

# Part II

## Survey Design and Analysis

### General outline of survey activities

This part of the book provides specific instructions for carrying out three different types of surveys: prevalence surveys; incidence rate surveys; and surveys to demonstrate freedom from disease. The design of each of the survey types presented is very different, but all livestock disease surveys share some features in common. Some of the activities that need to be carried out for every survey include:

Survey procedures

- Step 1:** Determine what question is being asked and how best to answer it.
- Step 2:** Identify the target population.
- Step 3:** Choose the right survey design.
- Step 4:** Decide if the survey is to use stratification.
- Step 5:** Calculate the best sample size.
- Step 6:** Plan field activities.
- Step 7:** Train survey teams.
- Step 8:** Conduct a pilot survey.
- Step 9:** Select the sample.
- Step 10:** Carry out field work.
- Step 11:** Collect information (from livestock owners and/or animals)
- Step 12:** Process specimens ready for analysis.
- Step 13:** Send the specimens to the laboratory.
- Step 14:** Check the data for completeness and accuracy.
- Step 15:** Enter the survey data and laboratory results into a computer.
- Step 16:** Check the data for mistakes during data entry.
- Step 17:** Analyse the data to calculate estimates.
- Step 18:** Report the data.

The following sections discuss several of these steps. The other steps are dealt with elsewhere.

- Step 1: Determine what question is being asked
- Step 2: Identify the target population.
- Step 3: Choose the right survey design.
- Step 5: Calculate the best sample size.
- Step 6: Plan field activities.
- Step 7: Train survey teams.
- Step 8: Conduct a pilot survey.
- Step 17: Analyse the data to calculate estimates.
- Step 18: Report the data.

## Step 1: The question

Every livestock disease survey aims to answer a question about the animal population. Developing and refining the question, and working out how to best answer it is an important first step in running a disease survey.

Usually, the question is first asked in very general terms. For instance: “Is the haemorrhagic septicaemia vaccination program working?”, or “How much hog cholera is there in the north of the country?”, or “Is there any bovine spongiform encephalopathy present in the country?”.

While these questions are a useful starting point, it is not immediately clear as to how to answer them. They need to be refined so that we can measure some quantity which will answer the question. For instance, the second question could be refined to specify how we can determine if the vaccination program is working. If the program is working, then the proportion of animals in the population with protective antibody titres should be higher than some target level. The new question is “What is the prevalence of pigs in the population with protective antibodies to hog cholera?” Once measured this value can be compared with some target level to assess success, and also compared with the level measured at different times in the past, to assess progress.

The process of refining a question and determining a measurable quantity that will answer the question may take several steps, and for most questions there are several different ways to answer them.

## Step 2: Population

Target population

In livestock disease surveys there are several different populations that we are interested in. The *target population* is the population that we aim to answer the question about. If the question is “Is there any rinderpest in the country?” then the target population is all cloven-hoofed animals in the country.

Source population

The *source population* is the population that was used to draw the sample from. If, to answer the previous question, a two-stage survey was conducting including all intensively farmed and village cattle and buffalo, then this would be the source population. Ideally, this is the same as the target population, but usually excludes some groups as they are impractical to sample. For instance, it may be impossible to capture and test wild buffalo with the resources available, so they are excluded from the source population.

It is important to remember that the results of the survey can be used to make inferences about the source population only, not the target population (if they are different). See page 20 for a discussion of inference.

### Step 3: Choosing the survey design

Usually, a well-considered and refined first question will point the way to what sort of study design is needed. When evaluating a vaccination program in pigs by assessing antibody levels, it is clear that a prevalence survey is necessary. When aiming to demonstrate that a particular disease is not present in a region or country, a survey to demonstrate freedom from disease is appropriate. Sometimes it is harder to decide whether to use an incidence rate or prevalence survey. See page 29 for a discussion of the role of the two types of surveys. An understanding of the procedures for carrying out the two survey types, explained in Chapters 7 (prevalence) and 8 (incidence rate), will also help in deciding which is most appropriate.

### Step 5: Sample size

Before starting a survey, you need to know how many units of interest (herds, villages or animals) need to be examined to answer the question. This is one of the more difficult parts of planning a survey. The larger the sample size, the more precise the results will be, giving you greater confidence in the answer. However, large surveys are more expensive and time consuming, so you need to compromise.

Each different survey design requires a different sort of sample size calculation and different information for that calculation. The methods for calculating sample sizes are explained in the following chapters. Some of the measures commonly used in sample size calculations are described below. The main factors which determine the sample size are: variance (amount of variation in the population), desired precision, and desired confidence level. Specific survey types may require other information for sample size calculation, as described in the following chapters.

#### Variance

Variance is a measure of how much variation there is in the population, or how much difference there is between individuals, herds or villages. Variance can be high, medium or low, and can be calculated based on earlier survey results. A population with a wide spread of values has a high variance, a population with a small spread has a low variance.

**Example:** Consider two different populations. Population 1 is all the cattle in a herd. Population 2 is all the pigs in the weaner shed of an intensive piggery. If we are interested in calculating the average age of the animals in the two populations, the spread of ages in the cattle herd (which includes calves, young and old animals) is much wider than the spread of ages in the weaner shed (in which all animals are within 2 months of the same age. When considering age, the cattle herd has a higher variance than the weaners.

Variance is an important factor when calculating sample size. In a population with a low variance, most animals are very similar. To estimate the average or a prevalence, selecting only a few animals will give a representative picture of what the overall value is. When the variance is high, more animals are needed, because each of the selected animals is likely to be quite different. Sample sizes need to be larger when the variance of the population is higher.

**Desired precision and confidence**

The precision of a survey indicates how good the estimate is (see page 24). This is usually measured by the width of the confidence interval (page 23). A very wide confidence interval suggests that we are not very sure about where the true value lies. On the other hand a narrow confidence interval indicates that we are pretty sure that the true value lies somewhere within a narrow range. This is a more precise estimate. The confidence level is a measure of exactly how sure we are that the value lies within the stated range or confidence interval. A 95% confidence level means that we are 95% sure that the real value is in the interval. If the study was repeated 20 times, then on average, we would be wrong once, but right 19 times.

By convention, we usually use a 95% confidence level, although 90%, 99% and 99.9% are sometimes used. When calculating sample size, greater precision (a narrower confidence interval) will require a larger sample size. The choice of level of precision (often expressed as half the width of the confidence interval, for instance  $\pm 5\%$ ) is often determined by what is practically possible.

**Step 6: Planning field work**

The tasks involved in planning the fieldwork for a survey vary greatly for the different types of survey, and the different situations in which they are carried out. It is hard to make generalisations. The list below is designed to act as a checklist of tasks that may be necessary for some of the survey types described in this book. This list does not include all the tasks that may be necessary, nor will each task be necessary in all situations.

- Planning checklist
- Establishing a plan with clear objectives
  - Obtaining official permissions
  - Obtaining, preparing and servicing vehicles
  - Planning schedule of village visits
  - Notifying villages
  - Reminding villages close to the visit date
  - Obtaining restraint equipment (e.g. ropes, halters, nose-grips, bleeding pole, pig-catcher)
  - Obtaining maps of the study area
  - Preparing the laboratory for the analysis of specimens
  - Preparing data recording sheets
  - Planning the order of village interviews
  - Training field staff
  - Testing interview technique, data recording sheets, and equipment with trial visits
  - Obtaining and setting up computers for data management or for use in the field
  - Training staff for computer data entry of survey data

**Step 7: Training the survey team**

Well-trained field staff will make the survey run much more smoothly, promote better relations with livestock owners, collect higher quality information, increase the confidence and motivation of the staff, and provide a strong resource for future work. A training program is essential in the preparation for any livestock survey. This book is designed to be used as a training resource, and Part III includes guidelines for trainers, lesson plans and suggested training activities.

## Step 8: Pilot survey

A pilot survey is a small survey that is carried out before the real survey. Pilot surveys are very useful for a number of reasons. They:

- help in the training of the survey staff,
- identify any problems in the questions or survey design,
- collect information on the population that can be used to calculate the sample size of the real survey more accurately, and
- identify if there are any unexpected responses, or different areas that need to be studied by the survey questions.

## Step 17: Analysis

Except for very simple surveys, most data analysis will need the use of a computer. All the survey designs described in this book can be analysed with computer software included on the accompanying CD. Specific instructions for the use of the software are included in the relevant chapters.

The software provides analysis for only the key measures described. For other questions that have been included in a survey, more general data analysis is required. Other software must be used for this, and Epi Info provides an excellent tool for this type of general data analysis. The book provides brief descriptions of key tasks in Epi Info. However to carry out the range of analysis that may be necessary, users should consult the Epi Info on-line manual, and become familiar with the Analysis program.

## Step 18: Reporting

The objective of a livestock disease survey is not to collect data or generate some results. The objective is to answer a question. The answer is then used to take some sort of action, usually to improve the animal health services, improve the health of livestock, and improve the lives of livestock owners and the population in general.

Survey results must be given to the people that are able to use them, or else all the work of the survey is wasted. It is beyond the scope of this book to cover all the aspects of report writing, and most veterinary staff are already experienced in this task. There are, however, a few key points that should be noted.

- After the completion of a survey, and even while the survey is still going on, data analysis and preparation of reports should be treated as a priority. Data that refer to the situation 6 or 12 months ago are of little use. The users of the information need to know what is happening now, so every effort should be made to produce reports as soon as possible after the field work is finished.
- Using computers for data analysis makes it very easy to produce a large quantity of figures, and perform complex types of analysis. When analysing data and producing reports, keep in mind that the objective of the survey was to answer one question. This question can usually be answered with one number. This is the most important information in the report. While other interesting data may have been collected at the same time, the report should make it very clear what the main finding is. Pages of numbers and complex analysis do not help people understand the situation, but are confusing and off-putting. Reports should therefore be kept as short and simple as possible.
- Information should be presented in a way that makes it easy and quick to understand. Rather than long lists of numbers in tables, it is much easier for the

reader to get the message if the same data are presented graphically. This may be as a pie chart, bar charts, line graphs, or, if possible, as maps.

- The information should be distributed to everybody who may need it, and everybody who participated in generating it. Field staff, local and village veterinary staff, as well as national veterinary staff should all receive reports, some of which may be more detailed than others. A mechanism for reporting the information back to the livestock owners is also important. This feedback is a great way to make all involved in the survey feel that they have achieved something useful, and they will be happy to help again in future.
- Thought should also be given to distributing the results internationally. Neighbouring countries will often find the information helpful in coordinating cooperative regional approaches to disease control. For the same reasons, the information should be sent to international animal health organisations, such as regional bodies, the Office International des Epizooties (OIE) and the Food and Agriculture Organization of the United Nations (FAO).
- If possible, the results should also be published in international journals. Apart from general interest, details on how the survey was carried out, and the results can help others planning similar surveys to better organise and carry out their work. A demonstrated ability to carry out high quality surveillance to internationally recognised standards greatly improves the international reputation of a country's veterinary authorities. This confidence can be a distinct benefit, especially in international trade issues.

# 7

## Prevalence Surveys

### Contents

Chapter 1: Introduction

#### **Part I: Background to Disease Surveys**

Chapter 2: General Principles of Animal Disease Surveillance

Chapter 3: Sampling

Chapter 4: Principles of Data and Specimen Collection

Chapter 5: Village Interviews

Chapter 6: Computerised Data Management and Analysis

#### **Part II: Survey Design and Analysis**

Chapter 7: Prevalence Surveys  
    Conducting a survey  
    Choosing the right design  
    Sample size  
    Stratification  
    First-stage sampling - the village or herd  
    Second-stage sampling - the animal  
    Data analysis

Chapter 8: Incidence Rate Surveys

Chapter 9: Surveys to Demonstrate Freedom from Disease

#### **Part III: Notes for Trainers**

Chapter 10: Guide for Trainers

Chapter 11: Lesson Plans

Chapter 12: Activity Sheets



## Prevalence surveys

Prevalence surveys aim to estimate the proportion of the population that had a particular disease or status, at a single point in time (see page 26). Prevalence surveys are the most commonly used way to gather information in livestock disease surveillance programs. This chapter describes a series of survey designs developed especially for developing countries. They are able to gather unbiased, reliable data, as quickly and as inexpensively as possible.

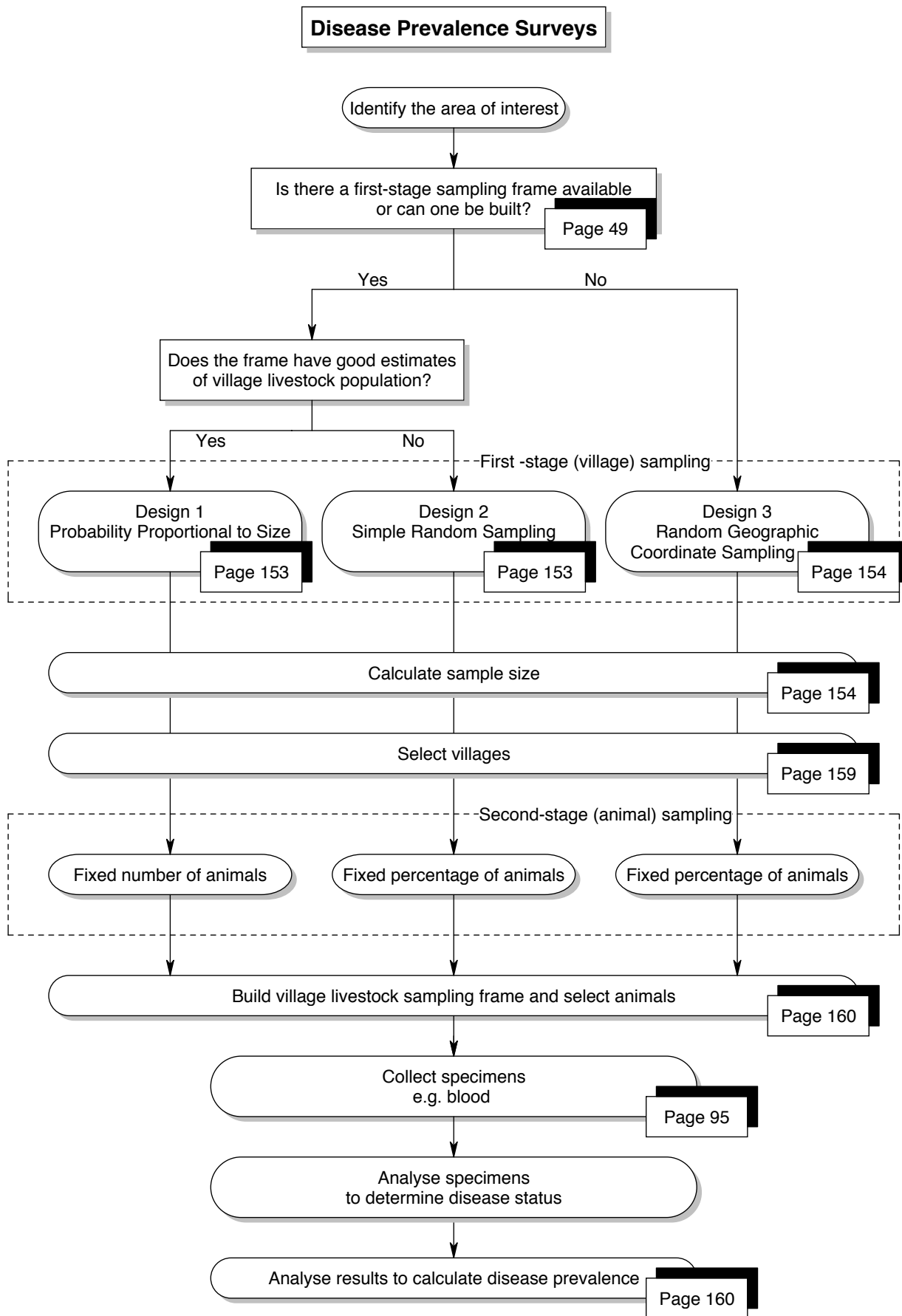
Prevalence surveys may be used to assess disease priorities and develop control strategies (determine how much disease there is in a population, where the disease is occurring), or to monitor the progress of a control program (e.g. determine the proportion of animals with antibodies due to vaccination). Both prevalence and incidence rate surveys measure the amount of disease in a population, but in different ways. The difference between the two measures is discussed on page 29, to help decide which one is best for a particular situation. Chapter 8 describes how to conduct an incidence rate survey. Surveys to demonstrate freedom from disease are similar to prevalence surveys in that they aim to identify diseased animals, but the design and analysis are quite different. These surveys are described in Chapter 9. While the three different survey types are dealt with separately, it is possible to collect information to estimate two or three of these measures during the one survey.

## Two-stage sampling

Surveys of large areas (national, province, state or district surveys) which measure prevalence at the animal level are difficult because of the lack of a sampling frame. To use random selection (and ensure reliable results) requires a sampling frame which includes every animal in the entire study area (see page 49). Building an accurate sampling frame of all animals in a large area is usually impossible. *Two-stage sampling* (page 64) avoids this problem by breaking the sampling into two steps. First, groups of animals are selected randomly (usually villages or herds). At the first stage, a sampling frame listing all the villages or herds (the first-stage units of interest) in an area is all that is required. Once the groups are chosen, each village or herd is visited, and a sampling frame of the animals in the group is constructed, and used to select animals (the second-stage units of interest) for the sample.

The strength of two-stage sampling is that animals need to be listed only for a small number of herds or villages, rather than the whole population. In addition, the field work is easier, as the survey team has to visit only a relatively small number of villages. If simple random sampling was used, there may well be one or two animals from very many villages, which would require a lot more travel. The weakness of two-stage sampling is that the survey design and analysis are more complex. The Survey Toolbox provides programs that make these jobs much easier.

This chapter is a guide to conducting two-stage prevalence surveys for livestock diseases as part of an active surveillance program in developing countries.



