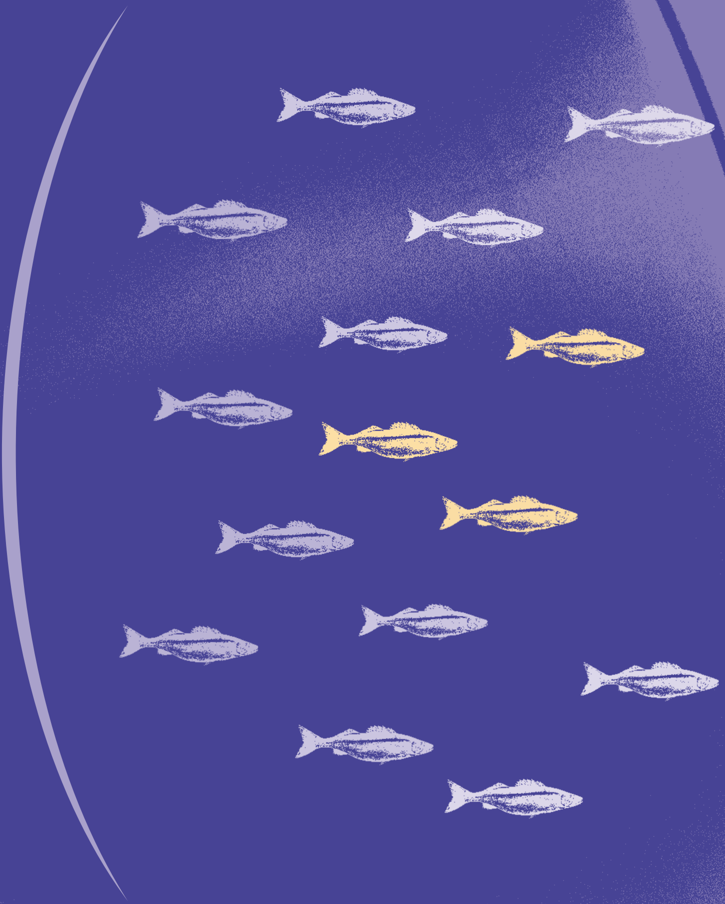
SURVEY TOOLBOX
for Aquatic Animal Diseases

A Practical Manual and Software Package



Angus Cameron

Survey Toolbox for Aquatic Animal Diseases

To Catriona, as always

Survey Toolbox for Aquatic Animal Diseases

A Practical Manual and Software Package

Angus Cameron

ACIAR
AUSTRALIAN CENTRE FOR INTERNATIONAL AGRICULTURAL RESEARCH

Canberra 2002

The Australian Centre for International Agricultural Research (ACIAR) was established in June 1982 by an Act of the Australian Parliament. Its mandate is to help identify agricultural problems in developing countries and to commission collaborative research between Australian and developing country researchers in fields where Australia has a special competence.

Where trade names are used this constitutes neither endorsement of nor discrimination against any product by the Centre.

Cameron, A. 2002. Survey Toolbox for Aquatic Animal Diseases. A Practical Manual and Software Package. ACIAR Monograph No. 94, 375p.

ISBN 1 86320 350 8 (print version)

ISSN 1447-090X (electronic version)
ISBN 1 86320 351 6

Contents

Appendixes

About the author

Angus Cameron is a veterinarian with a special interest in epidemiology, surveillance and information systems in developing countries. He graduated from the University of Sydney in 1988, and worked in dairy cattle practices in Victoria for several years. He completed a Masters degree in dairy cattle medicine and surgery through the University of Melbourne in 1992. Angus became a member of the Australian College of Veterinary Scientists in 1993.

Angus has lived and worked in Thailand and Laos, where he undertook his PhD studies in active surveillance and geographical information systems for animal health. This degree was awarded by the University of Queensland in 1998.

He now works as a veterinary consultant and is a Director of AusVet Animal Health Services Pty Ltd, providing epidemiological services to Australia and Southeast Asia. He regularly provides training, program design and other consultancies in aquatic animal health as well as livestock to countries throughout the region including Indonesia, Philippines, Vietnam, Thailand, Laos, Myanmar and India.

When not travelling, Angus lives and works in the Blue Mountains, New South Wales, Australia.

Acknowledgments

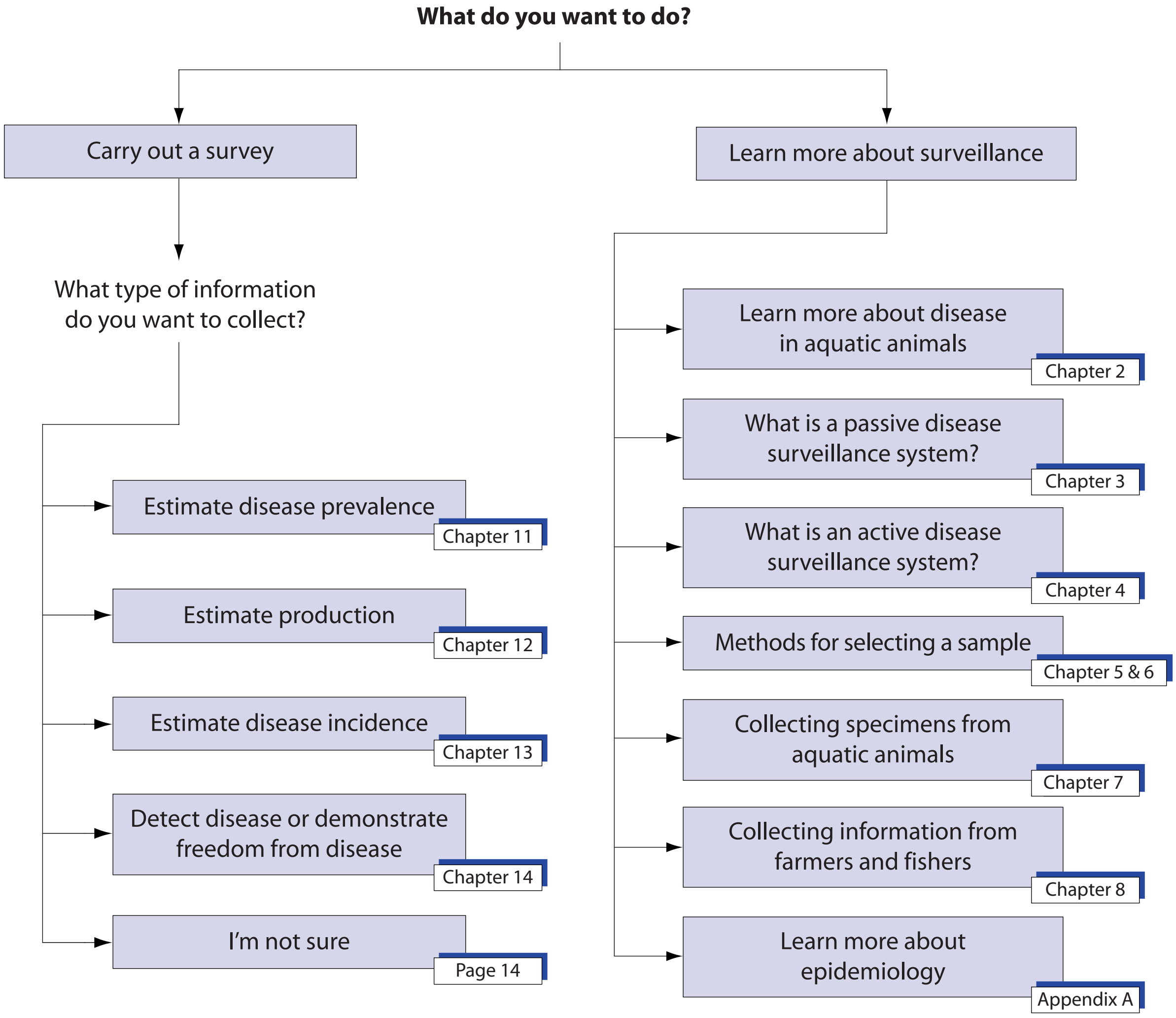
This book grew out of an earlier ACIAR publication, *Survey Toolbox for Livestock Diseases*, and I would like to thank all the people who contributed to that work.

Without the ongoing support of ACIAR, and in particular Barney Smith and Peter Lynch, it would not have been possible to complete this book. They deserve the thanks of developing country scientists around the world for funding, publishing and widely distributing books such as this and many others like it.

In writing this book, I have been fortunate to be able to draw from the expertise of a wide group of experienced aquatic animal health scientists, who have generously contributed their time and wisdom to make this book as useful and relevant as possible. Many of them gathered in Bangkok in 2001 to help plan the content and structure of the book, while others have reviewed drafts of the work: Melba Reantaso, Jimmy Turnbull, Phan Thi Van, Chris Baldock, Ian MacRae, Celia Lavilla-Pitogo, Barbara Nowak, Supranee Chinabut, Pornlerd Chanratchakool, Margaret Crumlish, Dan Feagan, Richard Friend, Huang Jie, Somkiat Kanchanakhan, Jim Lillie, and Sih Yang Sim.

There are yet others, scientists and farmers too numerous to name, who have had the kindness and patience to help teach me something about a range of different aquatic animal production systems.

Guide to the manual


What do you want to do?
Carry out a survey
Learn more about surveillance
What type of information do you want to collect?
Estimate disease prevalence
Chapter 11
Estimate production
Chapter 12
Estimate disease incidence
Chapter 13
Detect disease or demonstrate freedom from disease
Chapter 14
I'm not sure
Page 14
Learn more about disease in aquatic animals
Chapter 2
What is a passive disease surveillance system?
Chapter 3
What is an active disease surveillance system?
Chapter 4
Methods for selecting a sample
Chapter 5 & 6
Collecting specimens from aquatic animals
Chapter 7
Collecting information from farmers and fishers
Chapter 8
Learn more about epidemiology
Appendix A

1
Introduction and background

Purpose and scope of manual

Aquatic animals have provided an important source of food to humans for many thousands of years. First through the capture of wild fish, and later through aquaculture, aquatic animals have played a significant role in the diet of different people all over the world. Traditional fishing and farming systems continue to be used in many developing countries, but recently they have been supplemented by a massive growth in both highly efficient 'industrialised' fishing methods, and a dramatic intensification and commercialisation of aquaculture. The result has been a very large increase in aquatic animal production, meeting some of the increasing demands of a growing world population. However, increased production has come at a cost. Efficient fishing processes run the risk of rapidly overfishing existing stocks. Intensive farming systems use higher stocking densities and in some cases a variety of additives, contributing to environmental degradation and an increased risk of disease. Globalisation and increased trade have made the movement of aquatic animals around the world, either intentionally or unintentionally (e.g. in ballast water) simple and rapid. Pathogens move with these aquatic animals, and find themselves in new ecological niches. In some cases, this can result in disastrous outbreaks of disease.

In developed countries, with highly intensive production systems, new disease outbreaks can be extremely damaging to the fishing or aquaculture industries. However, due to the favourable economic situation, the damage is usually limited, and doesn't impact dangerously on the livelihood of a great many people. In developing countries, on the other hand, relatively more people depend on aquatic animal production for their food and income. A widespread disease outbreak has the potential to cause significant hardship to many people. The resources in such countries are often inadequate to either provide support for the people suffering from the consequences of a disease outbreak, or adequately support the research and disease-control activities to address the problem.

The prevention, control, and eradication of aquatic animal diseases depend on a good understanding of the disease and its distribution. Epidemiologically structured disease surveys are one of the most important tools to provide information about factors influencing the occurrence of disease, and its distribution. Conducting an effective aquatic animal disease survey in a developed country is often very challenging, even with adequate resources, but the many constraints faced in developing countries make the task even more difficult. One of those constraints is access to information about survey techniques suitable for developing countries, and appropriate for use in aquatic animal disease surveys.

This book aims to provide this information. It deals specifically with the collection of reliable, high quality information about aquatic animal diseases and production, using fast, inexpensive techniques that are suitable for developing countries.

The characteristics that make these techniques suitable for use in developing countries (inexpensive, rapid, reliable) also make them very attractive in other countries. It is therefore expected that this book will be of value to those studying aquatic animal diseases in any part of the world.

In this book, both 'aquatic animals' and 'disease' are defined broadly. Aquatic animals are considered to be those animals that spend at least part of their life cycle in water. This includes primarily fish, crustaceans and molluscs, but may also include amphibians. Although not specifically included in this definition, some of the techniques described may be applied even more broadly—to aquatic plants, for instance.

Concepts of disease and health are defined and discussed in Chapter 2.

Who is this book for?

This book is for people working in the area of aquatic animal diseases and production. The tools presented in this book will be valuable for anybody who needs to collect reliable information about aquatic animal diseases or production. The structure of the book allows it to be used at three different levels:

1. Planners
2. Trainers
3. Field operational staff

The roles and responsibilities of different staff vary from country to country, and individuals should determine what their needs are and use the book appropriately. In general terms, however, the responsibilities of the groups are as follows:

Planners are those responsible for the planning, coordination and conduct of disease surveys and disease-control activities on a large scale. They may be government aquaculture and fisheries staff responsible for aquatic animal health at the national level, but may also include government staff at different levels, research institute staff, development and research project staff (including NGOs), and possibly companies involved in aquatic animal research, test systems and product development. The main requirements of this group is a full understanding of the issues surrounding aquatic animal disease surveillance, and access to the information required to plan high quality surveys.

Trainers are those responsible for training survey staff in planning or field operations. Often they will be the same people as those listed above, but they may also include universities, and those responsible for in-service training. It is likely that much of the material in this book will need to be translated for use as student resources in non-English speaking countries.

Field operational staff are those conducting the actual survey fieldwork, collecting data during a survey. While they may be able to use sections of this book as a resource (particularly if translated), many field staff will generally need to be trained by another person, rather than through independent learning from this book. Field staff may not require an in-depth understanding of all the issues covered in this book, but they should have a detailed understanding of the operational skills required, such as sampling and specimen collection.

Notes for translators

The language used in this book is designed to be easy to understand by people speaking English as a second language. However, in many cases, sections of the book will need to be translated for use by survey staff who do not speak or understand English. It is hoped that the use of relatively simple language throughout the book will make the task of translation easier and more reliable. However, there are a few points that should be kept in mind by those involved in the translation of this book.

The ideal translator for a book like this is somebody who has three major skills: their first language is the language being translated into, they speak English very well, and they have a good understanding of aquatic animal health and epidemiology. Unfortunately, it is often difficult to either find such a person, or when such a person exists, they may not have the time to do lengthy translations. In practice, translations may need to be done by people with lower level English skills, or who are relatively inexperienced in epidemiology. If this is the case, extra care should be taken to avoid confusing the reader when the meaning is not clear.

While every attempt has been made to use simple language, some of the subject matter of this book is complex and technical, and explanations require the use of technical words. Often these words have both a general and a technical meaning, which are slightly different. For instance, the word ‘random’, in general use, means ‘without any reason or pattern’. In this book, it has a specific technical meaning that is defined. The word ‘haphazard’ is used to describe the general meaning instead. In many languages, there may be no one word to describe the idea of random, either in its general or technical sense, and certainly not two words with close but distinct meanings like ‘random’ and ‘haphazard’. In situations like this, using a simple dictionary definition of a word used in the technical sense is likely to result in incorrect and confusing text, and make it harder for the user to understand. A better solution to this problem may be to use the original technical English word (‘random’) untranslated, but provide a full and clear explanation of its technical meaning. Translators should not feel constrained to use a word-for-word translation, but be willing to insert extra text, sentences or whole paragraphs, to ensure that the intended meaning is made clear.

How to use this book and software

This book is accompanied by a set of software programs, collectively called Survey Toolbox. The book and software are closely integrated and are designed to be flexible, depending on the needs and preferences of the user.

Using the software

The purpose of the computer programs is to assist with the planning, implementation, and analysis of the surveys described in the book. Many survey procedures (such as random selection from a sampling frame, or the analysis of data collected from a two-stage seroprevalence survey) are either time-consuming or require extremely complex formulas and specialist statistical skills. In order to allow staff to carry out and analyse surveys without expert statistical assistance, all these procedures and formulas have been implemented in a set of computer programs, specifically designed for the surveys described in this book.

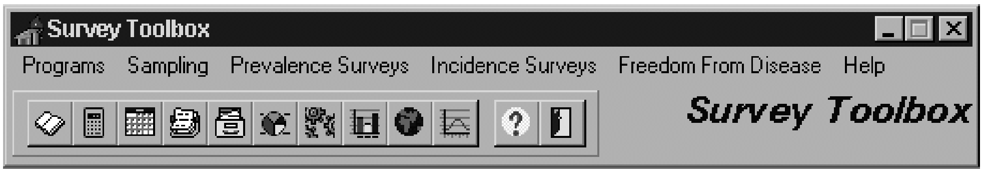
Each program performs a particular task and can be used independently. The programs do not provide a comprehensive set of data entry or statistical analysis tools. If further analysis or data manipulation is required, a separate database or statistical program, such as Epi Info[1], must be used. Only specialised procedures not readily available elsewhere have been included.

The use of each of these programs in Survey Toolbox is described in the book. When the use of one of the programs is discussed, a computer icon such as that shown appears in the left margin, and the program name appears in **Bold Type**. A listing of the name and use of each program, as well as a reference to the pages where it is described is given below, while a full description of the purpose, data input and outputs of each program is given in Appendix D.

Computer Programs

Program Name	Purpose	Page
Survey Toolbox	Main menu giving access to all programs	11
RGCS Win95	Selection of random geographic points	89
RGCS ArcView	Random points with map interface	90
Random Village	Random selection from a list of villages	97
Random Animal	Random selection of animals in a village	102
Prevalence	Sample size and analysis of two-stage prevalence surveys	188
True Prevalence	Convert apparent to true prevalence	191
Compare Prevalence	Compare the results of two prevalence surveys	192
Survival Size	Sample size for disease outbreak surveys	214
Survival	Survival analysis for disease outbreak surveys	218
CapRecap	Two-sample analysis for incidence rate estimates	225
FreeCalc	Sample size and analysis of surveys to demonstrate freedom from disease	233

Survey Toolbox is the main menu, shown below, and gives you quick and convenient access to all of the programs. It provides the main link between the book and all the programs.



Requirements

Windows 95

The programs have been written to run under the Windows 95 and later compatible operating systems (IBM-compatible computers). A full installation (including the electronic copy of this text) occupies 5 megabytes of hard disk space. If only the programs are installed, they occupy 2 megabytes.

MS-DOS

Some of the programs are also available in an MS-DOS version for older computers (as indicated in the list in Appendix C). These are located on the CD in the \MSDOS directory.

[1] Epi Info 6.04 (DOS) or Epi Info 2000 (Windows 95) are specialised programs for epidemiological analysis, produced by the Centers for Disease Control, Atlanta, GA. These programs are distributed free, and have been included on the CD version of the software accompanying this book. If you don't have the CD, you can obtain a copy through the Internet from <http://www.cdc.gov>. The use of these programs is described briefly in Chapter 6.

Installing

If you are installing from the CD, or floppy disk, insert the disk, click on the button, and select Settings | Control Panel. From the Control Panel, double click on the Add/Remove Programs icon, and click the Install button. The Windows Install program will guide you through the steps.

For the MS-DOS version of some of the programs, change to the \MSDOS directory (if using the CD), or insert disk 3 (if using floppy disks). Create a new directory on your hard drive called \ToolBox and copy all files into that directory.

Running

If using the Windows 95 version, the install program will create a new entry in your Start menu, called **Survey Toolbox**. Click on this to open a list of the different programs, then click on the program you want. You may want to place a shortcut to the Survey Toolbox menu on your desktop, to make it easier to access (consult Windows Help to find out how to make a shortcut).

Using the book

The book is divided into four parts.

Chapters 1 to 4 give an introduction to aquatic animal disease surveys, as well as a full discussion of active and passive surveillance. This is useful information for planners as well as trainers.

Chapters 5 to 9 give detailed information on the various techniques and procedures involved in disease surveillance, such as sampling and the collection and management of information. Planners designing surveys should review this information, as should those involved in training survey field staff.

Staff responsible for field work may wish to skip some sections but will find useful information on sampling aquatic animals, specimen collection, and interviews.

Chapters 10 to 14 provide a detailed description of the four main survey types described in this book. This is essential reading for those responsible for planning and organising disease surveys. Staff responsible for the field implementation of the survey should read the chapter on the relevant survey type, and some sections will also be valuable to field staff.

Chapters 15 to 17 are specifically targeted at those wishing to use this book as the basis for training of survey staff. It contains a guide to who should become a trainer, and effective teaching techniques. This is followed by a series of lesson plans, divided into three training courses. There is also a collection of activity sheets to help organise participatory training activities.

The appendices contain additional information that may be useful to some readers. For those wishing to understand more about epidemiology (particularly planning staff as well as trainers) there is an introduction to aquatic animal epidemiology, covering a much broader range of topics than just surveys. This is suitable for self-study, or could be used as the basis for a training course. There is also a glossary of statistical and epidemiological terms used in the text, example data collection forms which can be copied and used, and a list of the contents of the CD.

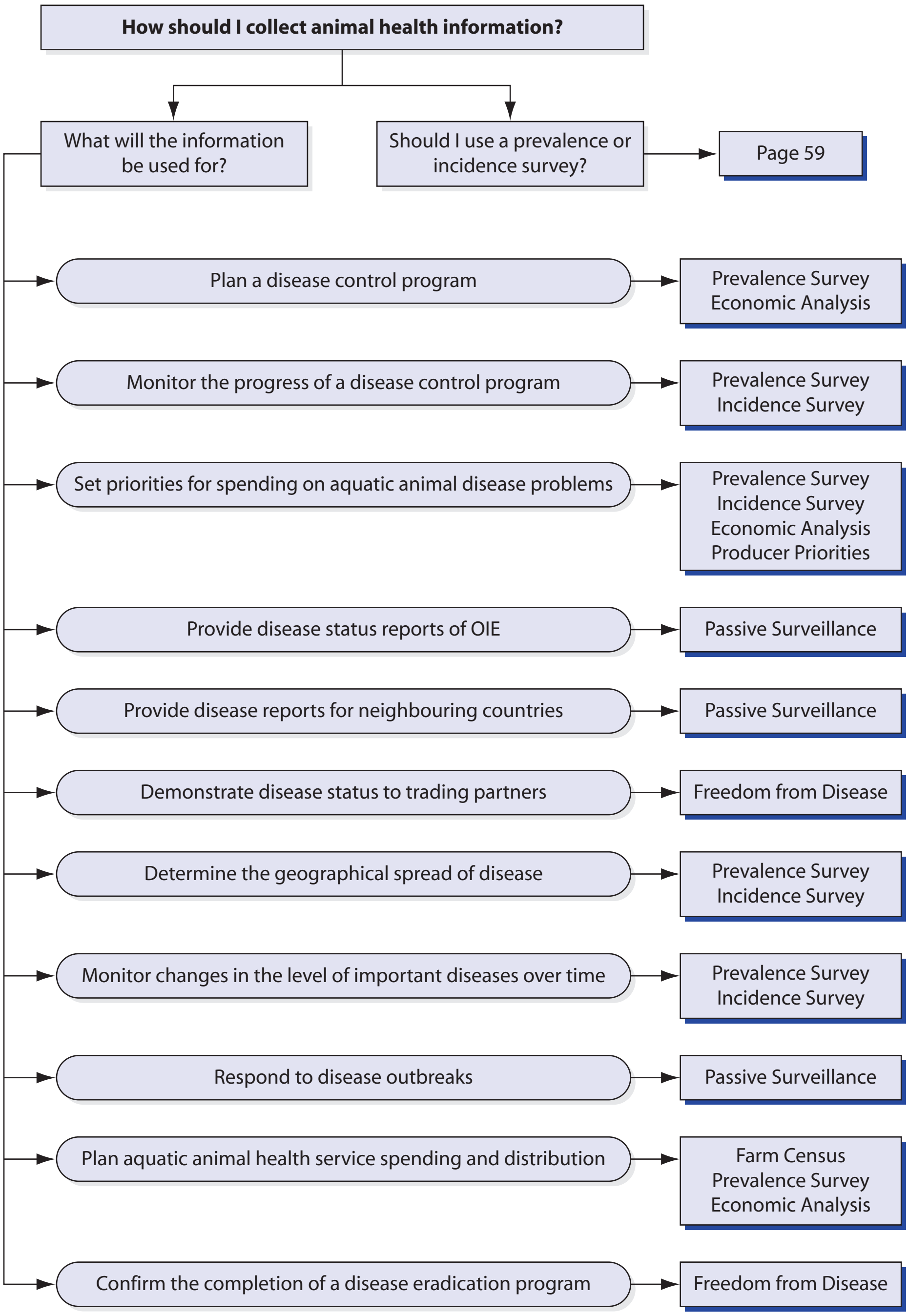
Now to find the information you need

The book is designed to be flexible to use. For those needing to understand the survey procedures and all necessary background in depth, starting at Chapter 1 and reading through to the end will provide a thorough understanding. Many of the target users will not require all the information, and so might like to pick and choose. There are three ways to do this:

- Use the contents to find the chapter you want, or the index to find the particular reference you want.
- Look at the flow chart on page 16 (also shown on the inside front cover). This will guide you to references to other flow charts or instructions throughout the book that help you decide what information you need or how to carry out a survey procedure. The flow charts refer to the page where you can find the information.
- Refer to the table below, which lists the recommended sections for different types of readers.

Chapter	Title	Planner	Trainer	Field staff
1	Introduction	✔	✔	
2	Aquatic animal health	✔	✔	
3	Passive surveillance	✔	✔	
4	Active surveillance	✔	✔	
5	Sampling principles	✔	✔	✔
6	Sampling applications	✔	✔	✔
7	Data and specimen collection	✔	✔	✔
8	Information from people	✔	✔	✔
9	Data management	✔	✔	✔
10	Survey design and planning	✔	✔	✔
11	Prevalence surveys	✔	✙	✙
12	Production surveys	✔	✙	✙
13	Incidence surveys	✔	✙	✙
14	Freedom from disease	✔	✙	✙
15	Guide for trainers		✔	
16	Lesson plans		✔	
17	Activity sheets		✔	
App A	Aquatic animal epidemiology	✔	✔	
App B	Glossary			
App C	Forms			
App D	Contents of CD			

✔ Should read this section
✙ Read this section only if conducting the relevant survey type


How should I collect animal health information?
What will the information be used for?
Should I use a prevalence or incidence survey?
Page 59
Plan a disease control program
Prevalence Survey
Economic Analysis
Monitor the progress of a disease control program
Prevalence Survey
Incidence Survey
Set priorities for spending on aquatic animal disease problems
Prevalence Survey
Incidence Survey
Economic Analysis
Producer Priorities
Provide disease status reports of OIE
Passive Surveillance
Provide disease reports for neighbouring countries
Passive Surveillance
Demonstrate disease status to trading partners
Freedom from Disease
Determine the geographical spread of disease
Prevalence Survey
Incidence Survey
Monitor changes in the level of important diseases over time
Prevalence Survey
Incidence Survey
Respond to disease outbreaks
Passive Surveillance
Plan aquatic animal health service spending and distribution
Farm Census
Prevalence Survey
Economic Analysis
Confirm the completion of a disease eradication program
Freedom from Disease

2

Aquatic animal health

Health and disease

What is a disease?

This book is about diseases in aquatic animals. There are many different diseases that affect different species of aquatic animals. Some diseases, like vibriosis, have names that indicate the cause of the disease, *Vibrio* bacteria. Others, like epizootic ulcerative syndrome (EUS) have names that simply describe the disease. There are other diseases that may not have a proper name at all, such as problems due to low oxygen levels, or build-up of ammonia in the water, or high water temperatures.

If we wish to study diseases, it is important to first have a clear idea of what we mean by a disease. In the examples above, some diseases are caused by a known pathogen, while in others the cause of the disease is mixed, uncertain, or unknown, and in others there may be environmental or management factors leading to problems. Can we call something a disease if we don't know what causes it? Can we call something a disease if the cause is due to management, or to environmental changes? In this book, all of these problems will be considered to be disease. A very broad meaning of disease will be used:

Disease is any abnormality of structure or function.

This means that whenever there is something abnormal about the animals, we can consider it to be a disease. For instance, if a fish looks normal and behaves normally, but is growing more slowly than normal, then that fish is diseased. We may describe the disease as 'slow growth rate'. This definition may be further expanded to *any condition impacting on an animal which may be deleterious to animal or human health*, including infection, contamination, residue or other condition.

Health is another term that is commonly used when discussing disease. Based on the definition above, health is simply the normal state of an animal, or the absence of disease. Determining if an animal is healthy or not doesn't just mean that we have to identify some physical abnormality, or a disease agent. In the example above, the level of production (growth rate) can be an indicator of whether an animal is healthy or diseased. Measures of production to indicate health status can be very useful and will be discussed in more detail later.

Causes of disease

The reason for studying disease is to try to do something about it—either to prevent it, cure it, or minimise its negative effects. In order to do something about disease, we must first understand something about the cause of the disease. In many cases, this seems simple. For example, the cause of the disease vibriosis is the *Vibrio* bacterium. Unfortunately, in aquatic environments, it is rarely as simple as that. *Vibrio* is often present in ponds in quite high numbers, but with little sign of disease. In order for it to cause disease, there needs to be something else going on; for instance external parasites causing skin damage, to allow bacteria to penetrate and multiply within the animal.

This is an example of a *multifactorial* disease. In order for the disease to occur there have to be multiple factors or separate causes: 1) the presence of *Vibrio* bacteria, and 2) the presence of external parasites causing skin damage. If only one of these causes is present, then the disease will not occur. Virtually every disease in aquatic animals is a multifactorial disease.

Vibrio + parasites = vibriosis

The different factors or causes that can lead to disease are called *component causes*. The presence of *Vibrio* in the water is one component cause for the disease vibriosis, and parasites causing skin damage are another component cause. There may be other component causes involved; for instance skin damage resulting from netting the fish, or an imbalance in the microbial flora. On their own, none of the component causes is able to cause the disease, but when they are combined, they can cause the disease.

Vibrio + parasites = vibriosis

Vibrio + netting = vibriosis

Vibrio + imbalance in microbial flora = vibriosis

If we wish to prevent or control disease, it is important to understand something about the component causes of the disease and the way they interact. In our example, there are now four component causes, *Vibrio* bacteria, parasites, and netting. We need only two of these component causes to cause the disease, *Vibrio* and parasites, *Vibrio* and netting or *Vibrio* and an imbalance of other bacteria. It is clear that *Vibrio* is necessary in all cases to cause the disease, but parasites, netting or bacterial imbalance may also be involved. To distinguish the importance of *Vibrio*, it is called a *necessary cause*. This means that without *Vibrio*, it is not possible to have the disease vibriosis. The other component causes are not necessary causes. If there are no parasites, it is still possible to get vibriosis, if the fish have been netted. Similarly, if there is no netting, you may still get vibriosis if there are parasites.

An understanding of component and necessary causes of multifactorial diseases gives us a better understanding of how we may control the disease. If we avoid netting the fish, but parasites are still present, the disease will continue to be a problem. Similarly, if we practise improved parasite control, but the skin is damaged by netting, there will be ongoing disease. In order to prevent the disease, we need to make sure that there are few parasites, and no skin damage from netting. By removing these two component causes, it will be possible to control the disease, without doing anything about the necessary cause—the presence of *Vibrio* bacteria.

In many cases, it may be very hard to identify an infectious agent causing a particular disease problem. However, if other factors or component causes can be identified, it may well be possible to control the problem without ever knowing exactly what the agent was. By removing other component causes you can prevent the disease from occurring.

Types of disease

Diseases are commonly divided into infectious and non-infectious diseases. Infectious diseases, like any other diseases, are multifactorial. However, at least one of the component causes (often a necessary cause) is an infectious agent or pathogen. There may be many other component causes, some of which involve infectious agents and some of which don’t.

Example

Marteiliosis (QX disease) of oysters is a multifactorial disease. One component cause is the presence of a protozoan parasite (*Marteilia* sp.—a necessary cause), but there are several other hypothesised component causes, including the time of year, density of stocking, position in estuary, water quality and sediment particle size. There is also thought to be an intermediate host for the parasite. The exact relationship between these component causes is still uncertain, but it is clear that, if there is an intermediate host, it can be thought of as a second 'infectious' necessary cause. That is, without the intermediate host, there can be no disease, and the parasite cannot complete its life cycle. In areas where the intermediate host is absent, all other component causes may be present, but there will still be no disease.

Non-infectious diseases are diseases in which none of the necessary causes are infectious agents. This means that it is possible to have the disease with no infectious agents involved. The causes may include nutritional, environmental or genetic factors, or a combination of these. There may or may not be infectious agents involved as component causes. Cold shock is an example of a non-infectious disease. Component causes may include low temperature (a necessary cause related to the environment), pond size, stage of growth, and presence of other disease agents, such as parasites, which may be weakening the fish. You can suffer from cold shock without parasites being present, but you can't if it doesn't get cold. Just as infectious agents can be component causes of non-infectious diseases, non-infectious disease factors are often component causes of infectious diseases.

Syndromes

A syndrome is a collection of signs and epidemiological behaviour that often occur together, and can be used to identify a disease.

By using syndromes to think about diseases, we can move away from the inaccurate idea that 'disease agent equals disease'.

Often, the name of a disease is based on one of its component causes (usually a necessary cause) that is an infectious agent, for example, vibriosis. When we have a good understanding of the disease process, and all the component causes, this may be an appropriate 'shorthand' way of referring to a disease. However, in many cases, aquatic animal diseases are much less well understood than diseases of land animals. It is often more appropriate to think in terms of a syndrome than a specific disease agent causing the disease.

A syndrome is a collection of signs and epidemiological behaviour that often occur together, and can be used to identify a disease. By using syndromes to think about diseases, we can move away from the inaccurate idea that 'disease agent equals disease'. Epizootic ulcerative syndrome (EUS) is the name of a disease, expressed in terms of a syndrome. When the disease was first recognised, the causal agent(s) were not identified, but there was a clear collection of signs and epidemiology that consistently occurred together. The name describes the main sign (ulcers) and the nature of the epidemiology (epizootic, or occurring as outbreaks). White spot syndrome of shrimp is another example, in which the main sign is used as the name for the syndrome. In this case, the virus which is a necessary cause for the disease was identified after the disease was named, so the virus was called white spot syndrome virus (WSSV).

When working with syndromes, it is important not just to have a name, but also a clear case definition, which identifies all the signs and epidemiological characteristics that are required to diagnose the syndrome.

Making a diagnosis

The purpose of studying a disease is to determine what causes it and how it can be cured, prevented or its effects minimised. One of the most important steps is making a diagnosis, which is the process of determining the health status, and identifying the factors that produced it.

There are many different levels at which a diagnosis can be made, depending on the information available. If only a small amount of information is known, then the diagnosis may simply describe the disease and associated factors, but if more is known, it may be possible to identify more of the factors in detail, including infectious agents if they are involved. Normally, the name of a disease is given as the diagnosis, but it is possible to give a diagnosis more generally, without specifying a particular disease name.

Making a diagnosis is a bit like solving a complex puzzle. There are many different clues, and these have to be pieced together to find the final solution. If only a few of the clues are available, it may be possible to give a general description of the problem.

Example

An oyster farmer reports that they have a problem with deaths amongst their oysters. Based on an investigation, we find that about 40% of oysters have died in the space of one month, during late summer. Examining the oysters shows that the digestive gland is a yellow colour. Based on this clinical and epidemiological picture, we may think that it is likely that the disease is marteiliosis (QX disease).

This is known as *clinical diagnosis* because it is based mainly on clinical examination and history. The diagnosis may be wrong, but this is our 'best bet' based on the information available at the time. It may be useful to provide not only a clinical diagnosis, but a *differential diagnosis*. This is a list of other diseases that may be responsible and, in this case, it may include gill necrosis caused by iridovirus, bonamiasis, or winter mortality disease (mikrocytosis). A differential diagnosis acknowledges that the exact cause is not known, but provides a great deal of information by listing the diagnoses that are considered possible. A differential diagnosis may also be ordered based on probability.

When narrowing a differential diagnosis, it is essential to use other information, or clues to the puzzle. One important piece of information is the location of the disease outbreak, and the types of diseases that are known to exist in the area. If there have been previous outbreaks of QX disease in the same area, then it is perfectly possible that this is the cause. On the other hand, winter mortality has only ever been described in Australia, so if these oysters are not in Australia, it is very unlikely. However, it is always important to remember that diseases can spread, and this may be the first report of an exotic disease. The fact that it has never been seen before should not mean that it can't be seen in the future. Winter mortality should be low on the list of diagnoses, but it is still a possibility.

In order to be more confident about our diagnosis, we need to gather more clues. One good way is to use laboratory testing. If some affected oysters were collected and sent to the laboratory for histopathology, we may find that there appear to be protozoan parasites present in the digestive gland, and that the gills are shrunken but otherwise normal. Bonamiasis, gill necrosis and winter mortality usually all show signs of damage to the gills, so this makes these diseases very unlikely. A further laboratory test may be performed; for instance the polymerase chain reaction (PCR) technique to detect the DNA structure of the protozoans. If this is found to match with *Marteilia* sp., we can be very confident that this is the parasite that is present. This is known as a *laboratory diagnosis*, because it is supported by laboratory information. In this case, the clinical, epidemiological, and laboratory clues to the puzzle all point towards the same disease—QX disease. We can therefore be reasonably confident in making a *definitive diagnosis* (meaning that we are completely certain of our diagnosis) that the disease is QX disease.

In this example, all the clues matched one particular disease. Unfortunately, this is not always the case.

Example

A farmer reports an outbreak of disease in her snakehead ponds. Five fish are collected and bacteriological examination performed. *Vibrio* sp. bacteria are found in all fish.

In this example, the laboratory diagnosis might be that the disease is vibriosis, because the disease agent was identified. However, this diagnosis is based on only one piece of the puzzle; it does not consider all the clues together. If the clinical and epidemiological picture were considered, it may be found that there were large sores and ulcers on the bodies of the fish, and that the farmer had suffered 60% mortality of his fish stock due to the disease. Furthermore, if different laboratory tests had been performed, for instance a fungal culture, *Aphanomyces invadans* may have been found. All these other clues are pointing to EUS, not to vibriosis. In fact, *Vibrio* sp. bacteria may invade damaged tissue caused by other diseases, but are not the primary cause of the disease.

> A diagnosis is made by considering all of the information available, including history, clinical, epidemiological and laboratory findings. Information that comes from a single laboratory test is *not* a diagnosis—it is just one more piece of information that may help make a diagnosis.

If we wish to interpret a diagnosis, it is important to know how the diagnosis was made, and therefore how much confidence we can place in it. If a diagnosis was made using just one or two clues, it is not as good as one made on the basis of a complete investigation in which almost all the clues were available.

- *Rumour or hearsay.* There is no direct clinical or epidemiological evidence, simply a report of what somebody else thinks. You can have almost no confidence in the reliability of this type of diagnosis, but the information is important, as it may be worth investigating further.

- *Clinical examination and epidemiological description*. A great deal of information can be gathered by careful discussion and examination of affected animals. This type of diagnosis is very useful, and probably the most common, but it is often not able to provide a high level of confidence about the specific disease agents. This is called a ***Level 1 diagnosis***.
- *Clinical and epidemiological information combined with simple laboratory tests, such as smears and histopathol*ogy. It may not always be possible to characterise disease agents exactly using these tests, but the nature of the pathological change is very clear. This type of diagnosis has a high level of confidence regarding the presence and nature of a disease process, but provides less specific information about the specific disease agent. It is called a ***Level II diagnosis***.
- *Clinical and epidemiological examination, supported by sophisticated laboratory tests, such as electron microscopy, or molecular techniques such as PCR*. This provides the highest level of confidence about the presence of a specific disease agent, but may not provide as good evidence about the presence and nature of the disease process. It is called a ***Level III diagnosis***.

Importance of aquatic animal diseases

Importance of aquatic animals

Aquatic animals, mainly wild-caught, have been an important food source for people for millions of years. Aquaculture has been practised for thousands of years. However, the rapidly increasing world population is putting enormous pressure on the environment to provide adequate food. The increasing population and improved technologies have led to a dramatic increase in production from capture fisheries over the last 100 years. As a result, the ability of wild stocks to be sustainably harvested is being threatened and, in some cases, these stocks have already been exhausted.

The dramatic growth in aquaculture over the last few decades has partly been in response to a fear of overfishing in wild fisheries. Aquaculture offers the possibility of producing large volumes of aquatic animals without depleting the rivers and oceans, and may therefore be a more environmentally sustainable approach to food production. However, aquaculture is also being widely promoted as a production system to help poor farmers in developing countries improve their livelihoods. The majority of global production occurs in developing countries.

Importance of disease

Diseases of aquatic animals are starting to play a major role in production levels. In traditional fishing or culture systems, production levels were low. Lower densities and low intensity production resulted in relatively low levels of disease. Increased intensity of production and movement of aquatic animals and their associated pathogens has resulted in an increase in disease levels.

When there is a disease problem in a wild population or a cultured system it may cause many types of loss:

- loss of animals due to deaths from the disease;
- reduced production through lower growth rates or decreased feed utilisation;
- costs of treatment or prevention;
- loss of investor confidence;
- loss or damage to brood stock may have major consequences; and
- damage to wild populations may result not only in the loss of a resource but also decreased biodiversity and a shift in the ecological balance.

What makes disease in aquatic animals different?

Environment

There are many difference between the fish, crustaceans and molluscs that are raised for food in aquatic systems, and the mammals and birds that are used in terrestrial systems. The most obvious and most important is the water that they live in, in contrast to the air that surrounds us and farmed livestock. In many ways, air and water are similar—for instance, they are both fluid, have winds or currents, and contain oxygen to sustain respiration. The heart of the difference between water and air is that water is more dense and, as a result, more buoyant—many things float in water, but few float in air.

Among the things that float in water are the animals that live in it. Fish and crustaceans (and even molluscs during part of their life cycle) are able to move about not only over the bottom surface, but in three dimensions throughout the water column. One of the implications of living in three dimensions is that it makes mobile aquatic animals hard to catch.

Another thing that floats in water is a range of suspended particles: soil and silt, algae, organic matter etc. This means that it is often difficult to see through water for any distance. Air too may be foggy or dusty, but visibility is usually much greater than in water. This means that aquatic animals can be hard to see.

Disease-causing agents are the third important thing that floats in water. Bacteria, viruses, fungi, parasites and all sorts of toxins are all mixed together with the water that the animals live in, and can pass from one animal to another with complete ease. On land, there are a few toxic gases and vapours, and some viruses can be spread in droplets or the wind, but most diseases need to be spread by direct contact, and can’t pass easily through the air. In enclosed water bodies, changes in one location are likely to affect all animals living in that water. In rivers and oceans, toxins and pathogens can easily spread large distances with widespread effects.

Another key difference about the aquatic environment is that it is not the natural environment for humans. Unlike the species that we catch or culture, we are limited to staying at the surface, or using expensive, inconvenient equipment to venture below.

Host

In addition to differences in the environment, there are some important differences in the aquatic animals themselves. The first is in their reproductive strategy. Because of the ability to move easily in three dimensions (which makes life easier for predators), and the ease with which disease spreads, life for young aquatic animals is very dangerous. In the natural environment, many do not survive. To increase the chances of survival, most aquatic animals produce massive numbers of offspring.

The second difference is that most of the species are relatively small. In aquaculture, this means that a single pond, cage or net will often contain a very large number of animals. In capture fisheries, it means that the population that we are working with is often very large.

The third significant difference is the physiology of the organisms. Mammals and birds are warm-blooded animals, relatively higher on the evolutionary tree than fish, crustaceans and molluscs. Aquatic animals are generally somewhat simpler. Bony fish have circulator systems, and immune systems that may involve the thymus, kidney, spleen, gut- and mucosa-associated lymphoid tissues, and antibodies in the form of IgM and IgH, but they lack the more specialised antibodies. They do not have discrete lymph nodes, nor IgG, IgA or IgE. Crustaceans and molluscs have a simple form of cell-mediated immunity but lack any antibodies or specialised organs.

Implications of differences for disease

The nature of the aquatic environment, and of the aquatic animals themselves, poses some challenges for trying to understand and respond to disease problems. The key implications of the differences identified above are:

- They are more difficult to catch.
- They are more difficult to see.
- Disease agents spread quickly and easily.
- They often gather or are cultured in large numbers.
- They are difficult to isolate.
- Disease is often difficult to detect and characterise.

To understand disease we must be able to identify disease in an individual animal, and to examine the patterns of disease in populations or groups of animals.

Identifying disease in an individual

In order to identify disease, we usually have to notice that something is wrong. The fact that water is rarely clear, and that there are large numbers of animals involved, means that many animals can be sick or dead without any obvious sign. In cultured shrimp, for example, healthy shrimp will usually eat sick or dead shrimp, so there is nothing left to see. Less severe signs of disease may be hard to identify, as they are not obvious when the animal is in the water. In order to see the signs clearly, we must first catch the animal and examine it closely. Catching animals is not easy, and catching a particular animal is almost impossible. Once an animal is caught, it can be examined, but catching it, and keeping it out of water to examine it often causes much more damage than the disease that it might or might not have. Animals are often killed simply by our looking for disease in them.

Even when animals are captured, they express relatively fewer external signs of disease than terrestrial animals, so it may be hard to determine if they are sick or not. It is often therefore necessary to use laboratory tests to determine if the animal is diseased. However, there are fewer laboratory tests available for diagnosing disease in aquatic animals than for terrestrial animals. For instance, antibody tests are rarely useful, due to the nature of the immune system of the animals. Viral culture is also extremely difficult due to a lack of suitable culture media. In general, laboratory diagnosis of disease is often limited to histology to detect changes in tissues, and direct visualisation of disease agents through light or electron microscopy.

Examining patterns of disease in a population

If identifying disease in an individual is difficult, understanding the patterns of disease in a population is even more difficult. One of the first major problems is knowing how big the population is. This is clearly a major difficulty in capture fisheries, but is also a problem in closed systems such as pond culture. For example, when shrimp are stocked into a pond, the post-larvae are very small and there are very many of them. The hatchery may sell bags said to contain, say, 2000 post-larvae, but this is only an estimate, and the real number may be more or less. As they grow, there are likely to be losses due to various diseases. However, even with larger shrimp, you can't count them by looking at the pond because 1) you can't see them, 2) they are in three dimensions, so you would lose track of them and 3) there are too many of them. Instead, the only completely accurate way to know how many have survived is to count them at harvest time. The problem is that it is still impossible to assess the impact of disease if we don't know precisely how many were really stocked in the first place. Estimation techniques can be used to overcome these problems. These include estimating the population based on feed consumption, or by sub-sampling from the pond as described later in this book.

Aquatic system health

That fact that aquatic animals live in water, and that there are often large numbers in relatively small areas means that they have a much closer connection to their environment than terrestrial animals. In fact, it is really not possible to consider either the animals or the environment in isolation. Anything that affects one will inevitably affect the other.

The interactions that take place within the aquatic environment are often much more complex than on land. For instance, in a fish pond, there are not just fish and water. There is an entire ecological system made up of millions of algae, protozoa, and bacteria, all in a dynamic balance. These microorganisms may provide the fish with food, and contribute to maintaining an appropriate environment, for instance by buffering the pH of the water. However, they may also be parasites or otherwise threaten the health of the fish. To maintain the health of the fish, and achieve good production, it is not enough to understand about the fish, or to understand about the microorganisms, nutrients, minerals, toxins and other factors present in the water. We must understand how these different elements interact, and the effect they all have on each other.

Because of this, it is often misleading to think about 'aquatic animal health' because it places the focus on the individual fish, crustacean or mollusc. Instead, it is more realistic to think about 'aquatic system health'—that is, the state of health of the entire system. If any aspect of a pond's ecosystem is 'diseased', for instance a crash in the algal bloom, it will inevitably result in disease in other organisms that share that environment, including the fish that are being cultured.

Traditionally, when trying to diagnose the cause of a disease, we examine diseased animals. The individual animal is the unit of interest, the thing that we study to make our conclusions. When we think in terms of aquatic system health, we are no longer trying to diagnose the cause of disease in an individual animal, but the cause of disease in an aquatic system. The unit of interest is no longer a sick fish, but a 'sick' pond. The tests we use on individual animals, such as skin scrapings or histology, must be combined with tests that we use on the pond, such as measuring pH, soil bottom conditions and turbidity. In many situations, being able to step back from the individual and consider the system is essential to understanding disease in aquatic animals.

Disease surveillance

Approaches to understanding disease

There are many ways to understand disease. The simplest is to look at a diseased animal, for instance a shrimp. A gross examination of the animal may reveal different signs that indicate what the problem could be. Simply looking at a sick shrimp can give us only so much information—for instance, that it has white spots on the carapace. Traditionally, biologists have adopted the approach of looking closer to understand what is the cause of the disease. Instead of just looking at the animal, the next level may be to use a microscope to examine the cells of the shrimp histologically. After that, we may use electron microscopy to look at subcellular structures, including viruses. It is possible using molecular techniques to examine the DNA and understand the problem at a molecular level.

This approach to understanding disease, moving from the animal, to the cell, to subcellular structures, to the DNA, and down to the molecule, is known as a downward-looking approach to understanding disease. The focus is always on something smaller to understand the biological mechanism of the disease.

Epidemiologists (scientists who study disease in populations, not in individuals) use the opposite approach to studying disease. Instead of starting with the animal and focusing downwards, epidemiologists start with the animal, and move upwards. For example, the next level up from the sick shrimp may be the pond. Instead of looking at an individual, we may ask 'How many of the shrimp in this pond have been affected by this disease?' The next level up may be the farm. A farm is made up of a number of ponds, and we may want to know which ponds have been affected and which have not. Moving further up, we may look at the local area, made up of a number of farms, then the state or province, then the whole country, or even the whole world.

At each level, an epidemiologist is trying to understand the *patterns of disease* in order to understand what component factors may be involved in causing the disease. For instance, when examining a pond, it may become clear that the larger shrimp are more often affected with disease than the smaller ones. This suggests that one component cause of the disease may be the size of the shrimp. Looking at the farm, it may be that older ponds with poorer pond bottom conditions tend to have disease more often than new ponds—a second possible component cause. Examining the area could indicate that farms drawing water from a particular water source are more likely to suffer from disease, suggesting that the water source may be another important component cause.

Both the laboratory-based downwards looking approach, and the field-based upwards looking approach are essential to properly understanding disease.

Surveillance

There are several tools that epidemiologists use to understand patterns of disease in populations. One example is the use of outbreak investigations or case studies. Surveillance is another tool for understanding patterns of disease in the population. Surveillance involves the *systematic* collection of information from a population. Monitoring is another term that is often used. There are several different definitions used for surveillance and monitoring. Formal definitions of surveillance and monitoring are:

Surveillance: Systematic collection, analysis and dissemination of information in order to support the claim that a specified population is free from a particular infection or disease; or to detect an exotic or new disease so that control action can be quickly instituted.

