with a target precision level at minimum cost. The level of precision for a chosen variable such as a mean or proportion can be defined according to a number of criteria, e.g. half width of the confidence interval, or standard deviation of an estimator as a percentage of point estimate.
- tabular comparison of a range of sampling designs on the basis of efficiency of estimates and cost of sampling.
- maximizing the expected net gain from sampling in a static context, or maximizing the present value of benefits less costs over time from sampling.

The survey cost function

A widely used cost function for two-stage sampling is the following formula (Yamane, 1967, p. 264; Snedecor and Cochran, 1989, p. 449; Levy and Lemeshow, 1991, p. 262):

$$C = C_0 + C_1 m + C_2 m \bar{n}$$

where c_0 are overhead costs

c_1 are costs per PSU, and

c_2 are costs per SSU.

This is a linear function in that the cost per PSU is assumed constant over all PSU, and the cost per SSU is assumed constant for all SSU within each PSU. For practical purposes, the overhead costs are sometimes omitted in sample size calculations, since they do not depend on m and $\bar{n}$.

In setting up an active surveillance program, costs would be incurred for:

- laboratory facilities to test blood specimens
- training of technicians in laboratory testing
- purchase or hiring of vehicles
- acquiring equipment such as chillers, specimen bottles, syringes, ropes and protective clothing
- obtaining sampling frames and drawing samples
- reporting of survey results.

For each village included in the sample, there would be costs involved in

- vehicle expenses incurred in travelling to and from villages
- wages for staff collecting specimens
- accommodation and meals for field team
- payments or gratuities to villagers assisting with the survey.

It is probable that other activities would be carried out at the same time as the survey, such as vaccination for FMD or other diseases. This would affect the time required in each village.

No allowance is made here for the cost of associated veterinary activities associated with the active surveillance program.

For each animal from which a blood specimen is drawn, there would be costs comprising mostly

- staff wages
- laboratory testing costs (materials and technician time).

Decisions have to be made as concerning the extent to which costs are allocated to the sampling program as distinct from normal activities of veterinary staff and facilities. An example of the kind of question which arises is whether any of the overhead costs of vehicles should be costed against sampling. On the one hand, purchase of vehicles could be regarded as a sunk cost independent of the sampling activities. On the other hand, additional use would increase vehicle depreciation, other routine activities involving use of vehicles may be carried out in conjunction with the survey, and active surveillance may become part of the regular activities of the veterinary agency. Similar issues arise with respect to placing a cost on time devoted to surveys by regular veterinary staff, as distinct from staff employed specifically for surveys.

From an opportunity cost viewpoint, vehicles and staff not engaged on the sampling program could be allocated to other desirable activities by the agency, or hired out to the private sector. The view is taken here that these resources should be charged to the sampling program at their full opportunity cost.

Designs based on cost constraints

Using the linear cost function and estimates of the population variance parameters, the sampling design in terms of m and $\bar{n}$ can be found such that the sampling efficiency is maximized for a fixed budget C, using the Lagrangean multiplier technique. For two-stage sampling, this involves determining an optimal sample size for one of the stages, then substituting into the cost function to find the optimal size for the other sampling stage. Sample size formulae derived by this approach are presented in Appendix A.

Use of simulation to compare designs

An alternative to deriving formulae for an ‘optimal’ sampling design is to simulate sampling

from populations with specified parameters, using random number generation procedures on a computer. This simulation approach may be regarded as a fallback method when an analytical approach to determining the sample size cannot be obtained. But in practice, it will often prove more flexible and easy to apply than an analytical approach even when the latter is available.

For any nominated sampling design (equivalent to a treatment in a simulation experiment), a large number of synthetic samples can be generated (replicates in experimental design terminology), and the sampling efficiency determined in terms of the estimated variance of the overall sample proportion of protected animals. A sampling cost function can be included to indicate total cost of any sampling design. Once the simulation method has been set up, and programmed to a computer, this approach allows a large number of alternative sampling designs to be evaluated in terms of efficiency and cost. As well, the parameters relating to the livestock population and sampling cost can be varied readily, to allow a comprehensive sensitivity analysis.