## Conducting a survey

There are 20 main steps in running a two-stage prevalence survey, as shown below. This description is based on a survey of village livestock, in which specimens are collected for laboratory analysis (e.g. blood samples) to determine the disease or antibody status of the animals. Some of the procedures will need to be modified slightly (simplified) if 1) no specimens are collected, and only clinical examination is used to determine the disease status of animals, or 2) herds rather than villages are selected in the first stage, in which case it may be easier to select animals randomly.

- Step 1:** Determine what question is being asked and how best to answer it.
- Step 2:** Identify the target population.
- Step 3:** Choose the right survey design. There are three different designs (PPS, SRS, and RGCS), based mainly on the way herds or villages are selected at the first stage of sampling. The choice depends on what sampling frame is available.
- Step 4:** Calculate the best sample size. A computer program is provided to help with this, but some knowledge of the disease and population is also needed.
- Step 5:** Decide if the survey is to use stratification, and if so, what basis will be used.
- Step 6:** Plan field activities, decide on interview questions, prepare data collection sheets, transport, restraint equipment, specimen collection and processing equipment.
- Step 7:** Train survey teams.
- Step 8:** Select the first-stage sample (herds or villages) using random sampling.
- Step 9:** Visit selected herds or villages.
- Step 10:** Conduct a village interview, to build an animal sampling frame for the village, and ask other questions.
- Step 11:** Select the second-stage sample (animals) using random sampling.
- Step 12:** Visit livestock owners, and identify selected animals.
- Step 13:** Restrain animals and collect specimens.
- Step 14:** Process specimens ready for analysis.
- Step 15:** Send the specimens to the laboratory.
- Step 16:** Check the data for completeness and accuracy.
- Step 17:** Enter the survey data and laboratory results into a computer.
- Step 18:** Check the data for mistakes during data entry.
- Step 19:** Analyse the data to estimate the prevalence.
- Step 20:** Report the data, providing feedback to livestock owners, local veterinary staff, national veterinary authorities, and perhaps international publications or organisations.

Some of the key steps are described in detail below.

### Step 3: Choosing the right design

There are three different survey designs for two-stage prevalence surveys. The choice of the best design to use depends on what sort of sampling frame is available for first-stage sampling – a sampling frame with population data, a sampling frame

without population data, or no sampling frame at all (see page 49 for a full discussion of sampling frames). The type of sampling frame determines how herds or villages are chosen at the first stage, and how animals are chosen at the second stage.

### Design 1 (PPS)

Probability proportional to size

In the best situation, a complete sampling frame is available, listing all herds or villages and including *reliable livestock population data*. This may come from data maintained by the veterinary services, who regularly update livestock population figures, or else a recent agricultural census. When this information is available, probability proportional to size (PPS) sampling may be used.

In Design 1, villages or herds are chosen at the first stage so that the chance of selecting one with a larger population is greater than one with a smaller population. When selecting animals at the second stage, a *fixed number* of animals is chosen from selected villages using simple random sampling.

**Example:** Design 1 (PPS) was used for a prevalence survey of village pigs. An agricultural census was carried out 1 month earlier, and the data were used as a sampling frame. Forty villages were chosen with probability proportional to pig population. Each of the villages was then visited, and a village interview of pig owners used to build a sampling frame for the village. In each village 15 pigs were selected by simple random sampling from the sampling frame.

This survey design is the most efficient, as it is able to make more accurate estimates for a given sample size than the other designs. It is also easier for the survey teams when doing the field work. Unfortunately, while a complete sampling frame might be available, it is quite uncommon to have complete up-to-date data on village or herd livestock populations for the relevant species. If the data are only a few months old, they will already be incorrect. If there are only small changes in the population, this doesn't matter too much, but if there have been large changes in some villages or herds, the population data can no longer be considered reliable, and Design 2 should be used.

### Design 2 (SRS)

Simple random sampling

When a good sampling frame, containing all the herds or villages in the area is available, but there are no reliable population data, simple random sampling (SRS) can be used at the first stage.

In Design 2, every village or herd has the same chance of being selected. At the second stage, a *fixed proportion* of animals are selected from the population using simple random sampling, instead of a fixed number as is used in Design 1.

**Example:** A survey of village chickens uses a Statistics Office list of all villages in a province as the sampling frame. No figures are available for the chicken population in the different villages. At the first stage, a sample of 40 villages is selected by simple random sampling. Each of these villages is visited, and the chicken owners are gathered for a village interview. A sampling frame is constructed, and in each village 5% of the population is selected at random for the sample. In a village with 252 chickens, 13 chickens are selected, in another village with 689 chickens, 34 chickens are selected.

Design 2 is reasonably efficient, but not quite as good as design 1. Because the survey teams don't know how many animals there will be in the villages or herds before they visit, they don't know how many animals will need to be examined or specimens collected. In a large village, there will be a lot of work, in a small village, not much at all. This makes planning the field work slightly more difficult. However, most of the time a sampling frame is available, and this is the survey design that should be used.

### Design 3 (RGCS)

Random geographic coordinate  
sampling

In the worst case, there is no sampling frame for herds or villages available at all. This is usually the case for nomadic herds, or when government structures and records have broken down due to war or other disasters. The only way to select a random sample of herds or villages at the first stage is to use random geographic coordinate sampling (RGCS).

In design 3, RGCS is used to select herds or villages (see page 65). At the second stage, a *fixed proportion* of the village population is sampled, just as with design 2.

The statistical efficiency of this survey design is similar to that of design 2, but the field work is much more difficult. This is because a lot of field work is needed before the actual survey, to select the villages or herds. For this reason, design 3 should be used only if necessary. Usually it will be possible to find a reliable sampling frame.

## Step 4: Sample size

The sample size is calculated using the computer program described below. For a two-stage sampling survey, the sample size is made up of the number of herds or villages to sample at the first stage, and the number or proportion of animals to sample at the second. It is possible to select fewer villages and more animals, and still get results of the same accuracy. This makes two-stage sampling very flexible, and allows the survey design to be adjusted to achieve results of a specific level of accuracy, but at a minimum cost. Calculating the best combination of first- and second-stage sample sizes requires a few different pieces of information, described below. When a survey is being carried out for the first time in a particular area, some of these numbers may not be known, and estimates have to be used. However, when a survey is used as an ongoing part of a surveillance system, detailed information is available from previous surveys, and very accurate calculations of the minimum cost sample sizes can be made.

### Survey costs

The costs of the different parts of the survey need to be known – the cost per animal, and the cost per herd or village. It is the ratio of these costs that determines the least expensive combination of first- and second-stage sample sizes.

The per-animal costs are mainly made up of costs for laboratory testing, and equipment, such as blood tubes, needles, serum tubes, etc. They may also include a cost for the salaries of the field staff, which is based on how much time it takes to examine or collect specimens from each animal. Per-village costs are usually made up of field staff salary and transport costs.

These costs are summarised in the table below. Other costs that do not vary with the number of animals or the number of herds or villages (e.g. the cost of obtaining a village sampling frame) are not included in the calculation.

Per-animal costs	Per-village costs
Blood tubes	Fuel
Serum tubes	Vehicle costs
Laboratory tests	RGCS costs
Staff salaries	Staff salaries

When conducting a survey in an area for the first time, it is useful to keep accurate records of the costs involved. These figures can be useful in planning future surveys. When no previous figures are available, the costs need to be estimated.

### **Variance**

In two-stage sampling, there are two populations that are being sampled, the herds or villages, and the animals. Each of these two populations has its own variance (see page 145). The amount of difference between different herds is known as the “between-herd variance”. The spread of difference between individual animals within the same herd is called the “within-herd variance”. When calculating the sample size for a two-stage survey, both these variances are taken into account by the computer program. These values are very hard to estimate, so either values from a previous survey need to be used, or estimates based on similar surveys in other parts of the world.

For example, seroprevalence surveys of foot and mouth disease antibodies of cattle and buffalo conducted in Southeast Asia have yielded within-herd variance estimates in the range 0.15 to 0.22, and between-herd estimates between 0.03 and 0.08. If no other data are available, these figures can be used as a starting point for initial estimates.

### **Population size**

The total size of the population is needed for some sample size calculations (depending on the survey design). Where full population data are available for every village, this is not a problem. However, where no data exist, the total population must be estimated. Fortunately, it doesn't matter too much if this estimate is not perfect. There are usually some records available for the population in an area.

### **Estimated prevalence**

One of the most difficult things to understand about calculating sample sizes for prevalence surveys, is that you need to know approximately what the prevalence is before you do the survey. For surveys held as a regular part of an ongoing surveillance program, earlier prevalence estimates will allow good estimates to be made. However, for the first survey in an area, guesswork will be needed.

### **Area of study area and selection radius**

For random geographic coordinate sampling (Design 3), an estimate of the total area of the study area is needed. These figures are often available through the National Statistics office. If digital maps of the study area and geographical information system (GIS) software are available, the size of the study area can be calculated easily. If only paper maps are available, you can still work out what the approximate area is, by drawing a grid of squares over the study area, and counting them.

An estimate of the anticipated selection radius is also needed.

Relative error

**Precision**

Precision is usually measured as the width of the confidence interval. A fixed width may be used, or, for very low or high prevalences, the *relative error* may be better. This is because as the prevalence gets smaller, we often want to measure it more precisely.

**Example:** Using a fixed width confidence interval of  $\pm 5\%$ , a survey resulting in a prevalence estimate of 50% would have a confidence interval of 45% - 55%. This is probably precise enough for most purposes because the difference between 45% and 55% is unlikely to be very important. If the prevalence was 5%, the confidence interval would be 0% to 10%. The difference between 0% and 10% is probably quite important, so we would often want to measure the value more precisely if the prevalence is low. If we used relative error, the confidence interval for a prevalence of 50% may be 45% to 55%, but the confidence interval for a prevalence of 5% may be 3% to 7%.

The relative error is a measure of the width of the confidence interval as a proportion of the prevalence, so the smaller the prevalence, the narrower the confidence interval.

For fixed-width confidence intervals, a value of  $\pm 5\%$  or  $\pm 10\%$  is commonly used. If a smaller value is used, the sample size will increase in size dramatically. A relative error of 0.1 will produce a confidence interval of about  $\pm 10\%$  if the prevalence is about 50%, but if the prevalence is 10%, the confidence interval will be about  $\pm 4\%$

**Confidence level**

The confidence level determines how confident we are that the true value lies within the confidence interval. By convention, a confidence level of 95% is used most of the time. This means that in one case out of 20, the true value may lie outside the confidence interval.

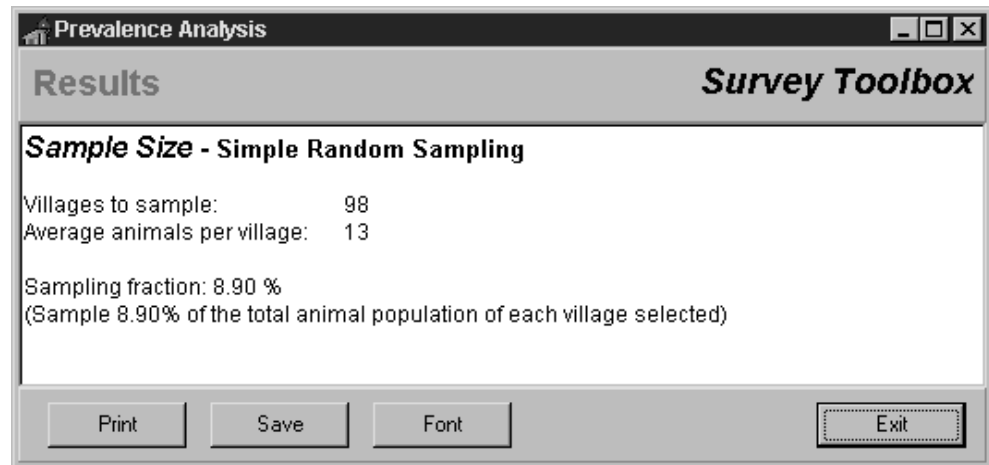
**Calculating the sample size**

The formulae for calculating the sample size for the three different survey designs are very complex, and can normally be calculated only by a trained statistician. In order to enable non-statisticians can do the calculations, the formulae have been incorporated into the **Prevalence Analysis** program, included on the CD. To start the program use the Windows **Start** menu, select Programs, Survey Toolbox, and Sample Size. To calculate the sample size required for a two-stage prevalence survey:

- Step 1:** Click on the Sample Size Calculation tab at the top of the window.
- Step 2:** Select the survey design to be used. In the First Stage Sampling Scheme box, select either Design 1 (Population proportional to size sampling - PPS), Design 2 (Simple random sampling - SRS), or Design 3 (Random geographic coordinate sampling - RGCS). See *Choosing the Right Design* on page 152 to help you decide.
- Step 3:** If you are unsure, click on the **Which one?** button for help deciding which design to use.



- Step 4:** Look at the Second Stage Sampling Scheme for advice on how to select animals at the second stage.
- Step 5:** In the Parameters box, enter all the parameters required, as described above. You must enter estimates for all the parameters. You need to decide on the required accuracy and confidence level yourself.
- Step 6:** If you have the results of a previous survey, you can use these to enter the first- and second-stage costs (cost per herd and cost per animal).
- Step 7:** If you have a data file from a previous survey, click on the **I don't know. Work it out for me.** button. This will open the data file, and allow you to analyse the data, to calculate the variance and prevalence estimates needed. It will also calculate an estimate of the population size. Be sure to set this yourself if the data came from a different population. See Data Analysis on page 160 for instructions on analysing data.
- Step 8:** When all the parameters are entered, click the **Calculate** button.
- Step 9:** The results will be displayed in a window. The best first- (herd) and second- (village) stage sample sizes will be shown. The second-stage sample size will be expressed as either a number (Design 1, PPS), or a percentage of the herd population (Designs 2 and 3, SRS and RGCS).



## Step 5: Stratification

Stratification almost always improves the accuracy of the survey, and usually makes the field work simpler (see page 45). When little information is available about the population, stratification is usually done by geographical area. For instance, a national survey may be stratified by state or province, or a provincial survey may be stratified by district.

Proportional allocation

In order to make sure that each area is properly represented, we usually want to select villages from each stratum proportional to the total number of villages in that stratum. This means that a district with more villages will contribute more villages to the sample than a district with fewer villages. This is known as *proportional allocation*. The number of villages to be selected from one stratum ( $n_k$ ) is equal to the total number of villages to be selected ( $n$ ) times the proportion of villages in the population ( $N$ ) that are in that stratum ( $N_k$ ).

$$n_k = n \times \frac{N_k}{N}$$

**Example:** In a survey, the first-stage sample size is 40, and the total number of villages in the study area is 480. There are 5 districts which are used for stratification. The number of villages in the first district is 120. The proportion of the total villages in that district is  $120/480 = 1/4$ . The number of villages to be selected from that district is therefore  $40 \times 1/4 = 10$  villages. A district with 80 villages would contribute  $40 \times (80/480) = 6.67 \approx 7$  villages.

When using stratification, you can generally use the same sample size that you would calculate without stratification, and divide it up between the strata with proportional allocation. The overall results will usually be slightly more precise than predicted, due to the stratification. Note that the estimates for the individual strata will be much less precise than the overall estimate, because the sample size in each stratum is much smaller than the overall sample size. If you require precise estimates for each stratum, calculate the sample size required for each stratum separately. You can then combine the stratum results to give an overall estimate (which will be very precise because of the large sample size).

## Step 8: First-stage sampling - the herd or village

The approach used for first stage of sampling depends on the survey design used. In all cases, however, sampling is done with replacement. This means that the same herd or village can be chosen twice. In this case, twice as many animals as normal are sampled from the village.

**Example:** The calculated sample size for a two-stage prevalence survey is 40 villages and 8% animals in each village. A good sampling frame is available, but no livestock population figures, so Design 2 is used. The villages are selected from the sampling frame using simple random sampling with replacement. One village is selected twice. The sample size is still 40, even though only 39 separate villages are visited. The village that was selected twice has a population of 145 pigs. Instead of the normal 8%, two samples (16%) are drawn from this population giving a total of 24 animals.



If a sampling frame is available on computer disk, and **Designs 1 or 2** are used, you can use the **Random Village** program to select the herds or villages (see page 50). Use the following steps to select the villages:

- Start the Random Village program.
- Click on the **Open** button and select the data file containing the sampling frame.
- Click on one or more Identification Fields to be displayed for the selected villages or herds (usually ID, name etc.).
- Under Number to Select, enter the total number of villages or herds (the first stage sample size).
- Enter the sampling type. If using Design 1 with probability proportional to size (PPS) sampling, click on Probability Proportional to Size sampling. You will then need to select the field from the table that has the size information (livestock population). If you are using Design 2, select Simple Random Sampling.
- Under Replacement, click With Replacement.
- If using stratification, click the check box next to Use Stratification?. You will then need to select the field that contains the information used for stratification. This will usually be a province or district code.
- Click on the **Select** button to select the random sample.
- The selected villages or herds are displayed. You can save them to a new table, or print them.

If the sampling frame is not available on disk, you can use the manual techniques described in Chapter 3 to do either PPS (page 48) or SRS (page 41) sampling.



If no sampling frame is available, and you are using **Design 3**, you can use the random geographic coordinate sampling program (**RGCS Win95**) to select random coordinates (see page 68). If you have access to a copy of the ArcView GIS program and a digital map of the study area boundaries, you can use **RGCS ArcView** to select coordinates (see page 69).

If using random geographic coordinate sampling, you will then need to screen the selected points (if possible), and then visit points to identify nearby villages. This process is described in detail in Chapter 3 (page 74).

## Step 11: Second-stage sampling - the animal

Once herds or villages have been selected, and the field work commenced, the second stage of sampling, selecting individual animals, can be done.

In all cases, the animals are selected without replacement, so that an individual animal can be tested only once. If Design 1 (PPS) is used, then a fixed number of animals is selected from each herd or village sampled. If Designs 2 or 3 are used, a fixed proportion of the total village or herd population is selected.