Monitoring: Systematic collection, analysis and dissemination of information about the level (e.g. occurrence, incidence, prevalence) of infections or diseases that are known to occur in a specified population.

In these definitions, disease is used in its widest sense, and includes the presence of residues or other abnormalities. Both activities are assumed to be ongoing in some manner whether continuous or intermittent. The difference between surveillance and monitoring is that surveillance is looking for a disease that you don't believe is present (trying to prove that it is not), while monitoring is measuring the level of a disease that is already known to be present.

Although there are some differences between the design of surveillance and monitoring programs, many of the activities involved are very similar. In this book, the word surveillance is used to encompass both surveillance and monitoring. Where a distinction is between the two is intended, this will be made explicit.

Surveys

Surveillance, from the definitions above, involves the systematic *collection*, *analysis* and *dissemination* of information on disease. Surveys, on the other hand, are simply tools for the *collection* of information.

Survey: An investigation using the systematic collection of information from a population that is not under the control of the investigator.

Surveys are therefore distinct from experimental studies, in which the investigator has control over various aspects of the population under study, for instance which type of food they are fed. In surveys, the aim is simply to observe what is happening in the real world. There are a number of different types of surveys, targeted at answering different questions about the population, including cohort studies, case-control studies and cross-sectional surveys. This book deals mainly with cross-sectional surveys, but the other types of surveys are discussed in Appendix A.

As suggested by the title, a large part of this book deals with surveys and survey techniques.

Information on the health status of aquatic animals is needed by a wide variety of people and organisations, including producers, government authorities at different levels, businesses, research organisations, and regional and international animal health organisations. Each uses the information for somewhat different purposes.

For the purpose of disease control, improving the health and productivity of aquatic animals, and thereby, the well-being of the people, information is needed to:

- identify what diseases exist in the country;
- determine the level and location of diseases;
- determine the importance of different diseases;
- set priorities for the use of resources for disease control activities;
- plan, implement and monitor disease control programs;
- respond to disease outbreaks;
- meet reporting requirements of international organisations (e.g. Office International des Épizooties, OIE); and
- demonstrate disease status to trading partners.

Approaches to surveillance

There are two broad approaches to surveillance that are commonly used. The first is called *passive* or *general surveillance*. Passive means that no special activity is undertaken by the collecting authorities to generate the information—it usually comes from the producers. The term general surveillance indicates that surveillance aims to collect information about all sorts of diseases, and is not targeted at a specific disease.

The second approach is the use of *active* or targeted *surveillance*. The term active indicates that the collecting authorities initiate the data collection, while targeted means that the surveillance is targeted at one or more specific diseases.

A comprehensive surveillance system usually makes use of both approaches, as each is appropriate for different situations.

Passive (general) surveillance

Passive or general surveillance typically takes the form of a disease-reporting system. If a producer notices a disease problem, this is reported and recorded in a systematic fashion. Passive surveillance is discussed in more detail in the next chapter.

Active (targeted) surveillance

Active surveillance differs from passive surveillance in that the main users of the information make active efforts to collect the information needed, or that the main purpose for the collection of the information is surveillance. As the collection of information is controlled by the users, it is possible to make sure that the information will be of appropriate quality.

The most practical way to achieve this is through the use of properly structured disease surveys. Surveys have two further advantages: they can be quick to conduct, and relatively inexpensive (compared with the cost of running an effective passive reporting system). Active surveillance and survey techniques are discussed in Chapters 4 to 14.

Active surveillance uses structured disease surveys to collect high-quality disease information quickly and cheaply.

3

Principles of passive (general) surveillance

Introduction

The main method of collecting information on aquatic animal diseases currently used in most countries is through a passive disease reporting system. When disease is noticed in the aquatic animals, the producer may contact the authorities, who may then either submit a disease report, or send a specimen to a diagnostic laboratory. These reports and/or the results of examination of the specimens provide information on what diseases are present in the country. Information is not collected about all diseases, or all cases of disease. Many countries have compulsory disease notification regulations to encourage the reporting of priority diseases, but other diseases are not reported.

This system is called a passive reporting system because the main users of the information (the aquaculture or fisheries services) take no action to initiate the collection of the information. The producer initiates the report, and the central authorities wait (passively) for the report to arrive. Passive surveillance also includes the use of information that was collected for a different purpose, such as disease diagnosis.

> Passive surveillance is a system in which the information users make no active efforts to collect disease information; they just wait for disease reports to come to them.

General surveillance is an expression which indicates that the users of this type of surveillance are interested in collecting all sorts of information, not just information on a specific disease. In this text, general surveillance and passive surveillance mean the same thing.

Advantages of passive reporting systems

There are several distinct advantages to the use of a general or passive surveillance system. The first is the ability to identify newly introduced exotic diseases. The large-scale unrestricted movement of aquatic animals, their products, and particularly seawater as the ballast for ships, means that there is a large risk of diseases (or infectious disease agents) being spread to areas that were previously free. Eradicating these diseases after they become established is extremely difficult (if not impossible, in many cases). It is much easier to eradicate a newly introduced disease when it is still very localised.

In order to do this, there must be some mechanism to detect the disease quickly when it arrives in a country. Targeted or active surveillance systems focus on collecting information about particular known diseases—for instance, a surveillance system for white spot syndrome virus (WSSV) in shrimp may conduct regular sampling of shrimp from different locations, and test them with a polymerase chain reaction (PCR) test to detect viral DNA. If WSSV is present, the test is likely to detect it. However, the test will not identify the presence of any other viruses, such as yellow head virus. Targeted surveillance systems are able to detect only the disease at which they are targeted. Even if we are interested in only WSSV, sample collection for the targeted surveillance system may occur only once a year, or perhaps once every 2 or 3 years. In this case, it may be years after the virus was first introduced before the surveillance system is able to detect it for the first time.

A general surveillance system is able to overcome these problems. It does this by depending on the producers to detect disease, instead of relying on the fisheries or aquaculture authorities to go out looking for disease, as is the case with active surveillance. Clearly, the producers are not just interested in one particular disease agent, such as WSSV. They are concerned about any disease which causes losses to them. They will therefore be able to notice and report the presence of yellow head disease just as easily as white spot disease.

Another advantage is that producers are in contact with their stock most of the time. They will generally be able to detect disease very quickly, instead of waiting until the next scheduled targeted sample collection exercises.

A similar use of passive reporting systems is to detect and respond to new or emerging diseases. In contrast to exotic diseases, which are known, but do not occur in a particular country, a new disease is one which has not been previously recognised. With increasing intensification of aquaculture, and increasing environmental pressures on wild fisheries, new diseases are being recognised regularly. This is partly due to improvements in laboratory technology. Detecting new diseases is similar to detecting exotic diseases. However, while it is possible for a targeted surveillance system to detect the first incursion of an exotic disease, it is not possible for such a system to identify a *new* disease, as we don’t know what we are looking for.

If disease reports are supported by laboratory diagnosis, then the information gathered through a passive reporting system may be quite accurate. It is therefore possible, using a passive reporting system, to identify which diseases are present in a country, and where those diseases are located.

A final advantage of passive surveillance systems is that they are generally able to meet the basic requirements of the international animal health organisation (Office International des Épizooties, OIE).

Disadvantages of passive reporting systems

Under-reporting

The most important problem is that of *under-reporting*. Even when regulations make it compulsory to report a particular disease, reporting depends on several people. First, the producer must recognise that the animal is sick, and then notify the local officer. The officer must, in turn, submit a report to the central authorities or a diagnostic specimen to the laboratory. In many countries, there are several more steps in which, for instance, the report is passed through provincial or regional offices as well, depending on even more people. The weakest link in the reporting chain is usually the producer, who may not recognise the disease, or may fail to report it for other reasons. The result is that not all cases of disease are reported. It is rarely possible to estimate the level of under-reporting (although Chapter 13 describes one method), so it is impossible to calculate the total number of cases of disease.

Example

Two estuaries, Sickfish Sound and Healthy Harbour, are trying to assess QX disease of oysters. Both have about the same number of oysters. The authorities in Sickfish Sound have decided to start a campaign in which free testing is provided to all oyster farmers. Healthy Harbour has decided that it will continue to charge producers for QX tests. After one year, the records of the diagnostic laboratories in the two estuaries are examined. Sickfish Sound, with free testing, has had 78 confirmed cases of QX. Healthy Harbour has had 35. Which area has the highest level of disease?

The higher number of diagnoses in Sickfish Sound suggests that there is more disease in this area. However, there are many other possible reasons why the number of diagnoses is higher. There may be more disease there, or there may be more people testing for disease, resulting in more diagnoses. Alternatively, in order to fund the free testing, Sickfish Sound may be using a different test that produces more false positive results than the test used in Healthy Harbour.

The information from the laboratories cannot therefore be used to conclude which area has the highest level of disease.

Passive reporting systems are often not able to provide information on the total amount of disease, because of under-reporting.

One consequence of the problem of under-reporting is that it is not possible to tell if a disease that has never been reported is present in an area or not. The absence of reports may mean that there is no disease, or it may mean that it is only present at a low level, that it doesn't cause producers many problems, or that there are other reasons why it hasn't been reported.

Passive reporting systems are not able to demonstrate that a disease is not present in an area.

Another problem with passive reporting systems is that there are many different reasons why a report may not be made when aquatic animals are affected by disease. These reasons can be different for different areas, or different types of producers.

Example

A particular developing country, Mountania, has large areas of mountainous terrain, and a very poorly developed road system. Some large-scale commercial freshwater fish farms exist in the plains close to the capital city, but most fish are raised by smallholder villagers in the mountainous regions. Examination of the disease report records indicates that a large number of reports of epizootic ulcerative syndrome (EUS) have been received from the large commercial farms near the capital, but virtually no reports have been received from smallholders in the mountainous regions. What can we conclude about the distribution of EUS in this country?

The records suggest that the level of EUS is much higher in commercial farms than amongst smallholders, and that there is more EUS on the plains than in the mountains. Clearly, this conclusion is not necessarily true. Intensive farms near the capital are likely to be in regular contact with the authorities and report almost every outbreak of EUS. Smallholders in inaccessible parts of the country may not know that they should report the disease, or may be unable to contact local officers to make a report. The patterns of reporting are not reflecting the patterns of disease distribution but differences in the efficiency of the reporting system in different places.

Passive reporting systems usually cannot provide representative information on the level of disease in the population, or of the geographical pattern of disease. More reports may come from one part of the population than another.

A third problem with passive reporting systems is that the size of the population that the disease reports relate to is generally not known. This makes the calculation of useful measures of disease, such as rates and proportions, impossible.

Rates are measures of the frequency of an event in a population. **Proportions** are measured by percentages.

Rates and proportions allow the comparison of disease figures from populations of different sizes. Two commonly used measures are prevalence (a proportion) and incidence rate, discussed under Measures of Disease (page 56).

Example

Two neighbouring provinces, one with a larger number of villages raising shrimp than the other, are trying to control loose shell syndrome (LSS). Field reports over the previous year record 36 village outbreaks of LSS in the larger province, and 24 in the smaller province. Which province has the highest level of disease?

Although there are more outbreaks in the larger province, there are also more villages raising shrimp, so you would expect more outbreaks. To compare the provinces, it is first necessary to know the total number of villages in each province. The problem is that some of these villages might not submit a report, even if there was an outbreak (perhaps because they are inaccessible, or perhaps because the head of the village has had an argument with the local fisheries officer, and refuses to talk to him). The 36 reports in the larger province have come from villages which would report an outbreak. The total number of villages which would report an outbreak (but didn’t have one) is required to calculate the rate, but this figure is unknown.

Passive disease reports are not reliable enough to be used to calculate rates or proportions.

These problems with passive reporting systems limit the value of information collected. This is because we have information only about disease *reports*, but we need information about disease *events*. Passively collected information *cannot* be used to:

- determine the level and geographic pattern of the disease;
- determine the importance of the disease;
- set priorities for the use of resources for disease-control activities;

- plan, implement and monitor disease control programs; or
- demonstrate disease-free status to trading partners.

However, the information *can* be used to:

- detect incursions of exotic disease;
- detect new or emerging diseases;
- respond to disease outbreaks;
- identify which diseases are in the country (but not prove that some disease is not in the country), if diseases are correctly diagnosed;
- identify where the disease is located (but not identify areas where there is no disease); and
- meet the basic disease-reporting requirements of OIE.

Components of a passive or general reporting system

When designing a new passive reporting system, or trying to improve an existing one, it is important to keep in mind the above advantages and disadvantages, and, based on them, the objectives of the system. For instance, a set of objectives may be:

- to rapidly identify incursions of exotic aquatic animal diseases;
- to detect new or emerging aquatic animal diseases; and
- to meet international reporting requirements.

To achieve these objectives, the system should be designed, as far as possible, to overcome the main limitations of a passive reporting system—specifically under-reporting and biased reporting. The key components of a passive reporting system are:

- farmer reports;
- local fisheries and aquaculture officers;
- disease report form/outbreak investigation form;
- diagnostic laboratory;
- data management and analysis; and
- reporting and feedback.

Farmer and fisher reports

The first and most important part of any reporting system is the participation of the producers—the farmers and fishers. In the case of aquaculture, this will be the farmers. In the case of capture fisheries, this means reports from people involved in fishing. There are many things that people can report, but normally disease events or disease outbreaks are reported. It is also possible to report other types of information, such as production data. In order for producer reporting of disease or production to work, the producers need to know:

- why they should report;
- what to report;
- when to report it; and
- who to report to.

Why report

In a passive reporting system, producers act as the eyes and ears of the aquatic animal authorities. They are the people that we are depending on to identify and notify us of disease problems. It is therefore very important that they understand their role in the system, and what benefits it may bring them. There needs to be considerable effort put into educating the producers about the need for a surveillance system, what part they play in the system, and what they can expect to get out of the system.

One reason that producers may wish to report diseases is because the government services can provide laboratory diagnosis of what is causing the problem. Another is that they may get assistance from extension staff to fix the problem. In many cases, once a problem has been noticed, it is too late to fix it, so the only advice that can be offered is how to avoid the problem in the future. Unfortunately, there is still very little information about how to avoid many of the problems encountered, as they are still not fully understood. These are the problems that require more research, and these are the problems that we need to hear about more often. Why should a producer report a problem if they know that the authorities are unable to help with that problem? In some situations, the fisheries officers also have the responsibility to enforce regulations. In this case, producers may be reluctant to have any contact, for fear of getting in trouble.

The only way to effectively encourage producers to assist in reporting all sorts of problems is to explain the purpose of surveillance information. For many problems, the reason the farmer reports is to get direct personal benefit—help in fixing the problem. When this type of direct personal help is not available, the producers need to understand that disease reports enable the authorities to understand how important the problem is, where it is, what sort of producers are suffering from the problem, and therefore possible solutions to the problem. Reporting all problems will therefore improve the ability of the authorities to assist producers with those problems, even if the solutions are not available yet.

> Producers should understand the purpose of a passive reporting system, and how it can help them.

What to report

Producers also need to be told what sort of problems they should report, and this depends on the purpose of the surveillance system. Most producers will be experiencing different sorts of problems all the time. Most of these problems are common, and easily resolved. It would be impractical for everybody to report everything that ever happens. Instead, it is more practical to ask producers to report *unusual* or *severe* problems. If they have never seen this type of problem before, it may be an exotic or new disease, and therefore worth reporting. If the problem is severe, then steps should be taken to control it before it causes too much loss. A general reporting system is therefore not concerned with collecting *all* disease information, but trying to collect the *important* information.

Some passive reporting systems move from being general reporting systems to targeted systems, although they still depend on the producer to make disease reports. This happens when the list of priority diseases that producers are asked to report on is short. These are often established in government regulations or legislation, and are known as *notifiable* diseases. Producers are required by law to make disease reports if they suspect that their problem is caused by a notifiable disease.

Most countries have a list of notifiable diseases, in the expectation that, if the disease occurs, this will increase the likelihood that the disease will be reported. Unfortunately, notifiable disease reporting suffers from the same problems as any passive reporting—in order for a producer to report the disease, they must first know that they should report it, as well as know what the disease looks like if it does occur. Even if they do recognise it, many producers may be afraid to report the disease if there are negative consequences to reporting—such as destruction of the infected crop without compensation.

Notifiable disease systems focus attention on identified important diseases, but still can't guarantee complete reporting. Nor will they assist with reporting of new diseases that aren't on the list of notifiable diseases. Although international trade regulations often require that a notifiable disease system exists, it should be very carefully designed, to ensure that it has the effect of encouraging improved reporting, rather than discouraging reporting.

When to report

Farmers also need to be told when they should make reports. This will vary with the nature of the disease. If a rapidly spreading exotic disease is noticed, the report should be made as quickly as possible. However, if there is a mild disease that is commonly recognised, there may be no urgency to make the report, even though reporting is still encouraged so that the level and impact of the disease can be recognised. Many passive reporting systems have two different reporting levels. The first is emergency reports for important diseases, which should be made as soon as possible. The second is for routine reports for other diseases, that are made on some regular schedule.

Who to report to

Farmers need to know whom they should contact in case of a problem. Usually this will be the local fisheries and aquaculture officer, but other people may be more appropriate in some situations. Contacting the authorities should be made as easy as possible—remember that the producers are helping the authorities by reporting, so it shouldn't involve any extra work for the producers. In some countries, the authorities set up an emergency disease hotline, a free telephone number that can be used to report diseases at any time.

Local fisheries and aquaculture officers

Once a farmer has recognised a disease problem, decided to report it, and contacted the right person, the responsibility will then usually lie with the local fisheries and aquaculture officer. As with the farmers, the officers need to be trained in the operation of the surveillance system, and clearly understand their role. One of their most important jobs will often be to establish a relationship with producers, and to explain the importance of disease reporting to them.

When a report is made, the officer should investigate the problem, and record it. They should therefore know how to conduct a disease-outbreak investigation, as well as the standard procedures for filling out disease report forms. It is important to note that neither the farmer, nor the local officer needs to be able to make a definitive diagnosis of the disease problem. It is often very difficult to identify the causes of a problem simply by looking at the animals and the environment. The main requirement is that the officer should be able to look at the problem, collect appropriate specimens if necessary, and record what was found in a clear and organised way, so that others with more specialised expertise will be able to make a diagnosis based on that description.

Another critical role for the officers is in providing feedback to the producers. This is discussed in more detail below.

Disease report form/outbreak investigation form

Local officers are often not disease experts. They may have training in production and extension, but health is rarely their speciality. It is often hard for them to know which things that aquatic animal health experts need to know in order to identify the causes of an outbreak. In some countries, local officers know that they are required to make a report. When an outbreak occurs, they sit at the office typewriter and prepare the report. However, because they have been given no details as to what should be in the report, the information that comes from one officer may be completely different from that given by another officer.

To overcome this problem, it is very important that all people responsible for making disease reports be given a standard report form, and instructed on how to fill it out. This means that they won't waste time trying to imagine what information should be included, they won't waste time recording a whole lot of irrelevant information, and they won't miss important bits of information. A well-designed outbreak report form can make disease reporting much simpler and faster, and ensure that the information collected meets the needs of the users of the surveillance system.

The design of forms is discussed in more detail in Chapter 10, but there are a few points to keep in mind for the disease report form:

Level of diagnosis

There are many different ways that we can learn about disease problems. For instance, we may receive a specimen at a laboratory, perform a PCR test, and get a clear result that a particular virus is present in the tissue. On the other hand, we may hear from a farmer that, in discussions with another farmer, they heard that there was a big outbreak causing high mortality in the next district. The process of making a diagnosis and the confidence that we can place on different types of diagnosis were discussed in Chapter 2 (page 21).

When designing a reporting system, we have to decide what level of diagnosis is necessary for a report. If we want to be very confident about all the information we collect, then all reports should be supported by laboratory confirmation, and the report form should also be a laboratory submission form. If, on the other hand, we wish to collect all available information about diseases, whether that information is highly reliable or not, then we may design to form to allow any type of information to be entered. This may include reports of rumours, or of clinical examinations that have not been supported by laboratory confirmation, as well as laboratory diagnoses.

If one of the objectives of a passive reporting system is to minimise the level of under-reporting, then the system should accept all types of diagnosis, from rumour to Level 3 diagnoses (supported by highly sophisticated laboratory tests). The form should include a space to record the type of diagnosis.

Simplicity

The disease report forms will be used by a wide range of people. Mostly they will be local officers, but others may be asked to complete forms. The form should therefore be as simple as possible, as well as being as short as possible. As a rule of thumb, the form should be no longer than one page.

If the form is too long, and requires too much detail, both producers and officers responsible will become bored or irritated by the form, and fill it out less often, or less well. It is much better to have a small number of details filled out accurately for most disease outbreaks, than a large number of details filled out inaccurately for only a few outbreaks.

Diagnostic laboratory

The diagnostic laboratory is an important part of a general surveillance system. As discussed in the previous section, not all disease reports must be supported with laboratory tests, but where this is possible, it increases the confidence we can have in the diagnosis.

In order to support a general surveillance system, laboratories should be capable of diagnosing a wide range of different diseases. A highly specialised laboratory that only works with a few diseases is of little use, as it is not able to provide information about the range of different diseases that occur, or to identify exotic or new diseases that arise.

Laboratories should be located relatively close to the main production areas, because many aquatic animal tissues break down rapidly during transport.

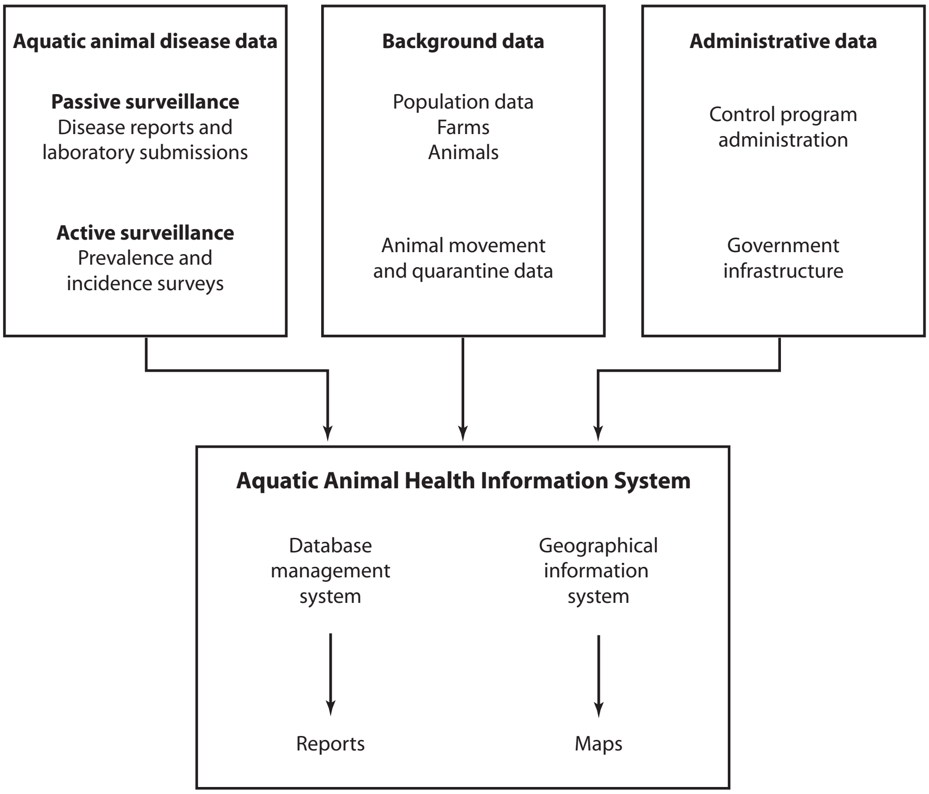
It is naturally very difficult for developing countries in particular to maintain a large network of laboratories capable of diagnosing all diseases, and distributed around all the major production centres. Many countries use a two-tier system to overcome this problem. The first tier is a network of laboratories with relatively simple, but general diagnostic capabilities—for instance, simple microscopy and possibly histopathology. These tests are able to characterise a wide range of diseases, and are relatively inexpensive. Histopathology slide preparation can even be contracted out to other nearby laboratories (such as human laboratories) to avoid the expense of maintaining the necessary equipment.

The second tier is a smaller number (perhaps just one or two) of specialised laboratories to assist in more complex diagnosis, that the first-tier laboratories cannot perform on their own. These second-tier laboratories may have more specific tests, specialised culture media, or access to equipment such as PCR machines or electron microscopes.

Data management and analysis

General surveillance systems, if they are working properly, tend to collect a very large amount of information. For this information to be useful, it has to be managed and analysed to provide the answers required by all the people involved in aquaculture and fisheries—the producers, extension staff, government policy makers, trading partners and so on.

Although it is possible to effectively manage large amounts of information using manual or paper systems, these are often very slow, prone to errors, and make data analysis very difficult. Efficient data management for a surveillance program therefore requires computerised systems. Data can be stored in a computer in many forms, such as text documents or spreadsheets. The best way to store and analyse disease information is to use a specialised database program. The diagram below gives an outline of the components of a typical aquatic animal health information system (AAHIS). A full description of the design and operation of a computerised, disease-information management system is beyond the scope of this book, but some of the important features will be highlighted below.


Aquatic animal disease data
Passive surveillance
Disease reports and laboratory submissions
Active surveillance
Prevalence and incidence surveys
Background data
Population data
Farms
Animals
Animal movement and quarantine data
Administrative data
Control program administration
Government infrastructure
Aquatic Animal Health Information System
Database management system
Geographical information system
Reports
Maps

Centralised data storage

All information on aquatic animal health should flow to one central location, and all the data should be stored on one computer system or network. It is possible for contributors (for instance at the district or provincial level) to maintain their own copies of the data, or to be given access to their own data on the central system. However, unless there is a single, central database, it is not possible to analyse the data effectively at the national level.

Disaggregated data

Data within the central database should be disaggregated. This means that the data represent information about the same unit as when they were originally collected, instead of being summarised to a higher unit. If information is collected about outbreaks of disease on farms, then details for each disease outbreak should be stored, instead of summary data about the number of disease outbreaks per district or province (summarised or aggregated data).

The problem with aggregated data is that they cannot be analysed in the same detail as disaggregated data. For instance, if we are interested in understanding the relationship between the sizes of farms and disease outbreaks, we may have information on the disease outbreaks, and the number of ponds in each farm in an area. Summarised data would tell us that farms have an average of, say, 4.2 ponds, and that there were 62 disease outbreaks in a given year. Disaggregated data would tell us, say, that farms with outbreaks had an average of 8.3 ponds, while farms that did not have an outbreak had an average of 3.2 ponds. This clearly suggests that there may be an important relationship between number of ponds and number of outbreaks, but this would be impossible to detect if only summarised data were available.

Distributed data entry

Data entry, or the process of taking data from paper forms and entering them into a computer, is a major task in an aquatic animal health information system. Instead of sending all report forms to a central location, it may be more efficient to spread the load of data entry by doing it at the provincial/state or district level. One advantage is that the amount of data to be entered by each person is much less, so the job is less boring and the accuracy of data entry may be better. Another advantage is that, if errors or missing data are found during data entry, it is much easier for provincial or district staff to contact the person who filled out the form and find out the missing data. The disadvantage of this system is that it requires more computers and staff who know how to use them.

Ancillary data

In addition to reports of disease outbreaks or other health (or production) problems, an AAHIS should include a variety of other ancillary data that will help with analysis and management of problems. Ancillary data should include at least some indication of the population, whether it be the population of producers of different kinds, or possibly the population of ponds. Other types of data that could be included in the system are:

- animal movement and quarantine data;
- data on control programs, such as inspections or treatments; and
- data on the fisheries and aquaculture infrastructure—the number of staff, offices, vehicles, boats etc.

As mentioned in Chapter 2, an effective surveillance system often requires a combination of active and passive surveillance. The information system should therefore be designed to manage information from active or targeted surveillance as well as passive surveillance.

Consistent coding systems

Examples of simple data analysis within an AAHIS include a list of reported diseases, and how frequently they have been reported, or a list of reports of a particular disease, broken down by the location of the farm. An example of a report on shrimp diseases follows:

Disease	Frequency
White spot	123
Loose shell	78
Loose Shell	69
Vibriosis	23
Vibrio	21
One month mortality	16
White Spot	12
Shell blisters	3
White gut	3
WSS	3
Zoothamium infection	3
Zoothamum infection	1
Total	355

This report clearly has a couple of problems. Several diseases appear twice, but spelt differently. To the human eye, it is clear that White Spot, White spot, and WSS are really all the same thing, but to the computer, they are entirely different. Looking at this table, it is hard to work out if loose shell syndrome is more common than white spot syndrome or not.

Similar problems can easily occur when identifying a location. The same village may have two different names, or two different spellings. When typing out the names of places, it is easy to make a typing mistake and get the name wrong. The solution to these sorts of problems is to have a consistent coding system built into the computer. When we talk about white spot, the computer automatically assigns a code of, say 233. The computer may also be programmed to recognise that White Spot and WSS should also be coded as 233. However, if we try to enter a disease called white spot, the computer will warn us that this disease is not known, and prompt for the correct name. When we produce a report, the computer searches out all diseases with a code of 233, and presents them all with the same name: white spot syndrome. Coding systems can make data entry faster and more reliable, and certainly make data analysis simpler and more accurate.

Automated analysis

There are many software programs available for the statistical or epidemiological analysis of disease information stored in a database. Some are more complex than others, some are very expensive, and some are free. However, with all of these programs, in order to analyse the data, you need to learn how to use the software to do the sort of analysis you want. Often the routine analysis of data (for instance, generating a report suitable for OIE), requires the user to do many different steps before the report is completed.

One of the great advantages of using a computer is that this process of routine reporting can be completely automated. All the different steps that are normally needed for producing the report can be stored in the computer, so that the report is generated simply by choosing a menu entry or clicking on a button. Most of the types of data analysis that are required for disease surveillance are routine, in that they are repeated at regular intervals. Automating these processes means that the system is faster and easier to operate, and can be done by somebody who doesn't need to know the details of how to generate complex reports from the beginning.

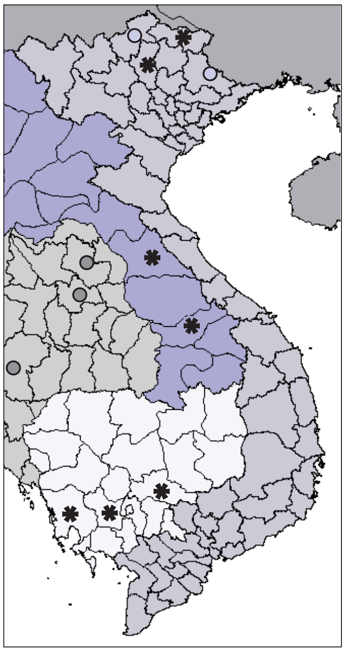
Computerised mapping

Another important advantage of the use of computers to store and analyse disease information is the ability to use computerised mapping systems to create maps showing the distribution of disease reports.

Understanding where disease is occurring, and the patterns of disease in time and space, can provide new clues about the causes of the disease, the way it is spread, and possible control mechanisms. As with other types of reports, it is possible to automate the process of generating maps, so that the computer does all the work.

One requirement for disease mapping is that the location of every disease event is properly recorded. This may be done at the provincial, district, village or farm level, by using a coding system such as that described above for diseases. For example, the computer may contain a standard list of every village in the country, along with its coordinates. When a disease report is received from that village, the data are entered, and the computer recognises the name of the village, assigning the correct code. To draw the map, the computer identifies the code of the village where the outbreak occurred, looks up the coordinates, and plots them on a map.

The map below shows the locations of disease outbreaks, identified at the provincial level, for part of Southeast Asia. This map was generated automatically as part of a regional disease-reporting system.



Ability for ad hoc analysis

Automated data analysis and automated mapping make routine reporting much faster and simpler. However, sometimes the automated reports are not able to answer all the questions. Any AAHIS should enable users to do specialised analysis of the data contained in it to answer specific or unusual questions.

Customised to local requirements

All countries are different, they all have different aquatic animal disease problems, and they all have different government structures responsible for managing aquaculture and fisheries. Any computerised information management system should be designed to meet the specific requirements of the users.

Reporting and feedback

What is the ultimate purpose of a surveillance system? While it may help us understand disease better, or provide information for international reporting, or improve our ability to make policy, the ultimate reason for a surveillance system is to benefit the producers. Until now, the description of a passive surveillance system has followed the movement of data from the producer to the local officers, to the laboratory and finally to a centralised computer system for storage and analysis. Up to this point, there has been no benefit to the producers at all.

There are two ways in which the providers of the information, the producers, can benefit from the system. The first is through appropriate analysis of the data, leading to correct conclusions about the causes of disease, and therefore better decisions about how to control or prevent disease. To achieve this, the results of the analysis must be delivered in an appropriate form to those responsible for making decisions about aquatic animal health.

The second way the producers can benefit from the system is in providing information from the system and the results of analysis directly back to the producers, so they can use it to make better decisions about their production and management. This information provided to the producers must be in a form that is appropriate and easy for them to use.

Providing information to producers is often difficult, but you can depend on the same systems that collected the information to provide the feedback. For instance, information can be passed from the central to the provincial, to the district and then to producer level, thereby providing feedback not only to the farmer, but everybody involved in the surveillance system.

Apart from providing direct benefit to the producers, feedback systems have two other very important purposes. The first is to continue to encourage data providers to play their role in the system. If producers or local officers continued to make disease reports, but never saw any evidence of how (or if) they were used, they would quickly become disheartened and stop making reports. Feedback demonstrates how important the providers are in the system and provides continual encouragement for ongoing disease reporting.

The second purpose of feedback is to ensure that the quality of data is good. If inaccurate data have been fed into the system, those at higher levels may not recognise it. However, when the data get back to the producer or local officer level, they will quickly see that the information is incorrect. This gives them a chance to correct it, and also acts as a stimulus to make sure information is reported more accurately in the future.

Sentinel surveillance systems

Various aspects of the passive or general surveillance system described above are designed to increase the level of reporting, and so overcome the main problem of all passive surveillance systems—under-reporting. However, no matter how good a passive surveillance system is, it is still very hard to get anywhere near complete disease reporting, or to avoid the biases caused by different rates of reporting in different areas and for different types of enterprises.

The distinctions between active and passive surveillance, and between general and targeted surveillance, are not always perfectly clear. Earlier, notifiable disease reporting systems were described. These are an example of a passive reporting system that is targeted instead of being general. In this section, we examine a different approach to surveillance—the use of sentinel producers—which represents a mix of active and passive surveillance to provide general surveillance data.

In a passive surveillance system, under-reporting occurs because only some producers report disease, and only some diseases are reported. To work perfectly, the system would have to collect all information from everybody, but this is impossible. An alternative approach is to attempt to collect almost complete information from a small number of producers, who are used to represent the rest of the producers. This is the concept of a sentinel surveillance system.

A sentinel system can complement the information collected from traditional active and passive surveillance systems. In a sentinel system, a relatively small number of producers work with the aquaculture and fisheries officers to keep a detailed record of disease and production information. The sentinel producers are chosen so as to give broad geographic representation. For instance, every district fisheries officer could be responsible for working with two local producers.

The advantage for the surveillance system is that detailed information can be provided from all parts of the country. The advantage for the producers is that they establish a close relationship with their local officer, who visits regularly to examine and analyse the data they collect, and make management recommendations based on the results. This relationship can also be used to assist with extension.

The scope of information collected from producers can vary, but may, for example, include key production statistics, such as numbers stocked, treatments used, and harvest details, as well as descriptions of disease or other unusual events. Although production data can be collected in other ways, sentinel producer systems offer a practical approach to collecting these data. They may be used for monitoring purposes, and for the early detection of disease problems.

Sentinel producer systems are therefore able to provide good information about the different diseases that are present, and the impact of these diseases in terms of production levels. If an unusual disease event occurs in a district, and most farms are affected, then the sentinel farm is likely to be affected as well. However, sentinel systems are not as useful for identifying rare diseases, or for detecting new or exotic diseases.

4

Principles of active surveillance

Requirements

In order to overcome the problems of passive surveillance, active surveillance must be able to:

- avoid problems of under-reporting;
- collect information which properly represents the true disease situation in the population; and
- gather data from a population of a known size to allow the calculation of rates and proportions.

Active surveillance

At the heart of the techniques used in this book is the use of active surveillance. As discussed in Chapter 2, traditional disease-reporting systems, such as compulsory disease outbreak notifications, or the use of laboratory submission data (passive surveillance) have a number of disadvantages. Under-reporting, expense and non-representative reports are some of the main problems. Active surveillance differs from a passive reporting system in that it uses surveys of a relatively small, representative sample of the population to gather specific information about that population. The key advantages of active surveillance are that the quality of information collected is usually better, the information reflects the true situation in the entire population, and it is often faster and cheaper to collect than with passive methods.

Passive surveillance

Despite their problems, passive reporting systems are an important source of disease information. Passive disease-reporting systems in one form or other are in place in virtually all countries, but relatively few countries make regular use of active surveillance, in spite of its advantages. This is partly due to the fact that appropriate techniques have not previously been available, and aquaculture and fisheries staff may not have been trained in the skills necessary. This book describes how to implement appropriate, active surveillance techniques and provides the skills needed to do so.

Disease surveys

Census

In order to produce *complete* reporting (and accurate measures of disease frequency), a passive disease-reporting system needs to gather information about every single case of important diseases in the country. To achieve this, every single animal must be regularly examined. This type of data collection is known as a *census*, where every member of the population must be checked. It is impossible for either fisheries and aquaculture services or the farmers themselves to achieve this detailed and regular examination of the entire population. This is why passive surveillance, census-based reporting always results in significant under-reporting.

Survey

Surveys are able to gather reliable information quickly and inexpensively because, instead of requiring a census in which the whole population is examined (by untrained farmers), only a small proportion (a sample) of the population is examined (by trained fisheries staff).

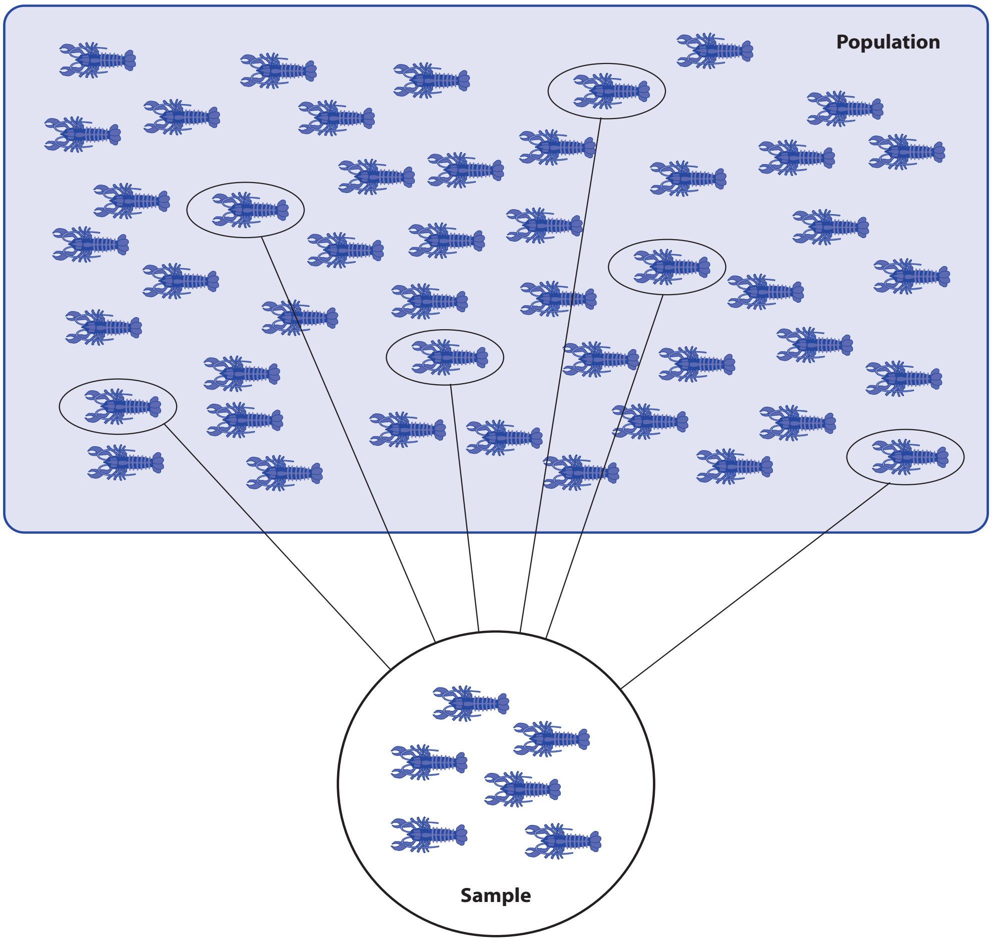
A census examines every member of the population. A survey examines only a small part of the population.

Populations, units of interest and samples

Population

This introduces two important concepts: the population and the sample. The population consists of all the things in a particular area that we want to know about. The population may consist of individual aquatic animals, but can also consist of other things (ponds, farms, boats, villages, producers). The things making up the population are known as the *units of interest*.

Units of interest



Example

We wish to know more about furunculosis in salmon (a bacterial disease of salmonids and other species), and decide to conduct a national survey to determine the level of the disease. The objective of the survey is to estimate the prevalence of furunculosis in the salmon of the country. The population that we are interested in is *all the salmon* in the country. The unit of interest is the animal (salmon).

Example

We are designing a survey to assess the economic impact of loose shell syndrome (LSS) in shrimp. The objective is to estimate the prevalence of ponds that have shrimp with LSS. To determine the level of disease, we are interested in a population consisting of *all shrimp ponds*. The unit of interest is the pond.

Example

We are studying the occurrence of marteiliosis (QX disease) in oysters. We want to calculate the proportion of estuaries that have experienced an outbreak of the disease in the past year. The population of interest is *all estuaries that have oysters*. The unit of interest is the estuary.

Example

A study is designed to assess the impact of epizootic ulcerative syndrome (EUS) in silver barb (*Puntius gonionotus*) in a reservoir. Researchers interview fishers at the reservoir to record how their catch has changed over the last year, and calculate the average loss per person. The population of interest is *all the fishers who fish in the reservoir*. The unit of interest is the fisher.

The sample is a small group of units (fish, ponds, estuaries, people) that have been selected from the population. Each element in the sample is then examined to collect disease information.

> A survey involves the examination of a small group (a sample) of elements (or units of interest) drawn from all the elements of interest (the population).

There is a problem with surveys. Once we have examined every element in the sample, we know exactly what the disease status of the sample is. However, we know nothing about the rest of the population that we have not examined.

Example

A national extension program has been started to support farmers with the control of *Lernaea* parasite in tilapia pond culture. We wish to monitor the effectiveness of the program, and conduct a survey to determine the prevalence of tilapia with *Lernaea*. There are eight million tilapia in the population. Instead of testing all the fish we take a sample of 200 fish, and examine them for parasites. We find that 36 fish (18% of the fish tested) have parasites. What proportion of the *population* has parasites?

There is no way of knowing what is the disease situation in the rest of the population, as we haven't examined the other 7,999,800 fish. It is possible (although very unlikely) that all of these animals have *Lernaea* and that in our sample of 200, we have tested the only 164 fish without parasites in the whole country.

Inference

If surveys give you a lot of information about a small number of animals, but nothing about the rest of the population, then what is the value of them? How can we use the results of a survey to learn something about the animals that haven't been examined?

The answer is *inference*. Inference is the process of estimating the true value of the disease status of the population, based on the results observed in the sample. In the above example, we could use inference to *assume* that the fish that were not tested were the same as ones that were tested. We therefore assume that the proportion of the population with parasites is *approximately* equal to 18%. This is the critical difference between the sample and the population. When we have finished the survey, we know exactly what the disease status of the sample is, but we can only estimate the disease status of the population.

> Inference is the process of assuming that the disease status of the population is similar to the disease status of the sample.

The danger with inference is that your assumptions can be wrong. If we have examined only 200 fish in the example, then it is very possible that the rest of the fish are quite different. The true proportion in the population may not be 18% but, for example, 53%.

Representative samples

While inference always runs the risk of being wrong, we can minimise this risk by ensuring that the sample selected is as similar to the rest of the population as possible. If the sample and the population are essentially the same (with respect to the characteristic of interest, or disease being measured) the sample is said to be a *representative* sample.

Example

Let us continue the above example. Most of the country's fish (seven million) are raised in the smallholder system in small ponds. There are only about one million fish in larger, more intensive farms. If all 200 fish in the sample were drawn from an intensive farm, they would be unlikely to be similar to the fish in village ponds. The sample would not represent the national population very well. A more representative sample would be one made up mostly of fish from village ponds with a small number from intensive farms. For instance, a sample consisting of 175 fish from village ponds and 25 fish from intensive farms would be more likely to be representative.

> A representative sample is one that is similar to the population. Inference is only valid when a representative sample has been chosen.

Bias

When, on average, the estimate from the sample is different from the true value in the population, the estimate is said to be *biased*. A single estimate from a survey will usually be slightly different from the true value, due to chance. However, if an identical survey is repeated many times, and the average result of the many surveys is different to the true value, the survey technique is said to produce a biased result. Bias can result from many different problems in a survey, most of which can be avoided through careful design.

Systematic error

Bias is caused by systematic error. Systematic error is an error that predictably causes the same type of error for each observation. For example, when weighing shrimp, if the scales used are incorrect, and always indicate that the shrimp are 5 grams heavier than they really are, then there is a systematic error in the results, and the estimate of the average weight would be biased. This is an example of *measurement bias*. Another important source of bias is *selection bias*, in which the sample selected is not representative, due to selecting animals which are systematically different from the rest of the population.

Bias is the difference between the average estimate from a survey and the real value in the population, caused by systematic error.

Selecting a representative sample is one of the more difficult tasks in any aquatic animal disease survey. One common situation that poses many problems is selecting a representative sample of farmers from within a village.

Example

Grouper farmers are being surveyed to assess the prevalence of viral nervous necrosis (VNN) in their cages. The fisheries authorities have been conducting a control program over the previous year, in which farm meetings were held to discuss ways to minimise the spread of the disease. There are 42 farmers in a village, and a total of 215 cages. Most producers have two cages, but a small number have up to 25 cages. Producers with only two cages are mostly located close to the village, but the larger farms are some distance from the village. How do you select a representative sample of 10 cages from the population of 215?

There are several methods that are commonly used in this situation:

- **Sample A**: On arriving in the village, either the head of the village or the village fisheries officer is contacted. This person then presents the survey team with ten cages.
- **Sample B**: The team goes to where most of the farms are, and moves from cage to cage, collecting specimens from every cage until they have ten specimens.
- **Sample C**: The team goes to one of the larger farms, and collects specimens from 10 of the cages in that farm.
- **Sample D**: The team wanders through the different farms selecting some cages from larger farms and some from smaller farms.

Each of these approaches is simple and practical, but in each case the sample selected is unlikely to be representative, and the survey results will probably be biased.

In Sample A, the survey team has no idea how the head of the village chose the cages. It is likely that they went to friends and cooperative farmers nearby. The farmers most likely to have attended the meetings and to be practising control strategies are those who are friends of the head of the village, and who live near the middle of the village. Farmers living outside the village, or who are not friends with the head are much less likely to have been to the meeting, or to be chosen during the survey. The result is that the sample is not the same as the rest of the population, and the results are likely to be biased.

Samples B and C suffer from similar problems. Sample B is not representative because the larger farms away from the village are not represented at all. In sample C, only one large farm is represented and none of the other farms. It is possible that this producer didn't attend the meeting, while most of the rest of the village did. The survey would indicate a high level of VNN in the grouper, while the true proportion in the village could be very low.

Sample D sounds better as both the larger cages on the edge of the village and the smaller cages at the centre would be represented. But there is still a serious danger of bias. This is because, even when trying to be representative, people tend to chose individuals for different reasons. For example, the survey team may subconsciously select mostly small cages and cages with easy access. This is because it is much easier to collect specimens from these cages. The problem is that larger cages may have more fish, and less accessible cages may receive less attention, and are therefore not representative of the population. The survey results would be biased, underestimating the true proportion of grouper with VNN.

Random sampling

There is in fact only one way to be confident that the sample chosen is representative of the population. To select a representative sample, we must ensure that every cage (unit of interest) in the population has the same chance of being chosen in the sample, regardless of its owner, location, size or any other characteristic. Sampling techniques of this sort are known as random sampling.[1] Random sampling has a number of other important advantages, and is discussed in detail in Chapter 5.

> Random sampling means that every element (unit of interest) in the population has the same probability of being selected in the sample. Random sampling is the only way to reliably select a representative sample.

Estimation

The aim of surveys is to determine some characteristic of the population (for example, the proportion of animals with signs of a disease, or the total number of fish in a lake). As only a sample of the population is examined, and inference used to make assumptions about the rest of the population, it is likely that the value measured from the sample will not be the same as the real value in the entire population. Random sampling is used to minimise this risk. Because we cannot know what the real value in the population is, we use inference to estimate it.

Example

A pond has 4500 shrimp. A survey designed to measure what proportion has tail rot is based on a random sample of 20. Fifteen of the 20 shrimp (75%) have signs of tail rot. We therefore estimate that the proportion with tail rot in the pond is 75%.

Sample size

When considering the results of a survey like this, it is important to have some idea of how good the estimate is. The number of animals selected in the sample (the *sample size*) is one of the most important factors that determines how close our estimate is likely to be to the true population value. In the above example, 20 shrimp were chosen. If instead we had chosen 2000 shrimp from the 4500 in the pond, it is

[1] To be slightly more precise, simple random sampling, in which each element has the same probability of selection, is just one of a group of probability sampling techniques. These share the feature that each element in the population has a known, non-zero (but not necessarily equal) probability of selection. Probability sampling is the only reliable way to avoid selection bias, and is required if estimates of population values are to be valid. Samples chosen with probability sampling may still be unrepresentative due to chance, but on average, they will be similar to the population in all ways.

clear that the estimate of the proportion is likely to be much more precise. On the other hand, if we had chosen only four shrimp for our survey, we would not have very much confidence that the result was accurate.

Surveys with large sample sizes produce more precise estimates.

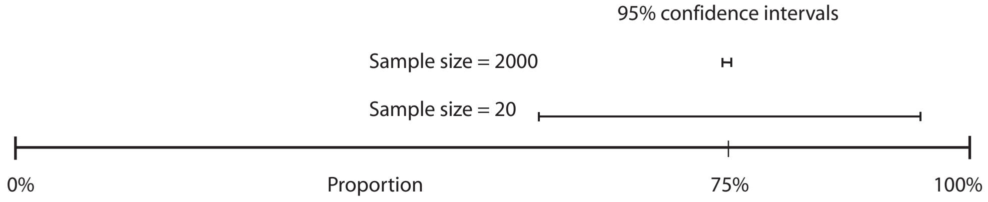
If we consider these three possible surveys of the shrimp in the pond, with three different sample sizes—4, 20 and 2000—it is still possible for each of the surveys to produce an estimate of 75%. However, we would be much more confident that the true population value was closer to 75% if we had used a sample size of 2000 rather than 4.