This simulation approach has been used in relation to sampling designs for active surveillance of protection levels against foot-and-mouth disease of cattle and buffalo in Thailand. The design problem is viewed as sampling a livestock population where owners are spread over a large number of villages *(M)*, each with N_i animals ($i = 1$ to M). A sample of m villages is to be selected by PPS sampling with replacement, then a constant number $\bar{n}$ of animals is to be selected by sampling without replacement from each sample village.

A computer program in the Q-BASIC language has been developed to carry out simulated sampling. Parameters of this program, which may be adjusted readily, and the default values, are indicated in Table 1.

Table 1: Default parameter values for simulated sampling experiments

Population proportion with positive titres		0.5
Between-village variance of proportion		0.06
Average number of livestock per village		200 head
Overhead cost of active surveillance sampling, c_0		100000 bt
Vehicle cost per village sampled	700 bt	
Labour cost per village sampled	600 bt	
Cost of accommodation etc. per village sampled	500 bt	
Data collection cost per village or PSU, c_1		1800 bt
Data collection cost per blood specimen	20 bt	
Laboratory analysis cost per blood specimen	50 bt	
Data collection cost per SSU, C_2		70 bt

These estimates are based on information provided by Cameron (1996), but with some modifications. In particular, an overhead cost of sampling has been added, and the data collection cost per village has been increased to provide a greater allowance for vehicle and labour costs. For convenience, it has been assumed that the proportion of protected animals across villages follows a normal distribution.

The program consists of nested loops in which levels of the two factors have been varied in defining treatments for the simulation experiment. The number of villages is varied between 10 and 40 in steps of 10, and the number of livestock per selected village is varied between 5 and 30 in steps of 5. The proportion of protected animals by village is assumed to follow a normal distribution with parameters as in Table 1. Each treatment is replicated 250 times. When drawing samples for each village, sampling without replacement is achieved by reducing total numbers of protected and unprotected animals according to whether each sample observation is a protected animal ($x_{ij} = 1$) or an unprotected animal ($x_{ij} = 0$). An example of simulation output is provided as Table 2.

Table 2: Sampling efficiency and cost for various two-stage PPS designs, as obtained from simulation experiment

Sampling design		Sampling efficiency		Total cost
m	n-bar	Est. Var (P)	1.96 SE	(1000 bt)
10	5	0.0098	0.1941	122
10	10	0.0082	0.1774	125
10	15	0.0069	0.1627	129
10	20	0.0061	0.1533	132
10	25	0.0056	0.1465	136
20	5	0.0046	0.1332	143
20	10	0.0038	0.1203	150
20	15	0.0032	0.1115	157
20	20	0.0034	0.1144	164
20	25	0.0027	0.1017	171
30	5	0.0031	0.1085	165
30	10	0.0027	0.1012	175
30	15	0.0023	0.0936	186
30	20	0.0024	0.0968	196
30	25	0.002	0.0885	207
40	5	0.0026	0.1002	186
40	10	0.0017	0.0805	200
40	15	0.0015	0.0770	214
40	20	0.0017	0.0806	228
40	25	0.0016	0.0785	242

Suppose the accuracy requirement for active surveillance is that the error in estimation of the proportion of protected animals in the population be not more than 0.10 or 10%. Table 2 indicates that a sampling design which includes a PPS first-stage sample of 30 villages (with replacement) and a constant second-stage sample size of 10 to 15 animals (without replacement) per selected village would on average yield meet this accuracy requirement. This table further indicates that

- provided the number of villages is sufficiently large (20 to 30 or more), increasing the number of animals per selected village is unlikely to have much impact on sampling error.
- while accuracy of estimation improves strongly as the number of villages included in the sample is increased from 10 to 20, increasing the number of villages beyond 20 to 30 yields only small increases in sampling accuracy.
- within the range of sampling designs of interest, the overall survey cost depends primarily on the number of villages included in the sample.

- if the estimate is to be made to within 10 percentage points, the classic {30,7} design would go close to providing would go close to achieving this accuracy target.

The simulation model may be used to carry out sensitivity analysis with respect to a number of parameters. Some examples are illustrated below.