If the animals are in a single herd, then a sampling frame may already exist, maintained by the owner. This can be used for simple random sampling done either by hand, or with a computer after first entering the data into the **Random Village** program. Alternatively, randomised systematic sampling (page 44) can be used if the facilities exist for putting all the animals in a sequence (e.g. handling yards and races).

In a village with multiple livestock owners, it is usually necessary to build a sampling frame first, and then select the random sample. Village interviews of livestock owners are an efficient way to build an accurate sampling frame and are described in detail in Chapter 5. The technique for selecting animals from this sampling frame using a manual method is described on page 55, and the use of the **Random Animal** program is described on page 58.

## Step 19: Data analysis

The analysis of prevalence data collected in a two-stage survey using any of the three designs is very complex. The formulae used are listed in Appendix B. The **Prevalence Analysis** program, which also calculates sample sizes for two-stage prevalence surveys, is included on the CD to analyse data from the three types of survey designs described.

Before analysis is done, the data must be entered into a computer, and stored in a file in either dBASE or Paradox format. Epi Info or another database program can be used to enter the data. The Prevalence Analysis program may also be used for simple data entry.

### Data inputs

The data files, data fields, and other information required for analysis depend on the survey design used and whether or not stratification was used. In all cases, an animal-level data file is required, with the disease status of each animal, and the village that the animal came from. The disease status may be a code which indicates Diseased/Non-Diseased, or a Yes/No field. It may also be a numeric value, such as an antibody titre. In this case you need to specify a cut-off value. Above this cut-off value, animals are considered to be positive, and below the value they are negative.



## Design 1 (Probability Proportional to Size)

### Without stratification

File 1 (Animal File)

- Disease status
- Village ID

### With stratification

File 1 (Animal File)

File 2 (Village File)

- |                  |              |
|------------------|--------------|
| • Disease status | • Village ID |
| • Village ID     | • Stratum ID |

## Design 2 (Simple Random Sampling)

### Without stratification

File 1 (Animal File)

File 2 (Village File)

Other figures

- |                  |                                |                                |
|------------------|--------------------------------|--------------------------------|
| • Disease status | • Village ID                   | • Total villages in study area |
| • Village ID     | • Village livestock population | • Total animals in study area  |

### With stratification

File 1 (Animal File)

File 2 (Village File)

File 3 (Stratum File)

- |                  |                                |   |
|------------------|--------------------------------|---|
| • Disease status | • Village ID                   | • Total number of villages in the stratum |
| • Village ID     | • Village livestock population | • Total number of animals in the stratum  |
|                  | • Stratum ID                   | • Stratum ID                              |

## Design 3 (Random Geographic Coordinate Sampling)

### Without stratification

File 1 (Animal File)	File 2 (Village File)	Other figures
<ul style="list-style-type: none"> <li>• Disease status</li> <li>• Village ID</li> </ul>	<ul style="list-style-type: none"> <li>• Village ID</li> <li>• Village livestock population</li> <li>• Village weight<sup>9</sup></li> <li>• Area fraction<sup>10</sup> (optional)</li> </ul>	<ul style="list-style-type: none"> <li>• Selection radius</li> <li>• Total number of random points used</li> <li>• Total area of study area</li> </ul>

### With stratification

File 1 (Animal File)	File 2 (Village File)	File 3 (Stratum File)
<ul style="list-style-type: none"> <li>• Disease status</li> <li>• Village ID</li> </ul>	<ul style="list-style-type: none"> <li>• Village ID</li> <li>• Village livestock population</li> <li>• Village weight</li> <li>• Stratum ID</li> <li>• Area fraction (optional)</li> </ul>	<ul style="list-style-type: none"> <li>• Stratum ID</li> <li>• Selection radius for each stratum</li> <li>• Total number random points used in the stratum</li> <li>• Total area of the stratum</li> </ul>

Some examples of appropriate Questionnaire file formats for Epi Info are shown below.

Demonstration Data Entry Form Prevalence Survey - Animal Data		
Tube Number:	#####	
Village ID:	#####	
District ID:	#####	(If using district for stratification)
Age:	##	
Sex:	<A>	(Single character, uppercase field)
Species:	_____	(E.g. cattle / buffalo, if necessary)
Antibody titre:	#####	(As reported by laboratory)
Disease status:	<Y>	(Calculated from antibody titre) (Use a standard cut-off value)

<sup>9</sup>The *village weight* is the total number of villages within the selection radius of the point used to select the village.

<sup>10</sup>The *area fraction* is the proportion of the area of the circle (as defined by the sampling radius around the random point used to select that village) which lies inside the study area. For most villages, this will equal 1, but for some near the boundary of the study area, it will be smaller.

Demonstration Data Entry Form  
Prevalence Survey - Village Data  
(For Design 2 (SRS) and 3 (RGCS) only)  
(Not required for design 1 (PPS))

Village ID: #####

Total population: #####

District ID: ##### (If using district for stratification)

Demonstration Data Entry Form  
Prevalence Survey - Village Data  
(For Design 3 (RGCS) only)  
(Not required for design 1 (PPS) and 2 (SRS))

Village ID: #####

Total population: #####

District ID: ##### (If using district for stratification)

Weight: ## (Total number of villages around the point)

Selection Radius: ##.## km (must be the same for every village in one stratum)

Total points: ## (total number of points used, including points with no villages. Same for every village in one stratum)

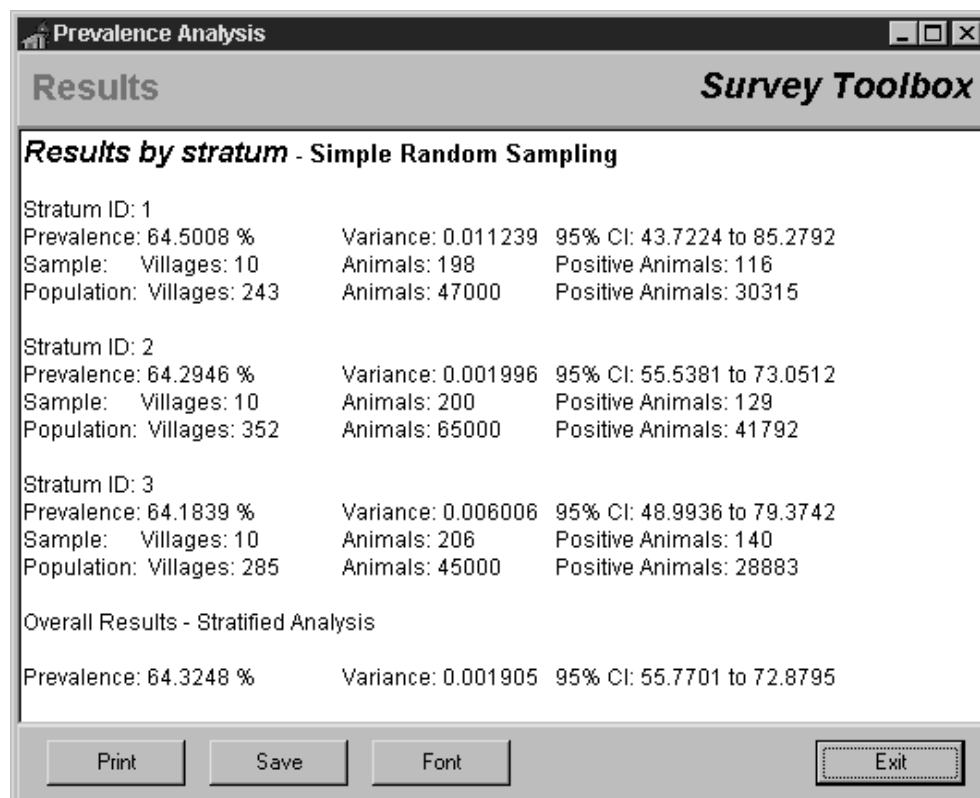
Study area: #####.## sq km (Total area or stratum area. Same for every village in one stratum)

## Analysing the data

To analyse the data use the following steps:

- Step 1:** Start the **Prevalence Analysis** program by clicking on the Windows **Start** button, selecting Programs, then Survey Toolbox and choosing Prevalence Analysis.
- Step 2:** Click on the Prevalence Data Analysis tab at the top of the window. The other tab is for sample size calculation.
- Step 3:** In the First Stage Sampling Scheme box select the survey design used, Probability Proportional to Size (Design 1), Simple Random Sampling (Design 2) or Random Geographic Coordinate Sampling (Design 3).
- Step 4:** If stratification was used, click the Stratification? check box.
- Step 5:** In the data fields, open the files and enter the data required for the type of analysis you are performing.
- Step 6:** In the Animal data box, Click the **Open Animal Data** button, and select the file with the animal level survey data.

- Step 7:** In the data fields box, select the fields in the database that contain the data for analysis. First, select the field that contains the disease status data.
- Step 8:** Then select the field which contains data identifying which village or herd the animal came from (First stage sampling units).
- Step 9:** Make sure the codes for disease status in the Status Codes box are correct.
- Step 10:** If displayed, enter data in the Village Data box. Click on the **Open Village Data** button, and select the fields required.
- Step 11:** If further information is required for stratification or random geographic coordinate sampling, enter the data required. Click on the **Open Strata Data** button, and set up the fields, or type in the parameters required.
- Step 12:** When all the fields have been entered, click on the **Calculate** button to analyse the data. A window will display the results, which can be either printed or saved to a file.



## Calculating true prevalence

The disease status of animals during a survey is assessed by means of a laboratory test, or by direct clinical examination. In both situations, it is possible to make a mistake in a few cases, and call some healthy animals diseased, and some diseased animals healthy (or alternatively, think that some animals with antibodies don't have them, and others that don't, do have them). There are two measures used to describe how good a test is at correctly determining the disease state of an animal: sensitivity and specificity (see page 33 for a full discussion).

Because most tests make a few mistakes, a few of the test results analysed could be wrong, which makes the estimate of the prevalence incorrect. Usually this error is quite small, but for tests that make mistakes more often, the error can be large.

If the sensitivity and specificity of the test are known or can be estimated, it is possible to correct for these mistakes, and convert the results of the analysis, the *apparent prevalence*, to the corrected result, the *true prevalence*.

The **True Prevalence** program on the CD carries out the calculations for you. When the results have been analysed with the Prevalence Analysis program, use True Prevalence to convert the result to the true prevalence, based on the test sensitivity and specificity:



- Step 1:** Start the True Prevalence program. Use the Windows Start menu, select Programs, Survey Toolbox, True Prevalence.
- Step 2:** In the Parameters box, enter the Apparent Prevalence, as reported by the Prevalence Analysis program.

- Step 3:** Enter the test sensitivity and specificity. The laboratory may be able to suggest figures for these, or else it might be necessary to search journals for published studies.
- Step 4:** Enter the sample size of the survey.
- Step 5:** Press the **Calculate** button.
- Step 6:** The true prevalence is shown, along with a confidence interval.

**Note:** The confidence interval is based on the assumption that the sample was selected by single-stage simple random sampling. For the two-stage surveys described in this chapter, the confidence interval reported will be smaller than the correct confidence interval.

Parameters	
Apparent Prevalence	14 %
Sensitivity	94 %
Specificity	88 %
Sample Size	560

Result	
True Prevalence	2.439 %
95% Confidence Interval	(0.000 - 5.944)

## Interpretation of results

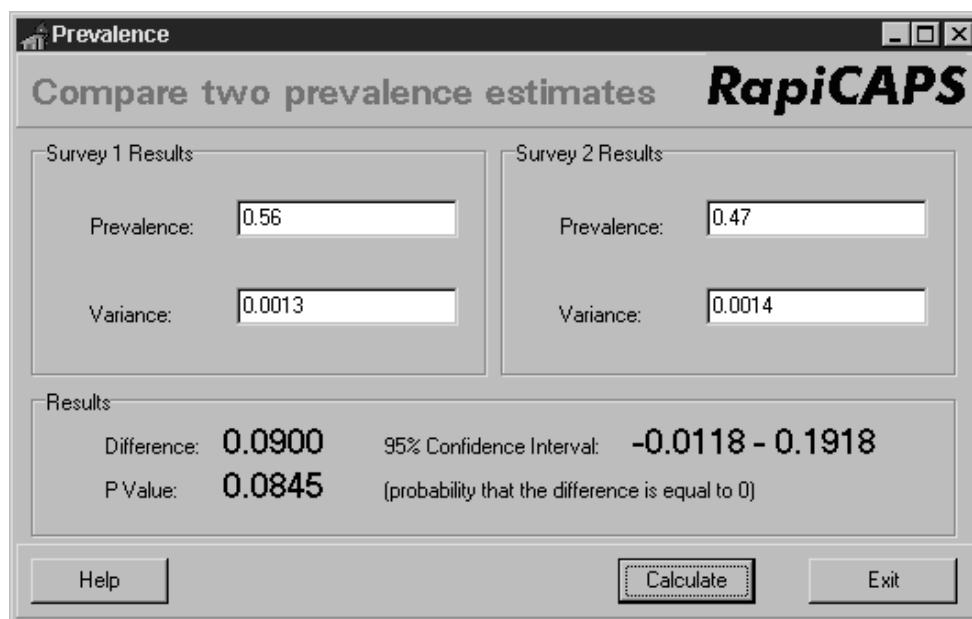
The key result from the survey is an estimate of the prevalence of the disease or state in the population. This is shown as a single figure, and a 95% confidence interval. The confidence interval can be interpreted to mean: "If the same survey were conducted in the same population many times, the confidence interval produced by the results would include the true prevalence of disease in the population 95% of the time." This can be loosely interpreted to mean that we are 95% confident that the true prevalence lies within the confidence interval.

If stratification was used, separate prevalence estimates and confidence intervals are shown for each of the strata. Note that because the number of animals sampled from each stratum is relatively small, the confidence intervals for the stratum estimates are usually very wide, indicating that our estimates are not very precise. The overall estimate is usually much more precise with narrow confidence limits.

### Comparison of two prevalences

When the prevalence estimates from two surveys have been calculated they can be compared to determine whether there is a real difference between them, or the difference is likely to be just due to chance. Monitoring changes in prevalence is an important way to evaluate the progress of a disease control program.

Use the **Compare Prevalence** program to compare two prevalence estimates. Click on the Windows Start menu, then select Programs, Survey Toolbox, Compare Prevalence.



Survey	Prevalence	Variance
Survey 1 Results	0.56	0.0013
Survey 2 Results	0.47	0.0014

Results	Value	Description
Difference:	0.0900	
95% Confidence Interval:	-0.0118 - 0.1918	
P Value:	0.0845	(probability that the difference is equal to 0)

- Step 1:** In the Survey 1 Results box, enter the prevalence and variance from the first survey, as reported by the Prevalence Analysis program.
- Step 2:** Enter the same figures from the second survey in the Survey 2 Results box.
- Step 3:** Click on the **Calculate** button.
- Step 4:** The results will be displayed.

The results show the difference between the two prevalence estimates, and a 95% confidence interval for that difference. In addition, they show a P value, which is a measure of the probability that the two prevalence estimates are in fact the same (the difference is 0). If the P value is very small, then we can be confident that there is a real difference between the two prevalences.

# 8

## Incidence Rate Surveys

### Contents

Chapter 1: Introduction

#### **Part I: Background to Disease Surveys**

Chapter 2: General Principles of Animal Disease Surveillance

Chapter 3: Sampling

Chapter 4: Principles of Data and Specimen Collection

Chapter 5: Village Interviews

Chapter 6: Computerised Data Management and Analysis

#### **Part II: Survey Design and Analysis**

Chapter 7: Prevalence Surveys

Chapter 8: Incidence Rate Surveys

Disease outbreak surveys

Introduction

Survey procedure

Sample size

Sampling (selecting herds or villages)

Data collection - the interview

Data management

Data analysis

Interpretation of results

Analysis of two data sources

Background

Data sources

Sampling

Data collection

Data management

Data analysis

Chapter 9: Surveys to Demonstrate Freedom from Disease

#### **Part III: Notes for Trainers**

Chapter 10: Guide for Trainers

Chapter 11: Lesson Plans

Chapter 12: Activity Sheets



## Incidence rate

Incidence rate is the number of the new cases of disease in a population at risk over a period (see page 28). Incidence rate measures the rate of spread of infectious diseases. It can also help distinguish between naturally acquired antibodies and vaccine-induced antibodies when used in conjunction with a prevalence survey.

One way to estimate incidence rate is to observe a group of animals for a long period of time, and to record which animals become affected with the disease. This type of incidence rate study is very slow and expensive, as every animal has to be regularly examined, and the study may last for many months or longer. These kinds of study have commonly been carried out in developed countries.