Confidence interval

When interpreting the results of a survey it is useful to have a measure of how precise the estimate is. This will tell us how much confidence we can have in the results. When we use random sampling, it is possible to calculate such a measure, called the *confidence interval.* A confidence interval indicates how close to the real population value our estimate is likely to be. All estimates from surveys should be reported with confidence intervals, so the users know how reliable they are.

The confidence interval of a proportion is a range of values within which we are confident the real value is likely to be. For instance, in our survey of 20 shrimp, with an estimated prevalence of 75%, the 95% confidence interval is 51–91%. This means that our best guess at the real prevalence is 75%, but we are 95% confident that, even if we are wrong, it lies between 51% and 91%. The real value is probably around 75%. It is possible, but much less likely, that it is closer to 51% or 91%. '95% confident' means that, if we did the same survey in the same way 100 times, even though we would probably get different estimates each time, the true value would lie within our confidence interval 95 times out of 100.[2] This confidence interval is quite wide. Although we think that the value is around 75% it could well be as low as 51%, which is quite a big difference.

Consider the same survey, using a sample size of 2000 instead of 20. If we observed 1500 shrimp with tail rot, then our estimate would still be a proportion of 75%. However, the 95% confidence interval would be 73%–77%. This means that we are 95% confident that the real value is between 73% and 77%. We can put a lot more faith in the second survey than the first as our estimate is much more precise. This is shown below.



[2] The probabilistic interpretation of a 95% confidence interval is as follows. If a survey using the same methodology and sampling strategy were used to study the same population many times, and a confidence interval calculated in the same way based on the result of each survey, the true population parameter would fall within the confidence interval 95% of the time.

By convention, the precision of an estimate is usually described by the 95% confidence interval. It is possible to calculate confidence intervals at different levels of confidence, such as 90% or 99%, but these are used much less commonly.

> A confidence interval indicates how confident we are that the estimate is correct. We can be 95% sure that the true population value lies within a 95% confidence interval. The smaller the confidence interval, the more precise the survey result..

Survey accuracy

In summary, the accuracy of the estimate from a survey is determined by two factors —*precision* and *bias*.

Precision; Random error

If the same survey is performed on a population many times, the answer will be slightly different each time. This difference is called random error, and is represented by the width of the confidence interval. If the differences between survey results are small, then there is low random error, and the survey results are very precise. Precision is determined mainly by sample size. A survey with a large sample size will have a smaller random error, and be more precise.

Bias; Systematic error

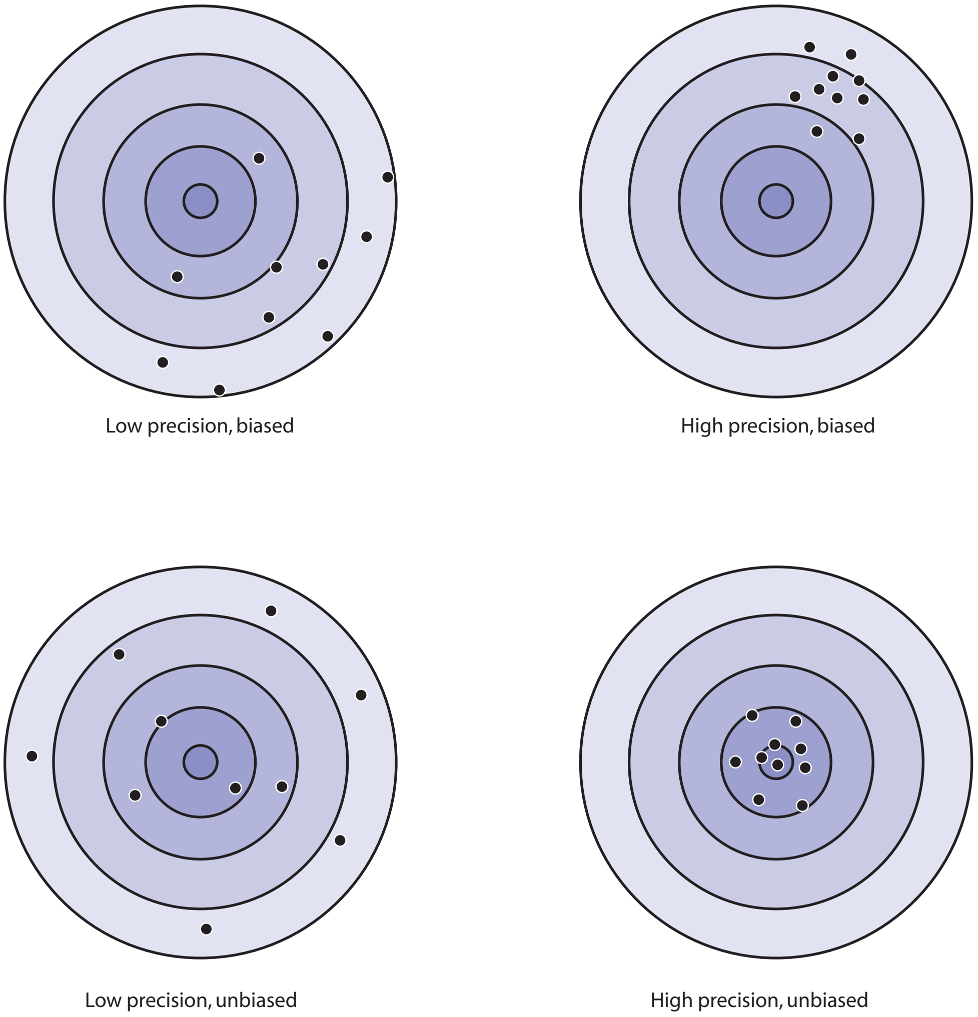
If a survey is repeated many times, and the result is always different from the true value by about the same amount in the same direction, this is called systematic error. Systematic error causes biased results, and this is controlled mainly through good survey design.

Example

Conducting a survey to estimate a value in the population is like shooting a gun to hit a target. If you are not a very good shot, you will hit all over the target. Your shooting is not precise, because there is a lot of random error. In order to overcome the random error, you need to use a lot of shots (use a large sample size) before you are likely to come close to the centre of the target. If you are very good, all the shots will hit very close to the same point. There will be very little random error, and the result will be much more precise. However, you can hit the centre of the target only if the sights on the gun are good. If the sights on the gun are not straight, then when you aim at the centre of the target, you will hit off to one side, and the result will be biased. If there is no bias, then you will hit the centre. This is shown in the figure on the next page.

The real population value that we are trying to estimate is the centre of the target. With each survey (or gun shot) we get a result which may be close to the true value or a long way away. If the average result of all the surveys (middle of the pattern of gun shots) is close to the true value (centre of the target) then there is no bias, as shown in the bottom two figures. If, however, the average result is centred away from the true result (top figures) bias is present, caused by systematic error (for example, incorrectly set gun sights).

If there is a wide spread of different results from different surveys, it is due to random error, as shown in the two figures on the left. A very good shot will have less random error and a more precise result, as shown on the right.



Low precision, biased

High precision, biased

Low precision, unbiased

High precision, unbiased

Measures of disease occurrence

In order to control disease effectively, it is first necessary to understand the distribution of the disease: how much disease there is, where it is, in what animals, and so on. Disease surveys are based on counting the number of animals, ponds, farms etc. with a disease, and the numbers without. These counts can then be used to calculate several different measures of disease, each describing the level of disease in a different way. The two most important measures of disease for active surveillance are prevalence and incidence rate.

Prevalence

Prevalence (sometimes called point prevalence) is a measure of the number of animals with the disease of interest at one point in time, as a proportion of the total number of animals in the population.

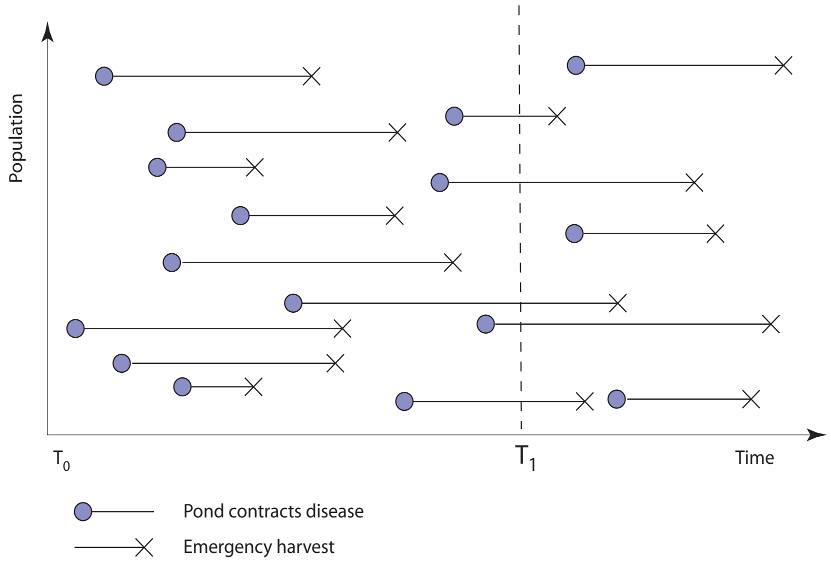
Example

An intensive shrimp farming area with 2000 ponds suffers an outbreak of white spot disease. The first ponds start to show signs of sick shrimp on 3 March. By 29 March many ponds have sick and dead shrimp. The local fisheries officer visits on 30 March. On that day, the officer counts 56 ponds showing signs of disease, and the producers report that a further 143 have already been emergency harvested, and 28 ponds had been diseased, but recovered. There are 1801 apparently unaffected ponds remaining. What is the pond prevalence of white spot disease in the area on 30 March?

The number of ponds with the disease is 56. The total number of ponds in the population is 1857 (2000 – 143 ponds that have already been harvested). The prevalence on 30 March is therefore:

Prevalence = 56/1857 = 3%

The figure below illustrates the idea of prevalence using a small part of the population over a period of time. Each line represents a single pond with disease. The line starts when the pond first becomes affected, and stops when it is harvested or recovers.



In the figure, prevalence is measured at time T1. At that time, the number of affected ponds is counted, giving a total of 5. If there were 16 ponds in the population at time To, and 6 had been harvested before time T1, the remaining population is only 10 ponds. The prevalence would be 5/10 or 50%. Note that prevalence is measured at a particular point in time, and there is no time period used in its calculation.

Prevalence is the number of sick animals, ponds or other units of interest at a single point in time as a proportion of the total population at risk at that time.

$$\text{Prevalence} = \frac{\text{Number of cases at one point in time}}{\text{Population at risk at the same point in time}}$$

Incidence rate

Incidence rate (specifically true incidence or incidence density rate) is a measure of the average speed at which the disease is spreading.[3] Incidence rate is the total number of new cases of disease divided by the total time that each animal in the population was at risk of getting the disease. For simplicity this can usually be calculated as:

$$\text{Incidence rate} = \frac{\text{Total new cases of disease during a period of time}}{\text{Average number of animals at risk} \times \text{Time period}}$$

In our example of the outbreak of white spot in the shrimp farms, we can use the figures to calculate the incidence rate, or rate of progress of the outbreak. If we use the length of time from the beginning of the outbreak (3 March) to the date of the visit (6 March) then the time period is 4 days (inclusive). The total number of new cases of disease during these 4 days is equal to:

- the 143 ponds that were emergency harvested,
- plus the 28 ponds that were affected and recovered,
- plus the 56 ponds that were affected at the time of the visit,

giving a total of 227 new cases of disease. The average number of ponds at risk can be calculated by taking the average of the number at the beginning of the time period and at the end of the time period. At the beginning (on 3 March), there were 2000 ponds. At the end (6 March) 227 of those ponds had either been harvested, or had already had the disease, and so were not at risk of getting the disease again. The population at risk was therefore 2000 – 227 = 1773. The average population at risk was (2000 + 1773)/2 = 1886.5. The incidence rate can therefore be calculated and expressed in several different ways:

$$\begin{aligned}\text{Incidence} &= \frac{\text{227 new cases of disease}}{\text{1886.5 days at risk} \times \text{4 days}}\\ &= \text{0.03 cases per pond per day}\\ &= \text{21 cases per 100 pond-weeks at risk}\\ &= \text{21 cases per 100 ponds per week}\end{aligned}$$

What do these numbers mean? The first number, 0.03 cases per pond per day means that if we look at one pond for one day, on average we will get 0.03 cases of disease. This is obviously meaningless when we talk about a single pond. The value for incidence rate can be multiplied by a larger number of ponds or a longer time period to make it easier to understand. The second figure, 21 cases per 100 pond-weeks, means that if we have 100 ponds, we could expect 21 of them to become infected in the space of one week, if the rate of spread of the disease stayed the same as it was during the first 4 days of the outbreak.

[3] Another commonly used measure of incidence is cumulative incidence, which is equal to the number of new cases of disease as a proportion of the total number at risk at the beginning of an observation period. At the individual animal level, it estimates the risk of getting the disease. Both incidence density rate and cumulative incidence depend on a measure of the total number of cases of disease in a particular time period, as discussed in Chapter 13. The text deals with only the incidence density rate, but the cumulative incidence can be calculated if desired.

We can use the diagram above to see how incidence rate differs from prevalence. To calculate the incidence rate, we need to count all the *new* cases of disease in a given period. The number of new cases of disease between time T_0 and time T_1 is 13. Some of these cases were harvested, some recovered, and some were still affected at T_1, but we are only interested in the number of new cases, not what happens to them. If there were 26 ponds in the population at time T_0 and 20 left at T_1 (two of which had already been affected and recovered, so were not at risk), then the average number at risk over the period was (26 + (20 – 2))/2 = 22. If the time period between T_0 and T_1 was 1 month, then the incidence rate is:

$$\begin{aligned}\text{Incidence} &= \frac{\text{13 new cases of disease}}{\text{22 at risk } \times \text{ 1 month}}\\ &= \text{0.59 cases per pond-month}\\ &= \text{59 cases per 100 ponds per month}\end{aligned}$$

Incidence rate measures the number of new cases over a period of time.

Prevalence versus incidence rate

Relationship between incidence rate and prevalence

Many other measures of disease exist, but prevalence and incidence rate are the most useful. The two are related, according to the duration of disease. A disease with a high incidence rate but of very short duration will have a relatively low prevalence. A disease of relatively low incidence rate with a long duration will have a high prevalence.

Example

A study is conducted of silver barb in small-scale pond culture to understand the impact of opportunist aeromonad septicaemia. The study population consists of all the fish in a pond. Because the occurrence of aeromonad septicaemia fluctuates due to a variety of causes, it is found that, over a period of one year, the incidence rate of the disease is high, with 95 cases per 100 animals per year. This means that almost all fish in the pond are affected by opportunist aeromonad septicaemia at some stage. However, the duration of the problem is usually relatively short, lasting only a day or less. At any one point in time, the prevalence (proportion of fish suffering from opportunist aeromonad septicaemia) is very low, at only 0.3%.

Example

We are interested in assessing the effects of *Lerneae* in common carp (*Cyprinus carpio*) in a reservoir. Most fish are infected when they are young, and keep the infection at low levels for much of their lives. The incidence rate (number of new infections) is relatively low at 8 new cases per 100 animals per year, as most animals already have the infection, and only young fish are susceptible to new infections. Because the duration of the disease is virtually lifelong, the prevalence is very high, at 97%.

The reason prevalence and incidence rate can be so different is because they are describing different aspects of the disease. If the size of the population is not changing, and the level of disease stays about the same, then we can estimate the prevalence if we know the incidence rate and the average duration of the disease.

Under certain specific conditions when prevalence is low (‹10%)
Prevalence

Example

Skin damage due to contact with the nets has become a problem in a salmon farm. The incidence rate of the disease is 5 cases per 100 fish per month. Affected fish usually recover in about 1 month (the duration of the disease). The prevalence of the disease is therefore about 5 cases per 100 fish per month times 1 month, or 5%.

Features of incidence rate and prevalence surveys

When planning a survey, we must decide which measure is most useful for a particular purpose. This is based both on the type of information that is collected and practical considerations.

Example

A survey is being conducted into the effect of algal crashes on disease in farmed shrimp. An algal crash occurs when a pond has a dense bloom of microalgae, and it all dies suddenly, in the space of a day or two. The researchers decide to do a survey to estimate the prevalence of algal crashes. They select a sample of 500 farms and visit each of them to examine the ponds to see if any are having an algal crash at the time of the survey.

At the end of the survey, the researchers have found only three ponds having algal crashes, or a prevalence of 0.6%. With such a low prevalence, it is very hard to determine if there is any relationship between crashes and disease. The researchers decide to use an alternative approach.

Example

Instead of a prevalence survey, they decide to measure the incidence of crashes. One approach would be to visit farmers, give them a recording sheet, and ask them to write down any crashes that occur. The survey team would revisit the farms 4 months later, and collect the data about the total number of ponds that suffered crashes. Unfortunately, there is not enough time to wait for the crashes to occur. Instead, the survey team decide to do a retrospective survey, and find out if there have been any crashes since the ponds were filled. To do this, they visit the farms only once, and take a sample of mud from the bottom of each pond. If there has been an algal crash, there will be a layer of black in the mud.

Using this approach, they discover that 40 out of 100 ponds surveyed have suffered an algal crash over the previous 6 months, or an incidence rate of 80 crashes per 100 ponds per year.

The difference in the two approaches is that they are measuring the same thing (algal crashes) using two different approaches—one with a very short duration (actually seeing the crash) and one with a long duration (a layer of black mud, indicating that a crash had happened at some stage in the past). For diseases or events with a very short duration, prevalence surveys are difficult. Incidence surveys may be able to give a much better measure of the level of disease, particularly if there is some way to find out if it has occurred in the past.

The following table compares the two types of surveys:

	Prevalence survey	**Retrospective incidence rate survey**
Cost	High	Low
Speed	Slow	Fast
Unit of interest	Animal, pond, village	Village/pond
Data quality (specificity)	Good	Moderate
Identifies	Disease	History of outbreaks

Diagnostic tests

Many disease surveys require the use of laboratory diagnostic tests to examine specimens collected from the animal. Examples are the use of laboratory tests such as ELISAs (enzyme-linked immunosorbent assays), histopathology, PCR (polymerase chain reaction) or biochemical tests.

Very few laboratory tests are perfect, although most tests give incorrect results only occasionally. When using a laboratory test as part of a disease survey, it is important to understand how accurate the test is, and what errors are likely to occur.

Sensitivity and specificity

The performance of a test is described by its *precision* (lack of random error, measured by the repeatability of a test) and *validity* (lack of systematic error). Validity is described by two measures—the *sensitivity* and the *specificity*. These terms have a special epidemiological definition which is different to the definitions used by laboratory scientists. The sensitivity of a test measures the proportion of truly diseased animals that the test correctly identifies as diseased. The specificity measures the proportion of non-diseased animals that the test correctly identifies as non-diseased.

Example

A new diagnostic test to detect infectious pancreatic necrosis virus (IPNV) in trout is being evaluated, and is used to test 20 fish. Ten of the fish are already known to be infected with the disease. The other 10 fish come from a disease-free area. When testing is finished, 8 of the 10 antibody-positive fish return positive results, and 2 are negative. Seven of the antibody-negative animals return negative results, and 3 are positive. The results are summarised in the table below.

		True disease status		
		Positive	Negative	**Total**
Test result	Positive	8 true positive	3 false negative	**11 test positive**
	Negative	2 false negative	7 true negative	**9 test negative**
	Total	**10 disease positive**	**10 disease negative**	**20**

Calculating sensitivity and specificity

Using these figures, we can calculate the sensitivity and specificity of the test.

$$\text{Sensitivity} = \frac{\text{8 true positive results}}{\text{10 disease positive animals}} = 80\%$$

$$\text{Specificity} = \frac{\text{7 true negative results}}{\text{10 disease negative animals}} = 70\%$$

> The sensitivity of a test is the proportion of truly diseased animals in the population which are correctly identified by the test as being diseased.

> The specificity of a test is the proportion of truly non-diseased animals in the population which are correctly identified by the test as being non-diseased.

Test sensitivity and specificity

When studying disease using any sort of test, it is very useful to know what the sensitivity and specificity of the test are. For instance, if you test a batch of shrimp post-larvae before stocking, to check if there is any mondon baculovirus present in the batch, the test used is unlikely to be perfect. If, for instance, the specificity is 95%, it means that, on average, out of every 100 disease-free shrimp, you would expect to get five positive test results (false positives). If the sensitivity is also 95%, you would expect to see 5 negative results for every 100 diseased shrimp (false negatives). If we test 200 post-larvae, and get 12 positive test results, what does it mean? We know that there is a chance of producing false positives, and we would expect about 10 false positives from a batch of 200. However, if there are 12 positive results, does it mean that some of those 12 results are false positives and some are true positives, or are they all false positives? On the other hand, we got 188 negative test results. Are all of these true negatives, or are there some false negatives amongst them? It is not possible to answer these questions definitely, because the true disease status of an animal is very rarely known. If we know the sensitivity and specificity of the test, we can, however, use probability theory to calculate how likely it is that the batch is infected or not. This will be discussed in detail in Chapter 14.

Consider another example: if we test for pancreas disease a single salmon from a cage, and the test gives a positive result, how likely is it that the fish is actually diseased? To answer this we need to know three things: 1) how common is the disease in the population (the prevalence), 2) how likely is it to be a false positive (specificity), and 3) how likely is it to be a true positive (1 – sensitivity). Without these three pieces of information it is not possible to determine if a positive test result really means that the animal is diseased. For instance, if the sensitivity is 96%, the specificity is 99% and the prevalence is 1%, then the probability that an animal that tests positive is actually diseased is only 50%. The formula for calculating the probability (the positive predictive value of a test) is given in Appendix A.

It is common to talk about the sensitivity and specificity of laboratory tests. However, any process that is used to determine if the true state of something is either positive or negative can be thought of as a test, and also has a sensitivity and specificity. For instance, a questionnaire to determine if producers are using antibiotics to treat disease in their ponds may be considered a sort of test. After answering questions, we get a test result—positive (the farmer is using antibiotics) or negative (the farmer is not). There is a chance that we might get the wrong answer, for all sorts of reasons. Maybe some farmers know that they should not be using antibiotics, and are reluctant to tell the truth. This would result in low sensitivity, or some false negative results. On the other hand, the question may be badly worded, and imply to farmers that they should be using antibiotics. Some farmers may be embarrassed that they don't and pretend that they do use them, giving a false positive result or low specificity.

Although it is rarely done, calculating the sensitivity and specificity of questionnaires or interviews can help avoid errors when the data are interpreted.

Calculating sensitivity and specificity

In the example calculation above, a number of animals with known disease status were tested, and the results of the test compared to the true disease state of the animals. This illustrates the principle of calculation of sensitivity and specificity but, unfortunately, it is rarely this simple in the real world. This is because, if virtually all tests make mistakes, it is very difficult to know the 'true disease state' of an animal. The other difficulty is that, while we often talk about sensitivity and specificity as fixed characteristics of a particular test, they can vary somewhat due to a number of factors. For instance, the stage of disease may affect the sensitivity, or the presence of other microorganisms that cross-react with the test may affect specificity.

Even though it is difficult to work out estimates of sensitivity and specificity, and they are not always accurate, these measures are extremely important if we wish to avoid making errors when analysing the results of tests. It is therefore often well worth while to spend some time and effort before a major study, working out the sensitivity and specificity of the tests that will be used. There are several different approaches that may be used.

Use of a 'gold standard'

The term 'gold standard' is meant to indicate a test that is able to determine the true disease status of an animal. In reality, there are virtually no such tests, so there is no such thing as a gold standard. In practice this term is used to describe the best test or combination of tests that is available. Evaluating the sensitivity and specificity using a gold standard is really just determining the relative sensitivity and specificity, compared with another test. If the so-called gold standard is much better than the test being evaluated, then there is little difference between the relative sensitivity and specificity and the true sensitivity and specificity. However, in some cases, the test being evaluated may be better than the gold standard. In this case, estimates of sensitivity and specificity will be incorrect.

An advantage of using a 'gold standard' is that animals from the population of interest can be studied to determine the test performance.

Animals of 'known disease status'

An alternative approach to the use of a gold standard is to test animals of known disease status. This generally means that samples are collected from animals in an area where the disease is known not to exist (either in the same country, or another part of the world), and tested to calculate the specificity. Sensitivity is determined by testing animals that are known to be diseased. These animals are either animals showing clear signs during an outbreak, or maybe experimentally infected animals.

This approach has the advantage of not depending on a potentially imperfect gold standard. However, there are a number of important disadvantages. First, it requires that we 'know' that some animals are not diseased, and some are. While we may be able to do this with a high degree of confidence, it is always possible to make a mistake. For instance, if we take samples from 'negative' animals in a country where the disease is 'known' not to occur, it is possible that a non-pathogenic strain of the disease does exist in that country, and could interfere with the test results.

The other disadvantage is that the animals tested are not from the same population as the animals that are to be studied. As sensitivity and specificity can sometimes vary between different populations, the estimated values, while correct for the populations tested, may not be appropriate for the population that you wish to study.

Modelling

A third approach to calculating sensitivity and specificity is the use of modelling techniques. A range of techniques exists, of varying complexity, the details of which are beyond the scope of this book. The principles behind two relatively simple approaches will be described.

Any mixed population of diseased and non-diseased animals can be thought of as two distinct populations, based on their disease status. If examined with a test that produces a multi-level outcome (such as an ELISA's optical density result), each separate population will produce a range of results. Some diseased animals will produce a low result, but most will have a higher result. Similarly, some non-diseased animals will have a high result, but most will produce a lower result.

If the results from a large number of individual tests are examined, they will produce a combined distribution that is made up of two separate distributions, one for the diseased population and one for the non-diseased population. The two component distributions can be modelled, usually using either a normal or log-normal distribution. Mixture population modelling is the process of estimating the parameters of the underlying distributions based on the observed combined distribution. This can be achieved easily using a spreadsheet's 'solver' function, by minimising the difference between a predicted combined distribution (based on the parameters defining the separate distributions) and the observed distribution. Once the parameters for the diseased and non-diseased population distributions have been estimated, the test cut-off can be used to calculate test sensitivity and specificity. An example spreadsheet implementing mixture population modelling is included on the CD.

Another simple modelling approach can be used when two different tests have been performed on a large number of specimens. This system is based on the principle that the population tested has a true, but unknown disease prevalence. This prevalence can be estimated if the sensitivity and specificity of a test is known. By choosing arbitrary values for sensitivity and specificity for each of the two tests, two estimates of the underlying prevalence can be made. Using a spreadsheet's 'solver' function again, the sensitivity and specificity estimates can be varied to minimise the difference between the two estimates of prevalence.

Combining tests

Often the performance (sensitivity and specificity) of a laboratory test is not as good as we would like it to be. One way to address this problem is to test each specimen with two (or more) 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 disease-eradication campaign, the specificity of the main test is not perfect, and when used to test very large numbers of negative animals may produce many false positive reactors. One approach is to test animals that produce a positive test result with another test (based on a different biological pathway). 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, and 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 re-tested. The result of this approach is to increase the specificity, but decrease the sensitivity. The overall sensitivity (Set) and specificity (Spt) 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 re-tested, and is 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.

Apparent versus true prevalence

If the sensitivity and specificity of a test are known, it is possible to correct for errors that occur because of imperfect tests. When a prevalence survey has been conducted, and the specimens analysed, the observed prevalence (*apparent prevalence*) may not be correct. This is because some of the diseased animals have been incorrectly identified as being negative, and some of the non-diseased animals have been identified as being positive. These errors can be corrected to give the *true prevalence* with a formula using test sensitivity and specificity.

The **True Prevalence** program performs this calculation for you. By entering the apparent prevalence, sensitivity and specificity, it will calculate the true prevalence. If you also enter the sample size of the survey, it will give you a 95% confidence interval. The program and example calculation are shown on page 191.

What survey to use?

This chapter has introduced the use of surveys for active surveillance to collect information on aquatic animal diseases, and discussed two important measures of disease. A third survey type, surveys to demonstrate freedom from disease, will be introduced in Chapter 14. The aquaculture and fisheries authorities and survey planners need to decide which type of survey or other data collection system is most appropriate for their needs.

To make this decision, you first need to decide what type of information is needed and what it will be used for. In some cases, this leads directly to the type of survey required. In others, it may be clear that the required information is already available from other sources and no survey is required. Different surveys not discussed in this book may be needed in some situations.

Example

You wish to export tiger shrimp and are involved in trade negotiations with another country. They are refusing to allow the import of live shrimp on the basis that they believe your country has white spot disease (WSD). Although there have been no clinical reports of WSD disease in the past 5 years, you need to be able to demonstrate that the disease has been eradicated. This situation clearly calls for a survey to demonstrate freedom from disease, described in Chapter 14.

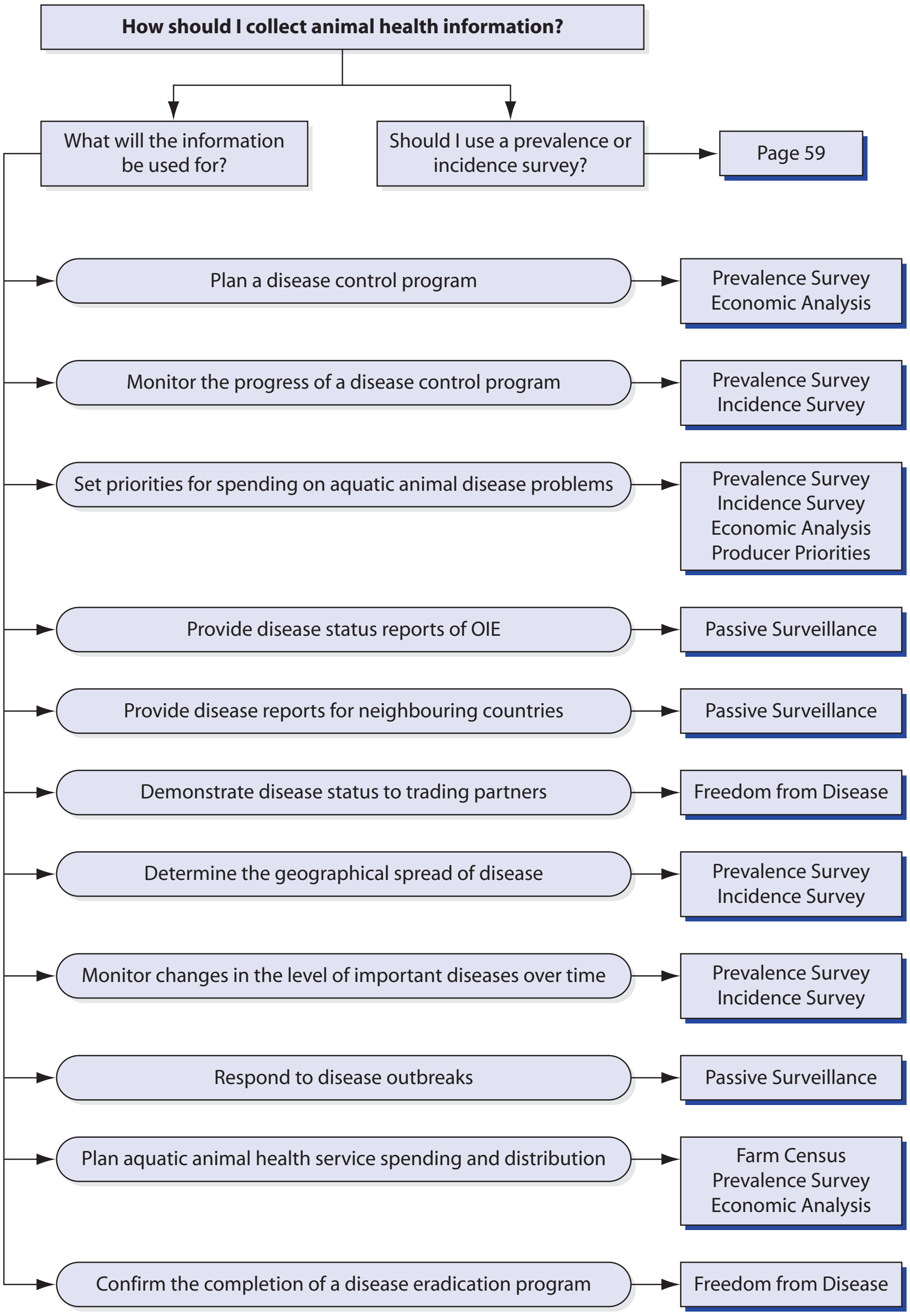
In other cases, the choice of survey type is not so clear. Several different types of surveys may be able to collect the information needed and in fact more than one type of information may be necessary. For some particular problems, passive data may be the most appropriate source of information.

Example

There are moves amongst neighbouring countries to establish a regional EUS eradication program, and you need to decide if the government will participate in this program and fund the control measures needed. To properly answer the question, you need to know, amongst other things, how much EUS is present; where it is; what the state of current control measures is; what the current impact of the disease is; what the cost of the control program would be; if eradication is feasible and, if so, how long it would take; and what the benefits of eradication would be.

Both prevalence and incidence rate surveys would be required to answer these questions, as well as economic studies (cost–benefit analysis) and other special-purpose studies. Passive reporting data may also be useful in making the assessment.

The flowchart below lists some of the possible uses of aquatic animal disease information, and suggests some ways in which this information may be collected. It is designed to act as a guide only, as the final choice of which survey to use will depend on many other factors.


How should I collect animal health information?
What will the information be used for?
Should I use a prevalence or incidence survey?
Page 59
Plan a disease control program
Prevalence Survey
Economic Analysis
Monitor the progress of a disease control program
Prevalence Survey
Incidence Survey
Set priorities for spending on aquatic animal disease problems
Prevalence Survey
Incidence Survey
Economic Analysis
Producer Priorities
Provide disease status reports of OIE
Passive Surveillance
Provide disease reports for neighbouring countries
Passive Surveillance
Demonstrate disease status to trading partners
Freedom from Disease
Determine the geographical spread of disease
Prevalence Survey
Incidence Survey
Monitor changes in the level of important diseases over time
Prevalence Survey
Incidence Survey
Respond to disease outbreaks
Passive Surveillance
Plan aquatic animal health service spending and distribution
Farm Census
Prevalence Survey
Economic Analysis
Confirm the completion of a disease eradication program
Freedom from Disease

5
Sampling principles

Sampling

When we wish to collect information from a population, it is rarely possible to test or examine every member of the population (i.e. conduct a *census*; see page 48). Instead, a smaller group (a *sample*) is selected from the population, and the members of this group are tested (i.e. a survey). *Sampling* is the process of selecting this group from the population. We examine members of the sample and use results to estimate some characteristic of the population from which the sample was drawn.

This chapter is divided into two sections. The first discusses the general principles of sampling, which can be applied in most situations. The second describes a number of specific applications of random sampling for aquatic animal health.

The need for random sampling

There are several different sampling techniques, and they can be divided into two groups: *probability sampling* and *non-probability sampling*. Probability sampling (random sampling) is the only way to ensure that the sample is representative of the population.

Non-probability sampling

Problems with non-probability sampling

In non-probability sampling, the probability of a member of the population being selected in the sample is not known, and some groups are more likely to be selected than others. This means that a sample selected using non-probability sampling is unlikely to be representative of the population, and the results of the survey will be biased. You should avoid using non-probability sampling techniques.

Convenience sampling

Convenience sampling is an example of non-probability sampling techniques. In convenience sampling, samples are selected because they are easy, quick or inexpensive to collect.

Example

A trout farm rearing fish in raceways, containing fast-flowing water, is surveyed to assess the proportion of fish with viral haemorrhagic septicaemia (VHS). A sample of 10 fish is to be examined. Convenience sampling is used to select the animals: a scoop net is used to collect the first 10 fish at the top of the raceway, because they are easy to examine, and you don't have to walk all the way down the raceway.

The problem with convenience sampling is that the sample rarely represents the population. In this case, diseased fish are weaker, and are therefore likely to be at the bottom of the raceway. A convenience sample of the first 10 fish at the top of the raceway is likely to have no affected fish, even if the prevalence of disease is high.

Purposive sampling

Another non-probability sampling technique is *purposive* sampling. In purposive sampling, elements in the sample are selected for some purpose. An attempt may be made to select animals which are judged to be typical of the group. Alternatively, when studying a disease, sick animals may be selected more than healthy ones. Even when 'typical' animals are selected, they are unlikely to represent the range of different animals in the population. Purposive sampling does not produce a representative sample.

Haphazard sampling

Haphazard sampling is a technique where elements are selected for no particular reason at all. It is designed to imitate random sampling. Unfortunately, when people select animals, there is always some unconscious reason for each animal being selected. For instance, the person selecting may think, 'I chose a big one last time, so I will pick a small one this time.' The sample is similar to a purposive one, and despite our best efforts, is rarely representative of the population.

> Non-probability sampling techniques are unable to reliably select a representative sample. The results from surveys using non-probability sampling are likely to be biased.

Probability sampling

The term *probability sampling* covers a group of techniques that includes:

- simple random sampling;
- probability proportional to size sampling; and
- random systematic sampling.

Simple random sampling

In *simple random sampling* (sometimes just called random sampling), every member of the population has the same chance of being selected. In the example of the trout survey, if simple random sampling were used, sick fish and healthy fish (or fish at both the top and bottom of the raceway) would have the same chance of being selected in the sample.

This is the first reason for using random sampling. Samples selected using random sampling are much more likely to be representative of the population than samples selected by non-probability techniques. This means that random sampling can avoid the problem of selection bias, and estimates of population values made from the sample are more likely to be correct.

> On average, random sampling produces representative samples.

The second reason to use random sampling is so we can calculate how reliable our survey results are. When survey results are used to estimate the true value in the population (e.g. the prevalence of sick fish), we use a formula to make the calculation. Similarly, a formula is used to calculate the confidence interval for the estimate, which indicates how confident we are that the results are correct. Each of these formulas is based on the assumption that the sample was selected using random sampling. If the sample was selected with a non-probability sampling technique, the formulas are no longer valid and the results may be incorrect.

Example

A survey was carried out to estimate the proportion of fishers using cast nets in an area. It was calculated that a sample size of 40 fishers (the unit of interest) would be interviewed from a village with a total population of 120 fishers. The survey team asked the chief of the village to take them to 4 different fishing areas, and interviewed 10 fishers at each area. Of the 40 fishers, 12 used cast nets. Can you infer the proportion of fishers in the village who use cast nets?

The sampling technique used was a non-probability sampling technique: it combined purposive sampling (4 different areas) and convenience sampling (letting the village chief choose). We know that 30% of the sample used cast nets (12 / 40 = 30%). We can *estimate* that the proportion of fishers using cast nets in the population is also about 30%. However, because random sampling was not used, there is no way to determine how likely that estimate is to be correct. We think the true value is 30%, but it could just as easily be 5% or 80%.

If random sampling had been used, we would be able to calculate exactly how likely the estimate is to be correct. The 95% confidence interval is 17–47%, which means that we are 95% sure that the true value lies between 17% and 47%.

> When random sampling is used, we can calculate how reliable an estimate is.

Random sampling techniques

Random numbers

Random sampling is based on the concept of randomness, and the use of random numbers. *Random numbers* are best explained by an example. Dice have six sides, numbered 1 to 6, and when one is rolled, each side has the same chance of ending on top. However, at each roll we never know which number will come up. What we do know is that if we roll the dice again and again, on average all numbers will appear equally often. Rolling dice is one way of generating random numbers (in this case, between 1 and 6).

Playing cards are another example of randomness. When we shuffle cards, we never know what order they are in. But it is possible to predict what will happen over a large number of games of cards. This is because each card in the pack has the same probability of being on top. That is how casinos are able to make money. They don't know if a particular person on a particular day is going to win or lose, but they do know that most of the time, more people will lose money than win money, because they can predict the average result of a large number of games.

When selecting a sample for a survey, we want to select members of the population in a way that will ensure that each member has exactly the same chance of being selected. There are many different ways of doing this, and the aim of this chapter is to explain some techniques that are useful to people conducting aquatic animal surveys in developing countries.

Physical randomisation

The examples of dice and cards are called *physical randomisation* techniques because we actually take physical objects and mix, shake, roll or shuffle them. This is one of the simplest approaches to selecting a random sample. The problem with dice is that there are only 6 numbers (although decimal dice exist, with 10 sides numbered 0 to 9). Blank cards are much more flexible, as shown by this example.

Blank cards for random sampling

Example

A large shrimp farm is surveyed to determine if it is infected with Taura syndrome. There are 30 ponds (the unit of interest) in the farm (the population), but only 8 are needed for testing. Each pond has a unique identification number. To select the sample of 8 ponds, we take 30 blank cards or pieces of paper and write the number of each pond on a card. The 30 cards are then shuffled well, and 8 cards selected. The ponds with the numbers selected are the ones to be included in the sample.

Problem with physical randomisation

This is an effective method for random selection. However, when the group is too large the method can quickly become impractical. For instance, imagine conducting a national survey of fishing practices in which 100 different villages are to be examined. If there are 24,200 villages in the country, this would require writing village names on 24,200 cards, shuffling and dealing out 100. Shuffling 24,200 cards could be difficult. For this reason, physical randomisation techniques are only used in small surveys. In larger surveys, random numbers are more convenient.

Random numbers

Random numbers are numbers that have been generated randomly, or by chance alone. This means that the chance of each digit being a particular number between 0 and 9 is the same. There are two sources of random numbers: *random number tables* and *computer-generated random numbers*.[1]

Example

A computer was used to generate a random number, 39024. The number has 5 digits, 3, 9, 0, 2 and 4. When the computer selected the first digit (3) the chance of it being any number between 0 and 9 was exactly the same. The number 3 was just selected by chance. For the second digit, 9, again, the chance of any number between 0 and 9 being selected was the same. The fact that 3 had been selected for the first number made no difference to which would be selected for the second number and so on. It was simply a matter of chance. It is as if the computer is rolling special decimal dice (with 10 sides) and recording each digit as it is rolled.

An example of a random number table is included in Appendix C, and its use is explained below on page 75. Various computer programs, including Epi Info, include random number generators. Both can be used equally well to select a sample using the following steps:

Selecting a random sample

Step 1: Make a list of all the members of the population. This list is called the *sampling frame*, and is discussed on page 81.

Example

A survey is planned of cultured tilapia disease problems in one district. The objective is to identify which disease problem occurs most commonly. There are 75 villages in the district (the population), and a sample of 10 villages is required. The unit of interest is the village. All the villages are listed by name (only the first five are shown).

[1] Computer-generated random numbers are really 'pseudo-random numbers'. They are a sequence generated by a formula, so that if you know the formula, you can predict what the next number will be. With real random numbers, this is not possible. However, for survey purposes, computer-generated pseudo-random numbers are just as good as true random numbers.

	Village name
	Nong Bone
	Sobtui
	Hang Chat
	Khounta
	Si Meuang

Step 2: Number each member on the list from 1 up to N, the total number in the population.

Example

The villages are numbered from 1 to 75.

No.	Village name
1	Nong Bone
2	Sobtui
3	Hang Chat
4	Khounta
5	Si Meuang

Step 3: Using a computer or a random number table, select random numbers between 1 and N. Select one random number for each element to be selected in the sample.

Example

10 random numbers are selected, between 1 and 75. The numbers are: 2, 5, 27, 42, 47, 52, 53, 57, 66, and 68.

Step 4: For each random number selected, find the corresponding element on the list. These are the ones to be included in the sample.

Example

Find the selected villages corresponding to the random numbers.

No.	Village name	
1	Nong Bone	
2*	Sobtui	Selected
3	Hang Chat	
4	Khounta	
5*	Si Meuang	Selected

Selecting random numbers using a random number table

Random number tables are a convenient source of random numbers. An example of a random number table is shown below and a full table is given in Appendix C. There are sets of numbers grouped into fives. In the above example, 10 numbers were selected to pick 10 villages from a total of 75. To use a random number table to select random numbers, proceed as follows:

Step 1: Choose a starting point and direction. You can start at the top of the table, or you can start anywhere in the middle. You can go across a row, or down a column. In this example, we will start at the top left number, and move across.

Step 2: Calculate the range for your random numbers. The numbers required in this example are between 1 and 75.

Step 3: Determine which digits to use from the numbers. The maximum number we want is 75, which has two digits. We therefore only need two of the five digits in each random number. To use the numbers efficiently, we can 'cut' them in half, and think of the first two digits (42) as the first number, and the third and fourth digits (53) as the second number. The last digit can be ignored.

Step 4: Search through the table for numbers in the required range. Any number between 1 and 75 is counted as one of our random numbers. Any number over 75 is ignored. Continue searching until enough numbers have been found (ten in this example).

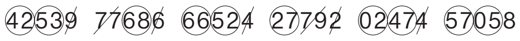
Example:

Using the table below, the first number is 42. This is between 1 and 75, so it is accepted. The second number is 53, and is also accepted. The next digit (9) is ignored. Moving to the right to the next group, the next number is 77. This is greater than 75 and is ignored. The next number is 68, which is accepted as our third random number. The last digit (6) is ignored. Continuing in this way we get a fourth (66), a fifth (52), a sixth (27), discard the next (79), a seventh (02, or 2), and eighth (47), a ninth (57) and a tenth (05, or 5).

Random number table

42539	77686	66524	27792	02474	57058	61530	76108	49436
27030	88085	84744	32591	57804	54790	24545	73422	23337
50253	66592	66151	18506	04391	35824	35397	32031	67780
54127	25147	79021	54189	43708	08178	82187	72106	53795

When using a random number table, it is a good idea to circle the numbers you select, and to cross off those that you discard. This helps you remember which numbers you select, and prevents you from using the same numbers again. You should always use new random numbers when sampling in a survey.



Selecting random numbers using a computer

While using a random number table is quick and simple, the job can be done even more conveniently using a computer. Various programs are available to select random numbers, and one is included in **Epi Info.** To generate random numbers using Epi Info, do the following:

Step 1: Start Epi Info, and select EpiTable from the Programs menu.

Step 2: Open the Sample menu, and select Random Number List.

Step 3: Enter the number of random numbers you want (in our example, 10).

Step 4: Enter the minimum value for the numbers (usually 1).

Step 5: Enter the maximum value for the numbers. This is equal to the total number of elements in the population (in our example, 75).

Step 6: Set the program to select either with or without replacement (see 'Replacement' below).

Step 7: Select the Calculate button. Epi Info will generate random numbers in the required range, and display them on the screen. They can then be printed or saved to a file.

To make sampling even easier, there are specialised computer programs that help you build the sampling frame, select the random numbers, and report the selected elements for you. Two such programs are included with the **Survey Toolbox**. The first, **Random Village**, is for selecting elements, such as villages or ponds, from a list. The second, **Random Animal,** is for a more difficult situation—selecting sampling units from a grouped population. Both are described later in this chapter.

Replacement

Sampling can be done in two ways: with or without replacement. Sampling with replacement is best illustrated with an example.

Example

A standard deck of 52 cards is shuffled, and one card is drawn from the deck. We record the value of the card, and then replace the card into the deck. The deck, still containing 52 cards, is shuffled again, and another card drawn. We record the value of this card, too, and return it to the deck. This is repeated until enough cards have been drawn. With each selection, the probability that any particular card will be drawn remains the same: 1/52.

Sampling without replacement

In contrast to this example, sampling without replacement means that as each card is drawn, it is kept out of the deck, and only the remaining cards are shuffled. One difference between the two techniques is that it is possible to select the same element twice when using sampling with replacement, but not when sampling without replacement. Another difference is that, when sampling without replacement, the probability of selecting a particular card changes with each selection. At the first draw, the probability of selecting a particular card is 1/52. However, if that card is not chosen the first time, the probability at the second draw is slightly higher (1/51), because there are fewer cards left. In some survey designs, it is better to sample with replacement, and in others, it is better to sample without. When the population is large, the difference between the two techniques is unimportant.

If you are sampling without replacement, using a physical randomisation technique (e.g. dice) or a random number table, you must be careful to check if a number has already been selected. If so, discard it and draw a new number.

When using **Survey Toolbox** to generate random numbers or do the sampling, you can simply tell the computer which method to use, depending on the survey design. The survey designs described in Chapters 11–14 tell you whether to use replacement or not.

Systematic sampling

Systematic sampling is an alternative technique when the elements of a population can be ordered in some way, but are difficult to identify and list individually.

Example

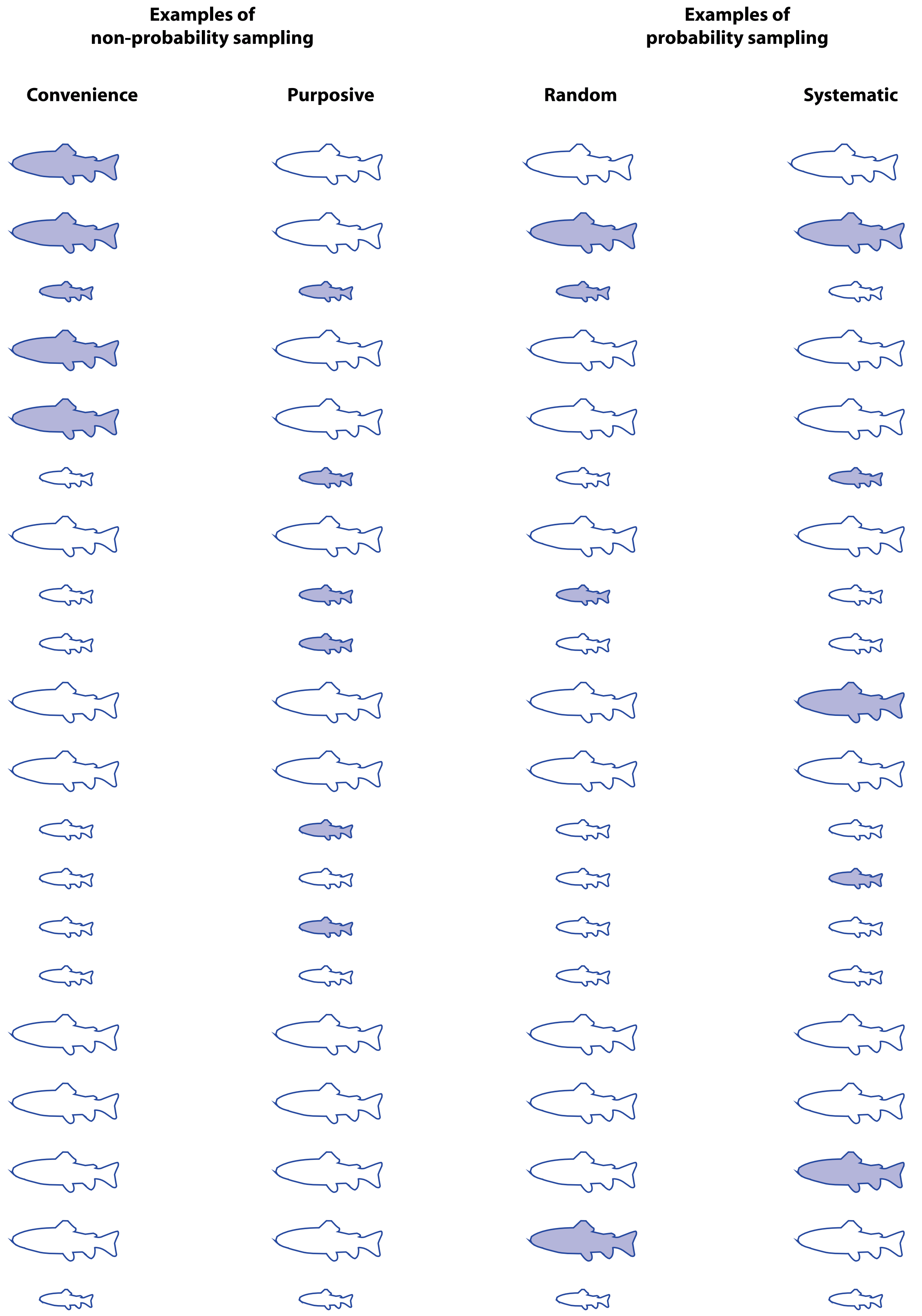
A survey is being conducted in a cage of 600 salmon to examine the level of parasites. If we wish to examine skin scrapings from 30 salmon, simple random sampling would require that each fish is first identified and assigned a number between 1 and 600. If the fish are not already identified (e.g. with tags), this process may be time consuming or impractical. Salmon are often graded at some stage during the growth cycle, so that larger and smaller fish can be kept separately. Using systematic sampling at the time of grading, prior identification of the fish is not necessary. The fish are caught one by one and measured. Every twentieth fish is then selected as it passes through, giving a total sample of 30 fish.