Increased variance of proportion between villages

If there is high variability between villages with respect to proportion of protected animals, then there is increased likelihood of obtaining extreme villages in the sample, and accuracy of estimation is reduced accordingly. This is illustrated in Table 3, for which the between-village variance has been doubled relative to that in Table 2, i.e. set at 0.12. It is apparent from this table that a larger sample - of about 40 villages and 10 animals per village - would be required to achieve a 10% accuracy requirement.

Higher population protection level

The worst case situation in terms of required sample size arises when the overall proportion of protected animals in the population is 50%. If the proportion is substantially higher, the sample size required to achieve a target level of precision will be reduced accordingly. This is illustrated in Table 4, which has been derived for a population protection level of 80%. In this case, a sample as small as 20 villages and 15 animals would meet the 10% error target.

Table 3: Sampling error for a range of two-stage PPS designs when between-village variance is doubled

Sampling design		Sampling efficiency		Total cost
m	n-bar	Est. Var (P)	1.96 SE	(1000 bt)
10	5	0.0125	0.2189	122
10	10	0.0113	0.2084	125
10	15	0.0107	0.2031	129
10	20	0.0095	0.1908	132
10	25	0.0082	0.1770	136
20	5	0.0063	0.1552	143
20	10	0.0052	0.1417	150
20	15	0.0049	0.1373	157
20	20	0.0049	0.1378	164
20	25	0.0042	0.1276	171
30	5	0.0039	0.1218	165
30	10	0.0036	0.1183	175
30	15	0.0033	0.1123	186
30	20	0.0037	0.1195	196
30	25	0.0031	0.1089	207
40	5	0.0033	0.1132	186
40	10	0.0023	0.0941	200
40	15	0.0023	0.0937	214
40	20	0.0026	0.0990	228
40	25	0.0025	0.0984	242

Non-constant sampling costs per village

In general, for a fixed overall sample size, the overall sampling cost will be lowest if the number of villages is small and the number of livestock sampled per village is large. However, the cost savings from reducing *m* and increasing $\bar{n}$ will be reduced if the sampling cost per village is a decreasing function with respect to number of villages, or an increasing function with respect to size of village samples. The first of these situations has been noted by Murthy (1967, p. 334), who divides the cost per PSU into two components, one a constant and the other depending on the number of SSUs in a selected PSU. The latter is made a function of the square root of *m* rather than of *m* itself, on the grounds that the more PSUs selected the less travelling between them will be needed.

Table 4: Sampling error for a range of two-stage PPS designs when overall protection level is 80%

Sampling design		Sampling efficiency		Total cost
m	n-bar	Est. Var (P)	1.96 SE	(1000 bt)
10	5	0.0066	0.1588	122
10	10	0.0062	0.1539	125
10	15	0.0055	0.1453	129
10	20	0.0048	0.1355	132
10	25	0.0044	0.1307	136
20	5	0.0038	0.1207	143
20	10	0.0027	0.1013	150
20	15	0.0026	0.0997	157
20	20	0.0024	0.0968	164
20	25	0.0020	0.0887	171
30	5	0.0023	0.0940	165
30	10	0.0019	0.0851	175
30	15	0.0017	0.0802	186
30	20	0.0017	0.0813	196
30	25	0.0015	0.0769	207
40	5	0.0017	0.0797	186
40	10	0.0013	0.0707	200
40	15	0.0011	0.0662	214
40	20	0.0012	0.0689	228
40	25	0.0011	0.0642	242

The cost per village assumed here is an average cost; villages which are more remote or have poorer road access may involve greater travel time and expense. For second-stage (village) samples of up to about 20 beasts, it would normally be possible in northern Thailand to obtain blood specimens from two villages per day. As village sample size increases, an increasing proportion of villages will require a full day for the veterinary team, leading to increased wage, accommodation, meals and transport costs per village. The increased time would arise not only because of the additional animals which must be bled, but also because negotiations with additional livestock owners and additional mustering would be involved.