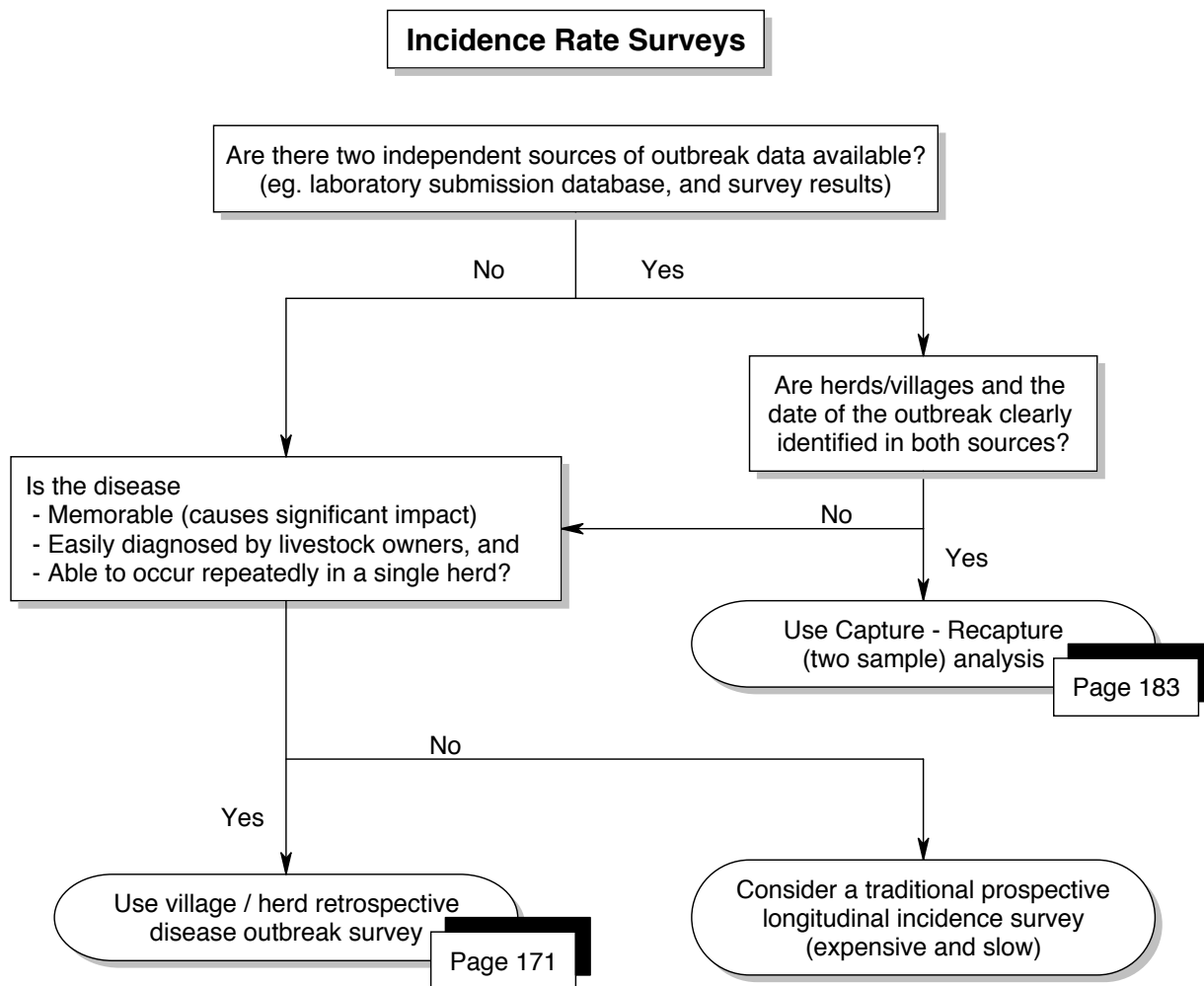
This chapter describes two alternative ways to gather information about disease incidence rate. The differences between these techniques and traditional studies are:

- the unit of interest is the village or herd, instead of the animal. This means that we are not interested in the number of animals that get a disease over a period of time, but the number of villages or herds that suffer an outbreak of disease over a period of time;
- the outbreak information is collected by using the memories of the livestock owners. Instead of starting a study, and observing the animals for a long time (a *prospective* study), livestock owners are asked about the disease outbreaks that have occurred over several years before the survey (a *retrospective* study).

For most major epidemic diseases, the disease is maintained through spread from one village or herd to another. It is rare that the bacteria or viruses are able to maintain themselves within the one herd, as animals either die, or develop immunity relatively quickly. Herd- or village-level incidence rate is therefore a more useful measure of the rate of spread of disease, or the effectiveness of a control program, than individual animal-level incidence rate. In a survey depending on the memories of livestock owners, it is also much easier to collect reliable information about herd- or village-level outbreaks than about disease in individual animals.

The first of the two techniques described in this chapter is the Retrospective Disease Outbreak Surveys technique. Village interviews are carried out to ask livestock owners about previous disease outbreaks. When the results are analysed, the measure of disease is not a traditional incidence rate measure, but can be used in the same way to assess the rate of spread of disease, and compare disease levels between different areas (or the same area at different times).

The second technique, Capture-Recapture technique, uses information on disease outbreaks that the veterinary authorities may have already collected. This technique uses two sources of outbreak data (such as diagnostic laboratory submissions or field disease reports), and combines them to estimate a traditional incidence rate measure. If two separate suitable data sources already exist, no field survey is necessary.



## Retrospective disease outbreak surveys

### Introduction

In this technique interviews with herd or village livestock owners are carried out to collect information about the date of the most recent disease outbreak. The reliability of the survey therefore depends on the livestock owners’ ability to correctly identify the disease, and to correctly recall the date that the outbreak started.

In order to ensure that the quality of information collected is high, the survey technique should be used only in appropriate situations. Firstly, it should be used to investigate only diseases which are:

Survey prerequisites

- *discrete and repeatable*: The disease must occur in outbreak form, last for a relative short time and it must be possible for it to occur more than once in the same village.

- *distinctive and well known*: Diagnosis is based entirely on the observations of the livestock owners. Diseases that are clearly different to other diseases, and have a dramatic and consistent clinical presentation are more easily diagnosed.
- *memorable*: The ability of livestock owners to remember the date of an outbreak depends on the effect the disease had on them. The more dramatic the disease, and the more disruption to the lives of the livestock owners, the more reliably it will be remembered.

Every effort is made to assist the livestock owners to accurately remember the date of the outbreak. A range of techniques available is discussed in Chapter 5.

The various strengths and weaknesses of the technique are summarised below.

Strengths and weaknesses of the Retrospective Disease Outbreak Survey methodology for estimating disease occurrence.

Strengths	Weaknesses
Rapid - collects data retrospectively	Data accuracy - depends on recall
Group interviews may be used to collect other data simultaneously	Limited to diseases causing significant impact and occurring in cyclic epidemics
Can be used for quantitative comparisons	Does not provide direct estimate of incidence rate
Inexpensive - no laboratory test or repeat visits	Requires staff training
	Depends of owner diagnosis
	No animal-level estimates

## Survey activities

The major steps in carrying out a Retrospective Disease Outbreak Survey are:

- Step 1:** Identify the question to be answered (disease and geographic area of interest).
- Step 2:** Identify the target population.
- Step 3:** Decide if the survey is to use stratification.
- Step 4:** Calculate the best sample size.
- Step 5:** Plan field activities.
- Step 6:** Train survey teams.
- Step 7:** Pilot survey.
- Step 8:** Select the sample.
- Step 9:** Visit selected herds or villages.
- Step 10:** Hold a livestock owner interview.
- Step 11:** Determine if the herd or village has ever had an outbreak of the disease.
- Step 12:** If so, determine the date of the start of the last outbreak.

- Step 13:** If not, determine the earliest date since which the livestock owners are confident that there has been no outbreak.
- Step 14:** Check the data for completeness and accuracy.
- Step 15:** Enter the survey data into a computer.
- Step 16:** Check the data for mistakes during data entry.
- Step 17:** Recode the data ready for analysis.
- Step 18:** Analyse the data.
- Step 19:** Report the data.

The key steps are described in detail below.

## Step 4: Sample size

Unlike the prevalence surveys described in Chapter 7, the unit of interest in Retrospective Disease Outbreak Surveys is not the individual animal, but the herd or village. This means that simple one-stage sampling can be used.

The sample size is calculated on the basis that the survey will be used to compare the rate of disease outbreaks, either in two different areas, or perhaps more importantly, in the same area at two different times. For example, if a Retrospective Disease Outbreak Survey was conducted last year, and then repeated this year, after the introduction of a disease control program, comparing the level of disease may indicate the success of the program.

The measure that is used to calculate sample size is median (or mean) time since the last disease outbreak. If all villages or herds in the survey have experienced the disease this is simply the average of the times since the last outbreak. When comparing the results of two surveys, if the median time since the last outbreak is large, only a relative small number of herds or villages are needed to be sure that this difference is not due to chance. If the difference in median times is very small, and the two groups are almost the same, many more herds or villages are needed to determine if the small difference is real, or just due to chance.

Use the **Survive Size** program to calculate the sample size needed for the survey. To start the program, use the Windows Start menu, select Programs, then Survey Toolbox, and Survive Size. Calculate the sample size using the following steps:



- Step 1:** In the box labelled Estimated Median Survival Times, enter the times for group 1 and for group 2. You can type in the times in any units (months, years, or days). These times are what you expect to see from the survey. Alternatively, you can think of these times as indicating the smallest real difference you want to be able to detect. If the difference is smaller, then your survey will not be able to reliably distinguish between them.

**Example:** An FMD vaccination program has been started in one part of a country where the disease is endemic. In order to monitor the progress of the vaccination campaign, it is decided to carry out disease outbreak surveys every year. Before the program started, the average time since the last outbreak for villages in the area was about 3 years. The veterinary authorities decide that lengthening this to 5 years is a good indication that the program is being successful, but anything less could be accounted for by year to year variation. When calculating the sample size for the survey,

they use 3 and 5 years as the median survival times for group 1 and group 2.

- Step 2:** In the Parameters box, enter the Significance you want. This is indicating how confident you want to be in the result<sup>11</sup>. Usually, you can leave this as 95%
- Step 3:** In the Parameters box, enter the Power you want. This is a measure of how well the survey will be able to determine if there is a difference between the two groups<sup>12</sup>.
- Step 4:** Click the **Calculate** button and the sample size will be displayed.

The screenshot shows a software window titled "Survival Analysis" with a sub-window "Sample Size Calculation" from the "Survey Toolbox". It contains input fields for "Estimated Median Survival Times" (Group 1: 3.5, Group 2: 5.5) and "Parameters" (Significance: 95%, Power: 90%). The "Results" section shows "Number of events (outbreaks) required in each group: 103". Buttons for "Help", "Calculate", and "Exit" are at the bottom.

The sample size is the number of herds or villages that have had an outbreak that need to be included in each group. During the survey, some villages may not have had an outbreak, so the overall sample size will need to be slightly larger than the number suggested by the program. When deciding on the sample size, you can either use your experience or judgment to estimate what proportion of herds or villages have never experienced an outbreak. Another approach is to carry out the survey, and continue to select more herds or villages until enough have been visited.

Most importantly, as with all sample size estimates, the figures should only be used as a guide, or rough estimate.

## Step 8: Selecting herds or villages

There are two ways to select the herds or villages: 1) simple random sampling (SRS), and 2) random geographic coordinate sampling (RGCS). Firstly, if a sampling frame (page 49) is available, the easiest approach is to use simple random sampling from that sampling frame. To select villages manually, use the procedure described on page 41.

<sup>11</sup>Significance is the probability that the results of the survey will indicate that there is no difference between the two groups when the two groups are the same.

<sup>12</sup>Power is the probability that the results of the survey will indicate that there is a difference between the two groups when there actually is.



If the sampling frame is available on computer, you can use the **Random Village** program to select the sample, as described on page 50, using the following settings:

- Sampling Type should be set to Simple Random Sampling (probability proportional to size sampling is not appropriate here).
- Replacement should be set to Without Replacement.
- Do not select stratification.

If no sampling frame is available, you can select herds or villages using random geographic coordinate sampling (RGCS). This technique is described in detail in Chapter 3. When using RGCS, be sure to record village or herd weights, for use during analysis.

## Steps 10 - 13: The interview

Village livestock owner interviews are discussed in detail in Chapter 5, including advice on techniques for collecting information about the date of the last disease outbreak (page 118). It is worth repeating the importance of establishing a censoring time for villages that have not experienced an outbreak (page 120).

In addition to these two dates, it is possible to collect other related information to help with the analysis. This may include:

- the number of animals in the village at the time of the interview;
- the number of animals in the village at the time of the outbreak (or the censoring time); and
- the proportion of those animals that were affected by the outbreak.

An example data recording sheet is shown in Appendix D.

## Step 15: Data management

When the field work is completed all the results need to be entered into a computer for analysis. See Chapter 6 for general advice on computerised data management.

Data may be entered using any database program that can export data to dBASE or Paradox format (including Epi Info). When creating the database table, the following fields are necessary:

- Village or herd identification
- Outbreak (yes/no, or code field indicating whether or not there has ever been an outbreak).
- Date of last outbreak (or censoring time). This may be included as a single date field if the day of the start of the outbreak has been estimated. Usually, it is only possible to recall the month of the outbreak. There are two solutions to this problem. Firstly, all outbreaks can be arbitrarily said to have started on 15<sup>th</sup> of the month. Alternatively, two numeric fields could be used, one for the month, and one for the year.
- Date of visit. This could be treated in the same way, with either a date field, or only month and year recorded in two separate fields.
- Time since outbreak. This field is left blank at data entry, and is calculated from the two dates later.

- If random geographic coordinate sampling was used to select the villages, there must also be a field to record the weight for each village.
- If the data from two areas or times are being compared, then the file needs to contain all the data from the two groups. If the data already exist in two separate tables, they can be merged into a single table using the merge procedure (available in most database programs, including Epi Info – see the Epi Info on-line manual). There must be a Group field, with a code to indicate which group the record belongs to. This may be a numeric code, text field, or a yes/no field.

An example questionnaire file for creating the table in Epi Info is shown below.

```

                Demonstration Data Entry Form
                Village Outbreak Survey

Village ID:      #####

Date of visit:  Month ##   Year ####

Had Outbreak?  <Y>           (Censoring variable)

Outbreak:      Month ##   Year ####
(or censoring time)

Time since outbreak:  ##.### (Calculated from visit date
                        and outbreak date)

Weight:        ## (Random geographic coordinate sampling only)
                (Number of villages within selection radius)

Group          #           (when comparing two groups only)
    
```

Other fields can be included for more complex analysis, such as livestock population at the time of the visit, population at the time of the outbreak, and proportion of animals affected. Analysis of these extra data requires more sophisticated techniques, and possibly specialised statistical programs which are not provided with the Survey Toolbox software. Complex analysis and software (described later) are not necessary to calculate the level of disease and compare two groups.

**Epi Info**

Before analysis, the dates recorded must be used to calculate the time since the last outbreak. The exact procedure depends on the database program being used, but most are similar. When using Epi Info, the procedure is as follows:

- Step 1:** Start Epi Info, and choose Analysis from the programs menu.
- Step 2:** Open the data file, using the Read command. If the file is in dBASE format, use Read \*.dbf. Select the file from the list.
- Step 3:** If the outbreak time and visit time have been stored as Date fields, calculate the time since last outbreak using the command: Time = Date1 - Date2

**Example:** If the date of the outbreak is stored in a field called OBDate, the date of the visit is VisDate, and the time since the last outbreak is to be stored in a field called Time, then type:  $\text{Time} = \text{VisDate} - \text{OBDate}$ . The result is the number of days between the two dates. If it does not already exist, you will have to create a new Time field, using the **define** command:  
**define time ###.**

**Step 4:** If the outbreak and visit times have been stored in separate month and year fields (for instance, OByear and OBmonth for the outbreak time, and Vyear and Vmonth for the visit time) use the command:  $\text{Time} = (\text{Vyear} - \text{OByear}) + ((\text{Vmonth} - \text{OBmonth}) / 12)$ . This will give the outbreak time in terms of years. If months are preferred multiply by 12.

**Step 5:** Save the new values to a new data file using the **Route** and **Write recfile** commands.

**Example:** To save the data in a file called obsurvey.rec, use the commands **Route obsurvey.rec** and then **write recfile**

## Step 18: Data analysis

### Basic analysis

The data are analysed using a special technique called *survival analysis*. This technique uses the times that a herd or village *survives* without experiencing an outbreak. The advantage of survival analysis is that it is possible to include data from herds or villages that have not had an outbreak.

The Survival program, included in the Survey Toolbox does all the analysis necessary. To analyse the data using the program:



**Step 1:** Start the Survive program. Use the Windows start menu, select Programs, Survey Toolbox, and choose Survive.

**Step 2:** Open the data file for analysis. Click the **Open** button, and select the file from the list. The file must be in dBASE or Paradox format. You can also use the program to create a new file containing all the data you need, by clicking the **New** button.

**Step 3:** You now need to tell the program which fields the data are in. In the Data Fields box, click on the arrow at the right of the Survival Times box. Select the field that stores the time since the last outbreak.

**Step 4:** Next click on the Censoring Indicator box, and select the field that indicates if the herd or village has ever had an outbreak.

**Step 5:** You then have to tell the program what the codes in the field mean. The computer displays the codes from the censoring field in the Censoring Codes box. Censored is the code for villages that have never had an outbreak. Uncensored is the code for villages that have had an outbreak. If the codes are the wrong way around, click on the **Switch** button to swap the codes.

**Step 6:** If the herds or villages were selected using random geographic coordinate sampling, click on the Weighted? checkbox, and select the

field that contains the village or herd weights. If you selected villages using simple random sampling, you can leave this unchecked.

- Step 7:** Select the type of analysis that you are performing. If you are analysing the results of a single survey, and not doing any comparisons, select Single Group Analysis. If data from two groups are in the data file, you can select just one group by clicking on the Select Group check box.
- Step 8:** If you are comparing two groups, select Compare Two Groups. You then need to say which field the group identifier is stored in. Click the Grouping Variable box, and select the field. The program will check the codes in the file, and display them in the Group Codes box. You can use the **Switch** button to swap the codes from group 1 to group 2.
- Step 9:** If comparing two groups, you may want to adjust for seasonal differences (see below). Click the check box to adjust for differences.
- Step 10:** When all the fields have been set, click on the **Analyse** button, and the results will be displayed. You can print or save the results of the analysis.

### Adjusting for seasonal differences

Some diseases have a clear seasonal pattern, with more disease at one time of year than another. If this is the case, the time of year that the survey is conducted will have an effect on the length of time since the last outbreak. If the survey is conducted just after the peak season for disease outbreaks, many herds or villages may have experienced an outbreak in the last two or three months. If the survey is conducted 7 or 8 months after the peak season, the time since the last outbreak is longer. However, this is not because the disease situation is different. It is just because the outbreaks occur in a clear season.

If there is evidence that the disease outbreaks do have a seasonal pattern, you may have to correct for this, so the analysis is not misleading. You only need to do this if the two surveys being compared were conducted at different times of the year. If they were conducted at the same time, or during the same months in different years, no adjustment is necessary.