Sampling interval

In this example 20 is known as the *sampling interval*. It is calculated as N/n, or the population size divided by the desired sample size. Systematic sampling is only a form of probability sampling if the first fish selected is selected at random. In our example, a number between 1 and 20 would be selected at random, say 15. The fifteenth fish measured would be the first fish selected, followed by every twentieth fish. If, however, the first fish is selected each time, there is no element of chance, and the technique becomes a form of non-probability sampling.[2]

The figure on the next page illustrates different types of probability and non-probability sampling techniques to select 5 fish from a tank. In the first, convenience sampling is used to pick the first fish. In the second, purposive sampling chooses the smaller fish, as they are easier to handle. The third uses random sampling, and the last uses systematic sampling with a sampling interval of 4 and a randomly selected first fish.

[2] Under most circumstances, random systematic sampling will produce a sample with very similar properties to simple random sampling. The same formulas used for simple random sampling may be used to analyse the results from random systematic sampling. There is a danger, however, if the population has some sort of cyclic variation that matches or nearly matches the sampling interval. For example, let us consider a study to estimate the average water temperature in hatchery tanks. Records are kept using an automatic recording device, which records the temperature every hour. Records for 30 days are available (720 measurements). To save analysing all the data, a systematic sample of 30 measurements is taken (giving a sampling fraction of 720/30 =24). A random number between 1 and 24 is selected, say 3, and the first record selected is the third record. After that, every twenty-fourth record is selected. This sampling scheme would result in the selection of records that were taken at exactly the same time every day for the 30 days. If record keeping had begun at midnight on the first day, each record would represent the 2.00 a.m. measurement for one day. The temperature at 2.00 a.m. clearly does not represent the range of temperatures throughout the day. Fortunately, such clear cyclic ordering in aquatic animal surveys is very unusual, and so is rarely a problem.



First five fish selected

Small fish selected for ease

Random numbers used to select fish

Second fish selected with random number, then every fourth fish selected

Stratification

Stratified sampling

Stratified sampling is not a sampling technique, but an approach that may be used with any of the sampling techniques discussed. It involves dividing the population up into separate, exclusive groups (or strata), and selecting a sample from each group (stratum). If random sampling is used within each stratum, then it is known as stratified random sampling. Stratification can be based on any characteristic of the population. A survey of village aquaculture may be stratified by production type, with pond culture in one stratum, and net culture or cage culture in a second. A survey of shrimp diseases might stratify by stage of production, breaking the population into hatcheries, nurseries and grow-out farms.

Reasons for stratification

There are three main reasons for using stratification. First, it enables us to calculate estimates not only for the whole population, but for each stratum as well. For a national prevalence survey, stratification by province may provide a much more useful picture of the distribution of the disease. In this case, stratum estimates will be less precise than the combined estimate, because of the difference in sample sizes. The second reason for using stratification is that it is often operationally much more convenient. If we are stratifying by area, the survey can be done in stages, one area at a time, which distributes the workload more evenly. The third reason is that stratification may produce more precise results by decreasing the amount of variation between elements within each stratum.

Example

A survey is conducted to estimate the prevalence of blood parasites (trypanosomes) in freshwater fish, in a country with three climatic zones: Zone A (tropical), Zone B (intermediate), and Zone C (cool). The vectors for the disease (leeches) are likely to be much more common in the tropical part of the country (Zone A) than in the other two zones. A national survey will find a lot of variability in the prevalence from one area to another. Stratification by climatic zone means essentially that three separate surveys are conducted, one in each zone. In Zone A, a uniformly high prevalence is likely to be found, while a low prevalence is likely in the other two zones. In each survey, the variability is much less. When the results of the three surveys are combined, the overall variability of the results is less, so the precision is greater.

To achieve this precision, the aim is to have the animals within each stratum as similar as possible with respect to the characteristic in question, and for animals in the different strata to be as different from each other as possible (low within-stratum variability, and high between-stratum variability). Before a survey we rarely know enough about the distribution of the characteristic in the population to ensure this, but we may know that the characteristic is linked to some other factor (in our example, climatic zone). Stratifying on this factor will therefore help to increase the precision of the survey.

Probability proportional to size sampling

In simple random sampling, every unit of interest in the population has the same chance or probability of being selected in the sample. Another probability sampling technique that is often useful is *probability proportional to size* (PPS) sampling. Instead of all units of interest having an equal chance of selection, in PPS sampling the chance of selection is proportional to some measure of the size of the unit.

Example

A survey of villages in a large province uses PPS sampling to select villages, based on the number of ponds in each village. The population is all villages in the province and the unit of interest is the village. Records of the total number of ponds in each village are collected and used as a sampling frame. Villages with a large number of ponds have a higher chance of being selected than villages with a smaller number of ponds.

PPS sampling requires reliable information on the size of each unit of interest in the population. When this information is available, PPS can be used for very efficient survey designs. Unfortunately, it is often difficult to find this sort of information.

It is possible to do PPS sampling in a similar way to that described above for simple random sampling; however, it is much more practical to use a computer to do the task. A program for this task, **Random Village**, is described on page 97. Using the computer program requires that the sampling frame is available on computer disk. If this is not the case, then it may be necessary to do PPS sampling by hand. Use the following procedure:

Manual PPS sampling

Step 1: The sampling frame must be a list identifying all the units of interest in the population, with data about the size of each of them (for instance, the number of ponds per farm, or the number of farms per village). Add another column to this list for the *cumulative total*.

Cumulative total

Step 2: In the new column, write down the cumulative total next to each item. The cumulative total is the size of the current item, plus the cumulative total from the previous line, as shown below.

Village name (unit of interest)	Number of ponds (size of unit of interest)	Cumulative number of ponds
Ban Dong	232	232
Ban Hai	89	3231 (= 89 + 232)
Sisakhet	144	465 (= 141 + 321)
Si Meuang	129	594 (= 129 + 465)

Step 3: The last line in the cumulative total column is the total number of ponds in the entire study area. In PPS sampling, instead of picking a random number representing a village, we pick a random number representing a pond, and then select the village that contains that pond. Using any of the techniques described previously, pick a random number between 1 and the total number of ponds in the population.

Step 4: Search down the cumulative total column until you find the last number which is equal to or greater than the randomly selected number. The unit of interest containing that number is the element selected.

Example

If the above list represents the entire population of villages in the study, the total number of ponds in the four villages is 594. Select a random number between 1 and 594, say 256. Searching down the list, the second village contains the number 256, and this village would be included in the sample.

Step 5: Continue until the required number of units of interest has been selected.

Sampling frames

Sampling frame

In random sampling, every unit of interest in the population has the same chance of being selected. In the techniques described above, this is achieved by using random numbers, and picking units of interest from a list. This list is called the *sampling frame*, and should contain every unit of interest in the population.

Example

A survey is conducted in a large ornamental fish farm to estimate the prevalence of tanks containing fish with fin damage. The farm has recently expanded, and 40 new tanks have been added to the existing 200. The farmer has a list of all the old tanks, each identified with a number, so this list is used as the sampling frame. Twenty tanks are selected using random numbers from the sampling frame, and the fish in these tanks are examined for signs of fin damage.

Clearly this survey has a problem with selection bias. It is not possible to infer the true prevalence of tanks with fin damage from the survey results, because the selection bias means that the sample is not representative of the population. The sampling frame does not include any of the new tanks, only the old tanks. The surface of the old tanks may have become more damaged and rough resulting in more fin damage to those fish kept in the old tanks. We are therefore likely to get misleading results, even though we used random sampling. This is because the sampling frame was incomplete, and did not include every tank.

A different problem can occur when a sampling frame lists the same units of interest more than once. In a village sampling frame, if one village is listed twice it has twice the chance of being picked of other villages. Another difficulty can arise in identifying the elements from the list. Sometimes there may by two ponds with the same number, or two villages with the same name. The ideal sampling frame is therefore a list that:

- contains *every* unit of interest in the population (no omissions);
- contains every unit of interest *only once* (no duplications); and
- *uniquely* identifies each unit of interest.

Sampling frames may also contain other information to help with more complex sampling schemes. One example is a village sampling frame that lists all villages in an area, but also includes information on the number of ponds or fishers in each of those villages (the size of the village). This extra information can be used for PPS sampling, described above.

Sources of possible sampling frames

Sometimes a suitable sampling frame already exists. When surveying villages, the government statistics office and many other government departments usually maintain lists of villages, often computerised with unique identification numbers.

These lists are very suitable for use as sampling frames. The statistics office or fisheries department may also maintain information on aquaculture establishments or fisheries, but to be useful this needs to be up to date, and give the population of each village instead of summary figures for districts or provinces.

Any sampling frame is likely to be imperfect, either missing a few members of the population, or not identifying others properly. A sampling frame doesn't have to be perfect to be useful. It is a matter of judgement to assess how good the sampling frame is, and whether the problems with it are likely to affect the results of the survey. For instance, if a sampling frame is missing 20% of the population, it may be better to try to find a better sampling frame. However, if members are missing in no particular pattern, results from a survey using the frame could be perfectly adequate. On the other hand, even if only a few are missing from the frame (say 5% or 10%), any clear pattern (e.g. the smallest or most recently established farms missing) means there is a significant danger of the biased results. Unfortunately, without thorough investigation it is often difficult to work out how complete a sampling frame actually is.

If there is no sampling frame

In many cases, no existing sampling frame is available. To carry out a survey using random sampling, it is then necessary to either:

- build a new sampling frame, by identifying all the units of interest in the population and creating a list;
- use a different sampling strategy, requiring a different type of sampling frame that is easier to obtain (such as two-stage sampling, described below; or
- use a specialised technique for random sampling with no sampling frame (see Spatial sampling, below)

Two-stage sampling

When a sample is being selected using simple random sampling, all animals need first to be identified and listed on a sampling frame. In systematic sampling, they need to be 'lined up' in some sort of sequence. When surveying very large populations (e.g. national-level surveys) both simple random sampling and systematic sampling are impractical, if not impossible. For example, it is not possible to create a list containing every single oyster in a country with a total population of 400 million oysters.

Two-stage sampling addresses this problem by gathering the units of interest into convenient groups. While a list of all oysters would be impossible to compile, a list of all oyster *farms* in the country might be more easily obtained (perhaps from registration records). In two-stage sampling, groups of animals (e.g. a number of oyster farms) are selected first, and then individual animals are selected from the selected groups. At the first stage, the population is all the oyster farms in the country and the unit of interest is the farm. At the second stage, the population is all the oysters on each selected farm, and the unit of interest is the oyster. At each stage of two-stage sampling, the sample is selected by random sampling. Stratification may also be used, usually at the first stage (for instance, stratifying farms by production type, size or district).

Advantages of two-stage sampling

Surveys using two-stage sampling have two distinct advantages. First, they are easier to plan because they don't require a complete list of all animals in the study area, only a list of the first-stage units. Second, they are more practical for the fieldwork team, as fewer sites need to be visited. The disadvantages are that the results may not be as precise as with simple random sampling, and the formulas for analysing the data can be very complex.

When designing a survey, there are often a number of levels that can be used for the first stage of sampling, such as province, district, village or farm. The best level to choose is the lowest level for which a readily available sampling frame already exists. For example, in a two-stage sample survey where the pond is the unit of interest, there are usually relatively complete lists of provinces, districts and villages available. The best choice for the first stage sampling unit would be the village, as this is the lowest level with a sampling frame. The second stage sampling would require a list to be compiled of all the ponds in the selected villages. If a higher first stage sampling unit had been chosen, such as province, the task of compiling the second stage sampling frame (all the ponds in the province) would be unmanageable.

Example

A national survey is planned to assess the impact of epizootic ulcerative syndrome on smallholder fish farmers. The population of interest is all the village fish farmers in the country, a total of 5.5 million. It is not possible to create an accurate sampling frame listing all these people. However, a list of all the 18,322 villages in the country does exist, maintained by the national statistics office and available on computer disk. We therefore use a two-stage survey. At the first stage, we use a computer to select 40 villages (the first-stage units of interest) from the village sampling frame. At the second stage, we select 10 farmers (the second-stage units of interest) at random from each of these 40 villages, to give a total sample size of 400 farmers. As no list of farmers is available for the villages, we use village interviews to create a sampling frame of all the farmers in each village.

Aquatic animal surveys often use two-stage sampling. Chapter 11 provides details of the design, implementation and analysis of data for a two-stage prevalence survey.

Spatial sampling

Spatial sampling describes a group of techniques that may be used to generate a random sample when there is no sampling frame. In simple random sampling, it is necessary to have sampling frame. In systematic sampling, no sampling frame is required, but the population must be able to be ‘lined up’ in some way. Spatial sampling has neither of these requirements. However, there is one important requirement: the population has to be relatively stationary. For example, spatial sampling may be used to select farms, ponds, or immobile animals such as oysters at the adult stage.

Spatial sampling is similar to random sampling, but instead of selecting individuals from a sampling frame, we select random *locations* from an area. The individuals to be included in the sample are those that are at the randomly selected locations.

Example

To select a sample of ponds in an intensive shrimp farming area, we obtain a map showing the location of all ponds, and draw grid lines across this map to divide it into 400 small squares. We number the squares from 1 to 400, and select a sample of 20 squares using random numbers. Each selected square is then examined on the map, to identify a pond that lies inside that square.

In this example, the procedure is very similar to simple random sampling. We have a sampling frame of 400 squares, and we randomly select 20 of those squares. However, the difference is that we are selecting squares, instead of ponds.

Example

A study is planned to assess wild fish stocks in a river. Fish will be trapped at a number of locations along the length of the river. To select the trapping locations randomly, the entire length of the river area under study is first measured from a map. Then, random numbers between 0 and the length of the river are chosen. These random numbers represent the distances from the bottom of the river to the randomly selected trapping locations. The distances are measured along the length of the river using the map, and the trapping locations marked.

This example is also very similar to simple random sampling, as it is a bit like picking random numbers from a list. The numbers in this case however are not necessarily whole numbers (eg the 6th pond), but should be fractional numbers (eg, a point 7.53 km upstream from the start of the survey area). Random systematic sampling could also be used, to pick, say, 20 points evenly distributed along the river. This is an example of spatial sampling in *one dimension*, along a line.

Example

A study is being conducted of the spread of amoebic gill disease in an estuary. The researchers want to know what the average concentration of the amoeba in the water is, and how much variation there is. To do this, they decide to take water samples from different parts of the estuary and test for the amoeba. Using a random number table, they generate pairs of random numbers, and interpret them as coordinates. These coordinate pairs indicate points randomly located throughout the estuary. The researchers take water samples from each of these points.

This third example of spatial sampling is different from the first two. Instead of squares or distances along a river, random points are selected in two dimensions. In the first example, the unit of interest was the pond and in the second it was the trapping site. In the third, the unit of interest is the water sample. When collecting water samples from an estuary, there is water everywhere, so any randomly selected point will have water available to sample. However, when selecting ponds using randomly selected squares, it is possible, and even likely, that some squares will have no ponds, and some squares may have two or more ponds. This is the main problem with spatial sampling. When sampling discrete objects (e.g. ponds, farms, fish, oysters) there is no one-to-one relationship between the random location and the unit of interest. This means that there isn’t one pond in every square, or one fish located at every random point.

In spatial sampling, there is often no one-to-one relationship between the unit of interest and the spatial sampling unit.

With continuous substances, such as soil, air or water, this problem doesn't exist, as there is generally an appropriate specimen at any randomly selected location. Villages and districts may be thought of as discrete objects, but if a map has the boundaries of villages or districts marked, and there are no gaps between them, we may think of them as continuous. If we chose a random point anywhere, it will be located in a district, or in the land area belonging to a village.

Example

A survey of districts is being conducted, and the districts are chosen using random coordinates. Pairs of random numbers are chosen, plotted on a map, and the districts containing those points are selected for the sample. The researchers use sampling without replacement, so each district can only be selected once.

While this is clearly a form of random sampling, there is one important problem. Randomly selected points will fall all over the country, and each point has the same probability of being selected. However, if one district is much bigger than another district, there are many more points located in that district, and it is therefore more likely to be selected than the smaller district. In simple random sampling, each district should have the same chance of being selected. In a spatial sampling scheme, larger districts have a higher chance of being selected—a form of PPS sampling. If for some reason you want larger districts to be represented more often in your sample, this doesn't cause a problem. But if you want to have every district with the same probability of selection, you can't use this type of spatial sampling.

When sampling units are different sizes, spatial sampling may result in larger units having a higher probability of being selected.

- In summary, it is difficult to use spatial sampling strategies to select a random sample when:
- the sampling units are not continuous (i.e. they have spaces between them);
- the sampling units are moving; or
- the sampling units are of different sizes.

Fortunately, there are different approaches that overcome these types of problems.

Example

We want to collect a random sample of fish from a floating net cage. The fish are discrete (they have space between them), and they are moving, but they are all approximately the same size. In order to use spatial sampling, we lift the net on one side of the cage to crowd the fish together on the other side. Once the fish are crowded closely together, there is no space between them and they can't move. This means that fish selected from random locations will be very similar to a simple random sample of fish.

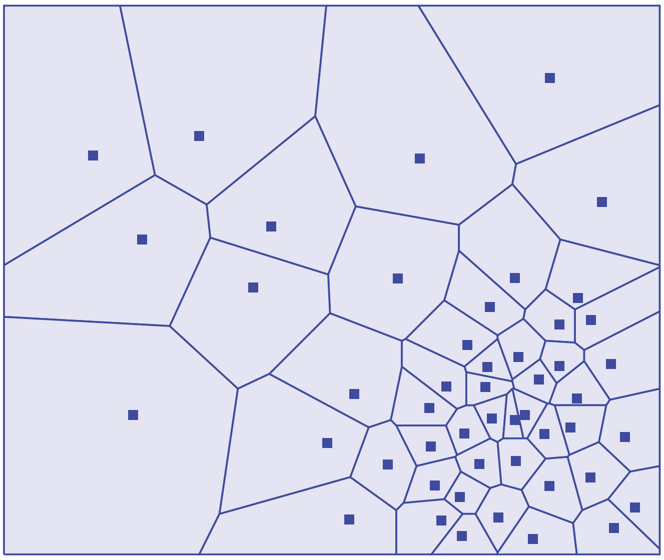
In this example, we have removed two of the constraints for spatial sampling: mobility, and gaps between units. However, we have still not managed to solve all the problems. When fish are crowded together, they will tend to be piled up on top of each other. Some will be on the surface, but others will be deeper down in the water. Up to this point, our discussion of spatial sampling has been limited to two dimensional sampling (i.e. from a flat surface). In this case, we need to consider the third dimension: how deep we go to select the fish. If we always select fish from the surface, there is a risk that the sample will be biased, because some types of fish (perhaps the smaller ones) are more likely to be at the surface. The solution is to choose not two random numbers (which describe a coordinate or a point on a flat surface) but three (which describe a point in three-dimensional space). The first two random numbers are interpreted as 1) the distance across the net from left to right, and 2) the distance across the net from front to back. The third is interpreted as the distance from the surface, or depth.

Spatial sampling may be either two-dimensional or three-dimensional.

This example has other, more important problems that are harder to solve. The first is that crowding fish together in a net is difficult, time consuming and likely to damage the fish. It may be very hard to get farmer cooperation if this type of approach is used frequently. The second is that crowded fish are less mobile than free-swimming fish, but they still move around a little. Looking up lists of random numbers and trying to locate a fish at a depth of 40 cm at a particular location in the net is likely to be extremely difficult in practice. Selecting a sample of, say, 40 fish may take quite a while, and result in a lot of damage to the fish.

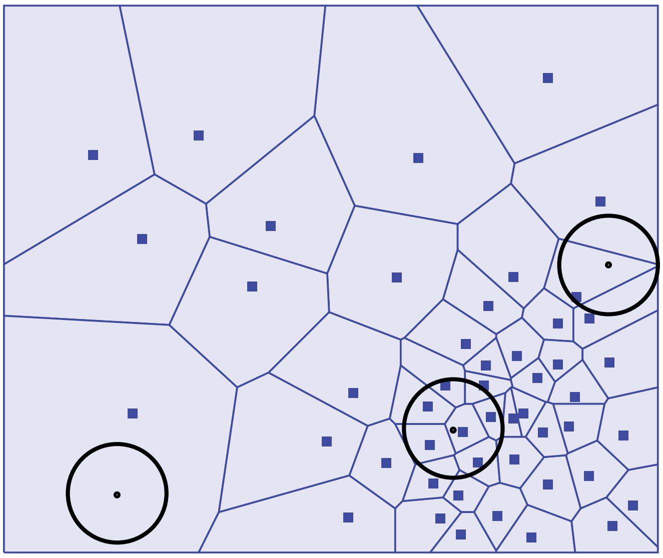
Instead of crowding, there is another possible approach to solving the problem of sampling units with gaps between them. Consider again the example of selecting ponds at random. Instead of dividing an area in to squares, let us choose random points. If a point falls inside a pond, then that pond is selected. However, if a point falls outside a pond (which most points normally will), what should we do? One option is to find the nearest pond, and select it. This is known as *nearest neighbour sampling*. While this sounds sensible, it may again lead to serious bias. Imagine two shrimp-farming areas, one very intensive with many ponds, and another with just a single pond with no other farms nearby. In the first, intensive area a random point needs to be very close to a particular pond for us to select that pond. If it is far away, there will be another pond that is closer. This means that only a small number of random points will identify any particular pond as the closest. In the second situation (with one pond), random points a very long way away will still select that one pond because it is still the closest. When we use this type of sampling in areas where there are some units close to each other, and some that are sparsely spread, the result is that the sparse units are much more likely to be selected than the densely spread units, because there are more random points close to the sparsely spread units.

This is illustrated by the diagram below {overleaf, opposite?}. Each point is a pond, and the shape around each pond marks all the points closer to that pond than to any other. The larger the area of the shape, the more likely the pond will be selected. In fact, the probability of selection is proportional to the size of the shape.



To overcome bias towards sparsely located points in nearest neighbour sampling, select only units that are within a certain maximum distance of a selected point. For instance, we could select ponds that are within 50 metres of the selected point. If there are no ponds within that distance, then a new point is chosen. This is known as *fixed distance weighted sampling*. This approach solves the problem of bias towards very sparsely located ponds, but introduces yet another problem. What if there are two or more ponds within 50 metres of the selected point? The short answer is to chose one of them randomly. When there is more than one pond, we can think of the one we choose as 'representing' all the ponds within 50 metres of the point. Because the one pond chosen is representing other ponds, we need to give more emphasis to the results from this pond. This is done by *weighting* the results from that pond during analysis.

In the figure below, three random points have been selected. Ponds within a fixed distance around each point are eligible for inclusion in the sample. For the lower left point, there are no ponds within the area. For the upper right point, there is one pond within the set distance (and one just outside), so the one pond within the distance is selected.



The lower right point has six ponds within the set distance. We select one of these six at random (e.g. by using dice), and include it in the sample. During analysis, we use weighting to give this pond six times as much importance as the ponds from the other areas. Weighting is discussed in the section on analysis.

Identifying locations

Spatial sampling involves the use of randomly selected squares or points to identify units for sampling. Once the locations have been selected, it is necessary to go to these locations to identify the sample. There are two main approaches to identifying the random locations that have been selected, depending on the scale.

Small area sampling

When a survey is being conducted of a small population in a relatively small area, identifying locations is quite easy. For example, consider a single pond. The pond is 70 metres long, 50 metres wide and 1 metre deep. We decide to use a cast net to sample fish from the pond at randomly selected locations (the use of cast nets for sampling is discussed in the next chapter).

We generate a set of 5 random points. This involves picking pairs of random numbers from a random number table. In each pair, the first number will be between 0 and 70 (the length), and the second number will be between 0 and 50 (the width).

When collecting the sample at the pond, the locations need to be marked. One approach is to use a tape measure on either side of the pond, and make a mark at the selected coordinates. A more practical approach might simply be to pace out the length and width of the pond, making a mark in the dirt every 10 metres. The person using the cast net enters the pond, and estimates the first location by lining up the marks on the side of the pond. For instance, if the first location is 28 (length) and 12 (width), the person in the pond would find the third mark on the length (30 metres) and move two steps back from there, then find the first mark on the width (10 metres) and move two steps across.

Large area sampling

Marking out a grid is easy when dealing with a pond, a net, a cage or even a whole farm, but is not possible if the area to be sampled is larger, such as a whole village, a district, a reservoir, an estuary or an ocean. In these situations some other method must be used. Instead of using relative coordinates (marking distance along or across a pond, using one corner as the zero location), we need to use absolute coordinates to identify our location on the surface of the earth. There are two main ways to do this.

The first method is simply to use a map. A good map of the survey area will have a grid marked along the edge of the map. This may be marked in a variety of different units—for example, geographic coordinates (degrees, minutes and seconds), or projected coordinates (metres, kilometres, feet or miles). By reading the range of coordinates on the map, it is possible to calculate the range for the random coordinates. For instance, the survey area may fall between 20 and 22 degrees north, and 104.5 and 106.5 degrees east. Random numbers (with a suitable number of decimal places, say 5) can be chosen in these ranges. These ranges actually describe a rectangular area on the surface of the earth. Often a survey area will not be rectangular, but defined by some sort of irregular boundary, either natural (the shores of an estuary) or artificial (the borders of a province or country). In this case, as each random coordinate is chosen, it must be checked to see if it falls within the survey area. If not, it is discarded and a new point is chosen.

The chosen points are plotted on the map. The task is then to use the map to navigate to the exact location marked by the point. This may be rather difficult, and depends on good map reading and navigation skills, as well as the ability to travel in the terrain in the study area.

The second approach to identifying random locations is to use a global positioning system (GPS) unit. This hand-held device uses a network of satellites to pinpoint a location on the earth, and makes navigation to a randomly selected location much easier by removing the need for good map reading or navigation skills. After you enter coordinates, the unit simply indicates the correct direction and tells you how far away the point is.

Creating random coordinates

Random coordinates can be selected using a random number table (as described on page 75) but are more conveniently selected using a computer. Two programs, one of which works with the ArcView GIS program, are included in the Survey Toolbox for the selection of random points.

RCGS Windows 95

The **RGCS** program included in the **Survey Toolbox** creates a number of random coordinates within a rectangular area. To start the program use the Windows Start menu, select Programs, then Survey Toolbox, then RGCS.

To select random points:

Step 1: Enter the boundary coordinates of the study area. You will need a map with a coordinate grid to find these figures. Max Y is the upper y coordinate, or the northernmost latitude of the study area. Min Y is the lower y coordinate, or the southernmost latitude. Min X is the left x coordinate, or the westernmost longitude. Max X is the right x coordinate, or the easternmost longitude. If the figures on the coordinate grid on your map are in degrees and minutes, convert them to decimal degrees before entering, by dividing the minutes by 60 and adding to the degrees.

Example

A survey is planned for a single state. A map of the state is obtained, with a coordinate grid marked in degrees and minutes. The latitude of the northernmost point in the state is read from the map as 14º34'. To convert this to decimal degrees, 14º34' = 14 + (34/60) = 14.5667. Enter this number into the Max Y box. Repeat for the other boundaries.

Step 2: Specify how many random points to select. When selecting from a sampling frame, each random number corresponds to a single member of the population, so the number of random numbers needed is equal to the sample size. With spatial sampling, some random points may have no units within the selection radius, and have to be discarded. Some of the random points will also fall outside the study area. For these reasons, you should select more random points than your sample size. The number of points should be equal to 2 or 3 times the sample size. Enter the number required in the Points to Generate box, or use the arrows to change the number.

Step 3: Specify how you want the results to be displayed in the Coordinate Type box. Cartesian coordinates are normal x,y coordinates, representing metres, kilometres, feet, yards or miles on a map grid. Degrees and decimal minutes are for latitude and longitude grids. Click on the button that corresponds to the units used on your map.

Step 4: Click on the Generate Points button to display a numbered list of points.

Step 5: Discard those points that fall outside the study area. Check each point, in order, against the map. Identify the location of each point, and check if it is inside the study area or not. If not, discard it using the Delete button.

Step 6: Save the list, or print it using the Print button.

RCGS for ArcView GIS v.3

Also included in the Survey Toolbox is a random point selection program, **RGCS ArcView**, designed to work with ArcView GIS version 3.[3] Those who have a copy of this program are able to load the extension file, and have ArcView select points within a specified area automatically. In order to use this program, you must have a copy of ArcView, and you need a digital map (theme or coverage) of the study area in a format that ArcView can read.

To install the extension, copy all the files from the CD directory \Survey Toolbox\AVRGCS to the extensions subdirectory for your copy of ArcView (usually c:\ESRI\Av_gis30\ArcView\Ext32).

Follow these steps to load the extension and select random points:

Step 1: Start ArcView. With the Project window active, select the File menu and choose Extensions.

Step 2: In the Extensions dialogue box, search down the list until you find 'Random Geographic Coordinate Sampling'. Click on the checkbox on the left to select it, then click OK to load the extension. An introductory screen and brief instructions will appear.

Step 3: Create a new view. Add a new theme, and load the digital map of the study area into the theme. The map may be of an area larger than the study area, but must have one or more polygons that describe the study area. For example, if you are conducting a survey of one province, a national map without provincial boundaries is not adequate, but you can use a national map showing the boundaries of all provinces. You could also use a map showing all districts within the province.

[3] ArcView GIS version 3, © 1992–1997, Environmental Systems Research Institute, Inc., 380 New York Street, Redlands, CA 92373 USA.

Step 4: Click on the Select tool, and select the study area. This may be represented by a single polygon (e.g. one province on a national map), or several separate polygons (e.g. all the districts making up one province).

Step 5: To start selecting points, open the Sampling menu and choose Select Random Points. Alternatively, you can click on the Run button, on the right of the button bar.

Step 6: The program asks how many points to select. Enter a number 2 or 3 times the sample size, to account for points with nearby units of interest.

Step 7: The program asks you to specify the selection radius. Enter the desired selection radius in *map units*. For example, if distances on the digital map are measured in metres, and you want a 2-kilometre selection radius, enter 2000. If the digital map is in geographic coordinates (degrees and minutes) you will need to convert the distance to degrees. One kilometre is approximately equal to 0.009 degrees (north–south).

Step 8: The program then asks for a filename to store the random points, first as a point theme, and then as a file in dBASE format. These files are stored in the default ArcView directory.

Step 9: Finally, the program selects the points within the selected study area, and displays them on the map. Each point is surrounded by a circle defined by the selection radius. The database file of coordinates is also displayed and may be printed.

Biased sampling

The whole purpose of random sampling is to avoid bias. Biased sampling, on the other hand, is an approach to sampling that tries to introduce a bias.

Example

A batch of shrimp post-larvae (PLs) is to be tested to check whether the shrimp are infected with white spot syndrome virus (WSSV). The batch contains 5000 PLs and a sample of 300 is required. Normal random sampling would involve the random selection of 300 PLs from the entire population (5000). Instead, the aim is to be sure that WSSV is not present. The surveyor decides not to select from all the PLs, but only from those that are most likely to have the virus. To do this, they first treat the batch with formalin for a short period. This stresses the PLs, and the weaker ones either die or become very lethargic. The normal PLs are returned to water, but 600 sick or dead ones are separated. A random sample is then taken from the sick and dead PLs and tested for WSSV.

The basis for this type of sampling is to improve our confidence when trying to demonstrate freedom from disease. Surveys of this kind are discussed in detail in Chapter 14. The theory behind this approach is that if there are infected PLs in the batch, they will be weaker. Selecting only from the weaker PLs means that we have a much higher chance of detecting the disease if it is present. This approach is often called *biased sampling* because the sample chosen is biased, in that it does not represent the overall population, and the members of the sample have a higher chance of having the disease.

If the purpose of the survey were to calculate the proportion of PLs with WSSV, this type of sampling approach would not be able to answer the question. The bias would give a prevalence that was higher than the true prevalence, because the sample is not representative of the population. However, in the example, the purpose of the survey was just to determine if the population was infected (yes) or not infected (no). Selecting sick and dead PLs increased our chance of finding disease if was present.

This approach has two major problems. The first is that it requires us to know something about the population that we are testing, before we actually test it. In the example above, we *assumed* that the sick and dead shrimp had a higher chance of being infected with WSSV than normal shrimp. It is possible that this assumption is wrong. Perhaps early infection with WSSV has no effect on PLs, so there is no difference, and the sick and dead PLs have other problems unrelated to WSSV. It is also possible that WSSV is in some way protective, and that infected PLs are able to withstand formalin treatment better. If this were the case, the results of our survey would be wrong: we thought we were biasing the survey towards having a higher chance of finding the disease, but instead we actually had a lower chance.

The effectiveness of this approach depends on how good our assumptions are about which animals are more or less likely to be infected. If our assumptions are supported by good scientific research (for instance, a study that has shown that WSSV-infected PLs are more susceptible to formalin treatment than normal PLs) then we can justify the approach. If not, we may be making a dangerous mistake.

> In order for biased sampling to be valid, our assumptions about which members of the population are more likely to be diseased must be correct.

The second problem is the idea that in biased sampling there is no need for random sampling. In our example, we needed 300 PLs, and there were 600 sick or dead PLs. We had to select a sample from a new population, not of 5000, but of 600. The way we select the 300 from the 600 is very important, because if we select the wrong way, the sample may not represent the population of 600, and we may introduce an new, unwanted bias. Clearly, we need some form of random sampling.

In fact, this type of sampling is nothing special or new. We are simply using traditional sampling approaches, but we have *redefined the population.* Originally, we were talking about a population of all the PLs in the batch. Now, we are talking about the population of all the sick and dead PLs from the batch after formalin treatment. Once we have redefined the population, the survey is the same as usual and we need to select a representative sample if we want to be able to make inferences about the population. In this case, if random sampling is used and the sample is representative, we can infer that, if there are no infected PLs in the sample, there are probably no infected PLs in the population (of sick and dead formalin-treated PLs). However, we can't use direct inference back to the original population—the 5000 PLs in the batch. In order to infer that there are no infected PLs in the entire batch, we have be confident that formalin treatment means we were much more likely to have infected PLs in the new population than in the old.

Biased sampling is really just redefining the study population. Inference back to the original population depends on our assumptions about the probability of different animals being infected.

Because biased sampling is not really a separate type of sampling, just a redefinition of the population, it is better not to use this term. Instead, you should make it clear what the different populations are, and what assumptions are involved.

6

Sampling applications

The previous chapter introduced some of the principles behind sampling. This chapter discusses ways in which these principles may be applied to the problem of selecting a representative sample in aquatic animal disease surveys. The first two sections discuss sampling farms, people, ponds, villages, districts etc., first when a sampling frame exists, and then when the population is grouped in some way. The last section in this chapter discusses the more difficult problem of sampling aquatic animals in a range of different situations.

Selecting a sample from a sampling frame

When a good sampling frame is available or has been constructed especially for the survey, the method of selecting elements from the sampling frame is quite simple, and has been described above in the section on random sampling techniques (page 72). A random number table or a computer can be used to pick random numbers, and these are used to identify members of the population.

When the sampling frame is available on computer, this job can be made much faster and simpler, by using a specialised program.

The Survey Toolbox includes one program, **Random Village**, to do this job. The program was designed to pick villages at random from a computerised list, but can be used to pick anything (farms, regions, districts, ponds, tanks, feed supply shops) as long as there is a computerised sampling frame listing all the members of the population. See Chapter 9 for more information about computerised lists (databases) and computer management of information.

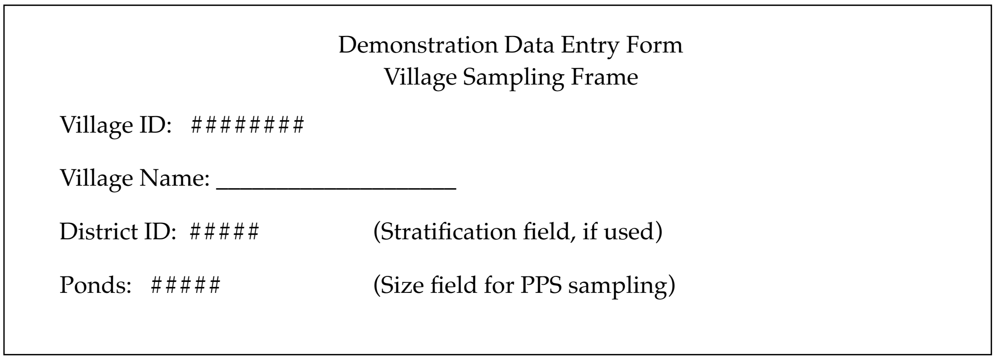
Requirements

To use a database file containing the sampling frame for random selection, the file must be in dBASE, Paradox or ASCII format. The file can contain any number of fields, but must contain at least one field with a unique identifier (name or ID number) for each element. Optionally, it can also contain:

- a field to be used for stratification (e.g. the district for a village sampling frame, the enterprise type for a farm sampling frame), again as a name or number; and
- a whole-number (integer) field to be used for probability proportional to size (PPS) sampling.

Epi Info

If you are building your own sampling frame, you can use Epi Info to enter and manage the data. Creating a new table using Epi Info is described on page 152. When creating the table, you need to create a data entry screen with field codes. An example screen suitable for creating a village sampling frame is shown at the top of the next page.



Demonstration Data Entry Form
Village Sampling Frame

Village ID: ########

Village Name: ____________________

District ID: ##### (Stratification field, if used)

Ponds: ##### (Size field for PPS sampling)

Selecting villages

Using an existing database or table as a sampling frame

To start the **Random Village** program, use the Windows Start button and Programs menu to bring up the list of programs, select **Survey Toolbox**, and choose **Random Village**. Alternatively, you can use the **Survey Toolbox** main menu.

When the program is running, select random villages by following these steps:

Step 1: Click the Open button to open the database containing the sampling frame. This will open the Open File dialogue, where you can change directories or drives to find the file you want. Select the file, and click the Open button.

Step 2: First you need to specify how many elements you want to select (the sample size). Enter the sample size in the Number to Select box, or use the arrows to change the number up or down. Chapter 5 explains how to work out the sample size you need.

Step 3: Select the sampling type. If using simple random sampling, you can leave it as it is. If you are using PPS sampling (see page 79), select Probability Proportional to Size from the Sampling Type box. Specify which field holds the information about the size of each element. Click the arrow at the right of the Size Field box, and select the field with the size information (e.g. number of farmers or ponds in each village).

Step 4: Now specify whether to use replacement or not (see page 76). Choose With Replacement or Without Replacement in the Replacement box, depending on the survey design. Usually, you will accept the default (leave it With Replacement selected).

Step 5: Next, specify if you want to use stratification (see page 79). If you want the sample stratified, click on the checkbox, and select the field that has the information for stratification.

Step 6: Now you must tell the program what information you want displayed. A list of all the fields in the database is shown on the right under Identification Fields. Choose the field that contains the unique identifier. By holding down the Shift key and clicking with the mouse, you can choose a range of fields; using the Control key and the mouse selects multiple individual fields. All the fields that are selected will be listed for the random elements, so pick all the fields that help you identify the element.

Step 7: Finally, you can choose to select from only a part of the list in the file. For instance, if the file contains a list of all villages in the country, but you wish to conduct a survey only in one province, you can instruct the program only to work with villages in that province. First click the Sample from Subgroup checkbox. Then, under Group Field, pick the field that contains the grouping information. In this example, you would chose the field with the province identifier. Next, choose the relationship. You can define a group as all elements that are equal to, greater than, or less than the value specified. In this example, we would chose 'is equal to' from the list. Last, enter the value under Group Identifier. In this example, you could pick the province name or number from the list to specify which province you want. As another example, if you wanted to survey only villages with fish ponds, you could specify Ponds as the Group Field, 'is greater than' as the relationship, and type in '0' as the Group Identifier. Only villages with fish ponds would be included in the sample.

Step 8: When all the settings have been made, click the Select button to have the program choose the random elements. You can click the Print button to print the information, or the Save button to write the information to an ASCII file or a new database.

There are several other options within the program. Instead of opening an existing database, you can click the New button to create a new database file for your sampling frame. You then need to enter the information into the computer yourself before you can select random elements.

If you do have a database with your sampling frame, but there are some mistakes in it, the Edit button opens the file ready for editing. You can change any information, and then return to the main screen to select random elements.

When random elements have been selected, there are two other buttons under the list: Font and Select Another. The Font button allows you to choose a different font to display the list. This is useful if the database has the names of villages listed in a non-English script.

To use the Select Another button, click one of the randomly selected elements in the list and then click the Select Another button. This will delete the selected element and randomly select a new one.

Warning: use the Select Another button only when absolutely necessary. The only reason for using this button is if it is impossible to survey the particular element. In a village survey, this may be because the village is inaccessible or if travel to the village is dangerous. Using this button means that the sample is no longer random, because each village doesn't have the same chance of being selected. Selecting a new element just because the first one chosen is inconvenient can lead to invalid results, wasting all the effort that was put into the survey.

Selecting a sample from a grouped population

A common problem encountered in aquatic animal disease surveys in developing countries is deciding how to select a random sample from a grouped population. In many countries, village farming systems make up an important part of the aquaculture industries. One village usually has many farmers, each with varying numbers of ponds, nets or cages. The cages used by different farmers, for instance, are often in relatively close contact, making it easy for contagious diseases to spread through the farms. Usually, from the point of view of a disease survey, all the animals in the village can be considered to belong to one large group, even if they are owned by many different producers. All animals are generally exposed to the same diseases, and are reared using similar husbandry techniques.

When conducting aquatic animal disease surveys of animals raised in the village system, it is usually more sensible to treat them as a single group, and to draw a simple random sample from the population of all cages belonging to the village. This is difficult, as the cages are owned by many different people, and there is usually no sampling frame available. Even if figures on the village aquaculture farmers are recorded, these are usually only collected once a year, and are often too out of date to be of any use. In addition, cages are rarely individually identified (e.g. with numbers).

A similar situation arises when the units of interest are grouped together. For instance, in an ornamental fish farm, we may be interested in the individual fish. These fish are not individually identified, but are grouped into tanks. Similarly, we may wish to do a survey of farmers, but we don’t have a list of all their names. Farmers may be grouped into villages.

To overcome these problems, a practical technique for randomly selecting a sample from a grouped population is described below. Note that this approach is different from using two-stage sampling (page 82), but both work with grouped populations. This technique allows you to use simple random sampling of grouped populations (only one level of random selection) while two-stage sampling has two stages of random selection. To illustrate the technique we will use an example:

Example

A survey is conducted in a village to evaluate the effectiveness of improved pre-stocking pond preparation in preventing diseases amongst farmed shrimp. The aim is to estimate the proportion of ponds in the village that had a successful harvest last season, and what pond preparation was used. The village has 48 shrimp farms. There are a total of 174 ponds, and we need a random sample of 20.

Building the sampling frame

Village interview

The first task is to build a sampling frame, listing and uniquely identifying every pond in the village. It is unlikely that any one person in the village would know exactly how many ponds each of the 48 farmers had, and any records are likely to be out of date. One option is to walk around the farms and either ask the farmers or count the ponds directly (that is, conduct a census of ponds in the village). Conducting a census is time consuming and it is easy to miss some ponds, but it may be the best approach in some circumstances. Another approach that is sometimes useful is to hold a village interview, to which all the shrimp farmers are invited, and to ask them how many ponds they have. A village interview with all the shrimp farmers requires some organisation, and may take a few hours to complete, but if this is possible, it is an efficient way to collect information for a sampling frame.

Village interviews are also extremely useful for collecting other types of information. These are discussed in detail in Chapter 8, along with guidelines on how to run an interview. Chapter 5 discusses the collection of information for a sampling frame, so only a brief description will be given at this stage. See page 81 for more information.

It is important to try to get as many of the village shrimp farmers as possible to attend the meeting, to make it easier to build a complete sampling frame. After explaining the purpose of the survey, each farmer present at the meeting is asked, in turn, what their name is and how many ponds they used during the previous season. We record this information on a sheet with the columns shown below (Appendix C provides a sample data-recording sheet).

Nº	Name	Ponds	Cages	Total	Cum. total	Selected
1	Lung Noi	5	–	5		
2	Tu Nyai	2	3	5		
3	Silipak	–	4	4		
4	Khamphone	8	2	10		

When the information has been collected from each farmer present at the meeting, collect information about those farmers not present. Ask the group to identify all the farmers who are not at the meeting, and to estimate how many ponds they had. This step may take some persistent questioning, and require prompts to help the farmers think of others not at the meeting. Experience has shown that village interviews are usually able to make a list that contains almost every pond in the village.

Selecting the ‘number’ of the units of interest

The list completed during the village interview serves as the sampling frame, but is different to the sampling frames discussed earlier. When drawing a random sample of ponds, a sampling frame will usually be a list of all ponds with an identification number for each. In this case, the list is a list of all farmers (identified by name and a line number), with the number of ponds used. This list may be used as a pond sampling frame (rather than as a farmer sampling frame) because each pond in the village appears on it (although they are not yet individually identified—we will solve this problem later).

Random selection of animals

The list can now be used to randomly select ponds. There are two ways of doing this: using a random number table, or using a computer. The computer technique is slightly faster and simpler, but requires a notebook computer to be available in the village during the survey. As this is not often possible, the manual technique using a random number table is described first.

Random number table

The technique for selecting random ponds is slightly different from that described previously (page 73) because the sampling frame is different. It is similar to the technique used for PPS sampling (page 79). To pick ponds, use the following procedure:

Step 1: On the data recording sheet, calculate the cumulative total number of ponds and write it in the column marked Cum. Total. The cumulative total is the total number of ponds in the village up to that point.

Example

The cumulative total for Farmer 1 is just the total number of ponds, 5. The cumulative total after Farmer 2 is equal to the number of ponds kept by Farmer 2 (5), plus the previous cumulative total (5), which equals 10. The cumulative total after Farmer 3 is 4 plus the previous cumulative total (10), or 14. This is continued to the last farmer. Note that the last number is equal to the total number of ponds in the village.

Nº	Farmer name	Total ponds	Cum. total	Random number	Ponds selected
1	Lung Noi	5	5		
2	Tu Nyai	5	10		
3	Silipak	4	14		
4	Khamphone	10	24		

The numbers in the cumulative total column are ID numbers for ponds in the village. Farmer 1 has ponds with ID numbers 1 to 5. Farmer 2 has ponds with ID numbers 6 to 10, and so on. These new pond ID numbers can now be used for random sampling.

Nº 1 (Lung Noi)					Nº 2 (Tu Nyai)					Nº 3 (Silipak)				Nº 4 (Khampone)									
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24

Step 2: Using a random number table, pick the first random number (see page 75 for instructions on using a random number table to select random numbers). The random number must be between 1 and the total number of ponds in the village, given by the last number in the cumulative total column. This number represents a pond that is to be selected. Find the owner of the pond from the list.

Example

In our example village, we need to pick a number between 1 and 174 (the total number of ponds). If we picked the number 12, we need to identify which farmer owns Pond 12. Search down the cumulative total column for the first number greater than 12, which is 14, at Farmer 3. This means that Farmer 3 is the farmer of Pond 12.

Step 3: Now we have identified the farmer, we need a way of identifying the individual pond.

Example

We have selected Pond 12 belonging to Farmer 3. This farmer has 4 ponds, and we need to decide which pond we want. The ponds belonging to Farmer 3 would be numbered 11, 12, 13, and 14 if we had a true list of individual ponds. If we want Pond 12, it is the second pond belonging to Farmer 3. A quick way to calculate this is to subtract the cumulative total for the previous farmer from the random number selected. In this case, we would take 10 (the cumulative total for Farmer 2) from 12 (our random number) to give 2. This means that we want the second pond belonging to Farmer 3.

Step 4: Record the number of the pond next to the farmer in the Selected column. Then repeat steps 2 and 3, selecting more random numbers from the random number table and finding the pond in the same way. Continue until enough ponds have been selected. If the same pond is selected twice, discard that random number and pick a new one. It is possible to pick several ponds belonging to the same farmer.

Example

Three more random numbers are selected: 17, 3 and 20. The ponds have been selected and recorded on the sheet below. Check for yourself to see how it was done.

No.	Farmer name	Total	Cumulative total	Random number	Pond selected
1	Lung Noi	5	5	3	3
2	Tu Nyai	5	10		
3	Silipak	4	14	12	2
4	Khamphone	10	24	17, 20	3,6

Computer program

The same procedure can be automated using one of the Survey Toolbox programs included with this book. The program is called **Random Animal** and can be started using the Windows Start menu, selecting Programs, then Survey Toolbox and Random Animal.

Step 1: Collect the information on the number of ponds used by each farmer in the village in the same way, and record it on the data recording sheet. (There is no need to calculate cumulative totals).

Step 2: Click in the Units column of the box on the left of the screen, and type in the number of ponds owned by Farmer 1. Press the Enter key or down arrow to move to the next line. The owner number is automatically entered for you.