The effect of an increasing cost per village as the number of animals is increased would be to favour a greater number of villages (PSUs) and a smaller number of animals per village. It would appear that for village samples larger than about size 20, it becomes difficult to complete all villages at the rate of two per day. It is probable that for samples of more than

about 50 animals, normally only one village could be sampled per day. This information can be represented in the following village cost function:

$$c_2 = \begin{cases} 1800, & \bar{n} \leq 20 \\ 1800 + 1800\,(\bar{n} - 20)/50, & 20 < \bar{n} \leq 50 \end{cases}$$

This modification to the cost formula has been included in the simulation model. However, sample sizes found to satisfy the accuracy requirement here involve 20 or fewer animals per village, so that the additional cost element should not need to be incurred.

Snapshot versus comparative estimates of protection level

If a survey is designed to estimate the level of protection at a given point in time, this level can be derived as a point or interval estimate. The interval estimate is obtained as

$$p - z\,S_P < \pi < p + z\,S_P$$

where π is the unknown the population proportion of protected animals

p is the proportion protected in the sample (the point estimate)

z is the standard normal variate (asymptotic to 1.96 as sample size increases)

S_P is the standard error as estimated from the sample.

As discussed in relation to the classic [30,7} design, under simple random sampling the proportion protected is estimated to within 0.10 or 10% if $S_P \leq 0.05$.

As distinct from this snapshot case, the interest may be in monitoring the progress of vaccination in terms of changes in the protection level over time. The change in proportion protected between two successive time periods may be derived as a point or interval estimate. The latter is given by:

$$p_1 - p_2 - z\,S_{p_1 - P_2} < \pi_1 - \pi_2 < p_1 - p_2 - z\,S_{p_1 - P_2}$$

where $\pi_1 - \pi_2$ is the unknown difference between population proportions protected in the two time periods

$p_1 - p_2$ is the difference between proportions protected as observed in the two

samples (the point estimate of the improvement or deterioration in protection level)

z is the standard normal variate

$S_{p_1 - P_2}$ is the standard error as estimated from the two samples.

If independent samples of villages are drawn each year, the variance of the difference between sample proportions is the sum of the variances for the individual samples

$$Var\,(P_1 - P_2) = Var\,(P_1) + Var\,(P_2)$$

The sample estimators are also additive, and the standard error which is the square root of this sum, will also be increased relative to the snapshot estimate for a single village. That is, a larger sample will be required to estimate the difference in protection level between years than is required to estimate the protection level in a given year, when both are to be estimated at the same level of precision. If the *same villages* are selected in each year, then this will increase precision with which the change in protection level is estimated, and partially eliminate the need for a larger sample.

Practical considerations in selection of villages and livestock

A reliable sampling frame of villages is available for the three provinces in the current trials. However, village sampling frames may not be available for other provinces or for applying active surveillance to neighbouring countries. This limitation may necessitate some form of area or grid sampling. A difficulty arises in that density of villages varies greatly between regions, and selecting villages nearest to randomly selected points on a uniform grid would lead to over-representation of small isolated villages. Over-representation of isolated villages could lead to downward bias in estimates of protection levels, and higher than average costs per sample member. A modification to grid sampling would be to divide the target area up into say high, moderate and low density areas in terms of livestock numbers, and vary the sampling fractions to each.

Selection of livestock within villages has given rise to a number of problems because some livestock owners do not present stock for testing (some fear testing may have adverse impacts such as abortion, some work fulltime outside their village), and it is difficult to avoid selecting the more prominent or co-operative livestock owners.

Further development of the simulation model as a decision-support package

This simulation approach to evaluation of two-stage sampling designs would lend itself to further development as a decision-support package for designing surveys for active surveillance into animal health or disease protection status. Enhancements to the method could involve

- setting up a computer package with a "user friendly" data input interface, in which default values are presented, but movement between them is possible by mouse or cursor so as to enter application-specific values.
- division of survey cost variables into a wider range of classes, e.g. several categories of overhead costs, and further division of costs per PSU and SSU.
- improved output format for sensitivity analysis, including graphical presentation.
- addition of an option to include economic returns from sampling and maximization of the expected net gain from sampling.