To adjust use the following procedure:

- Click on the Adjust for Season check box.
- In the Group Codes box, enter the month of the survey for groups 1 and 2.
- After setting up all the other fields, click the **Analyse** button to carry out the analysis.

### Complex analysis

The analysis described is usually adequate for most purposes. However, it is possible to do slightly more advanced analysis, to investigate the behaviour of the disease in more depth.

One problem when measuring herd- or village-level incidence rate is that not all herds or villages are the same size. A large village is likely to have more animals being brought in from outside the village, and this is one of the major risks for spread of disease. We could therefore expect that there would be more outbreaks in larger villages than smaller villages. When comparing two groups, if one group has more larger villages, and the other has mostly smaller villages, we would expect to see more outbreaks in the first group. This is due only to the size of the villages, not the overall level of disease. It is possible to adjust for differences in village size (livestock population), to give a fairer comparison of the amount of disease in the two groups. There are two ways to do this.

The first is use a different measure of time. Normally, we say that a village has been at risk of having an outbreak for a certain number of years. Instead, we could consider that each animal in the village has been at risk of getting disease for that period. The *animal-time* (number of animals in a village or herd, times the length of time since the last outbreak) will be greater for larger herds than small herds. Animal-time is an alternative time measurement that takes into account the size of the village livestock populations. Ideally, animal-time should be calculated on the basis of the average number of animals in the village during the time between the outbreak and the visit.

**Example:** A Retrospective Disease Outbreak Survey was carried out in 40 villages. The information collected was: 1) the time since the last outbreak (Time), 2) the pig population at the time of the visit (Vpop), and 3) the pig population at the time of the last outbreak (OBpop). In order to calculate the animal-time, the average number of animals in the village over that period was multiplied by the time since the last outbreak:  $((Vpop + Obpop)/2) \times \text{Time}$ .

Once animal-time has been calculated, the analysis can be repeated, using the animal-time field, rather than the time field. The results may show either an increased or decreased difference between the two groups, but either way, the difference will be taking variability in village livestock population into account, and therefore be somewhat more reliable.

The other approach to the analysis taking livestock population into account is to analyse the data using Cox's proportional hazards model. This is available only in sophisticated statistical software, and requires a good understanding of survival analysis and multiple regression models. If the software and technical expertise are available, village livestock population at the time of the visit and at the time of the outbreak can be included in the model.

## Interpretation of results

Unlike prevalence or traditional incidence rate estimates, the results of survival analysis are not expressed as a single number, but the estimate of a survival curve, or graph, which shows the disease experience of the entire study population. The Kaplan-Meier survival curve (named after the people that developed it) is a graph which shows the proportion of herds or villages that have 'survived' (not had an outbreak) for a particular time (measured backwards from the time of the survey). A population with fewer outbreaks or outbreaks occurring less often will have a greater proportion surviving for a longer period, so the curve will be closer to the top right of the graph. A population in which there have been recent outbreaks in most villages (from which we can imply that outbreaks are occurring frequently) will have a survival curve closer to the bottom left of the graph.

For both single and two-group analysis, the survival curves are displayed and can be printed. With experience, it is possible to interpret a survival curve. However, some summary measures which describe the curve are easier to understand.

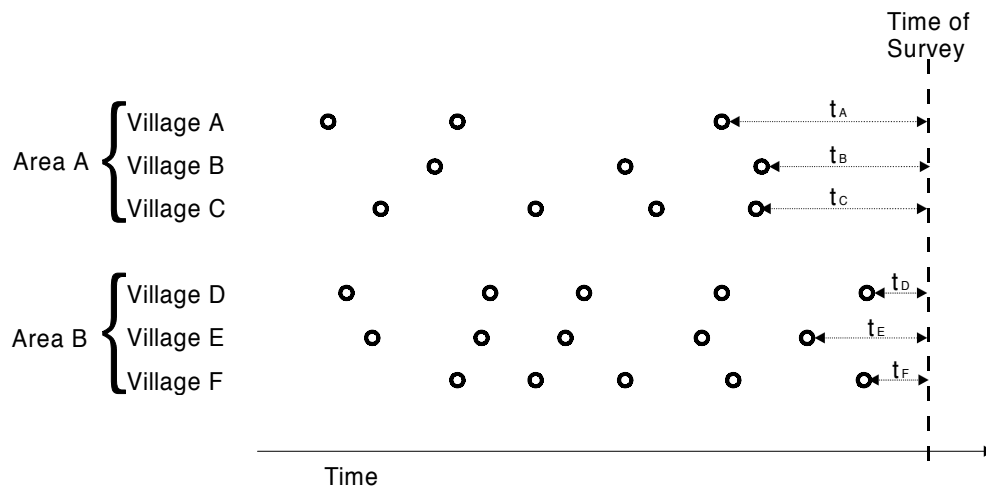
### Single group analysis

With single group analysis, the total number of observations and the total number of censored and uncensored observations are displayed. If weighted analysis is performed (because of the selection of herds or villages with random geographic coordinate sampling), these totals reflect the sum of the weights, rather than the true sum. The true totals are also displayed.

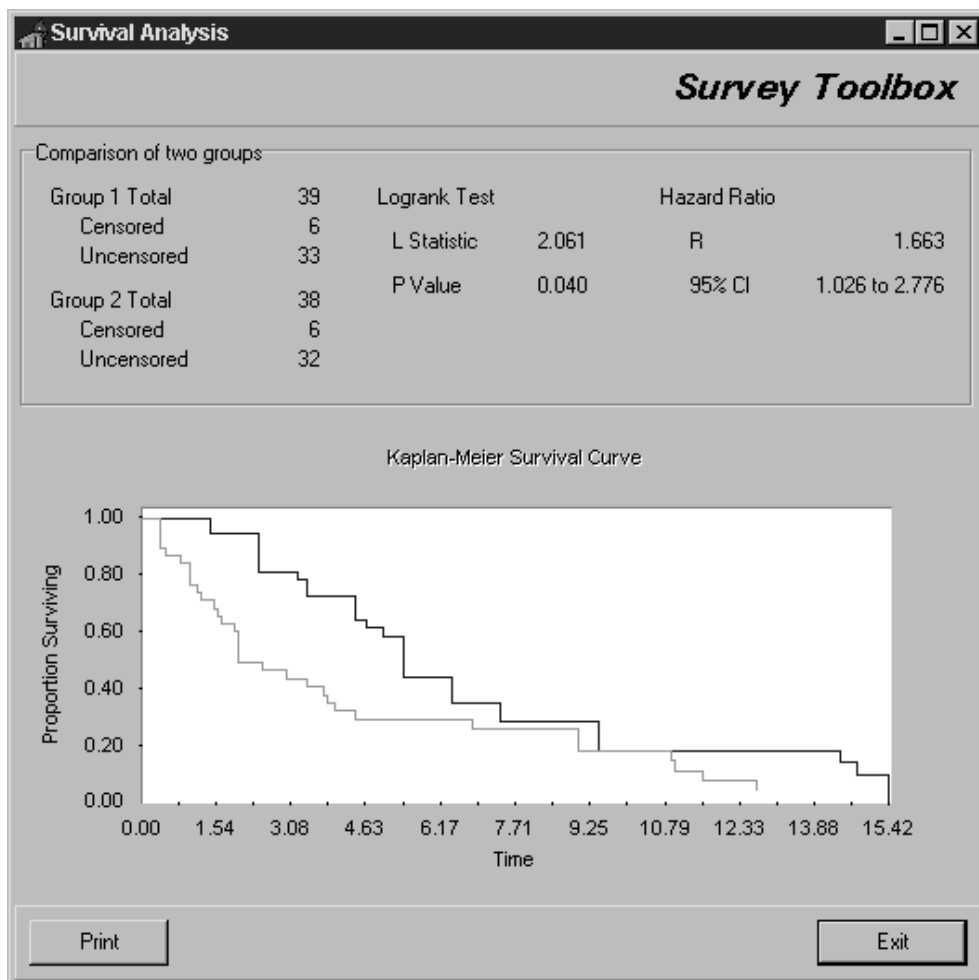
Median survival time

The *median survival time* is also shown. This is the time since which half of the herds or villages have not had an outbreak and half have. An estimate of the median survival time is used for sample size calculations (page 173).

The *mean* or *average survival time* is the average amount of time since the last outbreak for all villages. If the longest time is censored (hasn't had an outbreak, and is confident there hasn't been an outbreak for a long time), then the true mean can't be calculated. Instead, the time-limited mean is displayed, showing the mean of all times, up to a certain limit.



In the diagram, each circle represents an outbreak. The mean survival time for the villages in area A is given by  $(t_A + t_B + t_C)/3$ . The mean survival time for the villages in area B is given by  $(t_D + t_E + t_F)/3$ . The mean survival time for area A is longer than area B.



**Comparison of two groups**

The median and mean provide a basic description of the survival curve, and can be used for very simple comparisons. However, they really compare only the curve at one point. The reason for conducting a disease outbreak survey is generally to evaluate differences or changes in the disease situation. To achieve this, two groups of data must be compared.

When two groups are compared, a summary of the total number of observations in each group is presented.

Logrank test

This is followed by the *Logrank test*. This is a statistical test which compares the two survival curves to determine if there is any real difference between them, or if the apparent differences are simply due to chance.

The result of a Logrank test is given as a probability measure, called the P value. The P value is the probability that the two curves are in fact the same, or that any differences between the two groups is just due to chance. A small P value (less than 0.1 or 0.05) means that it is very unlikely that any difference is due to chance, so is interpreted as providing strong evidence that there is a real difference between the two curves. When the P value is small, there is said to be a statistically significant difference between the two curves.

The Logrank test calculates the probability that the two curves are the same. A low P value suggests that the curves are different.

Hazard Ratio

The Logrank test doesn't tell us what sort of difference there is, or how big the difference is. The key measure when comparing two groups is the *Hazard Ratio*. This is the ratio of the estimated hazard or risk of an outbreak in the two groups. A hazard ratio of 1 means that the risk of disease outbreaks in the two groups is about the same. A hazard ratio of 5 means that the risk of an outbreak in group 1 is 5 times greater than the risk in group 2. If group 1 is the population of pigs in a province two years ago, before the start of a vaccination campaign, and group 2 is the same population now, after 2 years of vaccination, a hazard ratio of 5 would mean that the risk of an outbreak amongst pigs was 5 times greater 2 years ago than it is now. This would provide strong evidence for an improvement in the disease situation. The hazard ratio is presented with a 95% confidence interval. Loosely speaking, we can be 95% sure that the true hazard ratio lies within this confidence interval.

The hazard ratio measures the ratio of the risk of an outbreak in two groups

## Analysis of two data sources

The Retrospective Disease Outbreak Surveys described above do not produce a traditional measure of incidence rate. This is because the total number of villages or herds that have suffered an outbreak in a fixed period of time can't be calculated (see page 28). This section describes a simple approach to estimating village or herd-level incidence rate which takes advantage of data that already exist.

The veterinary services in most countries maintain good records of disease outbreaks, particularly for important or notifiable diseases. This passively acquired data (compared with active surveillance – see page 14) can provide some picture of the disease situation. However, reporting is almost always incomplete, so any incidence rate estimate based on these data will be too low.

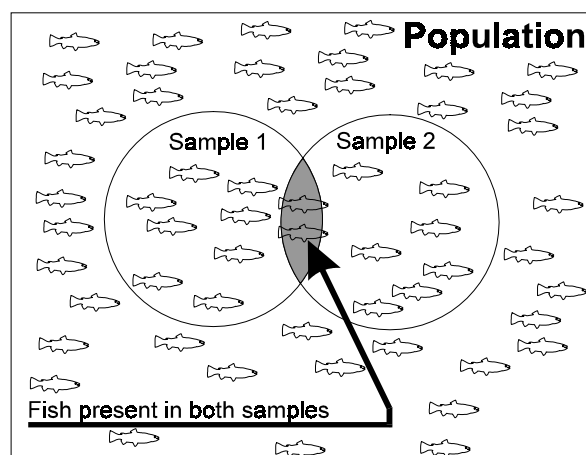
By combining this type of data with a separate, independent data source, it is possible to estimate how many outbreaks have been missed, and therefore what is the total number of disease outbreaks.

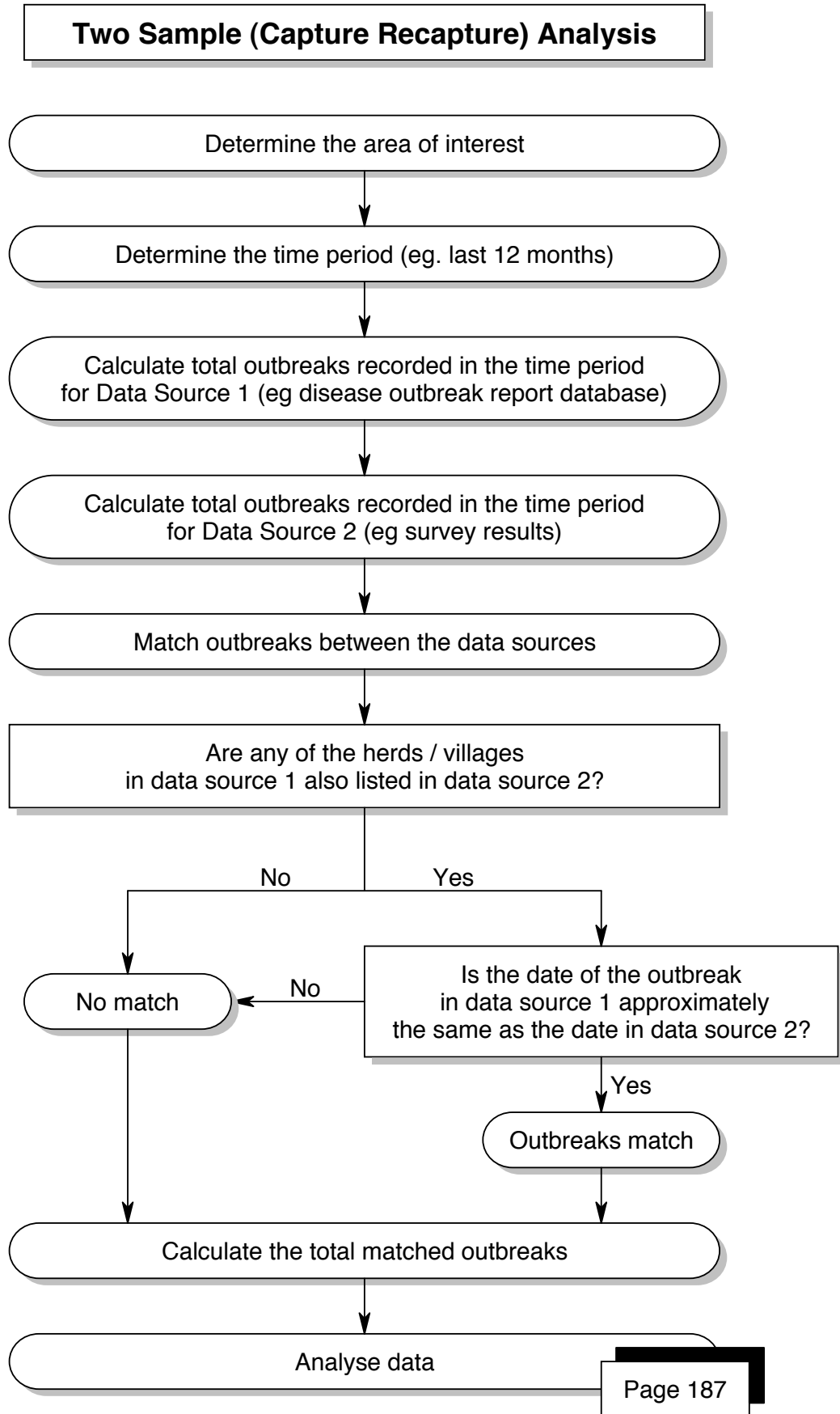
### Background

The technique is known as capture - recapture methodology. It was developed for wildlife studies, where it is very difficult to count every member of a population.

**Example:** A researcher wants to estimate the total number of fish in a lake. It is impossible to count all the fish so a different approach is used. First, for three days, the researcher catches as many fish as possible, and keeps a count of the total. Every fish that is caught is tagged, and then released into the lake. After three days, the fish are left to mix for 2 days. Then the researcher spends another three days catching as many fish as possible. Each time they catch a fish, they record if it has a tag or not. At the end, the researcher has three figures: the total number of fish caught the first time, the total caught the second time, and the total number caught both times (the tagged ones caught a second time).

This is shown in the diagram below. Sample 1 is all the fish caught the first time, sample 2 is the fish caught the second time, and the shaded area is the fish that were caught in both samples.





These figures can be used to estimate the total number of fish in the lake, using a simple formula:

$$Total = \frac{n_A n_B}{n_{AB}}$$

where  $n_A$  is the total in the first sample (A),  $n_B$  is the total in the second sample (B), and  $n_{AB}$  is the total occurring in both sample A and B.

If most of the fish caught the second time already had tags, that would mean that most of the fish in the lake must have been caught already, and the total is only slightly greater than the number of fish caught. On the other hand, if very few of the fish caught the second time had tags, that means that there are many more fish in the lake than were caught the first time, and the total population is much larger.

This same technique can be used to calculate the total number of village or herd disease outbreaks in an area over a specified time interval, and this can be used to calculate the incidence rate. The population is now not fish, but all village or herd disease *outbreaks* (not the villages that had them). The unit of interest is the outbreak (not the village). The first sample is provided by the records of the veterinary authorities of disease outbreaks. These records have 'captured' a certain proportion of all the outbreaks, but probably not all of them. A second source of information is used to 'capture' information about disease outbreaks in the same villages during the same period. The total number of disease records from the first and second sources, and the number of outbreaks that are in both sources can then be used to estimate the total number of outbreaks.