Step 3: Continue entering the number of ponds kept by each farmer, pressing the down arrow to move to the next. Make sure that the farmer number on the screen corresponds to the farmer number on the form, as this will be used to identify the farmer.

Step 4: If you make a mistake, you can go back and change it. You can use the buttons at the bottom of the screen to add or delete records, or move up or down the list.

Step 5: When all farmers have been entered, click the Select buttons to indicate whether you want to select a fixed number of ponds or a fixed proportion of the village population. This depends on the survey design being used (see Chapter 10).

Step 6: Enter the number of ponds or the percentage of the village population you want to select in the Number to Select box. You can type in the number or use the up and down arrows to change it.

Step 7: Click the Select button to randomly select the animals.

Step 8: The report window opens, listing all the ponds that you need to select for the sample along with their farmer numbers. You can print the list or save it to disk with the Print or Save buttons.

Step 9: Use the Select Another button to choose a replacement for a pond selected from the list.

Warning: You may need to select another pond if you are unable to examine the first selection, but you should do this only when it is absolutely necessary. When you select a replacement pond, the sample is no longer random, and the results may be biased.

Identifying selected ponds

Regardless of which method you use to randomly select ponds, the result will be a list of farmers and pond numbers that looks something like this:

Farmer 1	Pond 3
Farmer 3	Pond 2
Farmer 4	Ponds 3, 6
Farmer 8	Pond 7
Farmer 11	Ponds 2, 9, 16
Farmer 22	Pond 2

Based on these numbers, identify which individual ponds should be included in the sample using the following steps:

Step 1: Identify the owners of the selected ponds. If selection was done manually using a random number table, the farmers' names are already written on the list. If selection was done using a computer, it is necessary to copy the information back onto the data collection sheet, using the farmer number to identify the correct farmer.

Step 2: Random selection can take place during the village interview. At the end of the village interview, all selected farmers should be asked for their permission to allow the survey team to interview them, or examine or collect specimens from their ponds.

Step 3: For each farmer, visit the farm, or look at a map of the farm.

Step 4: If the ponds are already numbered, the numbers may act as identification numbers for the sample. If not, ask the farmer to count, out loud, all of the ponds. In this way, the farmer is assigning a temporary identification number to each pond. Check on the list for which pond or ponds are required, and note which ones are assigned the relevant numbers.

Example

Using the list shown above for Farmer 4, Ponds 3 and 6 are required. On visiting the farm, you find there are 10 ponds. The farmer starts counting the ponds out loud, with the closest pond as number 1, and the next pond as number 2. The next pond counted is number 3, which is one of the ponds required. As counting continues, another of the ponds is given the number 6, which is also required. When the farmer has finished, the survey team visits the ponds given numbers 3 and 6 for examination or specimen collection.

Step 5: Repeat the process for each of the selected farmers, until all ponds have been examined.

This technique can be used in a wide variety of situations, though some modification may sometimes be necessary. It may initially seem complex, but with good planning and good communication with farmers it is simple to carry out. A few points need special attention:

- Anybody can count the ponds to assign numbers, including a member of the survey team, but it is very important that the person counting *doesn't know which numbers are to be included in the sample*. If they know, they could (consciously or subconsciously) select ponds which are easier to examine, or introduce some other bias.
- Sometimes farmers' ponds are in separate locations, rather than all in the one place. Before visiting the ponds, ask the farmer how many ponds are in each group. The groups can then be assigned ranges of numbers to help decide which needs to be visited.

Example

Survey staff talk with a farmer before visiting her ponds. She says that she has 24 ponds in three groups. The first group of 7 ponds is a short way from the village. This group is assigned numbers 1 to 7. The second group of 12 ponds is at the farmer's house just on the edge of the village. These are assigned numbers 8 to 19. The other ponds are at her brother's house. These are given numbers 20 to 24. If Pond 14 is required, only the group containing that pond (the one at the house) need be visited. Count these ponds, starting at number 8. When using this system, it is important that the person assigning numbers to the groups does not know which ponds are to be selected.

- Sometimes ponds cannot be examined or specimens collected. This may be because they are too far from the village to reach, or for some other reason. In these situations, a different pond will have to be substituted. When a pond is replaced, the sample is no longer completely random, because the chance of the replaced pond being in the sample is zero. This should therefore be avoided whenever possible. However, for practical reasons some ponds will need to be replaced. When selecting a new pond, the same procedure as that described above should be used, using either a random number table or a computer. It is better not to simply pick another pond belonging to the same farmer.

Sampling aquatic animals

Principles and constraints

Much of the discussion of sampling to this point has been concerned with methods to select cages, ponds, tanks, farmers, villages, districts etc. This is because, in many surveys, these are the units of interest. Understanding 'aquatic system health' means that we often need to consider the entire system (pond, cage or tank) in order to understand disease. However, there are many situations when, in order to understand what is going on within a system, we need to look at individual animals. For instance, to determine if a batch of shrimp post-larvae (PLs) is infected with white spot syndrome virus (WSSV), we may choose to select a sample of PLs and test them with a polymerase chain reaction (PCR) test to look for the DNA of the virus. Selecting PLs from a batch of 2000 can be thought about in exactly the same way as selecting farms from a district. We are trying to examine a sample, and infer the state of the entire population based on the finding from that sample. For this to be valid, the sample must be properly representative of the population, and we know that the only way to reliably select a representative sample is by using some form of random sampling.

The same principles apply whether we are assessing the prevalence of epizootic ulcerative syndrome in a pond, trying to estimate the population of fish in a reservoir, or calculating the incidence rate of marteiliosis (QX disease) in oysters in an estuary. In each situation, we need to examine individual animals, and these animals need to be randomly selected in order to truly represent the population.

We have discussed a number of different types of random sampling that may be applied to different situations. In summary, these are:

- Simple and two-stage random sampling (based on an existing or new sampling frame);
- Systematic random sampling (based on the ability to 'line up' the population into a sequence; and
- Spatial sampling (using location as a proxy for the sampling frame).

When considering farms, ponds, cages, villages, districts etc, it is generally possible to obtain an existing sampling frame or create a new one. If not, systematic sampling may be used, and if this is still not possible, some form of spatial sampling is possible. The characteristics of populations that make these forms of sampling appropriate are:

- the size of the population is generally relatively small, or the population can be grouped into a relatively small number of subpopulations (the number of villages in a province, the number of farmers in a village);
- the populations remain relatively constant over time (they do not change rapidly);
- the populations are stationary (they do not change position in space);
- the elements in the population can be individually identified relatively easily (village or farmer names, pond or cage numbers); and
- the elements can be examined relatively easily (looking at a pond, measuring a cage, interviewing a farmer).

In contrast, in the case of most aquatic animals:

- they are often grouped in relatively homogenous large populations (thousands of fish in a pond);
- populations change rapidly (production cycles are relatively short, disease problems can cause very high mortality rates in a short time);
- populations are highly mobile (except for mature molluscs);
- it is impossible or impracticable to identify individuals in most populations; and
- it is difficult to examine individuals, as they are hard to catch and examination often causes severe stress or death.

Another consideration is that aquatic animals are kept in a wide variety of systems, ranging from the open ocean, reservoirs and rivers, to cages, ponds, raceways, nets, tanks and so on. The challenges and opportunities for random sampling animals in these different systems vary widely. It is impossible to lay out a standard approach that is applicable in every situation. Instead, this section attempts to clarify the objectives of random sampling of individual aquatic animals and the methods by which it may be achieved, and then to present some examples.

When planning a disease survey involving individual animal sampling, designing a sampling strategy that is both practical and obtains a good representative sample is one of the most difficult challenges. In many cases, it is simply not possible to meet the requirements of random sampling. Obviously, if we are not able to use random sampling, there is the risk that our sample will not be representative and that our survey results will be biased. The real task during survey design is to work out, within the practical constraints of the system, how to select a sample that minimises any possible bias.

A good understanding of the principles of random sampling provides us with a clear idea of what we are aiming to achieve and the 'right' way to go about it. A good understanding of the physical, cultural, economic and social constraints on conducting a survey gives us a good idea of what is possible. This will almost always fall short of what we know to be the perfect or right way to sample. Our task, in designing a survey, is to look at what is possible, and make it as close as we can to what is right.

When the perfect sampling approach is not possible, we may be able to determine what sort of biases are likely to be caused by the practical, imperfect strategy that we have to use. If the direction of any likely bias can be determined, this makes the results much more useful. For instance, if a sampling strategy was judged to be more likely to catch healthy fish than diseased fish, and a survey yields a disease prevalence of 18%, we could conclude that we don't know the real prevalence (because proper random sampling was not used), but the prevalence is certainly equal to or greater than 18%.

Strategies used to select a sample

An ideal sample is simply one that perfectly represents the population with respect to the characteristic of interest. This can only be reliably achieved using random selection. Simple random sampling requires a sampling frame that contains every member of the population once and only once, and uniquely identifies them. It also requires that selected members of the population can be caught and examined. In aquatic animal populations, this is very rarely possible.

Systematic sampling removes the need for a sampling frame, but requires that the population can be lined up, either physically or conceptually, and that animals can be selected at regular intervals.

The third strategy that may be used is spatial sampling. This identifies random locations instead of elements of the population, and depends on the animals being relatively stationary in that location.

Other non-probability approaches to sampling may be used when probability sampling is not possible. For instance, haphazard sampling is often able to select a representative sample, but is not as reliable as random sampling. Haphazard sampling is far better than convenience or purposive sampling, which should not be used.

Practical strategies employed in the field to get a representative sample generally attempt to overcome some of the constraints imposed by the characteristics of aquatic animal populations listed above, so that one of these four approaches can be applied.

Most of the techniques involve one or a combination of the following approaches:

- lining the population up so that systematic sampling can be used;
- decreasing the mobility of the animals; and
- using some form of spatial sampling.

Some management activities provide an opportunity for systematic sampling. When such an opportunity exists, this is usually the best way to select a good representative sample. This will be discussed below.

Decreasing the mobility of animals, and spatial selection strategies, are used in situations where systematic sampling is not possible. These approaches will be discussed in relation to specific production systems.

Sampling techniques and equipment

Farmers and fishers have been catching fish and other aquatic animals for many thousands of years. Around the world, a huge range of different tackle and equipment has been developed to assist with the capture of animals. Equipment includes nets, traps, lines and spears, with many variations on each. Poisons, electricity and explosives have been added to the list more recently. Many countries and many communities have traditional methods that have been developed and adapted to local conditions. It is impossible to examine each of these different types of equipment in this book, and only a few typical methods will be used as examples. However, there are two very important considerations when assessing equipment to use for sampling aquatic animals.

The first is that the survey team should know about, and fully understand, the common and traditional methods used for catching aquatic animals in the survey area. Before suggesting a new approach, it is worth thinking for a moment why the existing system is being used. In most cases, local tackle has evolved because it is the most practical and efficient approach to catching aquatic animals in local conditions. Understanding this, the survey team should have a very good reason for suggesting that something new be used instead. If the current system has worked successfully for hundreds or thousands of years, there must be something pretty good about it.

Respect and understand local tackle and equipment for catching aquatic animals.

The second consideration is to understand the differences between the aims of local producers and survey staff. Local farmers or fishers may have a range of objectives. For instance, farmers may wish to collect all animals during a harvest. Alternatively, they may use a series of partial harvests, in which case they are only interested in catching fish of marketable size. Similarly, fishers may be primarily interested in catching large fish, and actively avoid catching immature fish in order to ensure a sustainable livelihood.

This preference for large rather than small fish may have other, unintended effects. For instance, large fish may tend to be healthy fish while diseased fish tend to be smaller. The fish captured using traditional techniques may therefore not be representative of the population, but show a bias towards healthy fish. Any survey based on this type of sample would underestimate the level of disease. Conversely, some methods of catching aquatic animals may catch slower individuals more easily than faster ones. A sample collected in this way would be likely to catch more diseased animals, and the survey results would overestimate the level of disease in the population.

The objectives of farmers and fishers using different types of tackle are often not the same as the objectives of the survey team—that is, to collect a representative sample.

A decision on the appropriate tackle to use depends on an assessment of these two issues: the most practical and efficient way of capturing individual aquatic animals in the local environment, and the sorts of biases that this method might introduce.

Opportunities for sampling

The variety of species and management conditions used in aquaculture and fisheries means that different sampling approaches need to be used in different situations. However, within a single management system, there are sometimes good times and bad times to attempt random sampling. During the production cycle there are some management activities that provide an opportunity for sampling. At other times during the cycle, sampling may be much more difficult.

In general, these management activities provide situations where *systematic sampling* is possible because, in some sense or other, the population can be 'lined up'.

Management activities may make it possible to collect a good representative sample from a production system in which such sampling would normally be impossible. However, there are some constraints. The first constraint is the time at which sampling is carried out. When sampling during specified management activities, the survey team has no control over when the sampling will occur. The times of activities are determined by the producers. If this fits in with the requirements of the survey, there is no problem. However, sometimes it may not. For instance, a disease may express itself most strongly during the growing phase. During this phase, animals are left undisturbed. At harvest time, when sampling is easier, affected animals may have either died or recovered, leaving little evidence of the disease. Using harvesting as an opportunity to detect such a disease would not be appropriate.

The other problem of sampling during specific management activities is the need for an excellent relationship with the producers. Harvesting, for example, is a busy time, involving a lot of work, during which a farmer has a chance to earn some income after months or years of investment. If survey activities make it difficult for the farmer to do their work, or if the survey damages the animals or causes the farmer to lose income, the involvement of the survey team at harvest time will not be welcome. Establishing a good relationship with the farmer, understanding their activities and needs and ensuring that survey activities bring as much benefit and as little disadvantage as possible to them, are all necessary to take advantage of management opportunities for sampling.

Stocking

Stocking aquatic animals into an aquaculture system (or an enhanced capture fisheries system) provides a good opportunity to evaluate the disease status of the juvenile animals. Because the fry, fingerlings, post-larvae or spats are small and relatively less valuable than adult animals, they are easy to handle and cheaper to collect. Testing animals at this early stage of life is useful to check if any pathogens are being introduced to a culture system, but doesn't help if we are interested in studying the progress or impact of disease.

Let us consider the example of silver barb fry (there may be some differences from other species, but most of the following ideas can be adapted). Typically, fry for stocking are delivered from the hatchery or nursery in plastic bags, with a specified number of fry in each bag. Many farms may stock fish from different hatcheries, either because of availability or to spread the risk of receiving poor quality stock. In a survey, the question of whether to treat one batch from one hatchery, or all the batches from different hatcheries, as a single population depends on the survey question being asked (see Chapter 10 on Survey Design). However, each batch (which might be made up of one or more bags) should normally be considered one population.

There is more than one way to collect a random sample. Some will ensure that the sample represents the population well, but require more effort. Others will be more practical, but may not represent the population so well and be likely to produce biased results.

One simple approach is use a two-stage sampling scheme. After selecting bags from the batch, select some fry from each selected bag. While this is fast and practical, there are some problems. If the sample-size calculations are conducted using the methods outlined in Chapter 11, page 179), the number of bags will usually be high, so most or all of the bags will have to be sampled. The second problem concerns which fry to select.

Another, much better, approach is to treat the batch as a grouped population, and select a random sample using the approach described on page 96. This will produce a list for inclusion in the sample that specifies we need fry numbers 10, 38, 43, 52, 57 and 69 from Bag 1, fry numbers 8, 23 and 48 from Bag 2, and so on.

A third approach is collect a systematic sample from all bags. For instance, if there are five bags, each with 300 fry, and a sample of 50 is required, then the population is 1500 and the sampling interval is 1500/50 or every thirtieth fry. The first fry to select should be selected randomly from between one and 30.

Once the sampling approach and individuals to sample have been worked out, the next step is to select those individuals. There are various techniques to do this. The one most commonly used is to mix the bag a little, and use a scoop to select some fry. If 10 are needed from a bag, a small scoop can be used to collect some fry, with any excess being returned. If there are fewer, take another scoop. This approach isn't random sampling at all, but convenience sampling: it is very likely that the fry that are near each other are similar in some way. For instance, if the bag was stirred to try to mix the fry, the weaker ones would have moved towards the centre of the swirl. If the scoop was taken from there, it may be possible to collect all weak fry, and bias the results. Alternatively, a scoop from the edge may collect all strong fry, with the opposite bias.

Despite this warning, this is a quick and practical approach to collecting a sample. Where time and practicality are most important, it may be suitable. However, you must be aware of the potential biases that may be caused by convenience sampling.

A similar and equally practical approach is to assess the number of fry on the basis of the volume of water. If a bag contains 300 fry in about 1 litre of water, and we want a sample of 100 fry from the bag, we might mix the bag, and pour out 330 ml of water into a second bag. This should contain about 100 fry. Again, this is a form of convenience sampling, and is unlikely to even get the right number.

A practical, but much more time-consuming method can be used to implement either the systematic or grouped population random approach mentioned above. Often, when fry are delivered, the total number per bag is only an approximation. This is because the hatchery didn't count the fry but only estimated the number in each bag, or because some of the fry died on the way to the farm. To be sure of the number stocked (and therefore the survival rate), many farmers choose to count the fry before stocking. To do this, they often use some sort of small container, sieve, or perforated spoon that lets most of the water out and retains a known number of fry of a certain size. For instance, one spoonful might contain, on average, 10 fry. By transferring the contents of the bag to another container, one spoon at a time, the farmer can count the total number of fry reasonably accurately and quite quickly. In the process, the fry can be thought of as being 'lined up', and suitable for systematic sampling.

In our above example of systematic sampling, we wanted to sample every thirtieth fry. If the first random number chosen was 17, then 17 fry would be transferred, and then one selected. Subsequently, we would take one fry every three spoonfuls until the whole population was counted. This still leaves the question of which fry to take. This doesn't matter, as long as it is consistent. For instance, the one nearest the handle of the spoon may be chosen.

This approach, while being somewhat laborious, can provide a great deal of confidence that a truly representative sample will be chosen (unlike the convenience approaches mentioned above).

The other random approach mentioned above, based on a grouped population and simple random sampling, can be achieved using a similar method. Count all the fry from one container to another. This time, instead of selecting every thirtieth, choose fry corresponding to the randomly selected numbers. For instance, if the first fry to be collected is number 27, transfer two spoonfuls. On the third spoonful, count out 6 fry and take the seventh for the sample. This approach clearly takes longer, and is unlikely to produce a significantly better sample that the systematic approach.

Despite claims by suppliers, the numbers of fry in individual bags may vary considerably. The only way to check this is to count the fry as described above. Basing the sampling design on the claimed number per bag may cause problems if the fry number much more or fewer than claimed. For instance, systematic sampling will produce a total sample size smaller than expected if the number is much smaller than claimed. In this case, you might decide to test only those collected, or to collect more, until you reach the desired sample size. If collecting more fry, be sure to select fry distributed through the entire population in order to avoid bias. For instance, it is not appropriate to start again at the first bag, continuing to collect every thirtieth fry, as this means that the first bag is sampled twice, and the last bags only once.

If bags contain multiple species, the research question determines whether only one species is studied, both species are considered part of one population, or each species is a separate population.

Grading

Another good opportunity for sampling is during grading operations. In some species, such as macrobrachium prawns or salmon, animals are graded during the growing phase. This is done to divide animals by size or sex to ensure more even groups, providing less competition during feeding and more even growth rates.

Grading may be performed in several ways, such as manually, or using buckets or sieves with holes of a specified size to let smaller animals through. However it is conducted, grading involves handling every animal, and is therefore an excellent opportunity for systematic sampling. Sampling may be made even more efficient if it is stratified by size. This can be done by taking a systematic sample after grading rather than before.

Sampling at grading has an important advantage over most other management opportunities for sampling, as it occurs during the middle of the growth period, rather than at the beginning or the end. If disease is likely to be active at this time, sampling during grading is the most appropriate method of selecting a representative sample.

As with most other methods, the disadvantage of sampling during grading is that grading is a busy management activity, and very good farmer cooperation is required. Handling animals runs the risk of damaging them or increasing stress, so it should be kept to a minimum. Only those animals selected for sampling should be subjected to extra handling beyond that required for the grading process.

Vaccination

In some more intensive systems, animals are vaccinated for particular diseases, such as vibriosis. Vaccines can be administered in several ways, but some are given by injection, requiring the handling of fish. As with grading, this means that systematic sampling is feasible during the vaccination process.

Transfers

Some management systems require the transfer of fish from one cage to another for various reasons, such as cleaning or repairing cages. This offers another opportunity for systematic sampling.

One approach is to use a fish pump to pump the fish, in water, into the new cage. This effectively 'lines up' the fish, although they may be passed through the pump unevenly, so that several come at once, followed by none for a period. Sampling may be possible by counting (even approximately) the fish emerging from the pump, and capturing, say, every fortieth fish. With large numbers of fish, it may be simpler to catch single fish at regular time intervals, rather than trying to count fish as they emerge. Just as spatial sampling uses locations as a proxy for the sampling units, this approach uses time. Catching one fish every 30 seconds is not an exact form of systematic sampling, but it is still likely to get a representative sample.

Another approach to transferring fish is to pass them through a race. With a little thought and preparation, systematic sampling can be achieved in this situation as well. The technique used may differ according to different physical layouts, but in some cases using a scoop net to take fish from the outlet of the race may be adequate, counting the fish as they emerge.

Transfers may be done manually, by lifting a net to crowd the animals, catching them in containers (e.g. buckets) and transporting them to a new site. This approach is very convenient, as each bucket can be assumed to hold a roughly equal number of animals. Systematic sampling then involves the collection of one animal from every, say, sixth bucket (or alternatively, 2 animals from each bucket), depending on the sample size. Deciding which animal to chose from a bucket should, of course, be done randomly—simply picking one off the top may introduce bias (for example, if smaller animals tend to sit on top). This is discussed below, under Harvesting.

Molluscs may be transferred from one place to another during their growth cycle, and can be sampled systematically at that time. During transfers, they will usually be gathered into small groups, such as trays or on sticks. Each group may be assumed to have a roughly equal number. If time permits, use simple random sampling, based on the techniques for sampling from a grouped population described above (page 96). It may be faster to use systematic sampling, taking one oyster from a regular number of groups.

Harvesting

Harvesting provides one of the best opportunities for sampling aquatic animals, for the following reasons:

- Every animal is handled.
- Animals are fresh, but will soon die, so handling them or keeping them out of water will not cause damage or upset the farmer.
- Animals are often transported in groups (crates, buckets etc), which makes systematic sampling convenient.
- In some systems (e.g. shrimp) the product is graded at the time of harvest to determine the price. Grading is usually done by selecting a sample (for example 10 kg of shrimp for each 500 kg produced) and examining it closely. Samples for grading are often collected using haphazard sampling, but systematic sampling is possible without increasing the workload. A subsample of the graded shrimp may be collected using systematic sampling at the grading table.

Despite these advantages, there are a number of potential problems with sampling at the time of harvest:

- Only complete harvests give access to the entire population. Many systems use partial harvesting and the harvested animals are very unlikely to be representative of the entire population. Progressive harvests, where a small number of animals are taken out regularly, are another form of partial harvest common in smallholder fish ponds.
- Harvesting is a busy time, and survey staff must work closely with the farmer to fit in and avoid disrupting the work.
- Harvesting may take a long time, sometimes a number of days. To collect a good sample, the survey staff need to be present over the whole harvest.
- The decision to harvest might be made at very short notice, leaving little time for the survey staff to travel or prepare. Many farmers will do an emergency harvest if they suspect there is a danger of a severe disease outbreak.
- Except in the case of some emergency harvests, harvesting usually occurs when animals are mature. This may not be the time of interest for studying the disease.

Capture sampling

When a study requires animals to be sampled from a closed population (pond, tank, cage, net, reservoir etc), or an open population (river or ocean), but none of the above management opportunities is available, it is necessary to capture the animals. In smaller closed populations, in which every individual can be captured (for instance, by draining a pond or tank) you can use systematic sampling. This section deals with the more common situation where it is not possible to capture every individual. Instead, some form of tackle is used to capture a group of animals for the sample.

Capture techniques

An enormous range of capture techniques has been developed around the world, each designed to capture different species under different environmental and cultural conditions. Common methods include cast nets, dip nets, lines, traps and trawl nets. Less common capture systems include poison, explosives and electrofishing.

One of the important characteristics of fishing tackle and other capture systems is their selectivity, or their ability to capture different parts of the population. With almost any tackle, one is more likely to capture some animals than others. Naturally, this always introduces a bias into the sample captured using tackle, with the potential to undermine the validity of the survey results. For example, when capturing shrimp with a net, the diameter of the opening, rate of movement of the net and mesh size all play a role in determining which shrimp will be captured. Very small shrimp might pass through the mesh, larger healthy shrimp might be more able to escape the net, and less active or diseased shrimp might be less able to escape, and therefore over-represented. On the other hand, the way in which the tackle is used may also influence the bias. A net used at intermediate depths at the edge of a pond might capture mostly healthy shrimp, while one used in the middle or at the bottom might capture mostly weaker or diseased shrimp.

The result is that virtually all capture techniques are biased, and none is able to produce a reliably representative sample. This is because they are not able to meet the basic requirement of random sampling—that all members of the population have an equal chance of being selected in the sample. This poses a major problem for studying aquatic animal populations. The simple solution is to avoid sampling based on capture techniques, and use techniques described earlier in this chapter. However, when this is not possible, we must consider how to minimise bias and improve the representativeness of a capture sample.

Improving a capture sample

Minimising bias. One of the best approaches is to attempt to minimise any bias caused by the choice of capture method. This requires a good understanding of the tackle being used and the factors influencing the catchability of different parts of the population. For example, it may be possible to use a series of nets of differing mesh size, each able to capture different segments of the population. Nets with a very fine mesh size can capture the smallest fish, but larger fish can escape. Larger mesh sizes miss the smaller fish, but capture the larger ones.

Selecting the best capture system or combination of systems often requires the assistance of experts with knowledge both of fishing tackle and the biology and behaviour of the species being targeted. This should be combined with local knowledge of the area being sampled, to identify all potential habitats.

Regardless of the care taken, there is still a large opportunity for the resulting catch to be a biased sample of the source population.

Some capture techniques that are commonly used to sample aquatic animals are much worse than others. Shrimp are often sampled using a feed tray, but this will almost guarantee a biased sample: only actively feeding shrimp will be found in the sample, while sick shrimp will never be detected.

Spatial and temporal sampling. Aquatic animal populations regularly move from place to place, especially in open waters such as rivers, oceans and estuaries. These movements may be over relatively small distances (perhaps only changes in the preferred depth), or can involve migrations over thousands of kilometres. Even in a closed population, such as a shrimp pond, animals can be found in different locations at different times. Because of these movements of animals, sampling in a single location will not usually produce a representative sample.

A knowledge of the movement patterns of the species being studied, and the factors that control these patterns, can help control bias in a sample and even dramatically increase the efficiency of sampling. For example, understanding the migration patterns of fish along a river may mean that the entire population can be sampled at a single location at a particular time of the year, during the peak of its migration.

To capture a representative sample using spatial sampling (sampling at different locations) and temporal sampling (sampling at different times), two main issues must be considered. The first issue is the pattern and scale of movements. Some species may be found at different depths at different times, and may move distances of several kilometres, while others move over vast distances. The second issue is the factors affecting these movements. Long-distance migrations are usually related to the season, and have an annual cycle. Changes in depth or location may be related to day and night, or the phase of the moon, or other factors such as mating behaviour. An understanding of all these factors will help you identify the best location and time to sample animals.

While the use of a particular capture technique, such as a cast net, may not be a good approximation of random sampling, it is still possible to improve the representativeness of the overall sampling by selecting random locations and random times at which to use the capture technique.

Example

A survey of one species is being conducted in an estuary. Nets are used to capture the fish, and it is estimated that the net will have to be used 50 times to catch the required sample. In order to get the best sample, select points at random all over the estuary, as described in Chapter 5. Take one sample at each point. However, to take into account the possibility of parts of the population being in different places at different times, assign each point a random time (e.g. a random hour between 1 and 24), at which to sample that point.

This approach increases the chance of a more representative sample. It can be thought of as sampling three aspects: sampling the fish (using a non-random capture technique), sampling space at random, and sampling time at random.

With each of these three aspects, we can define a 'population' and sample from it. For the fish, the population is all the fish of the relevant species in the estuary. For space, the population is all the points or locations in the estuary. For time, the population is all the hours in a day. If we suspect seasonal movement patterns, we could sample from all the months in the year.

If we have a good understanding of the biology and distribution of the fish, we can further refine the definition of the population to make the sampling more efficient. For example, if we know that some areas of the estuary are more suitable habitats for the fish than others, we might narrow the 'population of points' that we sample from by sampling at random only from suitable habitats. Similarly, if we know the species is only present during one season, we might sample random weeks from that season.

Crowding. One common cause of bias in capture sampling is the ability of aquatic animals to move. When a cast net is thrown, some animals will be able to move out from underneath it, thereby avoiding capture. The animals that are not able to move quickly are often the weakest or smallest, which means that the sample is likely to be biased. One way to overcome this problem is to use crowding to restrict the mobility of the population.

Crowding can be used in closed populations, mainly of animals in cages and nets. It may also sometimes be used in tanks or ponds, when a large net is used to collect all the animals into a small area. Once the animals are crowded together, they are much less able to move. It is then possible to use a spatial random sampling technique, by picking random locations, and perhaps random depths, from the area crowded with animals.

Example

A net is used to crowd fish into one side of a pond. Three decimal dice are used to pick random numbers. The first is the number of metres from the left end of the net. The second is the number of metres from the bank into the net. The third is the depth in centimetres. Rolling all three dice chooses a specific random position and depth. Select the fish at this location for the sample.

The advantage of crowding is that the fish at the chosen location is not able to move away quickly, and is therefore somewhat easier to identify and catch. Large fish and small fish, healthy fish and sick fish, will all be kept pressed together, and so have a roughly equal chance of being selected.

The main disadvantage of using crowding is that it risks stressing and damaging the animals, and may therefore be unacceptable to the farmers. If crowding fish, do it as quickly as possible to minimise any damage. This calls for good planning and preparation, and selecting random locations before starting. In some cases, haphazard sampling is used rather than formal random sampling, as it is faster. However, remember that this always risks introducing biases in the results.

Subsampling. Subsampling is the process of drawing a smaller sample from a larger sample, and is particularly useful when sampling open populations using commercial capture techniques. Commercial techniques, such as trawling, can often capture very large numbers of animals quite quickly. If all the animals captured are included in the sample, either the required sample size will be captured from a single location in a short time, or, if many locations are sampled, the total catch will result in a much larger sample than is actually required. This is likely to increase the cost of the survey, particularly if some sort of laboratory examination is involved in the analysis. The solution is to take a subsample from the total catch and use it as the sample.

Subsampling may involve some work, but, as the total population to be sampled (the catch) can be handled, is it is possible to use either simple random sampling or systematic sampling to ensure a representative and unbiased subsample. For example, when a catch of fish is dumped onto the deck of a trawler for sorting, every twentieth fish could be selected for the sample, while the remainder are discarded or retained for sale.

Subsampling can also be applied to commercial catches after they have been brought to shore. Systematic sampling of a catch during unloading at the dock, or random sampling of fish at markets, are both ways of subsampling the commercial catch.

While formal random techniques can be applied when subsampling, it is important to keep in mind how the initial sample (the commercial catch) was taken. Random subsampling ensures that the subsample is truly representative of the population from which it is drawn (in this case, the commercial catch). The subsample is not representative of the population that the initial sample was taken from (the wild population). If the capture technique used for commercial fishing is biased (as it usually is), the catch is not a representative sample of the wild population, and neither is the subsample taken from the catch.

Understand bias. When a study is being conducted and the only available sampling method is to capture the animals, it is often impossible to ensure that the sample is not biased. When surveying aquatic animals in open waters (rivers, estuaries or the ocean) there are no management opportunities to allow systematic sampling and there is no ability to crowd the animals. The use both of carefully selected and applied capture techniques, and of random spatial and temporal sampling to get good coverage of the population, will help decrease bias, but it is often not possible to eliminate it. Even in a closed environment, such as a shrimp pond, if a capture technique such as the use of a cast net is required for practical reasons, it is likely that the results will be biased.

One way of dealing with this problem is to conduct studies allowing us to understand the bias, and take it into account when interpreting the results of the study.

Example

A study of shrimp is planned, and animals are to be sampled using a cast net. The researchers know that cast nets are more likely to capture slow, sick shrimp than fast, healthy shrimp, and that the results of the study are likely to be biased. In order to understand what the bias might be, they conduct a preliminary study. They use a cast net to sample several ponds, and estimate the proportion of sick shrimp. Immediately afterwards, the pond is harvested and a systematic sample used to estimate the proportion of sick shrimp. The researchers compare the two estimates from the two sampling techniques. They find that, by the (biased) cast net sample, about 9% of shrimp are sick. Using the (unbiased) systematic harvest sample, they find only 6% of shrimp are sick. They conclude that the bias caused by the cast net sampling technique results in an overestimate of the prevalence of disease, by about 50%. They go ahead with their cast net study, but when reporting the results, they adjust their estimates to account for this bias.

This example demonstrates that bias is only a problem if you don't know how big it is, and in what direction it is. If you can measure the bias caused by a sampling technique, you can adjust for it.

Redefine the population. As we have seen, it is almost impossible to draw a truly representative sample from an open population. If the sample is biased, we can't use estimates made from the sample to make inferences about the total population.

A practical solution to this problem is to redefine the population of interest. Normally, when studying a species, we are interested in all members of that species—big and small, young and old, sick and healthy—in order to understand the true effect of disease. However, if it is not possible to represent all groups of animals in our sample, we may ask ourselves 'What does our sample actually represent?' Often, the answer to this is 'catchable fish', or those fish in the population can be caught by the capture technique used. Our sample only contains a few fish, but, if we are careful about where and when we sample, it is likely to be representative of those fish that can be caught. Clearly, our sample does not represent any of those fish that can't be caught by our capture technique.

By redefining the population, we can overcome the problem of bias and inference. If we find that 24% of fish in the sample have a disease, we may confidently assume that about 24% of the catchable fish in the population also have that disease. Unfortunately we are still left with an important question: what about the others? How many uncatchable fish have the disease? It is rarely possible to provide a solid answer to this question.

Sampling less mobile animals

One last special case of sampling is worth considering. Most of the problems with sampling aquatic animals result from the animals' ability to move quickly in three dimensions. One group of animals, molluscs (at least when mature), either can't move and remain anchored in the one place, or generally move more slowly or over shorter distances. This offers a range of sampling opportunities that allow us to collect a good representative sample, both from farmed and wild populations.

Management opportunities

In many culture systems, molluscs such as oysters are moved from one location to another during different stages of growth. When oysters are handled for transport, they can be easily sampled with systematic sampling. Oysters are often located in trays or on sticks with a relatively even number in each group. For example, using systematic sampling, every hundredth oyster might be sampled. If a average tray contains 30 oysters, skip 3 trays and sample the tenth oyster (counting systematically across, then down) in the fourth tray. Skip the remaining 20 oysters in that tray, along with the next two trays, and sample the twentieth oyster from tray after that, and so on.

Spatial sampling

If wild molluscs are to be sampled, or cultured molluscs need to be sampled while they are in the water, you can use spatial sampling techniques. The following procedure may be adopted:

1. Define the area to be studied (e.g. an estuary or a single bay).
2. Identify and map areas suitable for the mollusc species being studied.
3. Generate random points in the study area, retaining only those that fall within areas suitable for molluscs. Assign each point a number.
4. Using a hand-held global positioning system (GPS) unit (or a very accurate map), visit each random point.
5. Establish a minimum search distance around the point (e.g. 5 metres). If molluscs can be found in this radius, select one at random. If no molluscs are present, move to the next point.

This approach is particularly suited to species such as oysters that are anchored to the one place, and are easily visible. For more mobile molluscs, or ones that are less visible (eg live beneath the sand), it will be necessary to search carefully within the defined area. Targeted use of traditional capture techniques may be of assistance.

Using this approach will give you a reasonably representative sample. You can improve the estimate from this sample by weighting the data according to the density of molluscs around each selected point.

7

Principles of data and specimen collection

General principles

Sources and quality of information

There are several sources of information, collected in different ways, about aquatic animal diseases. In general, the quick and easy ways provide data that is less complete or less reliable. If good quality data is needed, collecting it usually requires more time, effort and expense.

Existing records

The easiest way to gather data is to use data that somebody else has already collected. This is known as passive surveillance, and is discussed in Chapter 3. Records that may already exist include outbreak reports, laboratory submissions and population estimates.

The problem with using existing records is that the data were collected for reasons different from those that interest us. For instance, population data (the number of farms or ponds) may have been collected from the villages for the purpose of reporting total numbers within a province. It is therefore only available as provincial totals, but for surveys we need individual village totals. When we use existing records, we have no control over how the information was collected or how complete it is. We can use active surveillance to overcome these problems.

Interview data

We can obtain better quality and more relevant information by collecting it specially, to answer particular questions. A relatively quick and simple way of collecting information is to ask people who know. The quality of information collected through interviews depends both on whom we ask, and on how we question them. In aquatic animal disease surveys, there are a number of people that know about the disease situation. Producer know the most about the animals they deal with, but may not have very good knowledge or training about diseases. Village extension workers who have received basic aquaculture training may be more knowledgeable about diseases than the producers, but may not know specific details about individual farms or animals. Local fisheries officers usually have much better technical knowledge of disease and production, and an overview of disease problems in their area. Provincial officers are likely to be even better trained and more experienced, and have a good understanding of the disease situation in their provinces. However, at each level up the hierarchy, training and technical knowledge increases while contact with the actual farms and animals decreases.

Example

A survey is planned to find out which diseases have the greatest impact in the village smallholder system. Funds and time are limited, so a written questionnaire is sent from the central office to provincial fisheries offices throughout the country. The questionnaire is addressed to the provincial officer, and asks for information identifying the most important diseases in the villages. There are only 26 provinces, so the survey is easy to prepare, and all the responses are received back at the central office within 2 weeks. When the data is analysed, the results closely match the government's published list of priority areas for disease control.

The result is not surprising, but it may not be particularly useful. The provincial officers know the government's disease priorities, as do their district staff. Most of the disease reports or calls for assistance received will be related to one of these priority areas, because it is known that these are important areas for the government. Unfortunately, the producers may have other priorities. The survey failed to ask the people who really had the information it required. The same survey could have been conducted by asking district staff, village extension staff, or the producers themselves. Each of these options would be more difficult, involving progressively more people and expense; however, the quality of the information would be better.

In some situations, particular groups of people are the best source, as they are the ones that hold the key information. For instance, in evaluating funding needs for local fisheries offices, the staff of those offices are the ones who know best the work and expenses of the office, just as producers know best about the management of their production systems.

Collecting information from people has a few added complications, due to human nature. When asked a question, a person may give the wrong answer for various reasons:

- They might have forgotten about something that happened some time ago.
- They might not want to admit to not knowing the answer.
- They might lie deliberately, because they fear how the information will be used.
- They might not understand the question.
- The interviewer might not understand the answer.

To overcome these problems, interviewers should be know which problems can occur and use techniques to avoid them. Chapter 8 includes a detailed discussion about collecting information from people.

Examining animals

Although farmers know their animals better than anybody else does, they might not be skilled at identifying particular diseases. When we are interested in collecting reliable information about aquatic animal diseases, it is sometimes better to examine the animals directly, rather than simply consulting the owner. The advantage of directly examining animals during a disease survey is that we can collect data of much better quality: we no longer need to depend on the farmer to make a diagnosis.

A disease survey can be conducted by just a few survey teams over a relatively short period. It is therefore possible to use survey staff who are well trained and highly skilled in diagnosing diseases (perhaps more so than the district or provincial staff).

Collecting specimens

It is not possible to diagnose all diseases simply by carrying out clinical examinations of fish, shrimp or molluscs. Many diseases can only be diagnosed using laboratory tests.

Surveys are often not concerned with clinical disease, but with measuring the prevalence of subclinical disease or evidence of past exposure. In these cases, laboratory testing of specimens (such as tissue samples) collected from animals is necessary. The quality of information gathered through the specimen collection and laboratory testing is usually better than we can achieve through any of the other means. Often, laboratory testing of specimens is the only way to provide a definitive diagnosis. On the other hand, surveys using specimen collection and laboratory tests are often expensive and time consuming. When planning a survey, it is important both to evaluate the available resources and to determine the quality of the information needed. The four different approaches to collecting data are summarised below.

	Existing data	Interview data	Examining animals	Collecting specimens
Expense	Very cheap	Moderately expensive	More expensive	Most expensive
Speed	Very fast	Fast	Slower	Slowest
Up to date?	Often out of date	Usually up to date	Very up to date	Very up to date
Accuracy	Not good	OK	Good	Best

Collection of specimens for laboratory examination

The best specimens for laboratory examination are always live animals. This is because aquatic animal tissues tend to break down very quickly, and other microorganisms can invade. Even when an animal has been dead for only one hour, opportunistic organisms may make it difficult to determine the real cause of the disease. For diagnosis of a disease problem, selection a range of animals— a few apparently normal, some obviously sick, and some freshly dead animals. The transport of live animals depends on the species. As a general rule, animals should be kept in an environment as similar as possible to that from which they came (e.g. water should have the same salinity). If using sealed plastic bags, ensure that there is adequate oxygen for the journey.

Often it is not possible to collect and transport live animals. In such cases, we have three options: to preserve freshly killed animals for transport to the laboratory; to prepare specimens in the field to be sent to the laboratory; or to perform the necessary tests in the field.

Euthanasia techniques

Before aquatic animals are preserved, necropsy performed or specimens taken, the animals must be humanely killed. This should be done for fish and crustaceans, as for molluscs where possible. Various techniques are available, including pithing, anaesthetic overdose, decapitation, or administering a sharp blow to the head. In fish, the eyes will roll upwards when the animal is dead.

Pithing

Pithing involves severing the spinal cord from the brain using a sharp needle. Insert the needle into the soft tissue at the top of the head, and work it back and forth quickly to cut the spinal cord. While it causes some tissue damage, pithing is quick, simple, and doesn't affect any skin parasites.

Anaesthetic overdose

To euthanase animals using anaesthetic overdose, place them in water containing one of the following anaesthetics at the concentration listed.

Methane tricaine sulphonate (MS 222)	40–80 ppm
Benzocaine	100 ppm
Quinaldine	10–15 ppm

If the animal is not dead after about five minutes, add more anaesthetic.

Preservation techniques

Chilling

To chill specimens, place them in a waterproof container and place the container directly onto ice for shipment or into a refrigerator for storage.

Freezing

Put the specimen in a waterproof container that seals tightly, and place the container on dry ice or into a freezer. Frozen tissues should remain frozen until ready for examination in the laboratory.

Chemical preservation

Chemical preservation can be used when organisms do not have to be kept viable. Fish up to 50 mm long can usually be placed directly into the preservative. Larger fish should be opened along the abdomen or the length of the swim bladder before being placed into the preservative.

Blood sampling

There are two main ways to withdraw blood: venisection and venipuncture. Venisection is best for small fish. Anaesthesia is normally required before withdrawing blood.

Venisection

Immediately after euthanasia, place the freshly killed specimen on its right side and dry the caudal area. Sever the caudal peduncle using a sterilised scalpel or scissors. Let the first few drops of blood drain out before taking a sample, as this minimises contamination by other body fluids. To collect blood, insert a heparinised capillary tube directly into the artery. Blood will be drawn into the capillary tube. Place one drop of blood from the capillary onto a glass slide.

Venipuncture

Position the specimen on its right side. Insert a hypodermic needle below the lateral line at the level of the anal fin. The needle should pass through the skin and muscle until it enters the caudal artery enclosed in the haemal arches of the spine. Withdraw the syringe’s plunger slowly to draw blood from caudal vein. Vacutainers can be used in a similar fashion. Once the needle has been inserted and blood appears, attach the tube, breaking the vacuum and withdrawing the sample.

Preparing blood smears

If an examination for blood parasites is required, the sample must be collected from freshly killed fish before the blood clots and/or haemolysis occurs. Place a single drop of blood near one end of a clean glass slide. The drop of blood should not be too big or it will not smear evenly. Place a second slide on the side of the drop closest to the centre of the first slide so that it is touching the blood. The two slides should be at 45° to each other. Some of the blood should spread along the bottom of the second slide. If it does not, move the second slide slightly towards the drop of blood. Move the second slide along the first, away from the drop. The blood should move with it, leaving behind a smear.

Air-dry the smear by briskly waving the slide. If humidity is high, you can use a portable electric hair drier. Fix the smear immediately in absolute alcohol or methanol. Fixed smears can be stored at low humidity for several days or weeks before staining.

Data collection and recording

Survey data can be collected and recorded either by specially trained survey staff or by the producers themselves. Well-trained survey staff can carefully explain questions when necessary, and control the quality and completeness of data recorded. Recording sheets can be relatively simple, as the staff know what each section is for, and how it should be filled out.

Survey staff usually only visit a producer once or a few times. They can usually only collect information or specimens at the time of their visit, and cannot make frequent or regular observations and recordings. To overcome this problem, some surveys ask the producers to collect and record some or all of the data. When producers are required to record data, much more care must be taken in the design of data collection forms and in explaining requirements. However, this method allows continuous measurements (for example, of daily temperature or pH values) to be made and unpredictable events (such as disease outbreaks) to be observed and recorded as they happen.

Collecting data

Collecting information during survey field work is often difficult and exhausting work. There is nothing worse than to find at the end of the field work that the information wasn't recorded properly or that sheets have been lost, and all that hard work has been wasted. Before collecting any information, it is important to plan how the information will be recorded, and to have a well-organised system for keeping the records safe.

When planning the survey, decide the sorts of information to be collected and how it is to be recorded, and then design sheets with places to write all the information. Good design should make it easy to record data in the field and easy to use the data for analysis. Appendix D provides a series of sample data recording sheets for the types of surveys described in this book. The sheets can be copied and used directly, or modified for your specific needs. It is useful to be aware of a few general guidelines when using data recording sheets:

- Village/farm/fisher codes. Every sheet should be marked with the source of the information. This may be the name of the village, farm or fisher. It is also useful to use a numeric code to identify all the villages or farms in the survey. This prevents problems with misspelt names, and makes it easier to enter the information into a computer for analysis (see Chapter 9 for more on using computers for analysis). Other useful data to include on each sheet are the date and the survey team (if more than one team is used).

- Always write down information as soon as it is collected, and never trust it to memory. For instance, during a village interview, much information can be collected quickly. The person conducting the interview may be too busy talking with the producers to write down all their answers. In this situation, one member of the survey team should lead the interview, and another should be responsible for recording all the information.
- Make sure that each member of the survey team knows their responsibility. One person should be responsible for writing down the information, while others may be responsible for collecting it.
- Write the information clearly and carefully, using a clipboard to make recording easier. If the information is hard to read when being analysed, mistakes may be made.
- Design sheets to minimise the amount of writing required. If you know that only a limited number of responses is possible, list the responses on the sheet so that the correct one can be ticked or circled, rather than being written out. On the other hand, if asking an open question, make sure there is room to write the full answer.
- Keep the sheets safe and dry. At the end of each day, either store all completed sheets in a folder in a safe place, or send them directly for data entry. If working on or near the water, use waterproof paper and a pencil to record the data.
- Don't make a 'neat copy' of the data. Every time information is copied by hand, you risk making mistakes. Make sure that the information is clear enough to read the first time you record it. If you want a copy of the information (which is a good idea, in case the original gets lost), make a photocopy.
- When collecting blood or other specimens, label each tube carefully with a permanent waterproof marker. Use an identification number on the tube, and record the same number on the data sheets, so that it is clear where the blood came from.

Recording the data

As each specimen is collected, record the number of the container on a data recording sheet. Other basic information that should also be recorded to help in the analysis of the specimen includes:

- the village, farm or fisher from which the specimen was collected;
- the date of collection;
- an identifier for the owner;
- the species of the animal (if more than one species is being sampled in the survey); and
- other animal characteristics, such as length, weight, sex or age.

A sample data recording sheet is shown in Appendix C.

Producer record keeping

There are significant advantages in having the producers keep records. Producer record keeping allows continuous observations and records to be made without requiring survey staff to be present all the time, and therefore requires fewer survey staff. It allows unpredictable events to be recorded when they happen (such as the collection of suitable specimens from a disease outbreak).

However, there are some very important considerations to be taken into account when requesting producers to collect and record survey data. The first is the producers' ability to collect and record the required information: do they have the necessary training and skills? If, for instance, a farmer asked to record pond pH daily has to be able to use a pH test kit and record the results accurately.

The second consideration is the producer's willingness to be involved in the survey over a period. Record keeping involves increased work and time. If the producer sees no benefit to the activity, they are unlikely to continue doing it, especially if the survey extends over many weeks or months.

Designing producer record-keeping systems

There is no point asking producers to keep records or make observations when they don't have the skills, equipment or time to do so. This will only result in missing data or fabricated data, which are of no use to the survey. In deciding what data to ask the producers to collect, we must first understand what the producers are able to collect reliably. This will vary enormously, depending on who we are surveying. For instance, in a developing country, village farmers with single ponds and no formal education are unlikely to be able to make extensive observations or tests, to make complex calculations, or, if they are illiterate, to write the results in a data recording sheet. However, they may have time to make simple observations and to record them with appropriate, non-written systems. A commercial fishing crew on a large ocean-going trawler probably have the resources and education to collect much more detailed information and make precise complex records, but may not have the time to do so.

The most appropriate survey information to ask producers to collect is information that they already collect for their own purposes. For instance, many fish farmers are likely to keep records of food consumption over time, and it is easy for the farmers to record this information for a survey. They already have the required skills, and the task takes no extra time. However, even with such tasks, it is important to ensure that all farmers are recording the data in a compatible way. For instance, one farmer might measure the food in kilograms, another might count scoops (which may be larger or smaller than scoops used by other farmers), and a third might count the bags used. Some training or extra data collection is required to ensure that these different values can be compared.

When asking producers to make observations and records of things that they do not normally record, some training will always be necessary. The more complex the task, the more detailed the training will have to be. Measuring pH or using a Secchi disk to measure water transparency are relatively simple tasks, and most people can be trained in minutes to do them. However, the correct preservation of pathological specimens, including dissecting an animal to remove the appropriate tissue samples, is a much more complex task and may be beyond the ability of some producers.

When planning a survey, you will often have to compromise between the amount of data you want and the ability of producers to collect it. In any case, some training will be needed to ensure that the data is collected properly. Once collected, the data must be recorded, and again the approach used depends upon the producers' abilities.

Written records

If the people being surveyed are able to read and write, you can use written records. The considerations discussed above for the design of data recording sheets apply, but much more care must be taken to ensure that all questions or data items are very clearly and unambiguously explained, that the exact form of the answer required is clear, and that the effort required for recording answers is minimised.