5. Other Approaches to Choice of Sampling Design

Acceptance sampling

Another approach which has been suggested for monitoring population attributes with respect to target levels (e.g. effective protection against FMD) is acceptance sampling (Harrison and Tamaschke, 1994, Ch. 18). This technique would approach vaccination effectiveness from an hypothesis testing rather than confidence interval viewpoint. Once a sample of blood specimens is collected and tested in the laboratory, the number or proportion of non-protected specimens (*defectives*) would be compared against an *operating characteristics curve* for a given protection level in the population or *lot*. A statistical decision would then be made as to whether the population was of an acceptable quality level. Acceptance sampling provides an alternative way of examining sample data, which is useful when the objective is to determine whether a required minimum level of protection in a population has been attained.

Bayesian decision theory

Bayesian statistical decision theory provides a mechanism to combine existing and perhaps subjective knowledge about disease incidence with information obtained from active surveillance. Also, it provides a mechanism for combining statistical and economic

information to derive more meaningful criteria for assisting decision-makers in choosing sampling designs.

The Bayesian approach (Harrison and Tamaschke, 1984) allows *prior* information about the values of estimators (such as mean and proportion) to be combined with *sample* information derived from active surveillance sampling) to obtain *posterior* probabilities for random variables. This technique is used for revision of probabilities in the light of new information, and could be applied to estimates obtained in previous surveys or derived subjectively. By comparing the expected value of sample information (EVSI) with the cost of sampling (COS), this technique could indicate the expected net gain from sampling (ENGS) for various sample sizes, and hence shed light on the optimal sample size.

The major limitation of the Bayesian approach is the difficulty in obtaining estimates of the expected value of sample information which are needed to estimate the expected net gain from sampling. The analysis goes beyond statistical considerations into economic estimation. Returns from information will depend on how the information is used, the consequent improvement in animal health, and the improved trade access and changes in producer, consumer and trader economic surpluses.

The Bayesian approach outlined above has been considered in a static or one-period context. An extension would be to estimate the costs and returns from an information system such as active surveillance in a multiperiod framework. Discounted cash flow analysis could then be used to derive performance criteria such as net present value and internal rate of return from alternative sampling designs.

6. Concluding Comments

Sample surveys are an expensive but potentially powerful means of obtaining up-to-date information about animal health status. Statistical theory for multi-stage designs is highly complex. If estimates of population parameters concerning protection levels and of sampling cost components are available, it is possible to estimate sample sizes required to achieve accuracy targets. Further, simulated sampling provides rapid and effective method to compare the statistical efficiency and cost of a range of sampling designs, and to carry out sensitivity analysis with respect to parameters for which values are uncertain.

7. Acknowledgements

The financial assistance for this research under ACIAR Project No. 9204 is gratefully acknowledged. I wish to thank Mr Angus Cameron, Australian Volunteer Abroad working on this project for stimulating interest in sampling designs, and Mr Ranjit de Mel of the Queensland Government Statistician's Office for expert advice on sampling design theory.

8. References

Cameron, A. (1996), personal communication.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd edn., Wiley, New York.

Fredrichs, R.R. (1989), 'Simple Analytic Procedures for Rapid Microcomputer-Assisted Cluster Surveys in Developing Countries', *Public Health Reports*, 104: 24-35.

Harrison, S.R. and Tamaschke, H.U. (1984), *Applied StatisticalAnalysis*, Prentice Hall, Melbourne.

Harrison, S.R. and Tamaschke, R.H.U. (1993), *Statistics for Business, Economics and Management*, Prentice-Hall, Sydney.

Henderson, R.H. and Sundaresan, T. (1982), 'Cluster Sampling to Assess Immunization Coverage: A Review of Experience with a Simplified Sampling Method', *Bulletin of the World Health Organization*, 60: 253-260.

Lemeshow, S. and Robinson, D. (1985), 'Surveys to Measure Programme Coverage and Impact: A Review of the Methodology Used by the Expanded Programme on Immunization', *World Health Statistics Quarterly*, 38: 65-75.

Levy, P.S. and Lemeshow, S. (1991), *Sampling of Populations: Methods and Applications*, Wiley, New York.

Murthy, M.N. (1967), *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

Snedecor, G.W. and Cochran, W.G. (1989), *Statistical Methods*, 8th edn., Iowa State University Press, Ames.

Yamane, T. (1967), *Elementary Sampling Theory*, Prentice-Hall, Englewood Cliffs.