## Data sources

To use the technique, there must be two sources of data on village or herd outbreaks of the disease in question. Furthermore, these two sources must be independent. This means that they should be collected by different mechanisms, and the presence of a particular outbreak in one data source doesn't affect the probability that it will appear in the other.

The first data source usually comes from either disease outbreak reports, or the records of a diagnostic laboratory that is testing specimens from outbreaks. These are both good sources, that should already be available for analysis. The second data source usually comes from a village survey. To be valid, both data sources have to refer to the same time interval, usually one or two years. Only reports or specimens received in a clearly defined period should be analysed, and the survey should record outbreaks occurring only within that same period.

In order to be independent, the same people can't be responsible for collecting both data sources. For instance, the district veterinary officer is usually responsible for submitting outbreak reports. If the second data source comes from a survey in which district officers were asked about village outbreaks, the two sources would not be independent. This is because the chances of an officer remembering an outbreak for the survey are much higher if they have submitted a report on that outbreak.

The best type of second data source is to conduct a survey of a random sample of villages. In general, this can be combined with another survey, for instance a prevalence survey. If another data source already exists (such as the results of an agricultural census in which villages were asked about outbreaks of disease), then this could also be used, avoiding the need for any field data collection.

## Selecting the sample

If a survey is required, the villages or herds should be chosen by simple random sampling from a sampling frame. The **Random Village** program can be used (page 50), using the settings:

- Simple Random Sampling
- Without Replacement
- No stratification

If no sampling frame is available, it is not possible to reliably estimate the incidence rate of village outbreaks. This is because we need to know the total number of villages in order to calculate incidence rate, and without a sampling frame, this is not known.

## Data collection

The data collected are similar to that used in the village outbreak surveys described above. The difference is that instead of just the most recent disease outbreak, the aim is to collect information about the date of every disease outbreak that has occurred over a defined period (usually over the last one or two years). Keeping the time interval relatively short makes it easier for livestock owners to remember. However, if there has been more than one outbreak in the village or herd, livestock owners may find it difficult to reliably recall earlier outbreaks. The techniques described in Chapter 5 (page 118) can be used to improve the quality of data collected during a village interview.

## Data management

Three figures are required for analysis – the total number of outbreaks in the first data source (disease reports or laboratory submission records), the total number of outbreaks in the second source (usually the results of a survey), and the number of outbreaks that appear in both data sources. The first two figures are easily counted. To count the last, outbreaks identified in one source must be matched to outbreaks in the other.

### Matching disease outbreaks

Matching requires good information in both data sources on the village (a village name or identification number) as well as the date of the outbreak. The date reported on a laboratory submission or a disease report will usually not match the date recalled by livestock owners during an interview. This is because there are often small errors in memory, and also because specimens or reports may not be submitted at the beginning of the outbreak. When matching outbreaks, there has to be some flexibility in matching dates. This depends partly on the epidemiology of the disease. For instance, if a disease is very unlikely to occur more than once per year in a single herd, and outbreaks tend to last many months, it may be safe to assume that outbreak dates differing by as much as 6 months or more are, in fact, referring to the same outbreak.

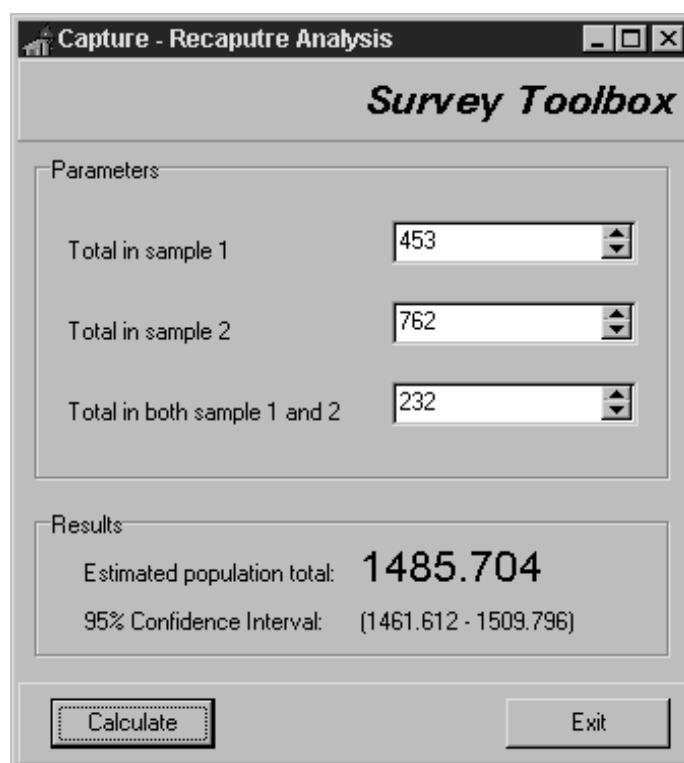
In general, some judgment will be needed for the matching process, and it is best done by hand. Some of the matching may be done with the help of a computer, but even this should be checked by hand.

## Data analysis



Once the totals have been calculated, you can use the **CapRecap** program to analyse the data. Start the program using the Windows Start menu, select Programs, Survey Toolbox, and CapRecap.

- Step 1:** Enter the total number of outbreaks identified in sample 1.
- Step 2:** Enter the total number of outbreaks identified in sample 2.
- Step 3:** Enter the number identified in both samples.
- Step 4:** Click on the **Calculate** button.
- Step 5:** The results show the estimated total, and a 95% confidence interval.



Section	Field	Value
Parameters	Total in sample 1	453
	Total in sample 2	762
	Total in both sample 1 and 2	232
Results	Estimated population total:	1485.704
	95% Confidence Interval:	(1461.612 - 1509.796)

The incidence rate is equal to the total number of outbreaks over the total number of villages multiplied by the time period (page 28). You can use the same formula with the limits of the 95% confidence interval to calculate a 95% confidence interval for the incidence rate.

**Example:** In a two sample study of foot and mouth disease outbreaks over a two-year period, 145 outbreak reports had been received by the veterinary authorities from a province containing 1293 villages. A survey of 85 villages was conducted, of which 47 had experienced outbreaks in the same two-year period. When matched, 36 outbreaks appeared in both data sources. Using the CapRecap program, the estimated total is 188 outbreaks, with a 95% confidence interval from 163 - 213. The estimated incidence rate is therefore  $188 \text{ outbreaks} / (1293 \text{ villages} \times 2 \text{ years}) = 0.073$ , or 7.3 outbreaks per 100 villages per year.

# 9

## Surveys to Demonstrate Freedom from Disease

### Contents

Chapter 1: Introduction

#### Part I: Background to Disease Surveys

Chapter 2: General Principles of Animal Disease Surveillance

Chapter 3: Sampling

Chapter 4: Principles of Data and Specimen Collection

Chapter 5: Village Interviews

Chapter 6: Computerised Data Management and Analysis

#### Part II: Survey Design and Analysis

Chapter 7: Prevalence Surveys

Chapter 8: Incidence Rate Surveys

Chapter 9: Surveys to Demonstrate Freedom from Disease

Herd or village surveys

Sample size calculation

Selecting the sample

Data analysis

Large area surveys

Sample size calculation

Sample size for minimum cost

First- and second-stage sampling

Data analysis

Other issues

Combining tests

FreeCalc options

#### Part III: Notes for Trainers

Chapter 10: Guide for Trainers

Chapter 11: Lesson Plans

Chapter 12: Activity Sheets



Consider these three different situations:

- When a disease control program is used amongst intensive farms (e.g. pigs or chickens, or intensive dairy or beef farms) one approach has been to use herd accreditation schemes. These schemes, targeted at farms which supply animals to other producers, involve testing of some farms to provide a guarantee that the herd is free from disease. This means that other producers can buy from that farm without the risk of introducing disease. As a result, the accredited farms are able to ask higher prices for their animals.
- Many national disease control and eradication programs are based on mass vaccination, and later test-and-slaughter. The aim of these programs is to eradicate the disease from the entire country. When this is achieved, all the costs of vaccination and other control measures, as well as losses due to the disease can be saved. However, at the end of an eradication program, the authorities need to be confident that the disease has indeed been eliminated.
- The development of export industries is an important way for developing countries to obtain foreign exchange and develop their economies. The export of livestock or livestock products is one area where agricultural countries have potential to develop an export industry. However, under the rules of the World Trade Organization, an exporting country may be asked to show that there is no risk of spreading livestock diseases to the importing country.

In each of these three situations, it is necessary to demonstrate that a population (either a single herd, village, district, province, or the whole country) does not have a particular disease. This chapter describes survey techniques that can demonstrate that a population is free from disease. In a way, this is the same as conducting a prevalence survey, and hoping that the prevalence is 0, but the theory behind the two types of surveys is quite different.

There are two problems when trying to show that there is no disease in a population. The first is that it is very hard to prove. If a herd has 342 animals, it is only free from disease if none of those animals has the disease. It is possible (although perhaps unlikely) that just a single animal is infected. If we take a sample from the herd and test the animals in the sample, we might, by chance, select the one infected animal, and be able to conclude that the herd is not free from disease. However, it is possible that we might not select that animal in the sample, and think that the herd is free from disease when it is not. The larger the sample size, the higher the chance that we will pick that one infected animal, but there is still a chance that we will miss it. The only way to be completely sure, is to test every animal in the herd. When trying to show that a herd of 342 animals is free from disease, this may be expensive, but not too difficult. When trying to prove that a country with 8 million animals is free from disease, it is impossible to test every animal.

The second problem relates to laboratory tests. In Chapter 2, the concepts of sensitivity and specificity of a test were discussed (page 33). Very few laboratory tests are perfect, and most make a small number of mistakes, calling some diseased animals non-diseased, and vice versa. This means that if you test a large number of animals, it is difficult to interpret the results. If there is one positive test result from 342 animals, is this animal really infected, or is it just that the test has given the wrong result? Are all the animals that tested negative really disease free, or are some of them diseased, and the negative test result was wrong?

A typical test may have a sensitivity of 95% and a specificity of 99%. The sensitivity means that for every 100 diseased animals tested, 95 of them will give a

positive test result, but 5 will give a false negative result. The specificity means that for every 100 disease-free animals, 1 will give false positive test results. This means that even if we test all the animals in a herd or all the animals in the entire country, we still can't be sure if they are all free from disease. Even if there is no disease, we will get some positive test results, because of the test producing false positives. If disease is present, we may also get some false negative test results, and miss the diseased animals.

These two problems mean that it is impossible to prove that a population is free from disease, as there is always the chance that we have missed an animal or that the test result is wrong.

It is impossible to *prove* that a population is free from disease.

Although we can't *prove* that a population is free from disease, if we test enough animals and take the performance of the test (sensitivity and specificity) into account, we are able to show that it is *very unlikely* that the population has infected animals. Surveys to demonstrate freedom from disease do not provide a guarantee, but they are able to say that the chance of the population having diseased animals is smaller than some acceptable level (say 5% or 1%).

If there could be a very small number of infected animals, then it is much harder to find them, and a bigger survey is needed. If the likely number of infected animals is high, then it is easier to find them in a survey, so a smaller sample size is needed. The results of surveys to demonstrate freedom from disease are therefore expressed as the chance that the number of infected animals is equal to or greater than some low value.

**Example:** A outbreak of rinderpest in cattle has occurred in a previously free area. Four neighbouring villages were affected, and the entire cattle population of the villages was slaughtered to eradicate the disease. All villages in the surrounding area (10 km radius) are being examined for clinical signs and serological evidence of infection with the disease. None of the cattle have been vaccinated, so if rinderpest did get into one of these villages, it is likely that it would spread very quickly, and affect a high proportion of the animals in the village, probably over 50%. It is very unlikely that only 10% or fewer animals would be affected by such a contagious disease. During the surveillance after the outbreak, surveys are carried out in each of these surrounding villages, to demonstrate that they are free from the disease. As it is impossible to prove that no animals had been infected at all, the surveys were designed to show that the chance that 10% or more of the animals had been infected with the disease was very low (less than 1%). If fewer than 10% were infected, it was less likely that they would be identified in the survey. Using this survey design all villages were declared free from disease, even though it was possible that some had as many as 10% of animals infected. This didn't pose any risk, because the highly contagious nature of rinderpest meant that if the disease entered a village, much more than 10% of animals would become infected.

Minimum expected prevalence

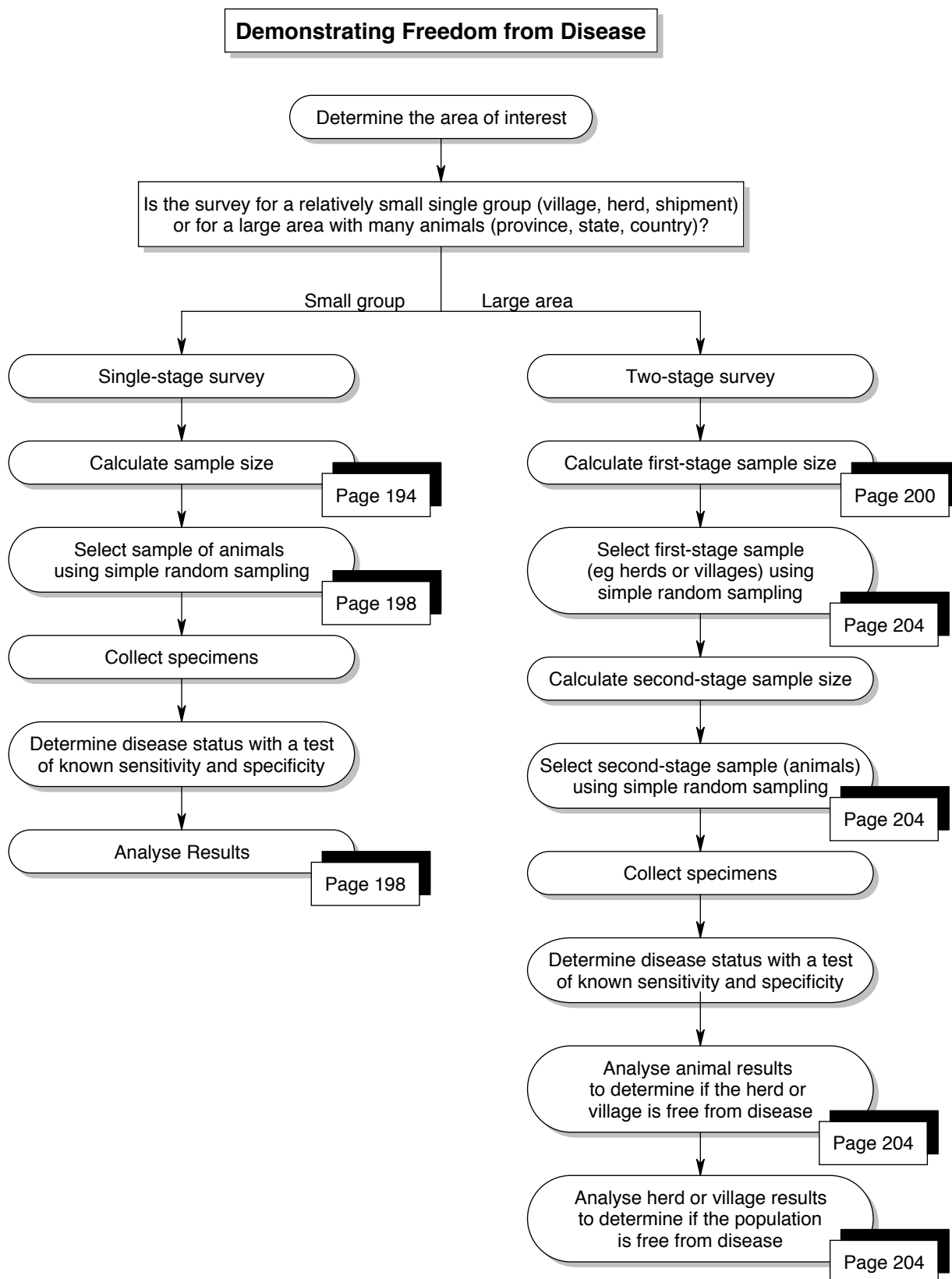
This example demonstrates the concept of the *minimum expected prevalence*. This is the minimum prevalence expected if a contagious disease was present in a herd. When conducting a survey, this is the lowest disease prevalence that the survey can be expected to reliably identify. If the disease is present in the population, but at a lower level, than the prevalence specified, then the survey may not be able to identify it. This level is based on a knowledge of the epidemic behaviour of the disease. For example, FMD may have a minimum expected prevalence of 30%, while Johnes disease may have a minimum expected prevalence of 5% or less.

Maximum acceptable prevalence

For diseases which are not highly contagious, the minimum expected prevalence may also be thought of as the *maximum acceptable prevalence*. If the level of disease in the population is less than this prevalence, then it is small enough not to worry about. This level always has to be greater than 0, because unless we have a perfect test and are able to test every single animal, we can't prove that the prevalence is 0.

As with any survey, it is easier to find diseased animals if the prevalence is higher. The measure of disease should therefore be that measure which gives the highest prevalence. As discussed in Chapter 2 (page 31), antibody status lasts much longer than clinical disease, and therefore has a higher prevalence. This is why it is much more common to base a survey to demonstrate freedom from disease on the use of serological tests rather than clinical examination of animals.

This chapter describes two different survey designs for demonstrating freedom from disease. The first is a survey conducted in a small population, such as a herd, village, or intensive farm. The second design uses a two-stage sampling scheme for surveys of larger areas (districts, provinces, states or countries).



## Herd or village surveys

The main steps in conducting a single-stage survey to demonstrate freedom from disease are:

- Step 1:** Determine what question is being asked. This involves specifying the minimum expected prevalence (maximum acceptable prevalence), and the probability levels which determine how confident we are of the results.
- Step 2:** Calculate the sample size.
- Step 3:** Select the sample using simple random sampling.
- Step 4:** Collect specimens.
- Step 5:** Process specimens ready for analysis.
- Step 6:** Send the specimens to the laboratory.
- Step 7:** Check the data for completeness and accuracy.
- Step 8:** Analyse the data to determine the probability that, if disease is present in the population, the prevalence is less than the maximum acceptable prevalence.
- Step 9:** Report the data.