Visual aids and definitions are often useful to assist producers when recording results. For instance, rather than asking producers to describe the colour of pond water, it may be better to supply standard photographs and to ask producers to simply write the number of the photograph that most closely matches the water colour. The same approach can be used to describe common lesions in fish. This saves time for the producer, and ensures that the results are relatively standardised.

Non-written records

When producers can neither read nor write, they can still keep records if the record keeping system is modified to suit their abilities. If producers are numerate (can write numbers), they can use a system with numbered pictures—such as a table with the date (day of the month) in one column and the picture identification number in the next.

If a producer cannot write numbers, there are other options for record keeping. For example, to record the amount of feed used each day we might set up a series of 30 tin cans. The farmer puts one pebble in a can for each scoop of feed used. Each day, the farmer starts a new can. At the end of the month, the survey staff visit to count and record the number of pebbles in each can.

Another approach can be used in surveys to examine productivity. If a farmer has a pond producing fish primarily for the family's own consumption, they might be asked to keep the tail of every fish eaten, and hang it up to dry. At the end of the month, survey staff count and measure the tails to give an estimate of the kilograms of fish produced during the month.

Most non-written recording systems are similar to these. A physical object is collected or organised in a way that allows survey staff to determine what the observation was. Marks on a chart, rather than physical objects like pebbles or fish tails, can be used to count things.

Maintaining cooperation

One of the most difficult problems when using producers to keep records is ensuring that the data is complete over the period of an extended study. Many producers may initially be enthusiastic, but over time the participation rate drops off. While it is rarely possible to overcome this problem entirely, especially in a long study, there are several things that can be done to help.

Before undertaking almost any activity, human beings do a quick cost–benefit analysis. They ask themselves, 'If I do this, how much will it cost, how difficult will it be, and what will I get out of it?' If the benefit does not outweigh the cost, they don't do the activity, and this applies to participation in a survey. If the cost is very small (such as answering questions for 10 minutes), they will probably be happy to participate even if they don't properly understand the benefit. To them, the benefit may be a chance to sit down and chat for a while, or to show how much they know about fish farming. However, if the cost of the activity is much greater (as in recording production on their farm every day for six months), there must be a much greater benefit to make them decide to do it. To get a high level of cooperation, we must to ensure that the cost–benefit ratio for participating producers is good enough.

We can do this by minimising the cost or by maximising the benefit. To minimise the cost, we must make participation in the survey as easy as possible. One way is to collect only data that producers collect anyway. Another is to design recording sheets so they can be completed with just a few ticks or numbers (see above).

The other side of the equation is to maximise the benefit to the producers who participate in the survey. This can be done through good communication of the real benefits of the survey, or by artificially creating benefits.

The most important part of this process is good communication with the producer. It may be sufficient to properly explain the benefits at the beginning of the survey, including the direct benefits to the producer and the potential benefits to the broader industry or community. Often, the benefits to the farmer include regular contact with skilled staff, who can offer advice and help sort out problems; a closer relationship with organisations that can offer support in the future; a better understanding of the production and disease problems in the producer's particular situation; increased respect within the community through participation in the survey; development of new skills through training required to conduct the survey work and so on. Artificial benefits may include payment or other incentives for participation, or the chance to win a prize.

Even if a producer decides that the benefits of participation outweigh the costs, there is still the possibility that they might change their mind during a long survey. Try to ensure that the producer recognises real benefits during the survey period, rather than after the survey is finished.

There are several common ways to achieve this. One is to organise frequent visits by survey staff, to collect data progressively and discuss any problems. This helps maintain data quality, but also gives the participant a feeling of importance. Providing regular feedback of results is also a valuable way to increase benefits to producers. If they can see how the information they provide is being used, they are much more likely to continue contributing to the survey.

When data submitted by a producer is incomplete or inaccurate, it is important to contact the producer to correct the problem. In this way, producers quickly recognise that others value their data highly and that they should, too.

Probably the most important of these approaches is to maintain regular contact with survey participants, to always recognise their importance in the survey process, and to treat them with the high degree of respect they deserve.

8
Collecting information from people

Approaches to collecting information

As we saw in Chapter 7, one of the fastest, most convenient ways to collect information is to ask people. Asking people about aquatic animal diseases is much easier and less expensive than collecting samples or examining animals. The problem is that the quality of information gathered by asking people is often not as good as that of information collected by direct examination of the animals.

This chapter discusses methods for collecting information from people, and ways in which the quality of this information can be improved. These methods differ in the choice of people from whom to collect information, the involvement of the survey staff and the ways the activity is structured. We might collect from one person at a time, or from groups of people. If we collect from groups, there are different ways to decide who should be in the group. The next section discusses the collection of information from groups.

Information can be collected directly by survey staff, usually in face-to-face interviews but sometimes over the telephone. Alternatively, the information can be recorded by the person being surveyed, without an interviewer present. This is usually done using a written questionnaire, but may in some situations be done using computerised questionnaires, either on a computer that is taken to the person or through the Internet. Self-completed questionnaires have to be very carefully constructed to make sure that everybody using them understands the questions properly and gives useful answers.

There is a range of approaches to structuring data collection from people, depending on the level of organisation. The self-completed questionnaire is one type of highly structured approach, in which every question and item of information is determined beforehand. With face-to-face or telephone interviews, we can use standardised questionnaires in which we plan beforehand the exact words used by the interviewer, so that all members of the survey team interview in exactly the same way. This helps overcome some of the problems associated with interviews. For example, if somebody doesn't understand a question, they may need to be prompted with examples of the type of answers expected. Prompting may bias the survey results, because people are likely to respond using the example provided. Standardised questionnaires use prompts or examples, chosen beforehand, that do not bias the survey results.

These formal and rigid approaches to collecting information pose some problems. First, they only find answers to the questions we ask. Often, when researching a new problem, researchers might not be aware of important information that could help solve the problem. Because they are not aware, they do not ask about it, and they will never find out. The second problem with structured questionnaires is that they may not collect the best quality information. This is because a structured questionnaire follows the pattern of 'ask a question, write the answer, ask the next question, and so on.' The interviewer is not required to think about the answer in detail, just to write it down, and the person being interviewed recognises this quickly. Because there is no discussion or consideration of the information, just a series of questions and answers, the process can quickly become boring, especially with a long questionnaire. There is a tendency to try to finish the process as quickly as possible, so answers may be given without proper consideration. The interviewer is not expected to question or discuss the answers, simply to record them and move to the next question.

Semi-structured interviews can be used to avoid some of these problems. In a semi-structured interview, the interviewer has a list of questions or data items, or a checklist of key points, all of which must be covered during the interview. However, the items are not dealt with using simple questions and answers. Instead, the interview takes the form of a discussion. The discussion can cover many topics, mostly related to the subject of the study, but the person being interviewed may want to talk about other issues or ideas. As each topic in the interviewer's checklist is covered the appropriate pieces of information are recorded, but the process might reveal new and possibly important unexpected pieces of information. One of the main advantages of this approach is that the interviewer actively discusses and thinks about the information being supplied. If the answer is unclear, or contradicts something that has been said before, there is a chance to discuss it further and clarify what the person actually means. The exercise is not boring for either person, and the quality of information collected may often be higher than with a structured interview. A semi-structured interview can be used to collect facts and figures, as well as opinions and other complex responses or ideas.

At the other end of the spectrum is the unstructured interview. In this approach, there is no prepared data collection sheet or checklist with various items to cover. Instead, there is simply a general topic to be discussed with a completely open mind and no preconceived ideas, with the objective of finding out what the person thinks or does. Again, this type of interview usually takes the form of a discussion. Information collected from unstructured interviews is often not suitable for quantitative analysis, as it is hard to ensure that comparable facts or figures are collected from every person interviewed. However, this approach is very useful when starting to investigate a problem about which you have no information. Records from an unstructured interview usually take the form of notes or a summary of the key points discussed. These may be read to look for patterns or suggestions that point the way to further formal investigations.

Participatory research is an approach to problem solving and a group of techniques that may be used so that communities and stakeholder groups can work together with researchers to control and participate in research to overcome problems facing the community. Participatory rural appraisal (PRA) or rapid rural appraisal (RRA) are related approaches. In participatory research, when trying to understand a disease and look for solutions, each of these different approaches may be used in sequence. The investigation may start with an unstructured interview, move to a semi-structured questionnaire once the main areas of investigation have been identified, and finally use structured questionnaires to gather specific quantitative data. Participatory research also involves the community in the analysis of the research data, interpretation of the results and decisions on how to implement the findings for the benefit of the community.

Information from groups

Village farmer or fisher group interviews are very valuable tools for collecting information from producers, and ensuring that the quality of the information the best possible. Group interviews are faster and more efficient, because all the people you need to ask are gathered together in the one place. The quality of the information is better because you can cross-reference information, comparing the thoughts and ideas of different people. It is easier to get reliable information about events in the past, because producers can help each other to remember things that one person may easily forget. The group memory of a village is much greater than the memory of an individual.

One of the advantages of village interviews is that many different types of information can be collected in a short time. The information collected at an interview depends on the aim of the survey, but some of the key types of information are discussed in this chapter.

These are just some of the reasons for using group interviews. Other benefits will be discussed later. However, to achieve these benefits interviews must be conducted skilfully, which requires practice and aptitude.

The type of interview depends on the type of information required. An unstructured interview, which may take the form of a guided discussion of disease problems, is useful to start to understand the main concerns of village fish farmers. A semi-structured interview has a list of topics or areas that need to be discussed, but there is still room for the producers to talk about issues not on the list. A structured interview has a clear list of specific questions that the producers are asked to answer. Structured interviews can quickly collect many different facts and figures. In practice, most village interviews will be a mixture of these different types—specific questions to determine key pieces of information (for instance the date of the last outbreak of a particular disease), and more general unstructured discussions to understand the problems facing the village.

Who should attend?

Decisions about who should attend group interviews or meetings depend on the objective of the survey. Clearly, the best people to interview are those who can best provide the required information, but it is not always clear who these people might be. Deciding is much easier if the objectives of the survey are clearly stated.

One approach is to invite everybody, for instance the whole population of a village. This ensures that everybody has a chance to contribute, but is usually impractical because too many people would be involved. Instead, it is usually better to identify particular people or groups of people.

If complete information is required (for example, if one of the objectives of the survey is to identify how many ponds or cages there are in a village), all people with ponds or cages should be invited. A pond usually belongs to a particular family, and to find out about that pond only one person from the family need attend. However, this will often be the head of the family (usually a man), and their perceptions of the health problems or management of the pond may be very different to those of their spouse or children.

The same problem arises with group interviews where complete population data are not required, and the aim is simply to get a representative idea of, say, the main problems in an area or common production systems. In these cases, men and women, who have different responsibilities, may have different perceptions of problems.

The solution is to try first to identify the different groups of people involved in aquaculture, and ensure that each has an opportunity to participate. People involved may variously be described as *resource users* or *stakeholders*. For instance, when considering the management of a community reservoir, we might include among stakeholders:

- fishers catching wild stock from the reservoir;
- farmers with nets or caged fish in the reservoir (including separately those responsible for decision making, daily management and feeding, harvesting and marketing the fish);
- those involved in trading or marketing the product of the reservoir;
- consumers who eat the fish from the reservoir;
- advisers and regulators who provide advice on managing the reservoir (perhaps separately including government, commercial and traditional advisers);
- suppliers who provide seed stock and equipment to either farmers or fishers; and
- others who use the reservoir for purposes other than aquaculture or fishing, who may be in potential conflict (e.g. potential polluters, or farmers using the water).

To get a full understanding of the issues, we should consult each of these groups. Pay careful attention to collecting information from groups that are often ignored or are less able to have their opinions heard, such as women, children and the poor.

Each of the groups may be interviewed separately during a series of meetings, and the results from each group compared to evaluate the accuracy of the results and discover how the perceptions or priorities of the groups differ.

Another useful approach is the *focus group*. This is usually a small group (6 to 8 people) selected to focus on a particular issue. The members of the focus group are often chosen to represent particular points of view (e.g. one member from each of the main stakeholder groups). Focus groups can discuss issues in detail and provide a great deal of insight into new issues. They are very useful before a more structured survey.

Collecting population information

One of the objectives of a group interview may be to gather complete population information, for instance about all the ponds or farmers in a village. Similarly, we may wish to collect information on the total catch by village fishers. In this discussion, we will use the example of village ponds. Ideally, all owners of ponds in the village should attend the interview. This is often not possible, but we should make an effort to ensure that as many of the owners as possible are present.

It is usually beyond the control of the survey team to determine which person from the household comes to the meeting. In some societies, such a meeting will be seen as important, and the head of the household, usually a male, will attend. In others, it may be seen as a waste of time, and lower-status members of the household asked to go instead. Alternatively, some people may be busy working when the meeting is held, so others have to go instead.

The best people to have at the interview are those who know best the health and management of the ponds. The head of the household may own the ponds, but it is perhaps the women or children who take most responsibility for caring for them each day, and who therefore have the best knowledge of diseases that may have affected the ponds.

The size of the group affects how well the interview can be run. Small groups are easy to manage but the 'group memory' and amount of discussion is also small. Large groups can be very difficult to manage and the comments of individuals lost in the general chatter. The ideal group contains between 10 and 20 people. When more people attend, one way to make the meeting run more smoothly is to split the group into two or more smaller groups and run separate interviews. This allows us to check the information gathered from one group against that gathered from another.

Organising the meeting

People involved in fishing or aquaculture need to know about the interview some time beforehand, so they can arrange to be present. The way this is done depends on the overall organisation of the survey, and the ease of communication with the village.

If communication with the village (by telephone, or preferably in person) is easy, contact the head of the village a week or two before the planned visit. Ask them to convene a meeting and invite all the relevant people. This initial contact is very important, and some brief explanation should be given of the purpose of the survey and the meeting.

Even if the village has been notified a week or two before the meeting, it is often a very good idea to remind the head about it the day before.

When access to the village is very difficult and more than one visit is not possible, the first contact with the village may be at the time of the interview. In this case, be prepared to notify producers individually and wait until most are able to come to the meeting.

The best time for the interview depends on the normal activities of the village, and on the other activities planned as part of the survey. An important point, which will be raised again later, is that the producers are assisting the survey by providing information. Make every effort to make producers' involvement in the survey as simple and enjoyable as possible, planning meetings for a time convenient to them, not for the survey team's convenience.

If interviewing fishers, for example, it is obviously best to avoid times when most are out fishing, although the most convenient time and place may be when most have just returned and a meeting can be held by the water. For farmers, harvesting is a very busy time, so this should be avoided. However, many may be absent at times when little activity is required. Understand the patterns of activity well, before deciding on the timing of the meeting.

Example

A survey is being conducted which involves both village interviews and the collection of specimens from a sample of the village ponds. The survey team could travel to the village early in the morning, and hold the interview before the farmers start work. The team could examine the fish during the rest of the day and

> return home in the evening. Another approach would be to travel to the village in the afternoon, hold the interview in the evening after the farmers finish work, and collect specimens the next morning, returning home or moving on to the next village during the middle of the day. If producers are happier to have a meeting in the evening, the second option is probably best.

Interviews may be held in a community meeting hall, a school, a place of worship, the home of one of the producers, an open space or by the water. Whichever venue is chosen, it should be quiet, with few distractions, so that the producers' responses can be clearly heard.

Who should be the interviewer?

There is an art to being an interviewer, and some people are more suited than others to the task. The person interviewer should:

- have a very clear understanding of the purpose of the survey, the order of the interview and the way information is to be collected;
- have a good technical knowledge of all the diseases being discussed and be able to answer questions on these diseases;
- speak fluently the same language and dialect as that spoken in the village (sometimes it is helpful if the leader of the interview has the same accent);
- be comfortable addressing the group and able to speak in a clear, loud voice;
- understand and respect the culture and social customs within the village, and make people feel relaxed;
- be aware of sensitive issues;
- value and respect the knowledge and skills of the people and let them know that their knowledge and assistance are appreciated;
- not be intimidating to the farmers or fishers, who should feel free to express their opinions;
- be able to elicit a response from the quieter or shier members of the group and encourage participation by all producers; and
- if interviewing women's groups, be a woman.

It is often difficult to find one person with all these qualities. When selecting and training survey staff, confident, intelligent, outgoing and sensitive people should be identified, trained and strongly encouraged. A skilful interview leader can have a strong impact on the quality of the information collected, as well as on the willingness of producers to participate in future surveys.

Getting good information

We conduct a survey to gather information to help answer a question or make a decision. If the information is not correct, the answer or the decision will be wrong. When information comes from laboratory analysis of specimens, there is still a chance of error but we have a pretty good idea of the likelihood of mistakes. When we deal with people, it is much harder to ensure the quality of information collected. This section discusses approaches to improving the quality of information.

Listening

One of the major roles of government officers is extension—providing advice and training to producers on animal health issues. During an interview, some survey staff with a government background may find it difficult to abandon this role as a provider of information, and have a tendency to interrupt producers to correct misconceptions, provide advice, or, at worst, lecture. The position of the survey team during the interview is that of student, not teacher, and its task is to ask questions and record the answers. For many staff, remembering to listen rather than to speak is one of the hardest things about an interview.

Despite this, a village interview offers an excellent extension opportunity. This is discussed below (Encouraging cooperation, page 140).

Encouraging participation

One of the most important advantages of using village interviews to collect information is the ability to question and collect information from many people at once. If many of the producers at the interview don't participate in the interview (don't offer their opinions or take part in the discussions) then this potential advantage of village interviews is not realised. A successful interview is one in which all the participants have had the opportunity to express their views and report their experiences fully, and it is up to the survey team, and particularly the interview leader, to try to ensure that this happens.

There are many reasons why some people may be reluctant to participate during an interview, but the main one relates to social status and local customs. At most meetings, there will be a range of people from different social levels. Amongst the producers, the head of the village will usually be a highly ranked person, along with other village members with official positions (a fisheries worker, for instance). The other producers will have their own ranking, perhaps related to age, the number of ponds or boats kept, or other criteria. In addition to the established social ranking within the village, another layer of ranking is introduced during a meeting to discuss aquatic animal health issues. Those members of the community with more experience or knowledge of health issues will be in a stronger position to participate than those with less knowledge. In this ranking, the survey team members will themselves usually be identified as animal health experts, and therefore of very high status. While this may be an advantage, it may also intimidate village members who are uncertain of their own knowledge of health issues and who fear exposing their ignorance.

One problem is that producers with a higher social status are more likely to express their opinions and speak for the rest of the village, while lower-status participants are less likely to speak, and certainly less likely to contradict or correct statements by village leaders. These generalisations do not apply to every society or culture, but the objective of ensuring the participation of all producers remains. A good understanding of the local culture and a sensitivity to status issues is an advantage, but some other techniques also help.

At the beginning of an interview, there is naturally some initial shyness or reluctance to speak. An activity, early in the interview, that breaks the ice and gets every producer speaking can help overcome these inhibitions. The form this activity takes depends on the objectives of the interview and on what is appropriate according to local customs. It may be a formal part of the interview (such as the collection of information on the number of ponds owned by each participant, discussed below or it could take the form of a game or activity specifically designed to relax the participants, get them speaking, and get them thinking about their animals.

Example

A competition is a good way to get people involved at the start of the interview. If you intend to build a village-pond sampling frame, you can use this activity to get people thinking about how many ponds they and their neighbours have. Divide the farmers into several teams (4 or 5 if there are enough people). Ask each team to try to calculate the total number of ponds in the village. Tell them that the team that gets closest to the true value will win a prize. Give them 5 minutes to think about how many ponds there are, and then ask each team to report its answer. Write the teams and their guesses on a board or large sheet of paper. You can then go ahead with building the sampling frame (page 99) to work out the real total number of ponds. The prize can be a net or some other simple item that assists with production.

This type of exercise starts people thinking about the number of ponds that each person has, including those people not present at the meeting. It also makes the next activity (asking each person how many ponds they have) much easier to understand—it is being done to find out who won the competition. It also underlines the fact that it is really just the total number of ponds that is required, and information on the number kept by each person is not going to be used for any purpose other than to calculate that total.

Another way to make it easier for all producers to participate is to try to minimise the perceived difference in status between the survey team and the people. This can be done by physically reducing the distance between the two. In meetings where participants sit on the floor or ground, the survey team should also sit on the floor or ground. In any case, where the size of the group permits, sitting in a circle is better than the traditional 'speaker/audience' seating arrangement. Reduce the verbal distance, too, by letting the producers know that you understand their problems and practical difficulties.

Directing specific questions to specific members of the group can also encourage participation. Often a few participants will do most of the talking. If a general question is posed, and one of these more vocal people answers it, the question can then be asked again, directed at specific people.

Example

One of the objectives of an interview may be to determine which are the most important disease problems amongst farmed aquatic animals in the village. First, we list all the diseases that occur in the village, and then have the group identify the most important. When the question is put to the whole group, a village leader answers that high mortality in fry is the most important problem. The leader of the interview then turns to one of the less vocal participants and asks 'Do you have a problem with fry mortality?', 'Do you have any other problems with your ponds?', and 'Which of these problems is most important?' This can be repeated for several of the other farmers, to confirm the initial response or to get a better picture of what diseases are occurring and how the producers rank them in importance.

In many societies, women may have different social status to men. Despite this, it is very important to encourage women to attend the interview, and to encourage them to participate. There are several reasons for this. While women may not be thought of as owners of the ponds, they are often responsible for most of the care of the animals and spend the most time with them—giving them a great deal of knowledge about problems that affect the animals. Another reason for encouraging the participation of women is that women represent a separate social network within the village, and are interested in, and have access to, different types of information from the men. For example, when we are building a village sampling frame (page 97), women are often able to provide better details of the number of ponds or cages owned by families not present at the meeting.

In many societies, women will sit in a separate group from the men during a meeting. If there are only a few women present, it may be difficult for them to make a contribution. However, if enough women are present, questions may stimulate some discussion amongst the women's group. A more confident spokesperson may emerge from the women's group to report on these discussions; if not, targeted questions to particular women should be able to discover their perspective. Alternatively, the women's group can be interviewed separately, preferably by a woman interviewer.

Another opportunity to collect valuable information from people who are reluctant to participate occurs after the meeting. It is often good to provide food or drinks, so that people will stay and chat at the end of the meeting. This is a much less formal and intimidating environment, and everybody is already thinking about the issues raised during the meeting. Moving around and talking to people at this stage might reveal new ideas from people who were too shy to express themselves during the meeting.

Language

The language used during an interview plays an important role in the participation of the fish farmers and in determining the quality of information collected. Clearly, the leader of the interview and the person recording details should be fluent in the local language and dialect. Speaking the same language and using the same expressions helps to reduce the distance between the survey team and the fish producers, and encourages better participation.

As well as the language and dialect, the *level* of language used should be appropriate to the people being interviewed. With well-educated, large-farm producers, it may be appropriate to use some technical words. However, the language used should normally be kept clear and simple, and avoid technical terms (something that trained fisheries staff sometimes find difficult). There is a fine line between using simple language and appearing to be condescending.

Disease names

The choice of words is particularly important when discussing particular diseases. Fisheries staff may be interested in a particular group of diseases, which they usually refer to by their technical name (often in English). The survey team will often think about diseases as particular entities, each with a particular separate cause. On the other hand, producers may not be aware of the specific causes of particular diseases, and think more in terms of disease syndromes. When fish or shrimp behave in a particular way, and show particular signs, they are thought of as having a particular disease. This disease syndrome might have a unique local name, or the technical name of a particular disease might be used.

Example

In a particular village, the fish farmers identify a particular disease as being very important. They describe how the disease affects grass carp, causing deep red erosions in the skin. They have a local name for this disease, but when asked by the survey team, they call it EUS. The local fisheries officer has visited the village to explain about epizootic ulcerative syndrome (EUS), and said that this was the technical name of the disease. There has never been any post-mortem or laboratory confirmation of the cause of the disease and, in fact, there are several different but apparently similar diseases that occur in the village, grouped together into the same syndrome. EUS is only one of these diseases.

The tendency of producers to talk about disease syndromes based on patterns of clinical signs, rather than specific diseases, should be kept in mind during the interview. Mistakes can be avoided by paying attention to a few points.

Don’t use the technical name for a disease when asking about that disease. If you know the local name, and understand what disease or diseases it truly represents, use that name. Better still, if you are interested in a particular disease, describe the signs of that disease or show pictures of animals showing typical signs, and ask the producers to tell you what it is called. Find out whether this name is used for a single disease, or if diseases are grouped together. If producers use a technical name, question them carefully to check that you are both talking about the same disease.

In some situations, the clinical signs and behaviour of the disease in the population are distinctive enough to be sure that the name given by farmers refers to a single disease. In other situations, it is not so simple.

When collecting and analysing disease information from aquatic animal producers, you need to remain aware of the type and quality of information being sought. When a disease syndrome, which may include several other diseases, is described, it is not possible to distinguish between the different possible causes in the analysis. If a survey was conducted to assess the incidence rate of EUS based on interview data, but the local name for the disease included several other diseases causing similar signs, we could not report that the incidence rate of village outbreaks of EUS was, for example, 12 per 100 villages per year, but rather that the incidence rate of outbreaks of ‘disease causing red erosions’ was 12 per 100 villages per year.

Persistent questioning

One rule of collecting information through interviews is that you should never be satisfied with the first answer. When a question is asked, the answer could be wrong, either for the reasons listed previously or because the experience of the person answering differs from that of the rest of the village. It is a good idea, therefore, to check and recheck every answer. Do this by asking the same question of several different people in several different ways. Each time, focus the question on some different aspect of the problem, and compare the answers. If there is some inconsistency, start a discussion to try to resolve it and come up with a consensus.

Example

A list of all fishers is being built to act as a sampling frame. All those present have reported the number of boats owned. The group is asked 'Are there any fishers who aren't here at the meeting?'. The group responds with three more names. 'Is that all? Are there any more?' One person responds that there aren't any more. 'Does anybody have a neighbour with a boat who is not here?' One person realises that they do, and the neighbour's details are recorded. 'Are there any people who live outside the village and go fishing?', 'How about on the road leading to the north?', 'On the road leading west?'

This type of persistent questioning, and prompting to help the producers remember information, can be continued until the survey team is convinced that it has gathered the best information possible.

Encouraging cooperation

The cooperation and goodwill of producers is essential for a successful survey. For many disease surveys, the only people with the information required are the farmers or fishers, and the only way to collect specimens is from their animals. If producers are unwilling to cooperate, the most valuable source of information on the diseases of aquatic animals has been lost. Without this information, any attempts to control these diseases may be much more difficult. The key role of the producers means that every effort should be made to ensure that they are happy to participate in the interview, and are happy to allow the survey team to collect any necessary specimens from their animals. Surveys that extend over a long period and require heavy producer participation require even more effort to ensure participation. These have been discussed in Chapter 7.

If no aquatic animal disease survey has been conducted in the village before, most producers are likely to be fairly happy to cooperate, perhaps mainly out of curiosity. However, the aim of the survey team should be to make sure that the producers are happy to cooperate the *next* time a survey is conducted, too. Although there may be no plans to resurvey the village soon, any ongoing disease control program will need to be monitored with regular surveys. If, at each survey, the producers become unwilling to participate again, as time goes by it will be more and more difficult to find cooperative villages.

The problem in ensuring future cooperation is that aquatic animal disease surveys are mostly designed to *take* from the village, not to give. During a survey, the survey team collects information and specimens and then leaves to move on to the next village. The benefits to the survey team are very great; they have the information and specimens necessary to understand the disease situation, and help manage disease control programs that will benefit the entire country. The benefits to the village and the producers who gave the information and specimens are not so clear. When the survey team leaves, perhaps the villagers have merely wasted a few hours at an interview and had to catch some fish for specimens, stopping them from working and upsetting their ponds. Some producers can see no direct benefit to themselves and could be reluctant to participate in future. The challenge to the survey team is to provide some direct benefit to the producers so they will be happy to assist next time. The best way to achieve this will depend on the situation and culture, but here are a few suggestions.

Explaining the objective of the survey

One of the easiest ways to make producers more cooperative is based simply on good communication—a means that is often forgotten. At the start of the interview, the leader should carefully explain the purpose of the survey, the village's role in it, and the benefits the village will get from participating. This should include the following points:

- The survey is conducted by the government (or other organisation) to collect information that will help solve disease problems throughout the country. The results will therefore benefit all the producers in the country (or province etc.), not just this village, but the benefits may not be noticeable for a while.
- The narrower objective of the survey can also be explained (for example, the collection of information to allow the government to decide which diseases are most important, so they can allocate more funds to trying to solve these disease problems).
- The survey is not working to directly assist *this* village or the other survey villages, but *all* villages.
- The village has been chosen at random to be representative of all the villages in the area.
- The information collected will only be used to try to solve aquatic animal health problems, and won't be given to anybody other than the fisheries authorities (this is to allay the fear of some producers that information may be used for taxation or other purposes).
- Explain how long the interview is expected to last, and what is expected of producers after the interview (if specimens are to be collected).

If the producers' expectations of the survey are realistic (i.e. that it is not aiming to directly benefit their village), they are less likely to be disappointed. If they are made aware of the importance and potential benefits of the survey at the wider level, they may be happier to assist for the general good.

Attitude

The attitude of the survey team towards the producers will influence the way producers feel about helping with the survey. The survey team should realise that the farmers are experts on their own animals and their animals' health, and that the information the farmers hold is very important. If the producers realise that their opinions and experience are respected, and that the team is grateful for their help, they are likely to feel better about participating and to be proud of their contribution.

Payment

Despite good explanations and a demonstration of respect towards the producers, it is still plain that there is no direct material benefit to them from the survey. In some circumstances, it may be necessary to provide some sort of payment to producers, so that they gain some benefit from participating in the survey.

Payment may take the form of money, perhaps as an inducement to come to the interview, or as a per-specimen payment for samples. While they are sometimes necessary, cash payment should be avoided where possible, for two reasons. First, it makes the survey more expensive, and the authorities in developing countries can rarely afford such extra expenses. Second, it builds expectations amongst the producers. If participating farmers are paid during a survey, any future unpaid survey of the village is very unlikely to get cooperation.

While cash payments are generally not a good idea, payments in other forms may be much more acceptable. For instance, the distribution of simple equipment to producers as a way to thank them for their assistance will usually build goodwill without the problems of cash. Other types of payment in kind that are able to directly benefit the health of village animals can also be used (e.g. free treatments). If local private businesses supply treatments to the village, handing out (unsustainably) free treatments might undermine the (sustainable) private system, and should therefore be avoided.

Information

The fisheries services of most developing countries may not have a lot of money to pay producers, but they do have information on diseases, and this is one way to provide real benefits to the producers in payment for their participation. A village interview is a good opportunity for the survey team to collect a lot of information from the producers, but it is also a very good opportunity for the producers to collect information from the survey team. In many villages, it may be quite uncommon for skilled staff to visit. After a village interview where disease problems have been discussed, producers are likely to be thinking about all the problems they have had with their animals. The survey team can provide the advice and information that the farmers require.

Providing information and answering producers' questions is therefore an important way to give some benefit to the producers. This is best done towards the end of the village interview, when all relevant disease issues have been discussed and all necessary information collected. Invite producers to ask any question relating to the health or production of their animals, for discussion or advice. If necessary, suggest some topics arising from the interview.

This part of the interview may not gather any more information for the survey, but should nevertheless be seen as an important component. Spend as much time as required to address all the questions raised. Government extension staff used to making lecture-style presentations should be careful to listen to the questions and address them specifically, rather than embark on a long, dry, prepared presentation on a topic.

Fun

If participants in the survey enjoy the experience, for whatever reason, they are likely to feel better about helping with the survey. One way to ensure this is to provide some form of entertainment or recreation as part of the survey visit. There are many ways that this could be done, perhaps achieving multiple purposes.

At the simplest level, all participants could be invited to share drinks or a meal, provided by the survey team, at the end of the interview. While enjoyable for the producers, this is also an opportunity to discuss issues in a less formal setting, and get a better understanding of the disease problems in the village. If the survey team stays overnight, the gathering could be extended to a meal and a party. Some other form of entertainment that also serves an extension purpose could be provided. For instance, the team could present a show conveying some important health themes in an entertaining way. Alternatively, a video or film with the same objectives could be prepared and shown during or after the interview. For this to be successful, the video must be entertaining, and not a dry lecture on disease control.

Example

The national fisheries service commissions a television studio and the cast of a popular television show to produce a special episode dealing with aquatic animal health issues. The characters are well known throughout the country, and the plot, as with every episode, is very entertaining. This episode is shown by survey staff at the end of the village interview, both to entertain and to underline important messages.

9
Data management

Computerised data management

In small aquatic animal disease surveys using only simple analysis, all the calculations can be done by hand or with a simple calculator. However, in large surveys or those with more complex analysis, the amount of data that must be managed and the types of calculations required make it much too difficult to work by hand. Computers make managing and analysing large amounts of data much faster and easier, and allow some types of complex analysis that would otherwise be impossible.

Using computers has other advantages. Once data has been entered into a computer, many different types of analysis can be performed and reports generated without the need to re-enter the data. When performing analysis using a computer, there is no need to remember complex statistical formulas, as they are coded into the computer program. Computers also make it much easier to avoid and correct mistakes.

Computers are therefore important tools for the aquatic animal disease surveys described in this book. The accompanying disk contains all the specialised software needed to carry out the particular types of analysis required for these surveys. Only a basic familiarity with computers is needed to use these programs. However, for more general data management and analysis, other software is required. A wide range of suitable programs exists, and it is best to use whichever you are already familiar with. This chapter discusses some general aspects of data management and analysis, and makes particular reference to the **Epi Info** program. This program has several very clear advantages: it can perform all the data storage and management tasks required, it is able to carry out a wide range of standard statistical procedures and specialised epidemiological analyses, and it is free of charge.

Principles of data management and analysis

This section gives a very brief overview of the use of computers for data management and analysis. If you have experience in the use of computers and computer databases, you may choose to skip to the next section.

Data and information

The aim of data analysis is to convert a large amount of *data* collected during the survey into a small amount of meaningful *information*. In effect, analysis tells us what the survey data actually means. To achieve this, the data is converted into a few easy-to-understand measures.

A computer is a tool to help us convert data into information, and its main advantage is in its ability to process a very large amount of information very quickly. When we use a computer, we are working with three separate components: the *hardware*, *software* and *data*.

Hardware

Hardware describes the computer itself. The computer is made up of several components with different roles.

First, there must a way of putting data into the computer, or *data input*. The keyboard is the main way this is done, through typing in the data. There must also be a way to store the data once it has been entered. Storage, or computer memory, comes in two forms. Disks (such as the hard disk inside the computer, floppy disks or CDs) can store data for a long time, and require no power when they are not in use. In contrast, the memory inside the computer (*RAM*, or random access memory) can store information only while the computer is turned on, and is used for temporary storage and doing calculations.

The most important part of the computer is the part that does the calculations. The *central processing unit* (CPU) is the 'brain' of the computer, and is a small silicon chip capable of performing a very large number of calculations every second.

Once the data has been entered, stored and analysed, we need to know what the results are. The parts of the computer for sending information out are called the *output devices*, and include the screen or monitor, and the printer.

Software

Software is the general name for computer programs. A program is a set of instructions that the computer is able to read and that tells it how to process data. For instance, the computer hardware doesn't know anything about statistics. To calculate statistics, we must first load a statistical program, which has a set of instructions telling the computer how to calculate the statistics. To calculate the average (mean) of a list of numbers, the program tells the computer to add the numbers and then divide the sum by the number of items in the list.

Types of software

There is an enormous range of software available, designed for many different purposes. Commonly used software performs several main tasks:

- *Word processors* are designed for working with words: writing letters, reports and other documents. They allow the computer to be used like a very sophisticated typewriter.
- *Spreadsheets* are programs for carrying out mathematical calculations, and are widely used for business purposes.
- *Databases* are programs for managing large amounts of similar data. Databases are the main programs to use for storing and managing the results of the aquatic animal disease surveys described in this book.
- *Statistical programs* use information that has been stored in a database, and perform a range of statistical calculations on the data.

Some programs can perform more than one of these functions. For instance, Epi Info can be used as a word processor, database and statistical program. Others have very specific functions, like the programs provided with this book; each of them performs just one specialised calculation or statistical analysis.

Epi Info

This chapter contains advice that is applicable to any computer program. However, examples and specific instructions are provided for carrying out procedures using Epi Info, as this should be available to all readers and has the ability to perform all the tasks required. Where instructions are given on using Epi Info, the paragraph is marked with the symbol shown on the left. Only very brief explanations of the Epi Info commands are given. For more details, use the online Epi Info manual.

Data

The third thing that is necessary when using a computer is the data. This is what the computer works with. In aquatic animal disease surveys, the data is all the facts and figures that are collected during the field work, or else produced by the laboratory when analysing specimens.

Data types

There are many different types of data, but most of these can be represented in the computer in just a few different ways.

Text

- Text. In computer language, words are called text or strings. A written description of something is data stored as text. For example, village names, the names of farmers, and the names of diseases that are important in a particular area are all text.

Yes / No

- Answers to questions can often only be yes or no. Many questions in aquatic animal surveys can be thought of as having yes/no answers, for instance 'Has there ever been an outbreak of EUS in this village?' Other types of information with only two values can also be thought of as yes/no data. When analysing smears for parasites, the results can be reported as Yes (parasites present) or No (no parasites detected).

Numbers

- Numbers are used to specify many types of data. Data stored as numbers can be divided into two groups:

Continuous data

Quantitative information

- Continuous data uses numbers to measure the value or quantity of something (*quantitative information*). Continuous data may take any value within a range. Examples of continuous data include age, weight, temperature and population.

Integer numbers and real numbers

Continuous data may be represented as *integers* (whole numbers, such as the number of fry in a tank, or the number of ponds on a farm), or as *real* numbers (fractional or decimal numbers, such as age, weight or temperature).

Categorical data

Qualitative information

- Categorical data uses numbers to describe what something is like (*qualitative* information). Categorical numbers do not measure the amount of something, but are used to classify into different categories. Categorical data can be divided into three types:

Nominal data

- Nominal data (or 'named' data). Numbers are used to represent different categories, usually identified by their names. Numbers are used as codes, and the different categories have no natural order. For instance, species is often represented by a code, so that 1 = catfish, 2 = tilapia, 3 = barb, 4 = shrimp.

Ordinal data

- Ordinal data (or 'ordered' data). Numbers represent different categories, but there is some natural order, so that 2 is greater than 1. For example, villages may be divided into small, medium and large, based on their fisher population. Ordinal codes could be used to classify the villages so that 1 = small, 2 = medium, and 3 = large.

Dichotomous data

- Dichotomous data. Only two values are possible. This is the same as Yes/No data, but represented by two numbers (e.g. 0 = No and 1 = Yes).

Dates

- Dates. Dates are a special type of number data, used for answers from questions like 'When did you first notice the tank was affected?'

Data storage

The way data is stored in a computer is very similar to the way data is stored on paper. Let us first consider a paper storage system for survey data.

Example

A survey is conducted to estimate prevalence—cages that have experienced problems with parasitic diseases in the past year. The survey collects some data from a number of randomly selected farmers, and collects specimens from randomly selected animals in each cage. The questions asked about each farm and the responses are stored in a file, with one sheet per farm. The data from the animals selected for sample collection is stored in a separate file, with one sheet per farm and one line per animal. These two files contain all the information collected during the survey.

In this example, the file for the farm data contains one sheet of paper per farm. The data from the same question for each farm is recorded in the same place on each form. All the data on each form relates to one farm only. In the animal data file, each line on the sheet contains the data from one animal. Each line extends across a number of columns, and each column contains the information about one aspect of the animal. A column, therefore, presents the same measure for all the sampled animals on one farm.

In a computer database, information is stored as a *table* in very much the same way. A table is stored as a file in the computer's memory, and data from different things is kept in different files (for instance one file for the animal data, and one for the farm data). Each table is made up of columns and rows. Each row holds information about one item (for instance one farm in the farm file, or one animal in the animal file). Each column holds only one type of data, which is the answer to a particular question about each farm or animal. For instance, in the farm file the first column might store the name of the farm, while the second records the number of animals, and so on.

In computer language, a row is called a *record*, and a column is called a *field*. When you create a new table for storing data from a survey, you need to say which fields (columns) you want to have, or in other words, what information you are going to store in the table. For each field, you also need to say what type of data will be stored there: text, numbers, yes/no, or dates. When you start putting data into the new tables, you will add a new record (row) for each farm or animal that you enter.

A collection of one or more tables with related information is called a *database*.

Data processing procedures

Once data collection has been completed, there are a number of steps before the data can be analysed. These are considered in more detail below.

Processing data for analysis

Step 1: Initial check for completeness and accuracy of data

Step 2: Data coding

Step 3: Creation of computer database

Step 4: Data entry

Step 5: Checking for errors and inconsistencies during data entry

Step 6: Recoding

Step 7: Converting data between different formats

Step 8: Analysis

Initial check for completeness and accuracy of data

Before any work is done on the computer, the data record sheets need to be carefully checked for any missing data or mistakes. Preferably, this should be done while the survey team is still in the field, so that the problem can be corrected. A quick examination of the sheets should show up any gaps where data has been missed. Finding other types of mistakes may require a closer look. For instance, a farm's size may have been recorded as 12 m^2 instead of 12 hectares, or a farmer stating that they have never had a disease outbreak has a record of the date of the last outbreak. If these types of problems are picked up early, the question can be asked again, or the person filling out the form may remember the true response and be able to correct it. If this is no longer possible, the answer may have to be left blank, and treated as missing data (see below).

Data coding

Coding data is the process of converting complex data into a simpler form that is easier to manipulate. Computers are designed to work with numbers, so using numbers as codes makes the work easier. Using codes also makes data entry faster and more accurate, and avoids inconsistencies.

Example

A survey was carried out of 40 farms and 20 animals in each farm, with a total sample size of 800 animals. Data about each animal is being entered into the computer. Each farm is identified by the farmer's name, and each animal must be identified with the farm that it came from. If the farmer's name is used, then the whole name (which is sometimes very long) has to be typed for each of the 800 animals. It is very easy to make a mistake when doing this much typing. If a mistake is made and the name of one farm is spelt in two different ways, the computer no longer thinks that they are the same farm, but two different farms. This will cause problems for the analysis, as the total number of farms is now 41 instead of 40. A better way to enter the data would be to use a numeric (nominal) code for the farm. The first farm is given the code of 1, the second is 2 and so on up to farm 40. It is much easier to type a short number than a long name, and there mistakes in data entry are much less likely.

Data dictionary

Codes can be used for any type of data that has a number of different categories (categorical data). Farm, village, district or province names are some examples, but disease, species and season may also be converted to numeric codes. Before coding data, a *data dictionary* has to be set up. This is a list of all the possible values on the data recording sheets, and the code to be used for those values. Sometimes this list will be easy to set up before the survey. For instance, when coding season in a tropical area you might decide to use 1 = Wet, 2 = Cool, 3 = Hot.

In other cases, you might not know all the different responses until after the data has been collected. Perhaps you have to code the responses to a question asking for the most important diseases of crabs in the area. In this case, after the data collection is finished, the data recording sheets should be checked for all the responses that were made, and a separate code assigned to each different disease that was mentioned.

Missing data

Missing data is a problem that should be considered at this stage. When setting up codes, it is a good idea to have one code for missing data as well. Missing numbers need to be treated very carefully to avoid mistakes. When recording information on data recording sheets, either leave missing data blank, or use a dash (—). Some programs do not allow you to leave a field blank when entering data, and will insert a zero instead. This is a problem because there is a big difference between knowing that a farmer had no outbreaks, and not knowing how many outbreaks they have had. Depending on the program that is being used, it may be necessary to use a special code for missing data in number fields. For instance, a field storing information about the number of ducks might have the total number, or, if the data is missing, –99. During analysis, any farmers with a code of –99 for the number of ducks can be excluded.

Epi Info

In Epi Info, if no data has been entered in a field, the field is automatically considered to be storing a missing value, and it is ignored during analysis. However, data imported from a different program may be treated differently. Missing data that has been given a special code (such as –99) can be excluded from the analysis using the Select command described on page 159 (for example, select outbreaks <> –99).

Creation of a table

Before we enter data into the computer, we must use a database program to set up a table to hold the data. A separate table must be created for each different grouping of information. For instance, if we have collected both farm and pond data, we need a separate table for each. This is because each record (row) in a table stores data about one kind of thing only, either ponds or farms, but not a mixture of both.

Defining fields

Within each table, we create a separate field for each different piece of data and specify the data type. For a pond table in a disease survey, the fields might look like this:

Pond ID:	Number (integer)
Farm ID:	Number (integer)
Date of visit:	Date
Area:	Number (real)
Outbreak?	Yes/No

Field width

For some fields (such as text) it is also necessary to say how wide the field is (how much text will be stored in it). For example, only one letter (M or F) would need to be stored in a sex field, so the width is one. For a field storing a farmer's name, we may need room for 20 or 30 letters.

Data entry form

When the basic table has been set up, most database programs allow us to set up a *data entry form* as well. This is a dialogue box, or 'screen', on the computer with places to type each piece of data for a single record. Instead of showing all the data at once as a table does, it shows only one record. When creating a data entry form, we should present the fields in the same order and with the same appearance as the paper data recording sheet. This makes it much easier to find the data during data entry, when we have to transfer the information from paper to computer.

Data checks

One of the advantages of using a data entry form is that it is possible to set up *data checks* that check the data during entry. Data checks can be of three types:

Range checks

- *Range checks* make sure that the value entered falls within a specified range (for example, that the days of culture in a shrimp pond is more than zero days, but less than 300 days). If we type a number outside this range, the computer displays a warning.

Allowed entries

- Similar to a range check is a list of *allowed entries*. For instance, when entering the sex of a fish in a text field, it must be either M or F. Setting M and F as the only two allowed entries will prevent mistakes.

Consistency checks

- *Consistency checks* compare two or more pieces of information to check for inconsistencies. For example, we might wish to record the number of successful harvests achieved on the farm. When the number is more than zero, the computer can check to make sure that the farm has been operating for an appropriate number of years (assuming, say, 2 crops per year). If the number of successful crops is more than the number of crops possible, the computer will display a warning and allow the user to correct the mistake.

Epi Info

In Epi Info, the processes of creating a new table and creating a data entry form are done at the same time. Use the following procedure:

Creating a new table

Step 1: Start Epi Info, and open Eped, the word processor. For detailed directions on how to use Eped, use the online manual.

Step 2: Design a data entry screen, which includes any text needed to describe fields or make data entry easier.

Step 3: Where data is to be entered into a field, insert field codes in the data entry screen. The main codes are:

_	(underline): text field (the number of underlines determines the size of the field)
#	(hash): number field (the number of hashes determines the size of the field). Insert a decimal point if required, e.g. ###.##
<Y>	yes/no field
<dd/mm/yy>	date field (day/month/year format)
<mm/dd/yy>	date field (month/day/year format)

Example

The questionnaire file for the survey shown above might look something like this:

Questionnaire file

```
Disease Survey Data Entry Screen

     Pond ID:          #####
     Farm ID:          #####
     Date of Visit:    <dd/mm/yy>
     Area:             ##.#
     Disease?          <Y>
```

Step 4: Save the file as a questionnaire file (.qes extension).

Step 5: Exit Eped, and select Enter from the Programs menu.

Step 6: When asked, type the name of the new data file that you want to create (usually the same name as the questionnaire file, but with a .rec extension).

Step 7: Select option 2, 'Create new data file from a .QES file'

Step 8: Type the name of the questionnaire file you just created in Eped.

Step 9: Enter 'Y' to indicate that everything is OK.

Epi Info will then create a new file according to the field definitions in the questionnaire file. Once the file is created you are ready to start entering data. However, you can also choose to set up some data entry rules and consistency checks within the file. Exit the Enter program, and use the following steps to set up checks:

Setting up data checks

Step 1: Select Check from the Program menu to start the Check program.

Step 2: Type the name of the newly created data file, or simply press Enter and select the name from the list.

Step 3: The data entry screen is displayed, ready to set up data entry checks.

Step 4: Specify different checks for all fields (see below).

Step 5: Press F10 to finish. Press 'Y' to save the changes to disk.

There is a range of different checks that can be performed on each field. Some of the important ones are listed below, but check the online manual for further details.

- The F1 and F2 keys will set the minimum and maximum values that may be entered (range checks). For many types of data, the minimum is 0 (no negative numbers allowed).
- Pressing the F3 key causes the data from the previous record to be repeated in the following record. For fields that have the same information for many records in a row, this can save a lot of time. If the data is different, you can simply type the new data instead.
- The F4 key sets the rule that data must be entered in the field before the record can be saved. This prevents errors in critical data that must always be present (e.g. ID codes).
- The F6 key can be used to set up a list of possible values (such as M and F for the sex field). No other entry will be accepted.

Data entry

Entering data into the computer is a time-consuming and boring task, and requires patience, accuracy and good skills on the keyboard. An experienced person can enter data quickly and very accurately, but it is always possible to make mistakes. Any mistakes made during data entry can lead to incorrect conclusions from the survey. There are various ways to try to avoid making mistakes.

Avoiding data entry errors

When entering a large amount of data, it is easy to lose concentration. Take a break every half hour or hour, and do something different. Alternatively, have two people work on the data entry and take turns.

Make sure the data recording sheets have been properly checked before data entry. Mistakes or numbers written unclearly make data entry more difficult. Correct these problems first.

Using a database program with built-in range checking, allowed-value checking and consistency checks will pick up mistakes as they are made, making them much easier to correct.

Double entry system

The best way to avoid mistakes during data entry is called the *double entry system*. Once all the data has been entered once, they are all entered a second time. During this second data entry, a special program compares the data being typed with the data that has already been stored in the computer's memory, and displays a warning if there are any differences. While it is easy to press the wrong key occasionally, it is very unlikely that the same wrong key would be pressed for the same figure during two data entry sessions. Ideally, the second entry should be done by a different person. An alternative double entry system uses two files that have been entered separately, and compares them for differences. Double entry is very good for avoiding mistakes, but requires twice as much work and takes twice as long.

Epi Info

Entering data

In Epi Info, the **Enter** program is used for data entry. Select Enter from the Programs menu, and type the name of the data file (or press F9 to select from a list of files). Select 1 (Enter or edit data), and type 'Y' for OK. A new data entry screen is displayed ready for entering data. A few tips might make the task easier:

- The number of the current record is shown in the bottom right corner of the screen. This helps you keep track of where you are up to.
- To edit data, use the F7 key to step back one record, or the F8 key to step forward one.

- To search for specific records, use Ctrl-F (hold down the Control key, and press F at the same time). You will then have to specify the information you want to search for (for instance a village ID number) and press F2 to do the search.
- Where a list of possible values has been set, you can press F9 to show the list, and use the arrow keys to select the value you want.