Appendix A: Optimal Sample Size as Yielded by the Lagrangean Multiplier Technique

Under this approach, the optimal two-stage sampling design can be obtained in terms of the values of m and $\bar{n}$ for which the variance of the sample proportion is minimized. This is determined for a given survey budget and given parameters in the sampling cost function. Optimisation is achieved by use of the Lagrangean multiplier technique. The formulae presented here follow Yamane (1967, p. 264).

These sample size formulae are based on the formula for the variance of the sampling distribution of proportions:

$$\widehat{V(\mu)} = \frac{1}{mN}\sum^{M} N_i(\mu_i - \mu)^2 + \frac{1}{mN}\sum^{M}(N_i - \bar{n})S_i^2/n$$

$$\text{where } S_i^2 = \sum\left(x_{ij} - \bar{x}_j\right)^2/(n_i - 1)$$

and the two-parameter cost model:

$$C = c_1 m + c_2\bar{n}$$

The formulae reported here are framed in terms of the population *mean*. However, they apply equally to the population *proportion*, which can be viewed as the mean of a particular type of population, i.e. on consisting only of binary or 0-1 variables.

When the second-stage samples are only a small fraction of the clusters (i.e. $\bar{n}$ << N_i), the second-stage sample size is obtained as (Yamane, 1967, pp. 266-268):

$$\bar{n} = \sqrt{\frac{c_1 S_{wti}^2}{c_2 S_{wtb}^2 - S_1^2/\bar{N}}}$$

$$\text{where } S_i^2 = \sum^{n_i}\left(x_{ij} - x_i\right)^2/(n_i - 1)$$

$$\text{where } S_{wtb}^2 = \sum^{m}\left(\bar{x}_\iota - \overline{x_{pps}}\right)^2/(m - 1)$$

and $S_{wti}^2 = \sum^{m} n_i S_i^2 / n$

To find *m*, the cost function is rearranged to

$$m = \frac{C}{c_1 + c_2 \bar{n}}$$

where *C* is the fixed survey budget (less overhead costs), c_1 and c_2 are the known cost coefficients and $\bar{n}$ is the optimal second-stage sample size per PSU.

ANIMAL HEALTH ECONOMICS

WORKING PAPERS IN THIS SERIES

1. Assessing Public Programmes for Control of Animal Diseases in Developing Countries: General Economic Issues with Thai Examples – An extended ISVEE Paper by C. A. Tisdell, S. Harrison and F. C. Baldock, August 2004.
2. Animal Health and the Control of Diseases: Economic Issues with Particular Reference to a Developing Country by C. A. Tisdell, September 1994.
3. Assessing the Approach to Cost-Benefit Analysis of Controlling Livestock Diseases of McInerney and Others by C. A. Tisdell, May 1995.
4. The Economics of Vaccinating or Dosing Cattle Against Disease: A Simple Linear Cost Benefit Model with Modifications by C. A. Tisdell and G. Ramsay, May 1995.
5. Livestock, The Environment and Sustainable Development with Illustrations of Issues from Thailand by C. A. Tisdell and S. Harrison, June 1995.
6. The Use of Serology to Produce Disease Information for the Economics Analysis of Disease in Extensively Grazed Cattle by G. C. Ramsay, R. J. Dalgliesh, F. C. Baldock and C. A. Tisdell, September 1995.
7. The Role of Animal Health Programmes in Economic Development by S. Harrison and C. A. Tisdell, October 1995.
8. Trends in the Thai Livestock Industry, Animal Health Implications and Thailand's Development: An Introduction by T. Murphy and C. A. Tisdell, October 1995.
9. Specific Livestock Industries, Livestock Diseases and Policies in Thailand: An Overview of Bovines (Buffalo/Cattle) by T. Murphy and C. A. Tisdell, October 1995.
10. Overview of Pigs and Poultry: Specific Livestock Industries, Livestock Diseases and Policies in Thailand by T. Murphy and C. A. Tisdell, December 1995.
11. Livestock and Livestock Health in Northern Thailand: A Socio-Economic Analysis of a Cross-Sectional Survey of Villages by T. Murphy and C. A. Tisdell, March 1996.
12. A Review and Extension of Economic Pest Control Model Incorporating Multi-Pest Species and Insect Resistance by R. Davis, April 1996.
13. Foot and Mouth Disease: An Overview of its Global Status, Control Policies and Thai Case by T. Murphy, August 1996.
14. An Overview of the Status of Cattle Tick *Boophilus microplus* in Queensland by R. Davis, August 1996.
15. A Review of the Economics of Controlling Diseases in Livestock and the Modelling of Control Policies by T. Murphy, August 1996.
16. Optimal Livestock Disease Control Models and Their Possible Application to Thailand by T. Murphy, August 1996.
17. An Overview of Trends in Development in the Thai Dairy Industry by T. Murphy and C. Tisdell, September 1996.
18. Cost-Benefit Analysis with Applications to Animal Health Programmes: Basics of CBA by S. Harrison, September 1996.
19. Cost-Benefit Analysis with Applications to Animal Health Programmes: Complexities of CBA by S. Harrison, September 1996.
20. Cost-Benefit Analysis with Applications to Animal Health Programmes: Spreadsheet Implementation of Discounted Cash Flow and Risk Analysis by S. R. Harrison, September 1996.
21. Cost-Benefit Analysis with Applications to Animal Health Programmes: Allowing for Project Risk in CBA in S. R. Harrison, October 1996.