## Sample size calculation

Calculation of the sample size for a survey to demonstrate freedom from disease is based on several different values.

### Test performance

The performance of the test being used plays an important role in the sample size. It is expressed in terms of sensitivity and specificity (page 33). If the test is not very reliable (either sensitivity or specificity or both are relatively low) then the sample size will need to be much higher. If there is a choice, the test with the best sensitivity and specificity (but particularly high specificity) should be used. See Combining Tests (page 205) for advice on ways to improve the specificity of a test.

Unfortunately, precise estimates for sensitivity and specificity are not always available for many tests. Another problem is that these measures vary somewhat depending on the population being tested, so that values published from a study in one part of the world may not be completely valid for the population being surveyed. If you don't know what the sensitivity and specificity of tests are, here is what you should try:

- Step 1:** Ask your laboratory people if they have conducted any studies in the local population to evaluate test performance.
- Step 2:** Ask them if they know of published figures based on other populations.
- Step 3:** Search the literature for published studies on the test. If more than one study is found, use the one that most closely matches your population.
- Step 4:** Contact leading experts with experience in using the test, and ask them for their estimates of the test performance.
- Step 5:** Organise a small study in the local population to measure test performance yourself.

If no reliable published figures are available, then estimates may be used. However, if the test is going to form the basis of important surveys, or be used as part of an ongoing control or eradication program, it is very important that its performance in the local population is well understood. It may be worthwhile to conduct a study to evaluate the performance of the test. This involves testing a number of animals that have a known status (some truly disease positive, and some truly disease negative), and directly calculating the sensitivity and specificity. Consult an epidemiology text or epidemiologist for advice on conducting this sort of study. Also, when finished, make sure that you publish the results, so that others can benefit from your work.

### Population size

You need to know the size of the population. Smaller populations require somewhat smaller sample sizes.

### Minimum expected (maximum acceptable) prevalence

The choice of this figure (explained earlier) is based on a knowledge of the disease, or practical limitations. The larger the prevalence chosen, the smaller the sample size and the easier the survey. With highly contagious diseases that are likely to spread quickly, it is safe to choose quite a high prevalence. However, with other diseases, the prevalence in a herd, if present, may be very low. Detecting a disease at low prevalence can be very difficult, requiring a large sample size. In the end, you may have to be content with a prevalence level that is based on the largest sample that can be afforded or practically tested, rather than on the biology of the disease.

### Type I and II error

Type I error

The *Type I error* (also called the  $\alpha$  (alpha) level) is the probability that the results of the survey will conclude that the population is not diseased when in fact it is. This is also known as the significance of the results, and is equal to 1 minus the level of confidence. The Type II error ( $\beta$ , beta) is the probability that the survey will conclude that the population is diseased, when in fact it is not. This is equal to 1 minus power. By convention, the Type I error is usually 0.05, and the Type II error is either 0.1 or 0.05. These can be adjusted to any value, depending on the importance of that type of error.

Type II error

**Example:** A piggery is being tested as part of a herd accreditation scheme. If the herd is found to have disease, then the owner will not be allowed to sell animals other than for slaughter. The owner is therefore very keen to make sure that the survey doesn't make a type II error, or conclude that the farm is infected when it is not infected. The owner would want to set the type II error level to be very low, to minimise the chance of this mistake. On the other hand, a client of the farm that buys pigs for breeding does not want to receive diseased pigs. The client would want to make sure that if the survey indicated that the farm was free from disease, this is in fact true. The client would want a very small type I error level to ensure that the farm is not declared free when it is actually infected.

The final decision on the error levels depends on a compromise between competing needs. The repercussions of the possible mistakes need to be taken into account as well, as in the following example.

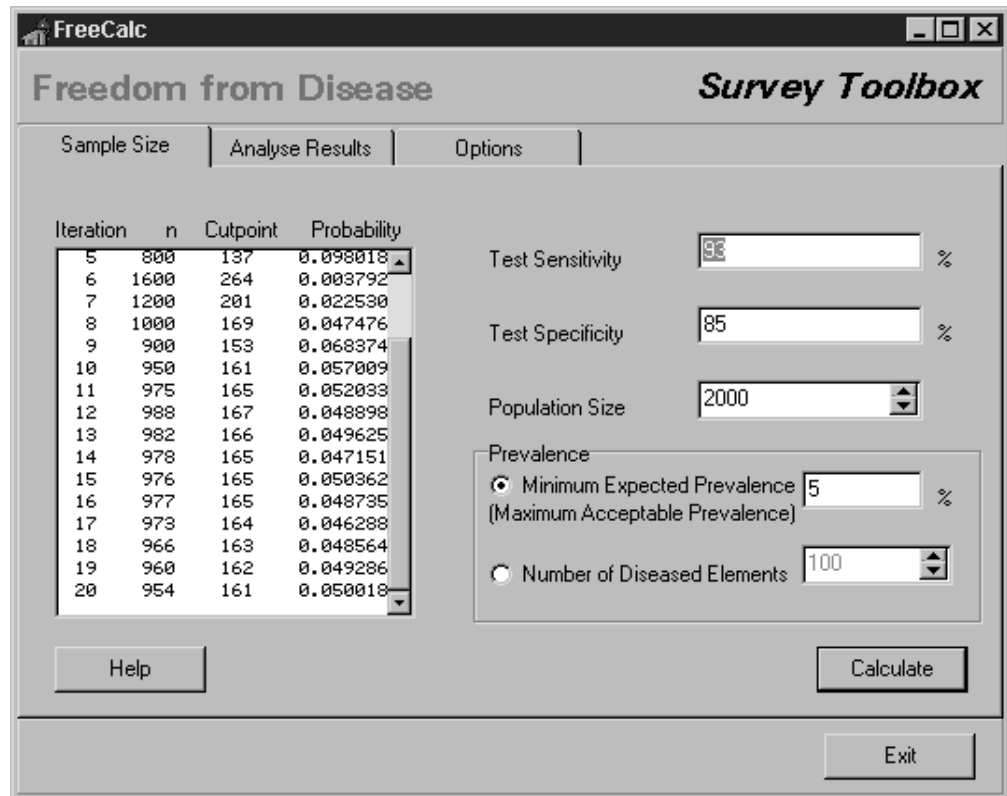
**Example:** After an outbreak of FMD, villages near the outbreak are being monitored. A survey is conducted in each village to determine whether it is free from disease. If it is free, quarantine is removed, and the village can trade again. If it is not shown to be serologically free, it is kept under quarantine. (If clinical cases are detected, the population of the village is slaughtered.) If the survey makes a type I error, and concludes that the village is free from disease when it does have disease, the consequences could be very bad. The disease could spread from that village to other parts of the country and the outbreak could start again, causing the death of many animals, and enormous expense. The probability of a type I error should be kept very very low. If a type II error is made, then the village is held under quarantine for a bit longer. This is inconvenient for the livestock owners in that village, but doesn't have a huge impact. The type II error probability could therefore be quite high.

### Calculating the sample size

When all these issues have been considered, you can use the **FreeCalc** program to determine what sample size is needed. Use the Windows start menu, select Programs, Survey Toolbox, then FreeCalc. FreeCalc is a program both for sample size calculation, and analysis of survey results.

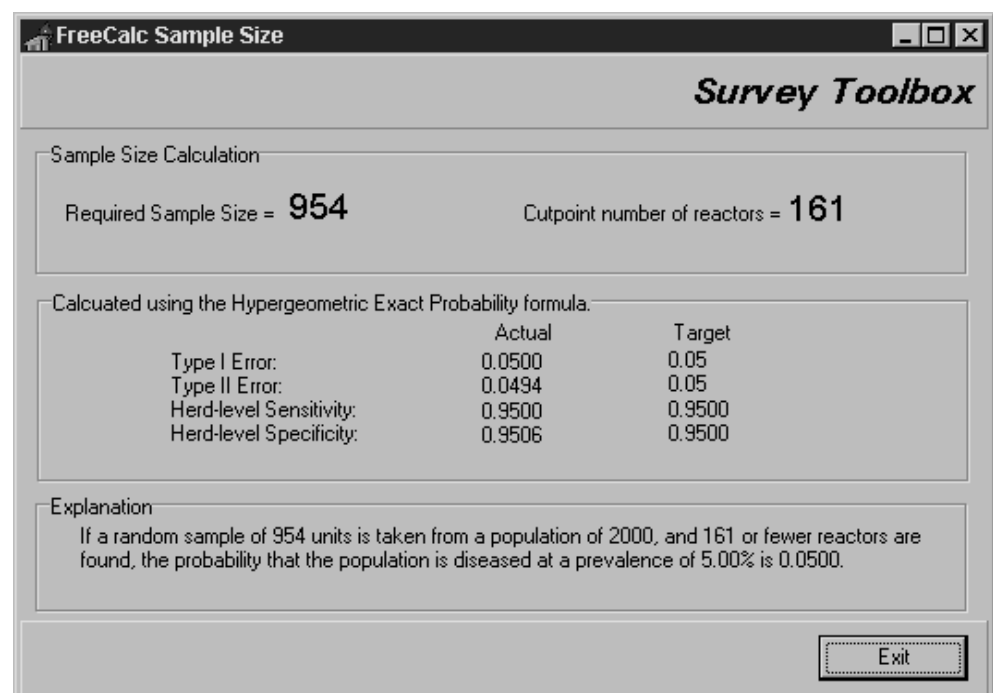


- Step 1:** Click on the Sample Size tab at the top of the window.
- Step 2:** Enter the test sensitivity and specificity as a percentage.
- Step 3:** Enter the size of the total population in the Population Size box.
- Step 4:** In the Prevalence box, enter the minimum expected prevalence (the maximum acceptable prevalence). This can either be entered as an estimate of the prevalence (a percentage) or as a direct measure of the number of diseased animals in the population. Click on the ratio button to choose if you want to enter the prevalence or the number of diseased animals, then type in the number. The equivalent value will be shown in the other box.
- Step 5:** Click on the Options tab at the top of the window.
- Step 6:** In the Formula for Calculation box on the left, you can usually leave this as Modified Hypergeometric Exact. The different formulae are discussed on page 206.
- Step 7:** In the Parameters box, enter the Type I and Type II error levels that you want to use. If you are unsure, leave them both at 0.05. (The other values in the box are discussed on page 207.)
- Step 8:** Click on the Sample Size tab again, and click on the **Calculate** button. As the calculation is taking place and the program is searching for the best sample size, the intermediate results are displayed in the box on the left.
- Step 9:** When finished, the results are displayed in a window.



The results show the sample size required to be confident that, if the disease is present, it is present at a level lower than that specified for maximum acceptable prevalence.

The results also show the ‘cutpoint number of reactors’. This is the number of animals that can return positive test results, and still let us conclude that the herd is free from disease. In other words, these are considered to be false positive test results. If we get fewer test-positive animals in the survey, we can still conclude that the herd is free from disease, but if there are more than this number, the evidence for being free from disease is not as strong.



## Selecting the sample

When the sample size has been calculated, you are ready to conduct the survey. Animals must be chosen using simple random sampling. If a sampling frame already exists, you can use the manual technique described on page 41, or, if it is available on disk, you can use the **Random Village** program (page 50). If no sampling frame exists, one must be made. If conducting a survey in a village, use the technique described for building a livestock sampling frame on page 115. If on a farm, it may be possible to select a systematic random sample, as long as there are some livestock handling facilities (page 44). See Chapter 3 for a full discussion of random sampling.

## Data analysis

When the sample has been examined, or specimens analysed in the laboratory, the key pieces of information required are the total number tested (the sample size), and the number that gave positive test results. You can then use the FreeCalc program to analyse the results

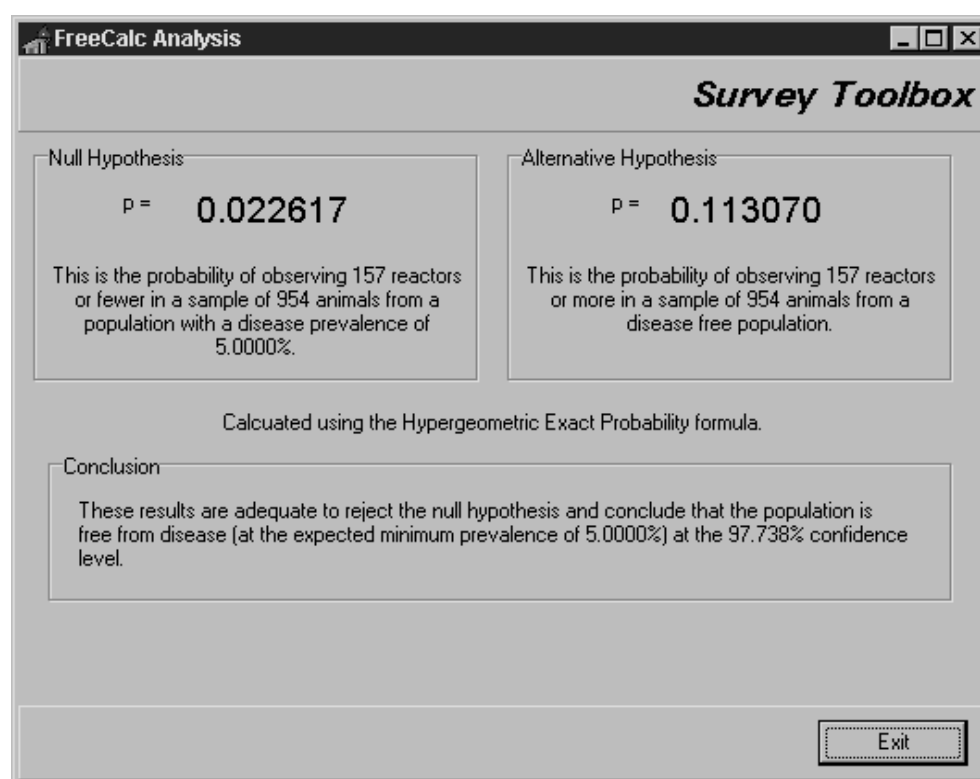


- Step 1:** Start the FreeCalc program (as described earlier) and click on the Analyse Results tab at the top of the window.
- Step 2:** Enter the test sensitivity, specificity, population size and prevalence as described for sample size calculation.
- Step 3:** On the left, enter the Survey Sample Size.
- Step 4:** Enter the Number of Positive Reactors from the test results.
- Step 5:** Click on the Options tab, and check the Type I and Type II error levels, to make sure they are correct.
- Step 6:** Return to the Analyse Results tab, and click on the **Calculate** button.
- Step 7:** A window is displayed with the results of the analysis.

The screenshot shows the 'FreeCalc' window titled 'Freedom from Disease Survey Toolbox'. The 'Analyse Results' tab is active. The 'Survey Sample Size' is 954 and the 'Number of positive reactors' is 157. The 'Test Sensitivity' is 93%, 'Test Specificity' is 85%, and 'Population Size' is 2000. Under 'Prevalence', the 'Minimum Expected Prevalence' radio button is selected with a value of 5%, and the 'Number of Diseased Animals' is 100. Buttons for 'Help', 'Calculate', and 'Exit' are visible at the bottom.

The results are displayed in terms of probabilities of the null and alternate hypothesis. The probability of the null hypothesis is the probability of observing this many reactors or fewer, if the population was diseased at a level equal to or greater than the specified prevalence. If this probability is small, we can conclude that it is very unlikely that the population is diseased. If the probability is large, then there is not enough evidence to conclude that the population is free from disease.

The probability of the alternative hypothesis is also shown. If this is small, then it is very unlikely that the population is free from disease. If it is large, then it is consistent with there being no disease in the population. If both the null and alternative probabilities are small, it suggests that the population is not free from disease, but the prevalence is less than the minimum expected prevalence specified. The conclusion is written at the bottom of the window.



## Large-area surveys

In Chapter 3, the problem of drawing a simple random sample from a large population was discussed. For surveys at the district, state, or national level, it is not possible to draw up a sampling frame which lists every animal in the entire population. In these cases, two-stage sampling (page 64) is much more practical. At the first stage, we need only a sampling frame that has all the herds or villages listed. For those herds or villages that are chosen, we can then build a sampling frame of animals.

The same approach is used for surveys of large areas to demonstrate freedom from disease. Two-stage sampling has the added advantage that it is able to account for *disease clustering*.

Disease clustering

Disease is not usually spread evenly through the population, but tends to occur in clumps or clusters. For instance, with Newcastle disease in chickens, most of the

animals in most farms and villages are not affected, and the overall prevalence in the population is very low. However, during an outbreak, a small number of villages or farms have a large number of birds affected, and the prevalence in those farms or villages is very high.

Two-stage sampling allows us to account for the fact that if a disease is present, very few villages or herds may be affected, but those that are usually have relatively high levels of disease. This is taken into account by specifying prevalence at two levels, the prevalence of infected farms, herds, or villages at the first level, and the prevalence of infected animals on farms at the second.

The procedure for conducting a two-stage survey to demonstrate the freedom of a large area from disease is as follows:

- Step 1:** Determine what question is being asked. This involves specifying the minimum expected prevalence (maximum acceptable prevalence) both amongst herds or villages and amongst animals within a herd or village.
- Step 2:** Calculate the first stage sample size (number of herds or villages).
- Step 3:** Select the sample using simple random sampling
- Step 4:** Build a sampling frame of animals within the selected herds or villages
- Step 5:** Calculate the sample size depending on the village or herd population.
- Step 6:** Select animals using simple random sampling.
- Step 7:** Collect specimens.
- Step 8:** Process specimens ready for analysis.
- Step 9:** Send the specimens to the laboratory.
- Step 10:** Check the data for completeness and accuracy.
- Step 11:** Analyse the data from each village or herd, and determine whether it is classified as diseased or non-diseased.
- Step 12:** When every village or herd in the sample has been classified, analyse these herd-level results to determine if the entire population is diseased or non-diseased.
- Step 13:** Report the data.