Double entry with the Epi Info Check program

There are two ways to use Epi Info for double-entry system data checking. The first is to use the Enter program, but select option 4 (Re-enter and verify records in the existing data file), instead of option 1. After entering the file name, you can enter data normally. However, if the data entered differs from that already in the file, a warning will appear.

Double entry with the Epi Info Validate program

The other way to check data is to use the Epi Info **Validate** program. Enter the data twice (into each of two separate data files). Then run the Validate program to compare the data in the two files for any differences.

No matter how much care has been taken during data entry, there is still the chance of some mistakes. All data should be checked after data entry to pick up mistakes, as described below.

Saving and backing up

Saving data to disk

While data is being entered, the computer usually stores the information in its RAM, the memory that only works when the computer is turned on. Saving the data means having the computer write the data to a file stored on a disk. Once the data is saved, the computer can be turned off without any risk of losing data. If the data has not been saved first, and the computer is turned off, all the data will disappear and the work of data entry has to begin again. This can happen when a power failure occurs, or when there is a problem with the computer.

Uninterruptable power supply

Where the power supply is unreliable, these problems can be avoided by using an *uninterruptable power supply* (UPS), which has a battery to take over when the power fails. Even if a UPS is connected to the computer, you should be careful to save the data onto a disk every 5 or 10 minutes. This will ensure that very little is lost if anything goes wrong.

Backing up data

Data that has been written to a disk is much safer than data stored in temporary RAM, but there can still be problems. Occasionally, a disk can develop errors, so that the data cannot be read from it. Although this problem is rare, it can be very serious if there is only one copy of the data. Using regular *backups* will overcome this problem if it occurs. A backup is a second copy of the data, stored on another disk. Usually the main copy of the data is stored on the hard disk inside the computer. Hard disks are very fast, and very reliable; however, they do occasionally develop problems. Backing up the data to an external (floppy) disk, and keeping the disk in a safe place means that the data is not lost if there is a problem with the hard disk. You should back up your data during data entry at least once each day. After data entry is finished, the data should be backed up whenever any changes are made.

Epi Info

In Epi Info, the data is saved to disk every time you change a record. The program asks 'Write data to disk (Y/N/<Esc>)?' every time. Answer Y to save the data.

To make a backup of the data onto a floppy disk, use your operating system to copy the file. In DOS, use the copy command (e.g. copy results.rec a:). When using Microsoft Windows™, use the File Manager or Windows Explorer to copy the files.

Checking for errors after data entry

Once all the data has been entered it must be checked again. If this step is neglected, a lot of time can be spent in analysing the data and producing reports. Often, during analysis, an unusual result resulting from an error in the data will be noticed. When this error is fixed, all the analysis has to start again because the data has changed. Rather than waste this time, it is much better to find the problems before analysis starts.

In contrast to checking for problems before data entry, checking after data entry is faster and easier, because we can use the computer to help us. The computer can search through all the data to find unusual values very quickly. Every field (column) should be checked separately, and then some fields may be checked together to assess consistency. Here are some useful techniques for using the computer to help check the data.

Epi Info

In Epi Info, the data can be checked using the Epi Info **Analysis** program. This is the same program that is used for general data analysis, and the same techniques described here are equally useful for that task. To start the Analysis program, start Epi Info, choose the Programs menu, and select Analysis. The Analysis screen is divided into two parts—the bottom part is for you to write commands to tell the computer what sort of analysis you want to do, and the top part reports the results of the analysis.

A few tips will help when using Analysis:

- To get a list of the different commands that can be used, press the F2 key. You can then highlight and select the command that you want.
- To find out how to use a command, type the command and then press F1. This will bring up a help screen explaining the command and its usage.
- Pressing F3 will bring up a list of the data fields in your table. You need to specify which field you want to work with in most of the commands. Highlighting the field in the list and pressing enter is a fast way to insert the field name in the command. Sometimes you need to enter more than one field name. You can select many fields by pressing the + key, and then insert them in the command by pressing Enter.
- Use the F4 key to examine the data as a table. Pressing F4 again will show one record as a data entry form. Press Esc to return to Analysis.
- As with all Epi Info programs, press the F10 key to close the program and return to the menu.

Before you can start working with a data file, you need to tell Epi Info which file you want. Use the Read command and type Enter. You can search through the list for the file you want. Only Epi Info (.rec) files are shown. To use a file in dBASE format, type 'read *.dbf'.

Counting records

The first thing to check is whether all the data has been entered, or whether some data has been entered twice. You can use the program to tell you the total number of records in the table; this should be the same as your sample size.

Epi Info

When you load a file in Epi Info, the total number of records is displayed at the top of the screen. Check that this matches your sample size.

Frequency tables

Breaking the data into different groups to see the total in each group, and the number of different groups, is a very good way to check for mistakes. This is done by creating *frequency tables* (sometimes called one-way tabulations).

Example

A survey of villages was carried out in a single province. Some villages from each of the province's 6 districts were included. A code has been used to identify which district each village is in. By producing a frequency table of the district codes, it is possible to check that each of the 6 different district codes has been entered properly, and the correct number of villages are in each district. The frequency table might look like this:

District ID	Count
1	4
2	6
3	2
4	5
5	8
6	5
13	1
Total	**31**

In this example, seven different district codes are reported, with code 13 standing out as different from the rest. This indicates that a mistake has been made with the district code for one village, where 13 has been entered, probably instead of 03. If the total number of villages in the survey was 30, instead of 31, then one village has been entered twice. The total number of villages in each district could then be checked to find out which district has too many villages.

To produce frequency tables in Epi Info, use the **Freq** command and specify which field you want the program to use. The above table was produced by typing 'freq districtid', and pressing Enter.

Epi Info

Frequency tables are useful for identifying missing data as well. A frequency table of the animals with lesions from a survey might look like this:

Lesions	Count
Y	124
N	32
Missing	4
Total	**160**

Frequency tables like this can be used for any text or categorical number field in which there is a limited number of possible alternatives (e.g. district codes, species, disease, sex etc.). If frequency tables are used for other numbers (such as population) a long unhelpful list of all the different population values is produced. For continuous values (integer or real numbers) there are two useful ways to check for errors.

Maximum and minimum

The first is to generate descriptive statistics for the field. In particular, the *maximum* and *minimum* values are useful.

Example

In a prevalence survey of shrimp, the age (in days) and estimated weight (in grams) of each animal were recorded. Summary statistics were produced for these two fields, with the following results: Age: Minimum 2 days, Maximum 1200 days; Weight: Minimum 3 g, Maximum 4000 g.

In both cases, the maximum is much larger than would be expected. These values probably result from accidentally pressing the 0 key twice instead of once, and really should have been 120 days, and 40 g. Check the original data recording sheets and fix the mistake.

Epi Info

The **means** command can be used to produce a range of summary statistics. Type 'means' and the field you want to use. The command usually produces a frequency table as well, which is not very useful. To just produce the results, type 'means fieldname /N'. The /N means 'no table'. This is an example of the results displayed when examining a field containing village pond numbers:

```
====>
means ponds/n
PONDS
   Total          Sum          Mean        Variance       Std dev       Std err
    410          41510        101.244     11268.943       106.155        5.243
  Minimum        25%ile        Median       75%ile        Maximum         Mode
   0.000         24.000        66.500      138.000        621.000        0.000
Student's 't' test, testing whether mean differs from zero
T statistic = 19.312, df = 409, p-value = 0.00000
```

Some of the information produced is not relevant, but Total (total number of records), Sum (sum of all the ponds in all villages), Mean (average number of ponds), Minimum and Maximum provide valuable information.

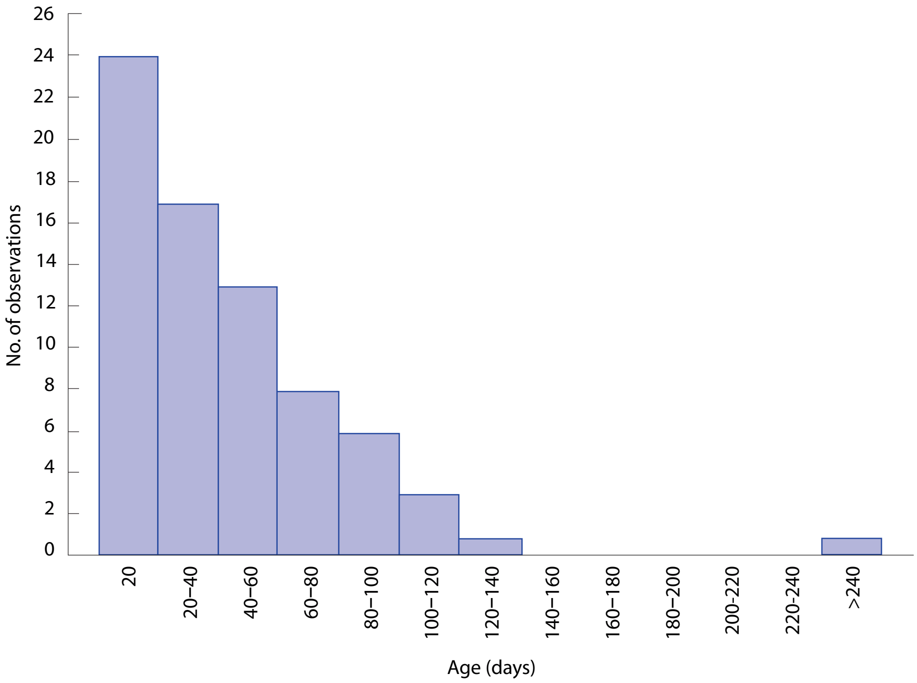
The other way to examine this kind of data is to draw a histogram showing the distribution of values. A histogram is a graph which indicates the number of records that have a particular value, or a value in a particular range. Any unusual values will be on the far right or left of the graph, and can be easily seen.

Example

After correcting the problem with the age data identified in the example of the prevalence survey of shrimp given above, a histogram was drawn, as shown on the next page. The histogram shows the number of shrimp that fall into different age categories, and shows that mostly young shrimp were included in the sample, with a few older shrimp. There is one shrimp whose age is more than 240 days. This is probably a mistake, and the data should be checked.

Histograms can be used for both categorical data (e.g. codes, sex, species) and continuous data (age, population).

To show a histogram in Epi Info, use the **Histogram** command, specifying the field to use. The accompanying graph was produced by typing 'histogram agegroup', where agegroup identified which age group the animal belonged to.[1]



Epi Info

Selecting subgroups

In addition to examining one field at a time, it is possible to check two (or more) fields at the same time, for consistency. One way of doing this is to select a group of records from the data that match a certain definition.

Example

A survey of farms has recorded species, and the diseases observed. A group of records was selected that identified farms with shrimp, but listed the disease as EUS. The definition for the group was: (Species = Shrimp) and (Disease = EUS). The number of records in the group was then counted to reveal that 3 records matched the definition. These three records were checked with the original to correct the mistake.

Epi Info

To select a group of records in Epi Info, use the **Select** command, specifying which records you want to use. After you use the select command, any analysis you perform will only work with the selected subgroup. To turn off the selection and work with all the records, use the Select command without specifying any records. Using several different Select commands selects only those records that match each of the commands. For example:

```
select species = shrimp      only shrimp are selected
select disease = EUS         shrimp with EUS are selected
means freq disease           frequency table of diseases (showing only shrimp
                             with EUS, including the number of records).
select                       turn off the selection - all records are active
```

[1] In Epi Info, only categorical data can be used to draw a histogram. If your file contains data on the age of shrimp, and you want to draw a histogram, you must first convert this into categorical data by grouping all the shrimp with a similar age together into groups. To do this, create a new numeric field called AGEGROUP, using the command define agegroup ##. You then need to fill the new field with a code that defines the age group of the shrimp. In the histogram, shrimp have been categorised into 2-week groups. Use the command: agegroup = 14 * round(age / 14) to create the categorical age group codes, grouped into 2-week brackets. To group into 4-week brackets, replace both 14s with 28.

Cross tabulations

Two-way tables, or cross tabulations, can be used to check two categorical variables as well. Using the same example to compare species and diseases, a two-way table might look like this:

Species	Disease	
	EUS	WSS
Shrimp	3	64
Carp	232	0

This table shows the number of records that match two different criteria. For instance, there are 64 records of shrimp with white spot but three records of shrimp with EUS. The three records with errors can easily be seen.

Epi Info

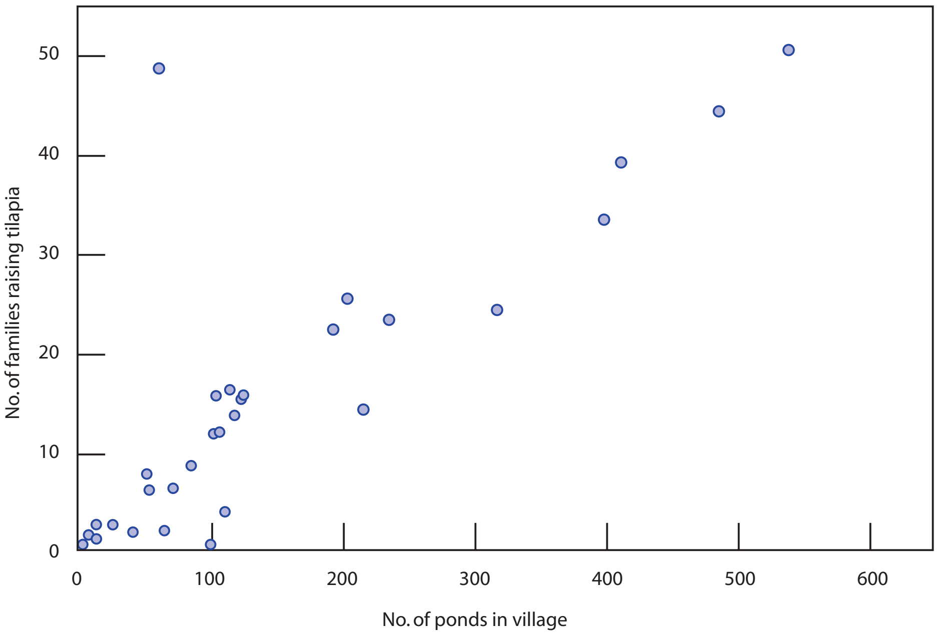
Use the Tables command to produce two-way tables, specifying the two data fields to include. The above table was created with the command 'tables species disease'. Note that the first field you specify is shown in the rows of the table, and the second in the columns.

Scatter plot

When two different types of continuous data are to be compared graphically, you can use a *scatter plot*. A scatter plot draws a point showing the value of one of the variables on the x-axis, and the other on the y-axis.

Example

Village data has been collected on the number of families raising tilapia and the total number of ponds in the villages. These two pieces of data are displayed on a scatter plot as shown below. Most of the points lie in a line close to the centre of the graph, indicating that villages with a small number of tilapia-raising families usually have a small total number of ponds, while villages with many families have many ponds. There is one point at the top left of the graph that is well away from the rest. This point indicates a village that seems to have many families that raise tilapia, but only a small total number of ponds. While this is possible (families may share ponds), it is clearly unusual, and may be a data entry mistake. The original data should be checked.



To draw a scatter graph in Epi Info, use the Scatter command, specifying which two fields you want to use. The graph above was produced with the command 'scatter ponds families'. The first field is shown on the x-axis.

Recoding

Data is not always in the best form for analysis after it has been entered. Recoding is the process of changing data from one form to another, to make it easier to analyse. Recoding after data entry is made much easier, as the computer can do all the calculations for us.

Example

In an incidence rate survey, the date of the last outbreak of yellowhead disease amongst shrimp ponds was recorded, along with the date of the visit to the farm. For analysis, we need to know how long ago the last outbreak was, not the date. This is calculated by subtracting the date of the outbreak from the date of the visit.

Recoding data is done by telling the computer how you want to change the data. It usually involves creating a new field. The computer calculates the values for this new field based on other data in the table.

In Epi Info, you must first create a temporary field to store the new coded value. Use the define command, specifying the name for the field and what type of data will be stored.

Epi Info

For instance 'define farmname__________' creates a new field with the name 'farmname'. This is followed by 10 underlines (_) which indicate that it is a text field, with room for 10 letters. Number fields are represented by hash marks (#) so the command 'define weight ###.##' would create a new field called weight, for storing numbers, with room for 3 digits and 2 decimal places. A date field is defined with 'define visitdate <dd/mm/yy>'.

Once the new data field has been created, you can tell Epi Info what data you want to be put in that field. Some examples of commands include:

```
define time ####            Create a new field
time = visitdate - eventdate  The time between two dates (in days) is
                            calculated and stored in 'time'.
define agemonths ###        Create a new field
agedays = agemonths * 30    The age, expressed in months, is converted to
                            the age in days.
define totalpop ####        Create a new field
totalpop = ponds + cages    The population of both ponds and cages is
                            calculated and stored in the 'totalpop' field.
```

Linking data

Sometimes all the data is not in one single table, but in two or more tables. Before analysis, the tables have to be linked.

Example

A seroprevalence survey was carried out in which 250 gill specimens were collected from grouper. During specimen collection, data on the sex, age and disease history of each animal was collected. At the end of the survey, this data was entered into a computer database. The gill specimens were sent to the laboratory for analysis. The laboratory uses a computerised recording system, and was able to send all the results from the gill tests on a computer disk. This means that a lot of time was saved by not having to retype the gill results, and typing mistakes were also avoided. However, the laboratory tested the gill specimens in a different order to that used on the data recording sheets. The two tables need to be linked.

Key field

To link two tables, there must be one piece of data, called the *key field*, that is the same in both tables. The computer uses this field to know which piece of data from one table belongs with which piece in the other table. The data recording sheet identified each specimen by a specimen number. The laboratory also identified each of the test results by the specimen number. The specimen number can therefore be used as the key field to link the two tables.

The actual process of linking depends on the database program being used. In Epi Info, link two tables using the **Relate** command, and specifying the name of the key field, and the name of the file to join. To join to files:

Step 1: Open the first file in Analysis, using the Read command.

Step 2: Make sure there is a key field with the same field name in both the open file and the file you want to link.

Epi Info

Step 3: Use the Relate command to link the second file.

Example

Using the above example, there are two data files, one called 'grouper.rec' and one called 'results.rec'. Both files have a field called 'specID'. After first opening the grouper file, the two files are linked using the command 'relate specID results'.

Making a link permanent

Once tables are linked, the data can be analysed. However, the link is only temporary, and when analysis is finished, there are still two separate tables. To make the link permanent, save the linked table to a new table. In Epi Info this is done in two steps: first define a new data file name (the Route command), and then write the information to the file with 'write recfile'.

For example, to save the two linked files to a new single file called alldata.rec, use the command 'route alldata.rec', and then the command 'write recfile'.

Converting data between different formats

If you use Epi Info for data management, you can also use it for data entry, checking, coding, linking and most of the analysis. However, there might be times when you need to use a different program for analysis. For example, there may be a special type of analysis that is required for particular survey types, such as those described in this book. The programs to carry out this analysis are included on the disk, but are not part of the Epi Info program. Alternatively, you may prefer to use a different database program to manage your data, and then need to use a specialised statistical program to analyse it. Sometimes data is provided by somebody else and may have been prepared using a different program. In all these cases, the data must be converted into a form that can be used by the different program.

Format

The *format* of a data file is the way it is stored on the disk. Different programs use different formats, as they store data in different ways. To use data in a different analysis program, you need to change the data format one that the analysis program can use. Commonly used formats include MS Access, dBASE, and ASCII.

Exporting and importing data

There are many different data formats, but fortunately there are a few standard formats that are used to move data between different programs. One example is dBASE files (with a .dbf extension). Many programs are able to both read and write data in this format. Writing data to a new format is called *exporting*, and is usually done through the Export or Save As menu of a program. Reading data from a different format is called *importing* and can be done through the Import or Open menu option of most programs.

Epi Info has two special programs, both of which can be accessed from the Programs menu, for converting data between formats. The import program reads data from four formats, including dBASE, and saves it as an Epi Info format (.rec) file. The export program is able to convert Epi Info data to 16 other formats (including dBASE). Epi Info is able to use dBASE files for analysis without conversion.

All the Survey Toolbox programs can read and save data in either dBASE or Paradox formats.

Analysis

Finally, when all the data processing is complete, the data is ready for analysis. Analysis of simple data can be done using the same tools and methods described above for checking data: frequency tables, counts, two-way tables (cross tabulations), simple descriptive statistics (maximum, minimum, average) and graphs (histograms, scatter graphs).

More complex data analysis is necessary for complex survey designs. The methods and programs for the analysis of data from the surveys described in this book are explained in Chapters 11–14.

Manual data management

Manual data management refers to the recording, storage and analysis of data using paper systems rather than computers. Most of this chapter has discussed how to use computers for this task. For large surveys and complex analysis, computers are almost essential, but for smaller surveys and simple analysis, manual data management and analysis may suffice.

Many of the principles used in computerised data management are the same when we use paper systems. The hardware consists of your brain and maybe a pocket calculator to do the calculations (instead of a CPU), and paper forms and a folder to store the data (instead of RAM and a hard disk). However, the types of data handled are just the same.

The data processing procedures in a manual system are, in principle, similar to those used by computers, namely:

Step 1: Initial check for completeness and accuracy of data

Step 2: Data coding

Step 3: Creation of calculation tables

Step 4: Transferring data to calculation tables

Step 5: Checking for errors and inconsistencies during data entry

Step 6: Recoding

Step 7: Converting data between different formats

Step 8: Analysis

The main differences are in the organisation of the paperwork, the use of calculation tables, and the approach to analysis.

Organisation of paper

In any survey, it is important to organise the data collection forms and other papers carefully. Each data collection form should be carefully protected during data gathering, and then stored in a folder where it can't be lost. If a data sheet is damaged, soiled or lost, all the work that went into collecting that data has been wasted. Data sheets should always be treated as if they are very valuable.

Having a separate folder for each type of sheet, and placing the sheets in an ordered sequence in the folders, helps ensure that none is lost.

Calculation tables

Depending on the design of the data collection forms, and the complexity of the calculations, some calculations may be able to be performed by reading data directly off the form. In some cases, it may be better to transfer the data to separate calculation tables (similar in concept to a structured database), to make the calculations easier. For example, perhaps you have a 7-page questionnaire, and you want to find the average of one figure on page 4, kilograms of fish produced. This means adding up all the values and dividing by the total. If there are 100 questionnaires, then you need to search for page 4 in each questionnaire, and add the number. Searching through the papers again and again can be very time consuming. An alternative approach is to design a table that contains a column for each different data item that you want to analyse, and a row for each record. Working your way through all the questionnaires in turn will allow you to copy the required data into the calculation table. Once you have finished, you can put the questionnaires away, and just work on the table. This will save sorting through all the questionnaires again and again for each new analysis.

The other advantage of calculation tables is that they give you somewhere to keep intermediate results of your calculations. For instance, if you want to calculate the average time since the last disease outbreak, but all you have is the date of the outbreak and the date of the visit, then the time for each farm has to be calculated first. On the calculation table, you could have three columns, one for the date of the outbreak, one for the date of the visit, and one for the time since the last outbreak. First, copy the dates from the data sheet, then, using the calculation sheet, work out the time between the dates for each farm. Finally, you can add up the times to work out the average.

A similar way to keep intermediate results is in the form of summaries at the bottom of each calculation sheet. Adding up 100 numbers can be difficult. If a mistake is made at any point, then you have to start at the beginning again. However, if each page of the calculation table contains, say, only 20 rows, adding up the 20 numbers is much easier. It is also faster to do it again if you make a mistake. At the bottom of each sheet you can have some rows for summary figures, including counts, totals, number of missing values and so on. After calculating the figures for each sheet, you can then add up the results for all sheets (for example, by adding only the five sheet totals together).

Example calculation table.

Farm ID	Visit date	Outbreak date	Days	Production

Count:		
Total:		

Another type of summary sheet is a cross-tabulation sheet. This is a table to work out how many records fit into certain categories. For instance, a survey may include questions on the types of cage used, and the survival rate of fish in those cages. During the analysis we may be interested to examine if there is any relationship between cage type and survival. A calculation sheet like the one below may be used:

Cage type	Survival rate			
	0–50%	50–70%	70–90%	90–100%
Net	|||	||||| ||||	||||| ||||| |||	||||| ||||| ||||| ||||
Bamboo	||||| ||||	||||| ||||| ||||| ||	||||| |||||	||||| |||

By examining each questionnaire, you can determine which box it belongs in. Make a mark in the correct box, for each farm, grouped in sets of five. Finally, count up the marks to work out the totals:

Cage type	Survival rate			
	0–50%	50–70%	70–90%	90–100%
Net	3	9	13	18
Bamboo	9	17	10	8

Calculating cross tabulations like this can take a lot of time. Plan carefully the way you want to examine your data, and try to work out the most efficient approach.

Analysis

The advantage of using a computer is that you don't need to remember the details of the different formulas for the calculations. Manual data-analysis techniques require a clear understanding of the formulas to perform the analysis. Simple analysis, such as the use of cross-tabulations, graphs, proportions and averages, are all straightforward, as the method and formulas are very simple. However, calculation of variance or standard deviation and confidence intervals is rather more complex. Analysis of two-stage or other complex survey designs using manual techniques is extremely difficult.

If anything more than a very simple analysis is to be attempted using manual techniques, you should use a statistics textbook to find the correct formula. For some types of calculation (e.g. standard deviation) there are several different formulas available, some of which are designed to be used to make manual calculation easier.

10

Survey design and planning

Survey activities

The following chapters provide specific instructions for carrying out four different types of surveys: prevalence surveys, production surveys, incidence rate surveys and surveys to demonstrate freedom from disease. The designs of each of the survey types presented differ greatly, but all aquatic animal disease surveys have some features in common. Some of the activities that need to be carried out for every survey include:

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

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

The question

Every aquatic animal disease survey aims to answer a question about the population. Developing and refining the question, and working out how to best answer it, are important first steps in running a disease survey.

Usually, the question is first asked in very general terms (for example, ‘Is the QX disease control strategy working?’, ‘How much EUS is there in the north of the country?’, ‘Is there any white spot present in the country?’).

While these questions are a useful starting point, it is not immediately clear how we might answer them. They need to be refined, so that we can measure some quantity and thereby answer the question. For instance, the first question could be refined to specify how we can determine if the control strategy is working. If the program is working, then there should be no expansion of the area infected with the disease. The new question is ‘What parts of the country had QX disease at the start of the program, and what parts have it now?’ Any difference in these areas can be used as a measure of success of the strategy.

The process of refining a question and determining a measurable quantity that will answer it may take several steps, and most questions have several possible answers.

Population

Target population

In aquatic animal disease surveys, we are interested in various populations. The *target population* is the population that we aim to answer the question about. If the question is 'Is there any white spot in the country?' then the target population is all species of crustaceans susceptible to white spot in the country.

Source population

The *source population* is the population from which the sample population was drawn. If, to answer the previous question, a two-stage survey was conducted including all farmed shrimp, this would be the source population. Ideally, this is the same as the target population, but it usually excludes some groups because it is not practical to sample them. For instance, it may be impossible to capture and test wild shrimp or crabs with the resources available, so they are excluded from the source population.

It is important to remember that the results of the survey can only be used to make inferences about the source population, not the target population (if they are different). See page 50 for a discussion of inference.

Choosing the survey design

Usually, a well-considered and refined first question will point the way to the sort of study design that is necessary. When evaluating a control program by measuring the proportion of affected farms, 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 59 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 11 (prevalence) and 13 (incidence rate), will also help in deciding which is most appropriate to use.

Questionnaires and data collection sheets

Well-designed forms on which to record the findings of the survey will make the collection, management and analysis of the data much easier. Designing forms is often a lengthy and complex task. The objective is forms that are easy and fast to use, and help maintain a high level of quality for the data collected. One of a number of approaches that might be used is as follows:

Identify the key data items required

The first step is to decide what information is required to answer the main survey question. If the aim of the survey is to calculate the prevalence of a disease, the information required is the total number affected and the total population. To estimate these from a sample, we simply need to identify each individual examined, and indicate whether it has the disease or not. A suitable data collection form would have two columns: specimen (number), and disease present (yes/no). Most surveys will be much more complex than this. If investigating a new or poorly understood problem, it may be better to start by using non-structured interviews with appropriate individuals or groups, in order to better refine the problem and get a clearer idea of the data needed to find solutions.

When listing the key data items, it is important to keep the following points in mind:

- Keep the list as short as possible. Only collect data that is essential to answer the survey question. Adding more questions just because they are interesting makes the survey more expensive and difficult, and will often result in lower-quality data. All too often, this extra data is not analysed, and has just been a waste of time.
- Make sure that there is a practical and reliable way to gather each data item. For instance, asking a farmer to estimate the proportion of sick shrimp in a pond may not provide any useful information, because of the difficulty of identifying sick shrimp.

Focus groups

Once the key data items have been listed, make an effort to understand how the people to be surveyed perceive those items. For instance, a questionnaire might be faster to complete, and more accurate, if a list of possible answers is provided to a question like 'What sort of fishing tackle do you use?' By discussing this with a focus group, it is possible to list all the likely responses on the form, so that only a tick is necessary. Without this step, the list may miss important fishing tackle used locally, and make the form harder for people to complete.

The use of focus groups during form design is similar to pilot testing before the full survey is conducted.

Design forms

The skills and training of the people who will use the forms determines the care with which they should be designed. If a small survey is being conducted, and the only person collecting data is the person who designed the survey, that person knows exactly what data is required. A simple sheet ruled into a number of columns with one-word headings is likely to be enough. On the other hand, if a questionnaire is to be completed by farmers or fishers without the assistance of survey staff, each question must be clearly and simply explained to avoid misinterpretation.

To be easy to use, forms and questionnaires should:

- Ask questions clearly and unambiguously.
- Make clear when a certain question needs to be answered or when it can be skipped.
- Make clear what the form of the answer should be (e.g. is the answer expected as a word or a number, are the units numbers of fish or kilograms of fish?).
- Provide lists of options where a limited number of responses is possible.
- Indicate if one or more selections from a list are allowed.
- Provide the opportunity to indicate another option, if none of the listed options is correct.
- Give adequate space for written responses.

Translations

If the survey is conducted in a language other than the one that it was planned in, or if it needs to be conducted in several different languages, the forms will need to be translated before use. Good translation of questionnaires is extremely difficult, and should be done very carefully to ensure that there is no misinterpretation. Ideally, it should be done by somebody who is fluent in both languages and has a good understanding of the subject of the survey, but this is not always possible. The best approach is to double-translate the questionnaire. This means translating the questionnaire to the language needed, and then having a second person translate it back to the original language. If the meaning of some of the questions has changed, they need to be done again.

Even a simple questionnaire may need to be translated many times before it is completely correct.

Sample size

Before starting a survey, you need to know how many units of interest (farms, villages, ponds, nets or animals) have 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 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 (the 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, farms 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 fish in a reservoir. Population 2 is all the fry in a bag supplied by a seed stock supplier. If we are interested in calculating the average age of the fish in the two populations, the spread of ages in the reservoir is much wider than the spread of ages in the bag of fry (in which all animals are within one or two days of the same age). When considering age, the reservoir has a higher variance than the bag.

Variance is important 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 the overall value. 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 53). This is usually measured by the width of the confidence interval (page 54). A very wide confidence interval suggests that we are not very confident about the true value. 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 were 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 ± 5%) is often determined by what is practically possible.

Planning field work

The tasks involved in planning the field work 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 sample-collection equipment (e.g. boats, nets, scoops, bags, cool boxes)
- Obtaining maps of the study area
- Preparing the laboratory for the analysis of specimens
- Preparing data recording sheets
- Planning the order of 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

Training the survey team

Training field staff well will make the survey run much more smoothly, promote better relations with producers, result in the collection 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 aquatic animal health survey. This book is designed to be used as a training resource, and Chapters 15 to 17 include guidelines for trainers, lesson plans and suggested training activities.

Pilot survey

A pilot survey is a small survey 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 any unexpected responses, or additional areas that need to be studied by the survey questions.

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 only provides analysis for the key measures described. For other questions 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. This book provides brief descriptions of key tasks in Epi Info, but to carry out the range of analysis that may be necessary, users should consult the Epi Info online manual and become familiar with the Analysis program.

Reporting

The objective of an aquatic animal disease survey is not to collect data or generate some results, but to answer a question. The answer is then used to take some sort of action, usually to improve the performance of government health services, improve the health of aquatic animals, and improve the lives of producers and the population in general.

Survey results must be given to the people who are able to use them, or 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 fisheries 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, treat data analysis and preparation of reports as a priority. Data that refers to the situation 6 or 12 months ago is 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 many figures and perform complex types of analysis. When analysing data and producing reports, keep in mind that the objective of the survey is 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 is presented graphically. This may be in the form of pie charts, bar charts, line graphs or, if possible, maps.
- The information should be distributed to everybody who may need it, and everybody who participated in generating it. Field staff and local and village fisheries staff, as well as national fisheries staff, should all receive reports, some of which may be more detailed than others. A mechanism for reporting the information back to the producers 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.
- Consider distributing the results internationally. Neighbouring countries will often find the information helpful in coordinating cooperative regional approaches to disease control. For the same reason, the information should be sent to international animal health organisations, such as regional bodies, the Office International des Épizooties (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 of how the survey was carried out and on the results can help others planning similar surveys to better plan and conduct their work. A demonstrated ability to carry out high-quality surveillance to internationally recognised standards greatly improves the international reputation of a country's aquatic animal authorities. This confidence can be a distinct benefit, especially in international trade issues.

11

Prevalence surveys

Prevalence surveys

Prevalence surveys aim to estimate the proportion of the population that has a particular disease or status, at a single point in time (see page 56). Prevalence surveys are the most commonly used way to gather information in aquatic animal 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 (e.g. to determine how much disease there is in a population, or where the disease is occurring), or to monitor the progress of a control program (e.g. to determine the proportion of farms practising a recommended management activity). 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 59, to help you decide which one is best for a particular situation. Chapter 13 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 their design and analysis are quite different. These surveys are described in Chapter 14. While the three survey types are dealt with separately, it is possible to collect information to estimate two or three of these measures during the one survey.

Prevalence surveys may be considered at two levels—small-area surveys and large-area surveys. When we survey a small population, such as the fish in a single pond, the farms in a village, or the fry in a bag from a hatchery, the population is relatively evenly spread and is not divided into distinct groups. This means that it is usually possible to develop a sampling frame for simple random sampling (or use systematic sampling). This type of approach to prevalence surveys uses one-stage sampling and is the simplest to plan and analyse.

Surveys of larger areas (national, province, state or district surveys) that measure prevalence at the animal, pond, or cage level are difficult because of the lack of a sampling frame sampling frame. Ensuring reliable results while using random selection requires a sampling frame that includes every animal in the entire study area (see page 81), but building an accurate sampling frame of all the animals in a large area is usually impossible.

Two-stage sampling

Two-stage sampling (page 82) avoids this problem by breaking the sampling into two steps. First, groups of animals or ponds (usually villages or farms) are selected randomly. At this stage, all we need is a sampling frame listing all the villages or farms (the first-stage units of interest) in an area. Once the groups are chosen, each selected village or farm is visited, and a sampling frame of the ponds or cages in the group is constructed and used to select the second-stage units of interest for the sample. For the rest of this chapter, the example of a two-stage survey with ponds as the unit of interest (second-stage sampling unit) and villages as the first stage sampling unit will be used. However, two-stage sampling can be used with any populations that are gathered into groups, for example:

- groups of post-larvae in tanks in a shrimp hatchery;
- groups of fish in fishing grounds in an ocean;
- groups of crustaceans in a reach of a river; and
- groups of wild oysters in the bays of an estuary.

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

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

Conducting a survey

There are 20 main steps in running a one- or two-stage prevalence survey, as shown below. This description is based on a survey of village aquaculture, in which specimens (e.g. tissue samples) are collected for laboratory analysis to determine the disease status of the animals or ponds. Some of the procedures will need to be modified slightly (simplified) if 1) no specimens are collected and only clinical examination or farmer interview is used to determine the disease status of animals, or 2) farms rather than villages are selected in the first stage, in which case it may be easier to select second-stage units 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 four different designs. A one-stage survey uses *simple random sampling* (SRS). There are three approaches to two-stage surveys (*probability proportional to size*, or PPS sampling; *random systematic sampling*, or RSS; *random geographic coordinate sampling*, or RGCS), based mainly on the way farms 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. Computer programs are 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 and prepare data collection sheets, transport, restraint equipment, specimen collection equipment and processing equipment.

Step 7: Train survey teams.

Step 8: Select the first-stage sample (farms or villages) using random sampling.

Step 9: Visit selected farms or villages.

Step 10: Conduct a village interview to build an animal sampling frame for the village and to ask other questions.

Step 11: Select the second-stage sample (e.g. ponds) using random sampling.

Step 12: Visit farmers and identify selected ponds.

Step 13: 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 farmers, 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 four different designs for prevalence surveys, one for one-stage surveys and three for two-stage 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 81 for a full discussion of sampling frames). In two-stage sampling, the type of sampling frame determines how farms or villages are chosen at the first stage, and how second-stage units are chosen at the second stage.

One-stage sampling

One-stage sampling describes the ‘normal’ approach to prevalence surveys. A random sample is collected from the population, using SRS based on a sampling frame containing every member of the population. Alternatively, other random sampling approaches can be used, as long as they are equivalent to SRS (as are RSS and some spatial sampling techniques). The process is called one-stage sampling because there is only a single stage of sampling, during which the units of interest are selected.

The main advantage of this approach is simplicity. Selecting the sample is simple, and analysing the data is very simple (compared to two-stage sampling). For this reason, one-stage sampling designs should be used whenever possible. One-stage sampling is also more efficient than more complex sampling designs. This means that, for a given sample size, the result of the survey is more precise if you use one-stage sampling rather than two-stage sampling.

In general, you should try to use one-stage sampling, unless the population of interest is too large to develop a sampling frame for simple random sampling or to use random systematic sampling.

Two-stage sampling

Design 1 (PPS)

Probability proportional to size

In the best situation, a complete sampling frame is available, listing all farms or villages and including *reliable population data* (e.g. the number of ponds in each farm). This may come from data maintained by the fisheries services, who regularly update population figures, or else a recent agricultural census. When this information is available, PPS sampling may be used.

In Design 1, villages or farms are chosen at the first stage so that the chance of selecting one with a larger population is greater than the chance of selecting one with a smaller population. When we select animals at the second stage, we use SRS to choose a *fixed number* of ponds or farms from selected villages.

Example

Design 1 was used for a prevalence survey of village carp ponds. An agricultural census was carried out a month earlier, and the data was used as a sampling frame. Forty villages were chosen with probability proportional to the number of ponds. Each of the villages was then visited, and a village interview of carp farmers used to build a sampling frame for the village. In each village 15 ponds were selected from the sampling frame by SRS.

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 farm populations (e.g. number of ponds) for the relevant species. If the data is even a few months old, it 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 farms, 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 farms or villages in the area is available, but there is no reliable population data, SRS can be used at the first stage.

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

Example

A survey of village fishers uses a statistics office list of all villages in a province as the sampling frame. No figures are available for the number of fishers in the different villages. At the first stage, a sample of 40 villages is selected by SRS. Each of these villages is visited, and the fishers 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 fishers, 13 are selected; in another village, with 689 fishers, 34 are selected.

Design 2 is reasonably efficient, but not quite as good as Design 1. Because the survey teams don't know before they visit how many animals there will be in the villages or farms, they don't know how many second-stage units 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

Random geographic coordinate sampling

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

In Design 3, RGCS is used to select farms or villages. RGCS is a complex sampling approach, and is beyond the scope of this book. For more information on RGCS, see *Survey Toolbox for Livestock Diseases* (Cameron 1999).

At the second stage of Design 3, just as in Design 2, a *fixed proportion* of the village population is sampled.

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 much field work is needed before the actual survey, to select the villages or farms. For this reason, Design 3 should only be used when necessary. Usually, it will be possible to find a reliable sampling frame.

Step 4: Sample size

One-stage sampling

As discussed in Chapter 10, the factors which influence sample size calculations are:

- the desired precision (how close we want to be to the true answer);
- the confidence level (how confident we are that the answer is close);
- the variance in the population (how different individuals are from each other); and
- the size of the population.

In addition to these factors, resources available, including time, money, staff and equipment, will usually play an important role in determining the most practical sample size.

In many other survey types, the variance in the population is difficult to estimate. However, in one-stage prevalence surveys, the variance is closely related to the prevalence. If we are able to estimate the prevalence, we can then calculate the variance for the sample size calculation. In practice, these calculations are done by a computer program.

There are a number of programs that can be used to calculate sample size for one-stage prevalence surveys. Two of these are included on the CD accompanying this book—EpiCalc 2000 (part of the Epi Info package) and **WinEpiScope**. These are also freely available over the Internet. For this discussion, **WinEpiScope** will be used.

Step 1: Start the WinEpiScope program from the Windows Start menu.

Step 2: From the menu, select **Samples | Estimate Percentage**.

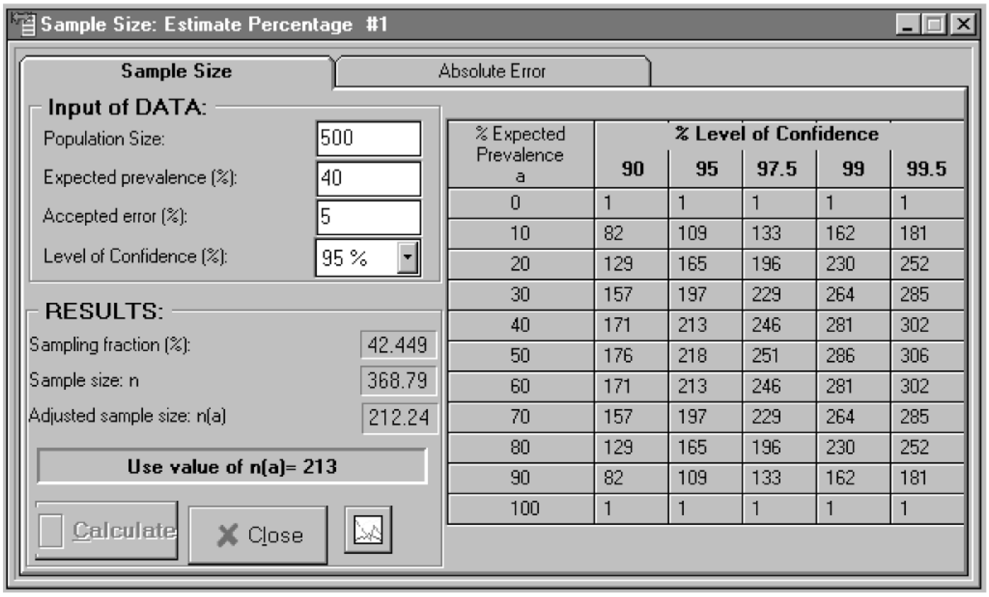
Step 3: Enter the values for calculation of sample size. In the first box, enter the population size.

Step 4: In the second box, enter the expected prevalence as a percentage.

Step 5: In the third box, enter the accepted error. This is the precision. For example, if you enter 5%, when the survey is complete (and if the prevalence is as expected) the confidence interval should be approximately ±5%.

Step 6: In the fourth box, select the level of confidence for the confidence interval. By convention, this should usually be 95%.

Step 7: Click the Calculate button.



Input of DATA:

Population Size: 500

Expected prevalence (%): 40

Accepted error (%): 5

Level of Confidence (%): 95 %

RESULTS:

Sampling fraction (%): 42.449

Sample size: n 368.79

Adjusted sample size: n(a) 212.24

Use value of n(a)= 213

% Expected Prevalence a	% Level of Confidence				
	90	95	97.5	99	99.5
0	1	1	1	1	1
10	82	109	133	162	181
20	129	165	196	230	252
30	157	197	229	264	285
40	171	213	246	281	302
50	176	218	251	286	306
60	171	213	246	281	302
70	157	197	229	264	285
80	129	165	196	230	252
90	82	109	133	162	181
100	1	1	1	1	1

The program calculates the sample size and produces a table of alternative sample sizes. The results panel reports three figures. The sampling fraction indicates the percentage of the population that needs to be included in the sample. The sample size is the calculated sample size, without taking the size of the population into account. If the population is relatively small, the sample size can be decreased. The last figure, adjusted sample size, shows that sample size adjusted for the population size. This is the sample size that should be used for the survey.

The table on the right shows the sample sizes that should be used for a range of expected prevalence and confidence values. Often the estimate of the expected prevalence used in the calculation is only approximate—after all, if we knew this answer precisely, we wouldn't be doing the survey. You can check the table to see what the sample size would be if your estimated prevalence is wrong. Note that the maximum sample size is at a prevalence of 50%, and gets smaller with higher and lower prevalences. If you are unsure of the expected prevalence, it is best to use a figure a little closer to 50%. This way, the sample size will be larger, but you can be more confident of getting a result with the precision that you want.

Two-stage sampling

For a two-stage sampling survey, the sample size is made up of the number of farms or villages to sample at the first stage, and the number or proportion of ponds to sample at the second. It is possible to select fewer villages and more ponds, 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 might not be known, and estimates will 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

We need to know the cost of the different parts of the survey: the cost per pond and the cost per farm or village. It is the ratio of these costs that determines the least expensive combination of first- and second-stage sample sizes.

The per-pond costs are mainly made up of costs for laboratory testing, and equipment such as containers, preservatives, needles etc. They may also include a cost for the salaries of the field staff, based on how much time it takes to examine or collect specimens from each pond. 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 according to the number of ponds or the number of farms or villages (e.g. the cost of obtaining a village sampling frame) are not included in the calculation.

Per-pond costs	Per-village costs
Containers	Fuel
Preservative	Vehicle costs
Laboratory tests	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 being sampled: the farms or villages, and the ponds. Each of these two populations has its own variance (see page 171). The amount of difference between different farms is known as the 'between-farm variance'. The spread of difference between individual ponds within the same farm is called the 'within-farm 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 we need to use either values from a previous survey or estimates based on similar surveys in other parts of the world.

Population size

We need to know the total size of the population for some sample size calculations (depending on the survey design). Where full population data is available for every village, this is not a problem. However, where no data exists 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 the approximate prevalence doing the survey. For surveys held as a regular part of an ongoing surveillance program, earlier prevalence estimates will allow you to make good estimates. However, for the first survey in an area, you will need some guesswork.

Precision

Relative error

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 ±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 0–10%. 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 45–55%, but the confidence interval for a prevalence of 5% may be 3–7%.

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 ±5% or ±10% is commonly used. If a smaller value is used, the sample size will increase dramatically. A relative error of 0.1 will produce a confidence interval of about ±10% if the prevalence is about 50%, but if the prevalence is 10%, the confidence interval will be about ±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 formulas for calculating the sample size for the three different survey designs are very complex, and can normally only be calculated by a trained statistician. In order to enable non-statisticians can do the calculations, the formulas have been incorporated into the **Prevalence Analysis** program included on the accompanying 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).

Step 3: If you are unsure which design to use, click the Which One? button for help, or see Choosing the Right Design on page 178 of this book.

Step 4: Look at the Second Stage Sampling Scheme for advice on how to select ponds 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 farm and cost per pond).

Step 7: If you have a data file from a previous survey, click the 'I don't know. Work it out for me' button. This will open the data file, and allow you to analyse the data in order 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 187 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, showing the best first- (farm) and second- (pond) stage sample sizes. The second-stage sample size will be expressed as either a number (Design 1, PPS), or a percentage of the farm 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 as well (see page 79). 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).

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 that 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 × 1/4 = 10 villages. A district with 80 villages would contribute 40 × (80 / 480) = 6.67 ≈ 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

One-stage sampling

In a one-stage survey, the first stage of sampling is the only stage. There are three approaches possible—SRS (simple random sampling), RSS (random systematic sampling), or spatial sampling. Sampling is conducted without replacement, which means that a single unit cannot be selected twice.

When SRS is used, it means that there must be a sampling frame. Sampling from the sampling frame can be done using random number tables, or, if the sampling frame is available on computer disk, using the **Random Village** software (page 96). Another option, when the population is grouped, is to use the **Random Animal** program (page 100).

When no sampling frame is available, but the population can be 'lined up', use RSS (page 77). Finally, if this is not possible, it may be practical to consider some form of spatial sampling (page 83).

Two-stage sampling

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

Example

The calculated sample size for a two-stage prevalence survey is 40 villages and 8% of ponds in each village. A good sampling frame is available, but no population figures, so Design 2 (see above) is used. The villages are selected from the sampling frame using SRS 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 ponds. Instead of the normal 8%, two samples (16%) are drawn from this population, giving a total of 24 ponds.

If a sampling frame is available on computer disk and you are using Design 1 or 2, you can use the **Random Village** program to select the farms or villages (see page 96). 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 farms (usually ID, name etc.).
- Under Number to Select, enter the total number of villages or farms (the first-stage sample size).
- Enter the sampling type. If using Design 1 with 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 (e.g. number of ponds). If you are using Design 2, select Simple Random Sampling.
- Under Replacement, click With Replacement.
- If using stratification, click the Use Stratification checkbox. 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 program displays the selected villages or farms. You can save them to a new table, or print them.

If the sampling fame is not available on disk, you can use the manual techniques described in Chapter 5 to do either PPS sampling (page 79) or SRS (page 71).

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.

Step 11: Second-stage sampling

In a two-stage sampling scheme, once farms or villages have been selected and the field work has begun, the second stage of sampling (selecting individual ponds) can be done.

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

If the ponds are in a single farm, a sampling frame may already exist, maintained by the producer. This can be used for SRS done either by hand, or with a computer after first entering the data into the Random Village program. Alternatively, RSS (page 79) can be used.

In a village with multiple farmers, it is usually necessary to build a sampling frame first, and then select the random sample. Village interviews of farmers are an efficient way to build an accurate sampling frame and are described in detail in Chapter 8. The technique for selecting ponds from this sampling frame using a manual method is described on page 99, and the use of the **Random Animal** program is described on page 102.

Step 19: Data analysis

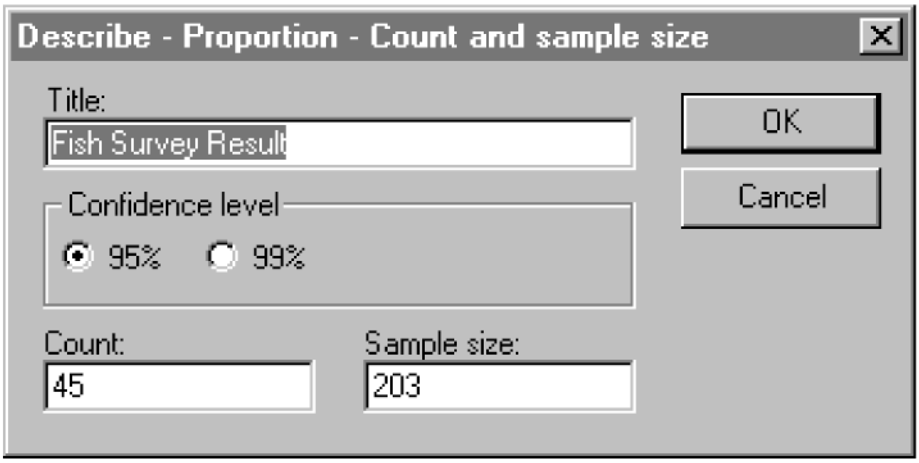
One stage-sampling

One of the main advantages of using a simple one-stage sampling approach is that the data analysis is very simple. The main figure that the survey aims to discover, the prevalence, can be calculated as (total positive / total surveyed). However, to know how confident we can be of the results, it is useful to calculate a confidence interval as well. Software is available to assist with this, including EpiCalc 2000 and WinEpiScope.