22. Cost-Benefit Analysis with Applications to Animal health Programmes: Valuation of Non-Market Costs and Benefits by S. R. Harrison, October 1996.
23. Cost-Benefit Analysis with Applications to Animal Health Programmes: Animal Health Programmes and Information Systems by S. R. Harrison, October 1996.
24. Women and Common Property Resources in the Management and Health of Livestock in Thai Villages by T. Kehren and C. A. Tisdell, November 1996.
25. Animal Health Information Systems by G. Ramsay, November 1996.
26. Collecting Animal Health Data for Cattle Properties in Extensive Grazing System by G. Ramsay, November 1996.
27. Sampling Considerations for Active Surveillance of Livestock Diseases in Developing Countries, November 1996.
28. On the Economics of Maintaining the Health of Livestock with Thai Examples by C. A. Tisdell, November 1996.
29. Economics of Investing in the Health of Livestock: New Insights? by Clem Tisdell, November 1996.
30. Macroeconomic Effects of Disease Control in the Thailand Livestock Sector – A CGE Analysis by T. Purcell, N. Karunaratne and C. Tisdell, January 1997.
31. The Thai Dairy Industry: Its Economic Evolution and Problems Raised by Land Rights and Cattle Diseases by T. Kehren and C. A. Tisdell, February 1997.
32. A Review of Economic Evaluations of Government Policies for the Control of Cattle Tick by R. Davis, February 1997.
33. Modelling to Predict Disease Incidence and Severity Using Age Specific Seroprevalence Data by G. Ramsay, March 1997.
34. Assessing the Effect of Vaccination on Disease Incidence and Severity by G. Ramsay, March 1997.
35. Calculating the Production Loss A Voided by Disease Control by G. Ramsay, March 1997.
36. Discounted Cash Flow Analysis of Diseases Control Programmes by G. Ramsay, C. A. Tisdell and S. R. Harrison, April 1997.
37. Private Decisions in Livestock Disease Control and the Value of Additional Information About Animal Health by G. Ramsay, S. R. Harrison and C. A. Tisdell, April 1997.
38. The Distribution of Benefits from Improved Animal Health Decision Making as a Result of the Collection of Additional Animal Health Information by G. Ramsay, C. A. Tisdell and S. R. Harrison, April 1997.
39. An Overview of the Occurrence of FMD in Thailand and Policies for its Control, by T. Kehren and C. A. Tisdell, September 1997.
40. Socio-Economic Status and Animal Disease Prevention and Treatment by P. Thani, T. Aphichatsarangkoon, B. Aekasing, L. Worachai, November 1997.
41. Production and Trade of Cattle, Buffalo and Pigs in Thailand by A. Smith and S. R. Harrison, December 1997.
42. Termination Report, Improved methods in diagnosis, epidemiology, economics and information management in Australia and Thailand by P. Black, December 1997.