## Sample size calculation

The sample size calculation has two steps - first calculate the number of herds or villages required, and then the number of animals from each herd or village.

In addition to all the measures required for sample size calculation for small populations, there is another important measure to be considered. For each herd that is tested, we need to decide whether the herd is to be classified as diseased or non-diseased. We analyse the results from animals within the herd to make this decision. However, this decision might be wrong. The probability of making a wrong decision is given by the Type I and Type II error levels.

When making a decision (or diagnosis) about an entire herd or village, the procedure (testing a sample of animals from the village) can be thought of as a test. This is often called *herd testing*. Just like any other test, its performance can be measured by sensitivity and specificity. The sensitivity of a herd test is the probability that a diseased herd will be classified as diseased. This is equal to 1 - Type I error level. If, when testing animals within a herd, we set the Type I error level

to 0.05 or 5%, then the sensitivity of the herd test is 0.95 or 95%. In the same way, the specificity of the herd test is equal to  $1 - \text{Type II error level}$ .

When we set the Type I and II error levels for determining the sample size within a single herd, we are in fact setting the sensitivity and specificity of the 'test' for that herd. With this in mind, we can go ahead and determine the sample sizes needed for two-stage sampling.

To calculate the number of villages or herds that need to be selected at the first stage, use the **FreeCalc** program:



- Step 1:** Click on the Sample Size tab.
- Step 2:** Under test sensitivity, enter the a value which is  $1 - \text{Type I error}$  used for selecting individual animals. For instance, if a Type I error level of 0.05 is used when selecting individual animals to test, this means that the herd test sensitivity is 95%.
- Step 3:** Enter the specificity in the same way. If the Type II error level for selecting individual animals is 0.1, then the herd-level specificity is 90%.
- Step 4:** Enter the size of the total population. This is the total number of herds or villages in the area being studied (not the total number of animals).
- Step 5:** In the prevalence box, enter either the prevalence or total number of disease-positive villages or herds which represents the maximum acceptable prevalence. Regardless of the disease in question, if the population is thought to be free from disease, then the proportion of positive villages must be set to a relatively low value (usually less than 5%). This means that the number of villages or herds that need to be tested will often be quite high.
- Step 6:** Click on the Options tab, and check the Type I and II error levels. These now measure the probability that the entire survey will make an error. See the discussion on page 195.
- Step 7:** Return to the Sample Size tab, and click **Calculate**.

The results will indicate how many villages or herds need to be visited. The procedure for selecting animals from each of the selected villages is the same as that described above for surveys of small populations on page 194. At this level, the sensitivity and specificity are measuring the performance of the laboratory test. The Type I and II error levels, which determine the herd test sensitivity and specificity, are set to the values mentioned in steps 2 and 3 above. The population size refers to the total number of animals in that village or herd. Because the sample size changes depending on the total population, and the population of every village is likely to be different, the second-stage sample size should be calculated for each village separately. If a portable computer is not available, the sample sizes for every possible village size can be calculated before the fieldwork, and written in a table for use in the field.

## Sample size for minimum cost

When using two-stage sampling, a survey can produce a result of the same accuracy by using a variety of combinations of first and second-stage sample sizes. For instance, if a small number of villages is selected and many animals from each village are tested, it is possible to get the same accuracy as if many villages are tested,

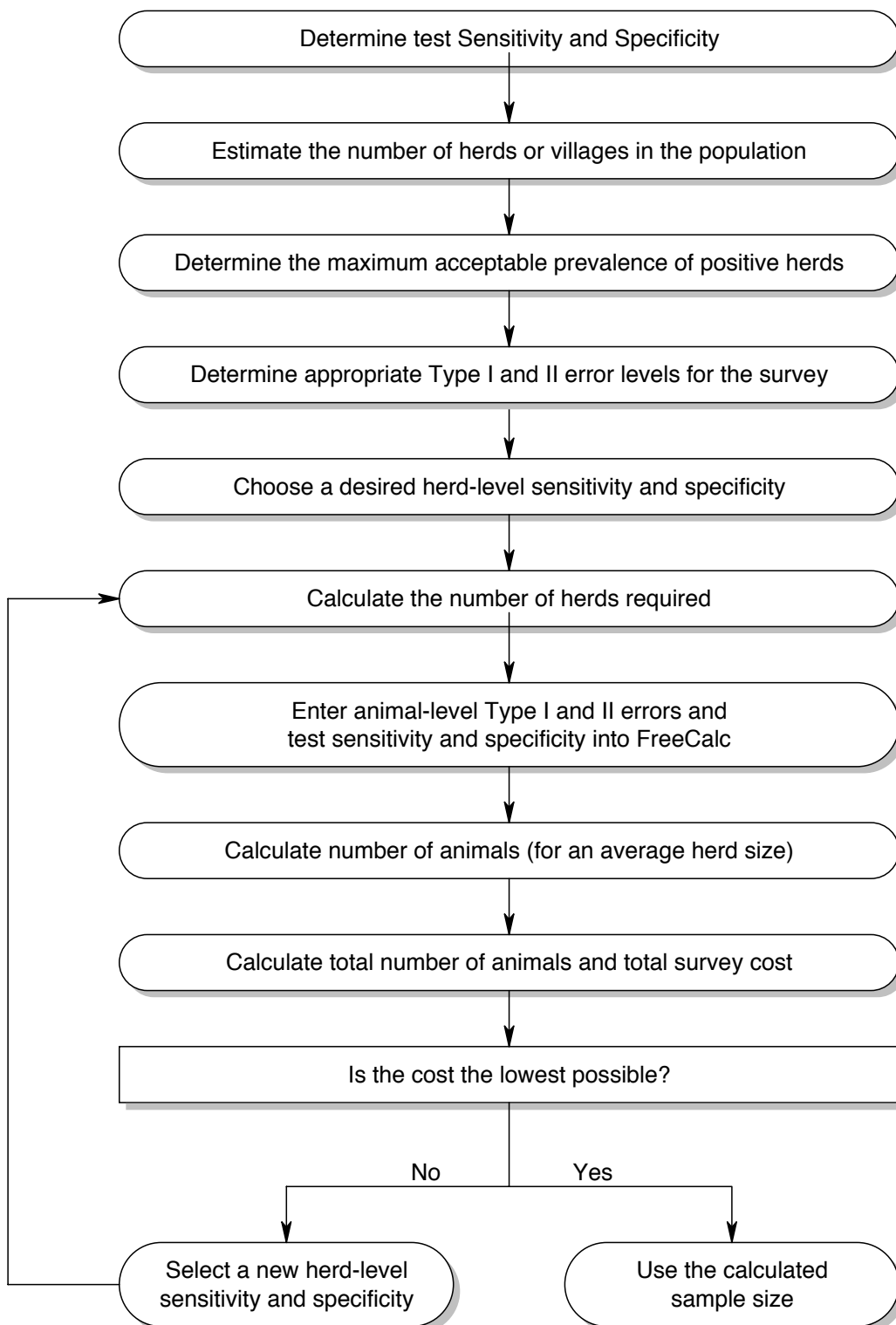
and only a small number of animals are tested from each village. By changing the Type I and II error levels used for selecting the sample size for the second stage (testing animals within a village or herd), we are also changing the sensitivity and specificity of the herd test (used when selecting villages or herds at the first stage). This enables us to produce a variety of different sample size combinations, all of which will provide the same level of evidence for freedom from disease.

This flexibility is one of the advantages of two-stage sampling, because not all the combinations will cost the same. The overall cost depends on how much it costs to test a single animal, and how much it costs to test a single herd or village. This was discussed in Chapter 7 (page 154). For prevalence surveys, there is a formula, used in the Prevalence program, to determine what is the cheapest combination. For surveys to demonstrate freedom from disease, the complexity of the calculations means that it is not possible to use a formula to work out the best combination.

Instead, it can be done using trial and error with the FreeCalc program. Use the following procedure to calculate the best combination of first and second-stage sample sizes:

- Step 1:** Determine the basic measures that we can't change. This includes the sensitivity and specificity of the laboratory test, the population of villages or herds (first-stage population size), an estimate of the average animal population of the villages or herds, the maximum acceptable prevalence of the disease amongst villages (first stage) and animals (second stage), and the overall Type I and Type II error levels for the survey (used when calculating first-stage sample size). You also need to know the cost of testing a single animals, and the costs associated with sampling a single village (see page 154).
- Step 2:** Pick starting values for the herd test sensitivity and specificity. The higher these values are, the fewer villages need to be tested, and the more animals need to be tested in each village. If they are very high, there may not be enough animals in some villages to achieve this level. In general, try to make the specificity as high as possible.
- Step 3:** Calculate the number of herds needed using the selected herd test sensitivity and specificity.
- Step 4:** Now use the same figures to calculate the second stage sample size. Set the Type I error to  $1 - \text{Sensitivity}$ , and the type II error to  $1 - \text{Specificity}$ . Change the sensitivity and specificity to those of the laboratory test, the population to the average herd or village size, and the prevalence to the maximum acceptable or minimum expected prevalence within the herd.
- Step 5:** Calculate the number of animals that need to be tested.
- Step 6:** Using the number of herds and villages, and the number of animals, calculate by hand the total cost of the survey, based on the cost estimates, and record the result.
- Step 7:** Now go back to calculating the first-stage sample size, but change either the sensitivity or specificity or both. Repeat the calculations in steps 3 to 6, and record the sample sizes and total cost of this alternative combination.
- Step 8:** Continue testing new values until you find the one that gives the cheapest cost.

**Calculating Optimal Sample Sizes for Two-Stage Surveys to Demonstrate Freedom from Disease**



## First- and second-stage sampling

At the first stage, villages or herds must be selected using simple random sampling from a sampling frame. Use a random number table to select from a written sampling frame or the **Random Village** program (described on page 50) to select from a sampling frame on computer disk. When using the program, set the Sampling Type to Simple Random, and select villages or herds without replacement. The sample may be stratified if convenient.

At the second stage, animals must again be chosen using simple random sampling, or, if possible, systematic random sampling. For selecting animals in a village, the technique described in Chapter 3 (page 55) can be used, either with the Random Animal program, or manually using a random number table. If a computerised sampling frame of all animals in a herd is available, you can again use the Random Village program, with the same settings.

## Data analysis

The data are analysed in two stages. First, the data from each selected herd or village are analysed to provide a herd result, indicating that the herd is either diseased or non-diseased. Use the same approach described on page 198 for small population surveys, and record the status of each herd or village. When analysing the results from each herd, be sure to enter the correct population size for that herd, and the correct Type I and II error levels selected for the second stage of sampling. The sensitivity and specificity should be those of the laboratory test.

When all herds or villages have been analysed separately, the population of herds or villages can be analysed. Use the FreeCalc program:

- Step 1:** Start FreeCalc and click on the Analyse Results tab.
- Step 2:** Enter the herd test sensitivity and specificity.
- Step 3:** Enter the total number of herds or villages for the Population Size.
- Step 4:** Enter the maximum acceptable prevalence in the Prevalence box.
- Step 5:** Check the Type I and Type II error levels in the Options tab. They should be the error levels for the overall survey, not for the second stage of testing.
- Step 6:** Return to the Analyse Results tab, and enter the Survey Sample Size. This is the total number of herds or villages that were tested (the first-stage sample size).
- Step 7:** In the Number of Positive Reactors box, enter the total number of herds that were classified as diseased.
- Step 8:** Click the **Calculate** button.

Although some of the herds were classified as being disease positive, the herd test was not perfect and may have made a small number of mistakes. This final stage of analysis calculates whether the number of positive herds can be accounted for by the errors in the herd test. If so, then it is possible to conclude that the population is free from disease. If the number of positive herds is too high, then the population must still be classified as diseased.

## Other issues

### Combining tests

Often the performance (sensitivity and specificity) of a laboratory test is not as good as we would like it to be. If the specificity in particular is low, then the sample sizes to demonstrate freedom from disease will be very high. Sometimes it might not be possible to achieve the confidence required by testing every animal in the population. One way to address this problem is to test each specimen with two different tests. Depending on the way the results are interpreted, this can dramatically improve either sensitivity or specificity (but not both).

**Example:** During the last stages of a tuberculosis eradication campaign, the specificity of tuberculin skin tests, although high, is not perfect, and when testing very large numbers of animals may produce many false positive reactors. One approach is to test animals that produce a positive tuberculin skin test result with another test (e.g. gamma interferon). If the second test is also positive, then the animal is considered infected, but if the second test is not, then the animal is considered uninfected. This type of testing increases the specificity of a test, but decreases the sensitivity (and therefore the proportion of false negatives).

#### Tests in series

To increase the specificity of a test, two tests can be used in series. The first test is applied, then, if the animal is positive, a second test is used. The animal is considered positive only if both tests are positive. Animals that test negative to the first test are not retested. The result of this approach is to increase the specificity, but decrease the sensitivity. The overall sensitivity ( $Se_t$ ) and specificity ( $Sp_t$ ) of the two tests combined (the "test system") will be:

$$Se_t = Se_1 \times Se_2$$

$$Sp_t = Sp_1 + Sp_2 - (Sp_1 \times Sp_2)$$

Tests can be used in series in a similar way to increase sensitivity. If an animal tests positive with a single test, then it is considered positive, but if it is negative, it is retested, and considered negative only if it produces a negative result in the second test as well. In this case the specificity decreases:

$$Sp_t = Sp_1 \times Sp_2$$

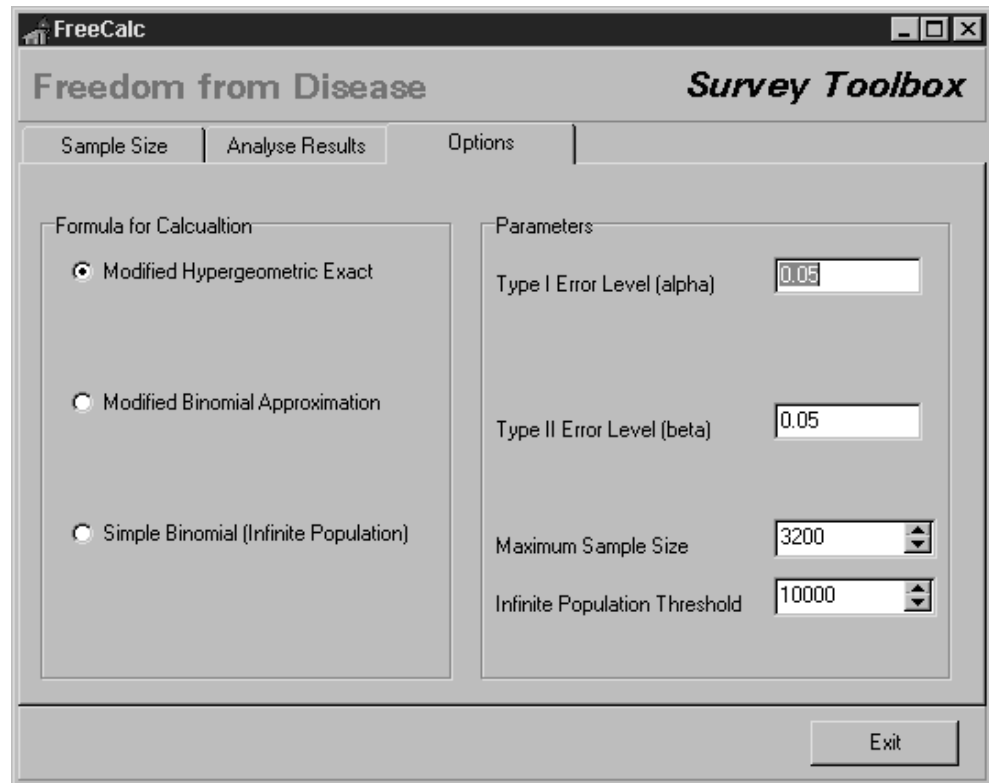
$$Se_t = Se_1 + Se_2 - (Se_1 \times Se_2)$$

#### Tests in parallel

Another approach is to test every sample with two tests, and base the final result on both the results. If both tests are positive or both are negative, the result is clear. However, when the two tests disagree, a decision must be made as to what the result is. If animals with conflicting results are considered negative, the sensitivity decreases and the specificity increases, as demonstrated by the equations in the first example above.

If conflicting results are interpreted as positive, then the equations in the second example above should be used, with a decrease in specificity and an increase in sensitivity.

## FreeCalc options



### Formulae

On the options tab of the FreeCalc program, there is a choice of three different formulae to use for the calculations.

The first is the Modified Hypergeometric Exact Formula. This formula calculates the exact probabilities for sample sizes and analysis of results. Under certain circumstances, this formula requires an enormous number of calculations, and can therefore be very slow. This happens when the sample size is large, due to poor test performance (especially low specificity) or a small maximum acceptable prevalence. Nevertheless, you should use this formula all the time, unless you find that calculations are becoming too slow.

The Modified Binomial Approximation Formula calculates the same probabilities, but uses an approximation, making the calculation faster. This formula still produces accurate results except when the sample size very large, relative to the population size. Use this formula if the Modified Hypergeometric Exact formula is too slow, and the sample size is less than half of the population size. Although much faster than the Exact formula, for very complex calculations (very large sample sizes) this formula can also become quite slow to calculate.

The Infinite Population Binomial Formula is the fastest to calculate. It assumes that the size of the population is infinite (or at least very much larger than the sample size). If you are working with very large populations, and the other two formulae are slow to calculate, use this formula. When the population is not very large, the use of this formula can lead to significant errors.

**Maximum Sample Size**

You can specify the maximum sample size for the program to calculate. If the sample size required is larger than this maximum, the program displays an error message, and stops the calculation.

**Infinite Population Size**

When the population is larger than a specified size, the program automatically uses the Infinite Population Binomial Formula, regardless of which formula has been selected. With very large population sizes, there is virtually no difference between the Exact and the Infinite Population formulae, so the faster of the two is used. You can enter the size of the population above which the faster formula will be used.