Using **EpiCalc 2000**, follow this procedure:

Step 1: Start EpiCalc from the Windows Start | Programs menu

Step 2: EpiCalc has two screens—the left for selecting the analysis, and the right for displaying the results. On the left screen, using the *right* mouse button, click on the top of the list to display the analysis menu. Select Describe, then Proportion, then Count and Sample Size. The data entry dialogue box will be displayed (see below).



Describe - Proportion - Count and sample size

Title: Fish Survey Result

Confidence level: ● 95% ○ 99%

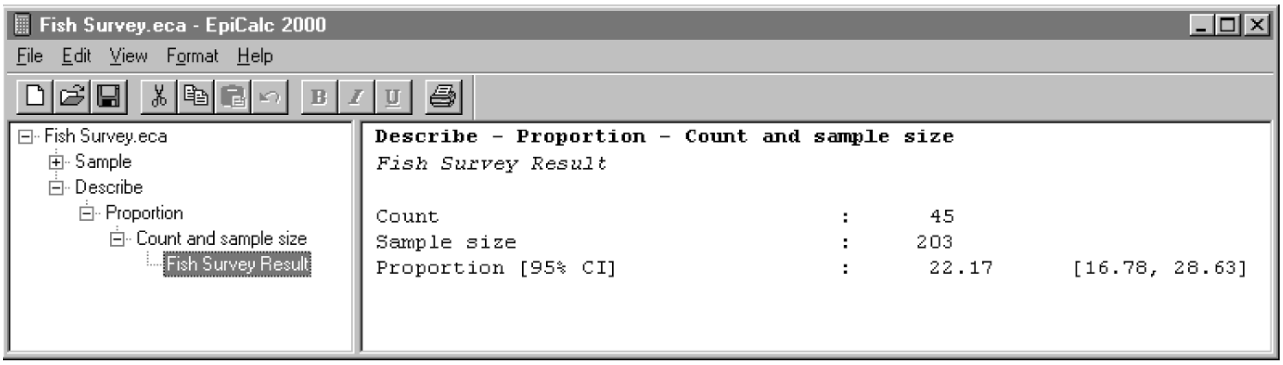
Count: 45

Sample size: 203

Step 3: Enter the required information. You can specify a title for the calculation, but you may leave this as it is, if you wish. Enter the confidence level (95% by convention), the count (number of positive test results) and the sample size. Click OK to perform the calculation.

Step 4: The results will be displayed in the right-hand screen (see following screen image).

In this example, the proportion is 22.17%, and the 95% confidence interval is 16.78% to 28.63%.



```
Describe - Proportion - Count and sample size
Fish Survey Result

Count                           :      45
Sample size                     :     203
Proportion [95% CI]             :      22.17      [16.78, 28.63]
```

Two-stage sampling

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

Before analysis, 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, a pond-level data file, with the disease status of each pond and the village that the pond came from, is required. The disease status may be a code for either diseased or non-diseased, or a yes/no field. It may also be a numeric value such as mortality rate, in which case you need to specify a cut-off value. Above the cut-off value, ponds are considered to be positive; below the value they are negative.

Design 1 (PPS)

Without stratification	
File 1 (pond file)	
• Disease status • Village ID	
With stratification	
File 1 (Pond file)	File 2 (Village file)
• Disease status • Village ID	• Village ID • Stratum ID

Design 2 (SRS)

Without stratification		
File 1 (Pond file)	File 2 (Village file)	Other figures
• Disease status • Village ID	• Village ID • Village population	• Total villages in study area • Total ponds in study area
With stratification		
File 1 (Pond file)	File 2 (Village file)	File 3 (Stratum file)
• Disease status • Village ID	• Village ID • Village population • Stratum ID	• Total number of villages in the stratum • Total number of ponds in the stratum • Stratum ID

Design 3 (RGCS)

Without stratification		
File 1 (Pond file)	File 2 (Village file)	Other figures
• Disease status • Village ID	• Village ID • Village population • Village weight[a] • Area fraction[b]	• Selection radius • Total number of random points used • Total area of study area
File 1 (Pond file)	File 2 (Village file)	File 3 (Stratum file)
• Disease status • Village ID	• Village ID • Number of ponds in village • Village weight • Stratum ID • Area fraction (optional)	• Stratum ID • Selection radius for each stratum • Total number of random points used in the stratum • Total area of the stratum

[a] The village weight is the total number of villages within the selection radius of the point used to select the village.

[b] 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) that 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.

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

```
                    Demonstration Data Entry Form

                    Prevalence Survey - Pond Data

Container Number:   ########
Village ID:         ########
District ID:        ########       (If using district for stratification)
Date stocked:       <dd/mm/yyyy>
Species:            __________
Mortality:          ######         (As reported by farmer)
Disease status:     <Y>            (Calculated from mortality)
                                   (Use a standard cut-off value)
```

```
                 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 checkbox.

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 Pond data box, Click the Open Pond 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: Select the field containing the data that identifies which village or farm the pond 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 the Village Data box is displayed, enter data. Click on the Open Village Data button and select the fields required.

Step 11: If further information is needed for stratification or RGCS, enter the required data. 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

During a survey, the disease status of ponds is assessed by means of a laboratory test, or by direct clinical examination. In both situations, it is possible to make a few mistakes, calling some healthy ponds diseased and some diseased ponds healthy. Two measures are used to describe how good a test is at correctly determining the disease state of an pond: sensitivity and specificity (see page 63 for a full discussion).

Because most tests are not perfectly reliable, a few of the test results analysed could be wrong, making 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. Using 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 it might be necessary to search journals for published studies.

Step 4: Enter the sample size of the survey.

Step 5: Click 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 one-stage SRS. For the two-stage surveys described in this chapter, the confidence interval reported will be smaller than the correct confidence interval.

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 (the point estimate) 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 ponds 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 if there is a real difference between them, rather than a difference 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.

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, we can be confident that there is a real difference between the two prevalences.

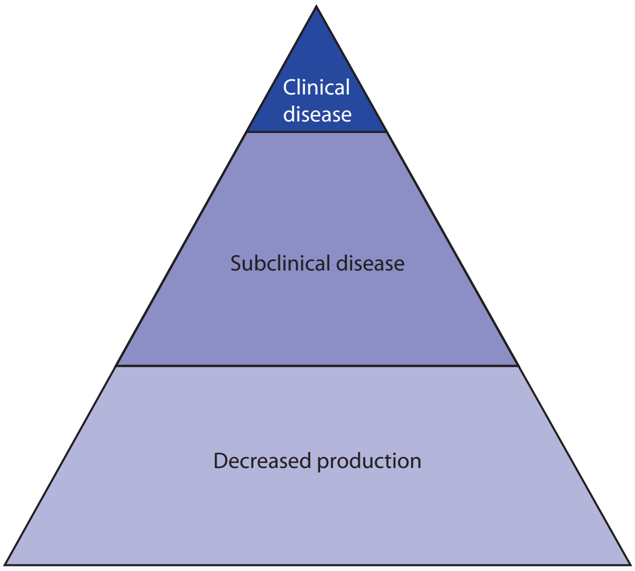
12
Production surveys

Disease and production

In Chapter 2, disease was defined as *any abnormality of structure or function.* It is the aim of any aquaculture or capture fisheries operation to achieve good, cost-effective production. Anything that decreases the production will reduce profit or the food available from the system.

Production can be influenced by many things. For instance, the feed supplied to a pond will determine how quickly fish can grow. However, disease (broadly defined as above) is perhaps the major factor that affects production.

To understand the way disease is exhibited in a population, it is useful to think of the 'disease pyramid', shown below.



When a population is diseased, or is suffering from some abnormality, the first way this is expressed is through *decreased production*. Examinations of individual animals will show very few specific signs of disease, but a large number of animals will decrease their growth rate or lose weight. As the disease progresses, some animals will show *subclinical disease*. Subclinical disease is a state in which changes due to the disease can only be detected through laboratory examination (such as histology or microbiology), with no signs apparent in a simple visual examination of the animal. Finally, when the disease has progressed further, a small number of animals will show signs of *clinical disease*. This is when it is possible to see changes in animals simply by looking at them, whether the changes are physical (e.g. ulcers, white spots, opaque eyes) or behavioural (e.g. swimming in circles, lethargy).

The disease pyramid shows us that, by the time we notice disease problems in a few animals, many more are already suffering from decreased production, and many have subclinical disease changes. The pyramid also shows that the earliest way to detect a disease problem is often through noting a decrease in production, rather than finding animals with signs of disease.

When disease is present, only a small part of the population usually shows clinical disease—most have subclinical disease, or decreased production.

When a disease problem begins in a population, it usually causes decreased production before subclinical changes. Clinical disease is only apparent after the disease has been present for some time.

Other chapters in this book deal largely with the detection of disease—mostly clinical disease, which can be detected by examining individual animals or interviewing farmers, but also subclinical disease, which can be detected by laboratory examination. Observing signs of disease and carrying out specific laboratory tests are extremely useful in understanding disease, because they can often give us some idea of the cause of the disease.

An alternative approach to detecting disease problems in a population is to measure the production of the system. Decreased production is a valuable way of detecting disease for two reasons: 1) we can detect disease earlier, because decreased production is often the first sign of a disease problem, and 2) we can detect problems more easily, because many more animals suffer from decreased production than show signs of clinical disease. This chapter examines the use of measures of production for disease surveillance.

Production indices

Measures of production in aquatic animals are very broad and contain many components. In a shrimp farm, the size, number, quality and growth rate of the shrimp are all aspects of production. If we want to use production to identify disease problems, we must be able to measure it. If there are many aspects to production, which ones should we measure and how do we interpret them?

A measurement of some aspect of production is called a *production index* (plural *indices*). There are many different production indices available, because there are many different components to production. If we wish to measure production to understand disease, we must first decide which production index or indices we should use. A suitable production index is one that is:

- objective (i.e. it is based on measurements rather than on perceptions);
- easily measured;
- relevant and informative;
- suitable for ongoing collection, so that changes can be monitored; and
- able to be compared with standards or targets.

Any sort of numerical measurement is only useful if we can interpret it. It is meaningless to say that the average weight of animals in a pond is 15 g, unless we know what the average weight is expected to be. This is why we need to have standards and targets for comparison. There are various ways to determine a standard. One is to use published or widely accepted figures—for instance, it may be accepted that the size of a particular species after 1 month should be 15 g. A second way is to calculate a standard based on local conditions. By visiting similar farms in the local area, and measuring weight at 1 month, it might be found that the average is actually only 10 g, and the best is 13 g. Using a standard of 15 g in this situation is clearly not realistic. In this case, it may be better to have two standards. If the weight is 10 g at 1 month, this can be considered average performance and there is unlikely to be a major problem. However, farmers may wish to know that it is also possible, under local conditions, to achieve 13 g at 1 month, so this can be a target to aim for. A third type of standard is not a single figure or target, but a set of standard values over time. For example, animals' feed consumption changes constantly as they grow. To assess feed consumption, it is common to use a standard curve, indicating what feed consumption can be expected at different stages of growth.

Indices may be made up of single figures, or, more commonly, be derived from a combination of two or more measurements. Examples of simple, single-figure indices are:

- length of individuals
- weight of individuals
- total number of individuals
- total amount of food eaten
- length of production period.

On their own, these figures are very difficult to interpret because they change constantly through the production cycle. It is more common to combine them with other figures to take into account changes over time. Examples of such *derived* indices are:

- *Survival rate.* This is the total number of individuals at a particular time, divided by the total number of individuals stocked, expressed as a percentage. For instance, if 2000 fry were stocked, and after 6 months, there were 1800 fish (assuming none had been harvested in the meantime), the survival rate over the first 6 months would be 1800 / 2000 = 90%. Survival rate is equal to 1 – (mortality rate), so in this example the mortality rate over the first 6 months would be 10%.
- *Growth rate.* The growth rate measures the average growth per individual over time. Growth is based on size, which can be measured in terms of either length or weight. Usually growth rate is expressed as grams per day. Growth rate is usually not constant throughout an animal's life, so the average growth rate is used.

Example

Fry are stocked into a pond at 5 g. After 6 months (180 days), the average weight of the fish is 250 g. They have grown an average of 245 g over 180 days. The average growth rate is 245 / 180 = 1.36 g per day.

To compare growth rates, use figures covering a similar period. For instance, the average growth rate, measured after 1 month, is likely to be different from the average growth rate after 6 months, but this is because of the stage of growth of the fish. However, a difference in growth rate between two ponds, both measured at 6 months, is likely to be due to management or disease factors.

- *Total production.* The total production of a system is often measured by the weight of the fish or shrimp it produces. One pond may produce 550 kg of shrimp, while another produces 850 kg, measured at the time of harvest. Multiple factors may influence the total production and interfere with comparisons, but the main factor is the size of the pond. Production is therefore often expressed in terms of production per unit area (the surface area of the pond). If a small pond has an area of 0.3 hectares and produces 550 kg of shrimp, the production is 550 / 0.3 = 1833 kg per hectare.

  Total production can also be influenced by the inputs into a system. If one farm feeds a large amount of very high quality food, we can expect its fish to faster and larger than those from another farm, which feeds cheap, poor quality food. However, the better food is much more expensive. Comparing production on the basis of costs allows comparison between these two systems. If the first farm spent $520 to produce 1400 kg of fish, the cost of production was 520 / 1400 = $0.37 per kg. If the second farm spent $130 to produce 460 kg of fish, the cost of production was 130 / 460 = $0.28 per kg. The smaller farm may be producing less fish, but it is doing so more efficiently. This is the basis for an economic analysis of production systems.
- *Food consumption.* This is one of the best indicators of disease, because it can be measured frequently and it changes quickly when a disease problem is present. Food consumption is the total amount of feed eaten in a pond, tank or cage, and can be expressed in various ways in order to make comparisons. In the simplest situation, comparisons are made with the same pond or cage, at different times. Food consumption can change for a number of reasons:
  - A change in the size of the animals. Smaller animals eat less food, so food consumption will increase progressively over the production cycle as the animals grow larger. Comparisons in food consumption need to take this into account.
  - A change in the number of animals. If there is a disease problem, and many animals die, there are fewer animals left to eat the food, and consumption will decrease. However, if the amount of food being fed is less than required, the animals may be smaller than they should be. If some animals die, the other animals may eat the extra food, allowing them to grow faster. The net effect is that there is no change in food consumption, despite having some deaths, as the surviving animals grow faster.
  - A change in the animals' consumption. With many disease problems, one of the first effects of disease is a decrease in consumption. There may be no initial change in the number or size of the animals, but if many stop eating, or eat significantly less, the overall consumption of the pond or cage will decrease.

  The most practical way to monitor food production over time is to draw a graph, plotting the amount of feed consumed each day or week. This allows two types of comparison. First, we can compare the feed consumption curve to a standard curve, as described above. This indicates if overall consumption is more or less than expected. In the other type of comparison, we examine the shape of the feed consumption curve. Normally, we would expect it to rise steeply during the early stages of production, and then a little more slowly as production continues. A sudden change in the shape of the curve indicates that something unusual has happened. For instance, there should usually be no drop in feed consumption, so if it falls there is a problem. Similarly, consumption should normally increase gradually all the way through the production cycle. If it levels off, or even changes slope so it increases more slowly than expected, it signals a problem.

- *Length of production period*. This index is easy measure, as it requires only the stocking date and the harvest date. It can also be easily compared to locally calculated standards. The production period can change length for a number of reasons, often including disease. In a disease outbreak, farmers will often carry out an emergency harvest, resulting in a shorter than average production cycle. Alternatively, if there is low-level disease, the animals may grow more slowly than expected. Farmers then have to delay the harvest to allow animals to achieve their full size, resulting in a longer than normal production cycle.
- *Other indices.* The indices described above suit various production systems. However, one of the criteria for useful production indices listed at the beginning of this chapter was ease of measurement. In many systems, the above indices are not easily measured. For instance, survival rate requires a knowledge of the total number of animals stocked, and the total animals at some other point in time (for example, at harvest). In a system that uses either progressive stocking (adding more stock at different times) or progressive harvest (removing some fish at different times) or both, it may be very difficult to determine how many fish are present, how long they have been there, and how large they are. In this situation, most of the production indices listed above are no longer useful.

  In situations like this, it is necessary to use some other production index. There are so many situations and ways of measuring production that it is impossible to list appropriate indices for all cases. One of the challenges of using production to assess disease is to determine, for a particular situation, which index is appropriate. Some creative thinking is often required to make up a new index.

Example

In a system using progressive stocking and progressive harvest, we may need new index. One possible index might be the number of fish stocked per fish harvested. This is similar to the survival rate, and will reflect the survival rate closely if measured over a long time. For instance, if the farmer stocks 800 fry and harvests 380 fish in 1 year, he or she has stocked 2.1 fry per fish harvested (which is roughly equal to a survival rate of 47%). Another index may be the total production in kilograms per year. These indices are subject to variation due to many different factors, and may easily change if the farmer's policy changes. For instance, if the farmer wants more income from the pond and increases the stocking rate, for a time the number of fry stocked will increase and the number harvested will decrease. To account for these sorts of changes, we must collect other information on the farmer's management strategies.

A further problem is introduced when production is not the main purpose of an aquaculture system. A farmer's pond in a village may be stocked with fish, but the farmer doesn't intend to harvest them all at once for profit. Instead, the pond may be seen as a sort of savings bank. When unexpected expenses arise, such as doctor's fees or children's school expenses, some fish can be taken from the pond and sold as required. In this case growth rate, food consumption and other indices are irrelevant. The main characteristics of the fish that interest the farmer are size and longevity. As this type of system has few inputs, and only irregular outputs, monitoring disease based on production is unlikely to help us identify disease early. The farmer's observations of signs of disease are probably the most practical means of detecting disease problems.

- *Production indices in capture fisheries*. The discussion up to this point has dealt with aquaculture, but the same principles can be used for capture fisheries, whether they exploit reservoirs, rivers or oceans. As there are usually no inputs to such systems, production measures related to outputs. The most commonly used measures are the total production, measured by the weight of fish caught or their size and number, and the 'capture effort', which is a measure of activity undertaken in order to catch the fish. A simple measurement of the total weight of fish caught from a reservoir may not be a good indication of their number or state of health in that reservoir, because the capture effort may be more intensive or less intensive. If three fishing boats catch a total of 130 kg of fish per day, this is very different from 40 boats catching a total of 800 kg per day. To compare these situations, we might use the number of boats as a measure of capture effort. The figures would then be 43 kg per boat, compared to 20 kg per boat. This suggests that there are more fish available in the first case.

  The use of production per unit of capture effort as an index provides a practical (but very approximate) way to estimate the number of fish available. However, it does not indicate whether a change in numbers caught is due to a disease problem or to over-fishing. Surveys, using the sampling techniques described in Chapter 5, must be used to investigate possible disease problems.

Uses of production indices

Production indices can be used as an alternative to, or in addition to the other sources of information discussed in this book, such as passive surveillance systems, prevalence and incidence rate surveys, and surveys for freedom from disease.

Measures of production differ from the other, more direct measures of disease, in that they are useful in many ways. In particular, production indices can be used by farmers to improve their management and productivity. Many medium and intensive farmers keep various records of their farm's performance and regularly calculate production indices, and so are able to recognise and deal with problems as they arise. Production indices are therefore useful both to the individual farmers (to monitor farm performance) and to fisheries and aquaculture authorities (to monitor production at higher levels, and to assist in detecting disease problems).

The fact that many more intensive producers regularly maintain records and calculate basic production indices means that this information is readily available for other purposes. On smaller farms, keeping records of production is less common, and those records kept are often less detailed.

Measuring production indices

The production indices discussed above are all measured at the pond, cage or farm level. They can be used at this level to monitor disease, or they can be used at higher levels, such as the village, district, province or national level.

Measuring indices at the farm level

Measuring production and calculating a production index for a single pond use the same methods as other surveys discussed in this book. You could use a census (measuring the weight of every fish in a pond at harvest time) or a sample (collecting a smaller number of fish and weighing them). If you use a sample, the result of the measurement will only be correct if the sample is representative, which requires random sampling (or a close approximation). This type of sampling approach is necessary for any measurement (e.g. weight or size) of individual animals, or even to estimate the total number of animals. Some measurements are made at the pond level, such as the length of the culture period or the amount of feed provided on a given day. Since pond-level measurements are single figures, they require no survey.

Collecting the information needed to calculate production indices can be done either by the farmer or survey staff. One of the advantages of measurements of production is that they are usually quick and simple, and often form part of regular farm record-keeping activities. There is often no need for survey staff to maintain the records, as farmers are able to do it themselves.

Where farmers do not currently keep suitable records, they may be taught how to, in order to collect the information necessary. Under normal circumstances, asking farmers to maintain records (for instance, for a prospective disease incidence rate survey) can present problems, as it is a new task and places an extra burden on them. However, keeping records for production indices produces information that is directly useful to the farmer in helping them make routine management decisions. Once farmers experience how easy it is to maintain suitable records, and how useful they can be when analysed appropriately, most are happy to keep such records for their own purposes, regardless of the needs of fisheries authorities. Approaches to keeping records at the farm level are discussed in detail in Chapter 8.

The way in which information is collected for various indices depends on the type of information. Generally, a farmer maintains a record sheet on which all relevant data is recorded. This can be used to calculate production indices when required. For instance, on a shrimp farm, a record sheet may have information on the number stocked, daily food consumption, a weekly estimate of the average weight of shrimp, the length of the culture period and the total weight of the harvest.

Food consumption is simply a record of how much food is used in the pond. This is only meaningful if some mechanism is used for determining how much food the shrimp are eating. Typically, this is done using a feed tray, containing a set amount of feed and suspended in the water. This is checked at a fixed time after feeding. If food remains in the tray, the amount fed is decreased. If all the food disappears quickly, the amount fed is increased.

Regular measurement of the average size of animals allows frequent calculation of growth rate. This is, in fact, a small survey. The population is all the shrimp in the pond. Using a sample taken using one of the sampling methods discussed in Chapter 5, we can take measurements and calculate the average size.

Measuring indices at higher levels

In some production systems, in some areas, many producers measure and record a range of information at the farm level, including various production indices. These are valuable, and assist with farm management and the early detection of disease in different ponds or cages.

These indices can also be useful for disease surveillance at higher levels, for example at a village, district or national level.

Example

Not all diseases have spectacular effects on animals. While many may result in obvious lesions (ulcers in epizootic ulcerative syndrome), changes in behaviour (swirling in swirling disease) or a high mortality rate (white spot syndrome), others may be much more subtle, causing no obvious signs but resulting in slower growth and poorer production. A single farmer, monitoring the growth of their own fish, might notice the change, but, without any obvious cause, consider it 'just one of those things'—normal variation in production from year to year. However, if many farmers in the same area are experiencing a similar decrease in production, this may provide a clue that a new disease is present and having a significant effect. Detecting this disease depends on collecting production indices from many farmers over a wide area.

If farmers maintain their own records of production indices, conducting a survey simply involves asking farmers to report the figures. For such a survey to be meaningful, it is important to ensure that all farmers are using the same production index and measuring it in the same way. For instance, some farmers may routinely measure growth rate at grading, while others measure it at harvesting. The growth rate at grading measures growth over the early part of the animal's life, while the growth rate at harvest measures the rate over the entire life, and will usually be somewhat lower. If these growth rate measures are combined, some farms will appear to have lower growth rates than others, not because of a disease problem but because of the way the index was measured.

If surveys of production indices are used to try to detect changes due to disease, their effectiveness depends on the extent and distribution of the disease, and the way the data is analysed. For instance, if a severe disease has caused a major drop in production in just a few farms in one part of the country, a national survey comparing the national mean production before and after the disease outbreak is unlikely to show any significant difference at all; most farms are unaffected, and the drop in production in a few farms is diluted by all the other farms in the survey.

However, if the data were analysed by district, the results could be different. Most districts would be unchanged, but the district affected by the disease could show a significant drop in production, because the affected farms make up a large proportion of all the farms in the district.

Disease is therefore easier to detect using production index surveys, if the survey is analysed using units that can capture the extent and distribution of the disease. The distribution of the disease might not be determined simply by geography. A disease may affect enterprises of one type, such as hatcheries, and leave most other farms relatively unaffected. Analysing the data from all farms together, including both hatcheries and grow-out farms, would probably make the change difficult to detect. However, if the data were analysed for hatcheries and grow-out farms separately, the effect in hatcheries would become obvious.

Calculating means

In the other surveys discussed in this book, the characteristic of animals or other things being measured is usually disease state. This can take two possible values: diseased or not diseased. This is known as a dichotomous variable, meaning that it can take only two values. Many other surveys involve dichotomous variables, such as surveys to determine the proportion of farms using a particular treatment: farms either use the treatment, or they don't. Dichotomous variables are usually summarised by using a proportion.

In contrast, most production measures involve continuous variables—that is, they are numbers that can take a range of values. Examples are length, weight, amount of feed etc. Continuous variables are summarised using the mean, or average.

The mean is calculated by adding up the values for each individual in the sample, and dividing by the total number of individuals in the sample.

Example

A farmer wishes to calculate the average length of fish in their pond. They collect a sample of 5 fish, and measure each one. The lengths of the fish in centimetres are 8, 6, 9, 6 and 5 cm. The mean of this sample is calculated as (8 + 6 + 9 + 6 + 5) / 5 fish = 34 / 5 = 6.8 cm.

The information required to calculate the mean is therefore 1) the sum of the measurements, and 2) the total number measured. Sometimes, when we need to measure many animals, we use a faster approach.

Example

When calculating the average weight of shrimp, weighing each shrimp individually and adding up the weights is too slow. Instead, a group of shrimp is weighed together, to find the sum of the weights of the shrimp with just a single measurement. Once the group of shrimp is weighed, the shrimp are counted. If a container of shrimp weighs 1.3 kg, and there are 54 shrimp in the container, the average weight of the shrimp is 1300 / 54 = 24 grams.

When values can't be measured in one go, like this, each value has to be measured separately, added up and divided by the total number. This is the approach we take when calculating the average of production indices for a village, district, province or country. When a very large number of values is involved, this task is much easier and less subject to error when done by computer. The values can be entered into a spreadsheet or database, and the computer calculates the mean, and any other figures that are required.

The mean is one of the most commonly used statistics, so any database or spreadsheet is able to calculate it easily. Using **Epi Info**, you first need to create a database to enter the information. An example database structure for a simple survey of basic farm production indices is shown below.

```
                         Farm Production Index Survey


    Farm ID:                          ___________
    {Survival} Rate:                ###.##%
    {Growth} Rate:                   ##.# grams per day
    {Culture} period:               #### days
    {Total} production:             ####.## kg per hectare
```

Once the data is entered (see Chapter 9) you can use the **Analysis** program to perform the calculations:

Step 1: Start Epi Info

Step 2: From the Programs menu, select Analysis

Step 3: Open the data file. If the file has been saved as ‘prodn.rec’, use the command ‘read prodn.rec’

Step 4: Calculate the means. If the survival-rate field is called ‘survival’, use the command ‘means survival’.

Step 5: A frequency table will be produced, followed by information similar to that shown below:

```
  Total       Sum        Mean      Variance    Std dev     Std err
   11         280       25.455     227.273     15.076       4.545
 Minimum     25%ile     Median     75%ile      Maximum      Mode
 10.000      10.000     30.000     30.000      50.000      10.000
Student’s ‘t’ testing whether mean differs from zero
T statistic = 5.600, df = 10, p-value = 0.00023
```

The first three values provide the information required. The first is the total number of observations, the second is the sum of the values, and the third is the mean, or average.

Calculating the variation

When reporting the results of a survey, whether it is a national survey, or a farmer assessing the average size of their own shrimp, it is important to understand how good the survey is, and how accurate the result is. The two most important factors that influence the accuracy of a survey result are the sample size, and the amount of variation in the population.

When examining production indices, the variation in the population is important for another reason. If a farmer stocks 300 fry at the same time, and grows them for 8 months, it is better if all the fish are approximately the same size. If some fish are much smaller than others, there are some possible implications. First, the larger fish will compete for food more effectively, so the smaller ones won’t get a chance to grow. Second, the large variation may be caused by a disease problem affecting some of the fish at some stage.

Three figures are commonly used to describe the variation in a population and the precision of the survey results: the standard deviation, the confidence interval, and the coefficient of variation.

Standard deviation

The standard deviation is a measure of how much variation there is in the population. It is a measure of the average difference (deviation) between each value and the mean. The variance is another commonly used measure: it is simply the square of the standard deviation.

The standard deviation (and the variance) are calculated automatically by Epi Info when using the Means command, and are shown in the sample output above.

If the data is spread in the typical way (i.e. if it has a normal distribution), we can use the standard deviation to work out how wide the spread is. About 68% of the values will fall within ±1 standard deviation, just over 95% will be within ±2 standard deviations, and 99.7% will be within ±3 standard deviations. This means that if the standard deviation is very small, most of the data is very close to the mean. If it is very large, the data is spread out and there is a lot of variation between values.

Confidence interval

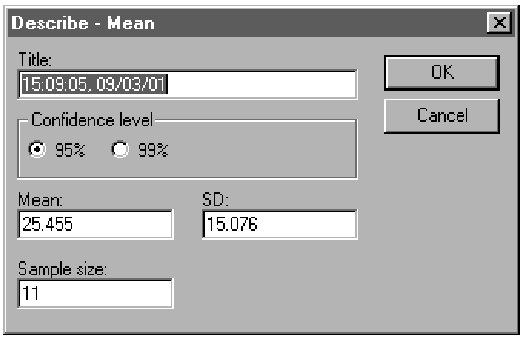
The confidence interval for a mean is used in exactly the same way as a confidence interval for a proportion. If a survey were repeated many times, a 95% confidence interval would contain the true mean 95% of the time. A narrow confidence interval indicates that the survey is able to estimate the mean very precisely. A wide confidence interval indicates that the estimate of the mean may be quite different from the true value.

As with proportions, the confidence interval can be calculated by a range of software. Using **EpiCalc**, follow these steps:

Step 1: Start the **EpiCalc** program, and use the right mouse button on the left pane to bring up the menu.

Step 2: Select Describe, then Mean.

Step 3: The data entry dialogue box will appear.



Describe - Mean

Title: 15:09:05, 09/03/01

Confidence level
● 95% ○ 99%

Mean: 25.455

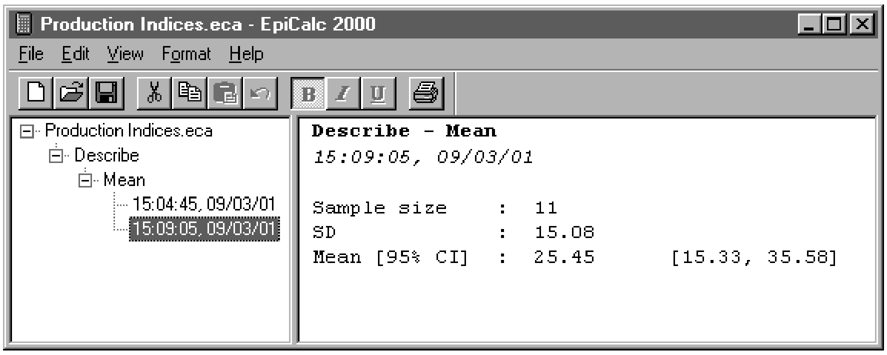
SD: 15.076

Sample size: 11

Step 4: Enter a title, if desired, and set the confidence interval width to 95%.

Step 5: Enter the mean and standard deviation calculated previously.

Step 6: Enter the sample size. Click OK to calculate the results, which will be displayed in the right side of the window:



```
Describe - Mean
15:09:05, 09/03/01

Sample size    :  11
SD             :  15.08
Mean [95% CI]  :  25.45      [15.33, 35.58]
```

Coefficient of variation

The third measure of variation is used to compare the variation between different populations. For instance, when fish are very small, with a mean length of 3 cm, the standard deviation is likely to be quite small as well. When fish are large, the standard deviation will be much larger. However, the relatively smaller standard deviation in small fish may be just as important, or more so, than the large standard deviation in large fish. The coefficient of variation allows us to compare these two different situations. It is calculated by dividing the standard deviation by the mean.

Example

A group of young fish with a mean size of 3 cm has a standard deviation of 1 cm. Larger fish with a mean size of 15 cm have a standard deviation of 4 cm. The overall variation is much greater in the large fish than the small fish, but this does not take the mean into account. The coefficient of variation for the small fish is 1 / 3, or 0.33, while the coefficient of variation for the large fish is 4 / 15 or 0.27. The larger coefficient for the small fish indicates that they are more variable, relative to their mean size.

Surveys for production

Surveys to measure production indices are similar in structure to surveys to measure prevalence, described in Chapter 11. One-stage sampling surveys are relatively simple, but two-stage sampling is more complex and may require the assistance of a statistician to properly plan and analyse.

Sample size

For one-stage surveys, plan the sample size. This can be done using **WinEpiScope**:

Step 1: Start the program and choose Samples, Estimate Mean from the menu.

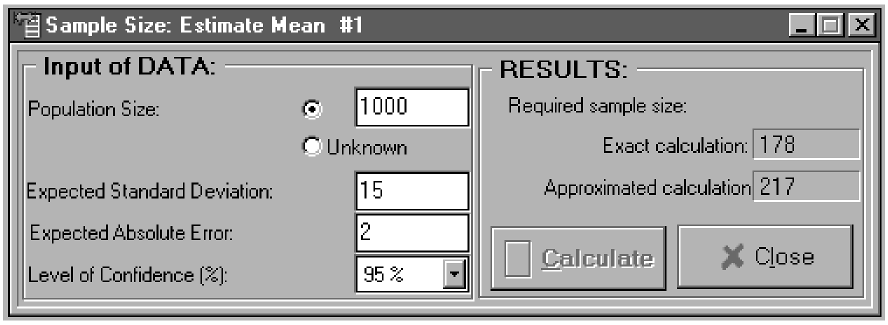
Step 2: In the input window, enter the population size (or unknown if it is large and you don’t know its size).

Step 3: Enter the expected standard deviation. As with all sample size calculations, we must know the amount of variation in the population if we want to know how many to sample.

Step 4: Enter the Expected Absolute Error (the precision). If we expect the mean to be 25, entering a value of 2 means that we should be able to estimate the mean with a confidence interval of ±2 (i.e. the mean will be between 23 and 27).

Step 5: Enter the confidence level (use 95% by convention).

Step 6: Click the Calculate button.



Population Size: ● 1000
○ Unknown
Expected Standard Deviation: 15
Expected Absolute Error: 2
Level of Confidence (%): 95 %

Required sample size:
Exact calculation: 178
Approximated calculation 217

The program produces two figures: an exact and an approximate calculation. The exact figure takes the size of the population into account, and is close to the approximate figure except for small populations.

13
Incidence rate surveys

Incidence rate

Incidence rate is the number of the new cases of disease in a population at risk, over a period of time (see page 58). Incidence rate measures the rate of spread of infectious diseases or the average rate of onset of non-infectious diseases. To calculate the incidence rate, it is necessary to know three things:

- the number of new cases;
- the time over which the population was at risk;[1] and
- the total size of the population.

The time for the study is usually clearly identified, and the population at risk is often either known or can be reasonably well estimated. The difficult part of incidence rate studies is finding the total number of new cases. There are two main approaches to solving this problem. The first is known as a *prospective* study. *Prospective* studies (literally, 'looking forward') start at a defined time, and identify cases of diseases as they occur. After a period, the study finishes, and the total number of new cases is counted. The second approach is to use a *retrospective* study ('looking back'). Information is collected at one point of time, about cases of disease that have already occurred.

The traditional way to estimate incidence rate using a prospective study is to observe a population for a long period, and to record which individuals become affected by the disease. The units of interest may be individual animals, but can also be the pond, cage or farm. This type of incidence rate study is often slow and expensive, as every individual has to be regularly examined, and the study may last for many months or longer. This prospective approach to incidence rate studies is described below.

This chapter also describes two alternative ways to gather information about disease incidence rate. These techniques differ from traditional studies in two respects:

- The unit of interest is the village, farm or pond, 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, farms or ponds that suffer an outbreak of disease in the period.
- The outbreak information is collected by using the memories of the farmers. Instead of starting a study and observing the animals for a long time (a *prospective* study), we ask farmers about the disease outbreaks that have occurred over several years before the survey (a *retrospective* study).

Most major epidemic diseases are maintained through spread from one village or farm to another. It is rare that the bacteria or viruses are able to maintain themselves within the one farm, as animals either die, or develop immunity, or are harvested relatively quickly. Farm- 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

[1] The time at risk should correctly be calculated as the sum of the time at risk for all individuals in the population. For instance, if a pond becomes diseased only one week into a two-year study, it has been at risk for only one week, whereas a pond that never has a disease outbreak is at risk for the full two years. Often, this detailed information is not available, so the average time at risk is used. This is the total time of the study (say, two years) times the average number of individuals at risk during the study period.

memories of farmers, it is also much easier to collect reliable information about farm- or village-level outbreaks than about disease in individual animals.

The first of the two alternative techniques described in this chapter is the *retrospective disease outbreak survey* technique. Village interviews are carried out to ask farmers 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 within the same area at different times).

The second method, the *capture–recapture* technique, uses information on disease outbreaks that the fisheries 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 suitable, separate data sources already exist, no field survey is necessary.

Prospective incidence rate surveys

Introduction

The traditional approach to calculating incidence rates is to conduct a prospective survey. The advantage of these studies is that they do not rely on memory or records to identify cases of disease. The main disadvantage is that they often take a long time to carry out. A retrospective study can collect information about the previous two years during a single half-hour interview. With a prospective study, collecting the same information would take two years. The other disadvantage is that the sample must be kept under observation for the entire period of the study, which may be difficult.

Collecting information in prospective incidence rate surveys can be done in two ways, either with continuous observation of the population, or duplicate surveys.

Continuous observation

In this type of study, the sample is kept under continuous observation for the entire survey period. This means that any new cases of disease will be observed when they occur, and recorded with the date of occurrence. It is very rare for a survey using continuous observation to use survey staff to monitor the individuals. This is because it would take too many people too much time to check each individual regularly. A more practical approach is to give the responsibility to somebody else.

Example

A prospective survey of epizootic ulcerative syndrome (EUS) in silver barb is being conducted in fish ponds in a province. The sample consists of 200 ponds, selected at random from all the ponds in the province. These ponds are owned by 185 farmers (a small number of farmers have two or more ponds in the survey). At the beginning of the survey, each farm is visited and the purpose of the survey explained to the farmer. The farmers are given recording sheets, to indicate the date of any EUS outbreaks in the survey ponds. The survey continues for one year, during which farmers are visited monthly, to check progress, collect data sheets and encourage ongoing participation.

In this example, the farmers are used to collect data, because they can observe the ponds almost every day and note any outbreak of EUS. Survey staff only visit occasionally, to make sure everything is all right.

Example

A shrimp farming area is served by a single diagnostic laboratory. There is a high awareness amongst local farmers of yellow head disease, and when an outbreak occurs, almost every farm sends a specimen to the laboratory to confirm the diagnosis. A survey is done, based on the results from the tests done by the laboratory.

In this second example, the laboratory is made responsible for recording outbreaks of the disease. This is even simpler to organise, but suffers from some major problems. First, it is assumed that every farm suffering an outbreak of yellow head will send specimens to the laboratory. If this is not the case, the incidence estimate will be lower than the true value. The second problem is estimating the size of the population. If a list of all shrimp farms in the area is available, and all farms use the laboratory, the size of the population may be known. However, if some farms don't submit, or if they use another laboratory, they aren't part of the population at risk.

Duplicate surveys

When continuous observation, either by survey staff or by somebody else, is not possible, duplicate surveys may be able to gather the information required. A survey is conducted at the start of the survey period to find out the disease status. At the end of the survey period, a second survey is conducted using exactly the same sample. The disease status is checked again, and any individuals that have acquired the disease between the two surveys are identified as new cases.

A duplicate survey is much easier, as it is really just the same as two prevalence surveys (except that individuals have to be carefully identified and matched). The only difficulty is that, at the second survey, it must be possible to tell that an individual has had the disease sometime since the first survey. This is possible when:

- the disease lasts for a long time, or leaves physical signs that can be detected for a long time after it resolves;
- the disease and species make antibody testing possible (as antibodies can indicate past exposure to the disease); or
- the duration of the survey is relatively short, so that a disease will not appear and resolve before the second survey.

If one of these conditions is met, the duplicate survey approach will be less expensive and simpler than continuous observation.

Survey activities

The activities undertaken during a prospective incidence survey are much the same as those conducted in any survey. In a survey using farmer continuous observation, there will be an extra step of recruiting and training the farmers. Unlike a prevalence survey, where a single interview may be used, a prospective study that depends on the farmer to record observations places much greater responsibilities on the farmer. The survey team has to ensure that:

- the farmer knows what they are meant to do;
- the farmer knows how to do it; and
- the farmer actually does it, and keeps on doing it during the entire survey.

One of the best ways to ensure that farmers cooperate, record data well, and continue to participate during the survey is to help them to understand why the survey is being conducted, and what benefits participation will bring to them. Regular follow-up visits will also help to maintain farmer commitment.

> Farmer training is an important part of continuous observation surveys using farmers to record data.

In duplicate survey studies, the procedures are almost exactly the same as for prevalence studies. However, the important difference is that both parts of the duplicate survey use exactly the same sample, and individuals in the sample have to be clearly identified. During the first survey, the population is examined to determine the current disease status. If an individual *is* diseased, it is not at risk of becoming diseased, so it must be excluded from the survey—only the non-diseased individuals are included. At the second survey, the same individuals must be examined to determine if they are currently diseased, or have been diseased in the past (since the first survey).

Sample size calculation

The calculation of sample size for incidence rate studies is based on the same procedure as used for prevalence studies. The factors to consider are:

- the size of the population;
- the expected number of individuals that will become affected during the survey period, as a proportion of the sample size;
- the precision desired (half the width of the confidence interval); and
- the confidence required.

Once these figures have been estimated, you can use WinEpiScope to calculate the sample size, as described on page 180.

Example

You are studying a defined population of 2000 crayfish over a period of 3 months, and expect that 30% will become infected by crayfish plague in that period. You wish to calculate the incidence rate, with a precision of ±5% with 95% confidence. Using WinEpiScope indicates that you would need to use a sample size of 278 crayfish.

Data analysis

The calculation of the point estimate of the incidence rate is relatively simple, using the following formula:

$$\text{Incidence rate} = \frac{\text{Total new cases of disease during a period of time}}{\text{Average number at risk } \times \text{ Time period}}$$

Based on the example above, if 65 crayfish contracted crayfish plague during the 3 months of the survey, we could calculate the incidence rate as follows:

Example

Total new cases = 65
Average number at risk = (Number at risk at start of survey (278) + Number at risk at end of survey (278 – 65))/2 = (278 + 213)/2 = 245.5
Time period = 3 months
Incidence rate = 65/(245.5 × 3) = 0.0883 new cases per crayfish per month.

This is often expressed in terms of years, or larger numbers of animals:
Incidence rate = 0.0883 × 12 = 1.06 new cases per crayfish per year
Incidence rate = 1.06 × 100 = 106 new cases per 100 crayfish per year

Calculation of confidence intervals is also done in the same way as in prevalence surveys, but you must be careful to use the right figures. The confidence interval should be based on the proportion of the sample that becomes affected during the study period, before any conversions are made into more convenient units. For instance, in our example above, you would calculate the confidence interval based on 65/ 245.5 (rounded to 246). Using EpiCalc as described on page 180, you would get a point estimate and confidence interval of 0.2642 (95% CI 0.2112 to 0.3248). These figures are the number of new cases per crayfish per three months. When converting to different units, you need to convert both the point estimate and the confidence interval. For instance, this is equivalent to 106 new cases per 100 crayfish per year with a 95% confidence interval from 84.5 to 129.9.

Retrospective disease outbreak surveys

Introduction

In this technique, interviews with producers (farmers or fishers) are carried out to collect information about the date of the most recent disease outbreak. The reliability of the survey therefore depends on the farmers' ability to correctly identify the disease, and to correctly recall the date that the outbreak started.

To ensure that the quality of information collected is high, this survey technique should only be used in appropriate situations. It should only be used to investigate diseases that are:

Survey prerequisites

- *discrete and repeatable*—the disease must occur as an outbreak, last for a relatively short time and be able to occur more than once in the same area;
- *distinctive and well known*—since diagnosis is based entirely on the observations of the farmers or fishers, diseases that are clearly different to other diseases, and have a dramatic and consistent clinical presentation, are more easily diagnosed; and
- *memorable*—the ability of farmers and fishers 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 farmers and fishers, the more reliably it will be remembered).

Every effort is made to assist the farmers and fishers to accurately remember the date of the outbreak. Chapter 8 discusses a range of available techniques.

The various strengths and weaknesses of the retrospective disease outbreak survey technique for estimating incidence are summarised below.

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 on farmer diagnosis
	• No animal-level estimate

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 farms or villages.

Step 10: Hold a farmer interview.

Step 11: Determine if the farm 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 farmers 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 11, in which the unit of interest may sometimes be the individual animal, the unit of interest in retrospective disease outbreak surveys is the farm, village or pond. This means that simple one-stage sampling can usually 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 farms 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 relatively few farms 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 farms 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, Survey Toolbox, 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), as long as you use the same units for both groups. 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

A vaccination program for channel catfish virus 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 has been decided to carry out disease outbreak surveys every year. Before the program started, the average time since the last outbreak in the area was about 3 years. The veterinary authorities decide that a lengthening of this average time to 5 years would indicate 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 (indicating how confident you want to be of the result).[2] 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.[3]

Step 4: Click the Calculate button and the sample size will be displayed.

The sample size is not the number of farms or villages to include in the survey. Instead, the number presented is the number of farms or villages that have *had an outbreak*, that need to be included in the survey. As some farms may never have had an outbreak, the total number of farms to include in the survey should be somewhat larger than the number produced by the sample size program.

The actual sample size can be calculated by estimating the proportion of farms that have never had an outbreak. Alternatively, it is possible to simply carry out the survey and keep selecting new farms until the required number of farms with outbreaks has been identified.

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

[2] 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.

[3] 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.

Step 8: Selecting farms or villages

Farms or villages should be selected using simple random sampling (SRS). To select villages manually, use the procedure described on page 73.

If the sampling frame is available on computer, you can use the **Random Village** program to select the sample, as described on page 97, using the following settings:

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

Steps 10–13: The interview

Farmer interviews are discussed in detail in Chapter 8, including advice on techniques for collecting information about the date of the last disease outbreak. It is worth repeating the importance of establishing a censoring time for farms that have not experienced an outbreak.

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 ponds on the farms at the time of the interview;
- the number of ponds on the farms at the time of the outbreak (or the censoring time); and
- the proportion of those ponds that were affected by the outbreak.

A sample data recording sheet is shown in Appendix C.

Step 15: Data management

When the field work is completed, all the results need to be entered into a computer for analysis. See Chapter 9 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 farm identification.
- Outbreak (yes/no, or code field indicating whether there has ever been an outbreak or not).
- 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. All outbreaks could arbitrarily be said to have started on 15th of the month, or 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 later calculated from the two dates.

- If the data from two areas or times is being compared, the file needs to contain all the data from the two groups. If the data already exists 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 online manual). There must be a group field, with a code to indicate which group the record belongs to. This may be a numeric code, a text field, or a yes/no field.

Below is an example of a questionnaire file for creating the table in Epi Info.

```
                         Demonstration Data Entry Form

                              Farm Outbreak Survey

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

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

Epi Info

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 the command '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 stored in VisDate, and the time since the last outbreak is to be stored in a field called Time, type: Time = VisDate – 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: Time = (Vyear – OByear) + ((Vmonth – Obmonth) / 12). This will give the outbreak time in terms of years. If you prefer months, 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 is analysed using a special technique called *survival analysis*. This technique uses the times that a farm or village *survives* without experiencing an outbreak. The advantage of survival analysis is that it is possible to include data from farms or villages that have not had an outbreak.

The **Survive** 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 is 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: Click the Censoring Indicator box, and select the field that indicates if the farm or village has ever had an outbreak.

Step 5: Now you must 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 farms or villages were selected using random geographic coordinate sampling, click the Weighted checkbox and select the field that contains the village or farm 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 is in the data file, you can select just one group by clicking the Select Group checkbox.

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 checkbox to adjust for differences.

Step 10: When all the fields have been set, click 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. This may be due to the biology of the disease, with environmental factors acting as component causes of the disease, or be due to management factors (for example, if the disease usually occurs at 4 weeks after stocking and most farmers stock in March, most of the disease will occur in April). 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 farms or villages may have experienced an outbreak in the past two or three months. If the survey is conducted 7 or 8 months after the peak season, the time since the last outbreak will be longer. However, this is not because the disease situation is different, but because the outbreaks occur in a clear season.

If there is evidence that the disease outbreaks have a seasonal pattern, you may have to correct for this, so that 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 checkbox.
- 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 farm- or village-level incidence rate is that not all farms or villages are the same size. A large village is likely to have more stock 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 in 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 only due to the size of the villages, not the overall level of disease. It is possible to adjust for differences in village or farm size (i.e. number of ponds), 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 pond in the village has been at risk of getting disease for that period. The *pond-time* (number of ponds in a village or farm, times the length of time since the last outbreak) will be greater for larger farms than small farms. Pond-time is an alternative time measurement that takes into account the size of the village.

Ideally, pond-time should be calculated on the basis of the average number of ponds 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 included the time since the last outbreak (Time), the number of ponds at the time of the visit (Vpop) and the number of ponds at the time of the last outbreak (OBpop). In order to calculate the pond-time, the average number of ponds in the village over that period was multiplied by the time since the last outbreak: ((Vpop + Obpop) / 2) × Time.

Once pond-time has been calculated, the analysis can be repeated using the pond-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 into account differences in village pond numbers and will therefore be somewhat more reliable.

The second approach to taking account of pond numbers in the analysis is to analyse the data using Cox's proportional hazards model. This is only available in more sophisticated statistical software, and requires a good understanding of survival analysis and multiple regression models. If the software and technical expertise are available, the number of ponds in the village 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 as a survival curve, or graph, which shows the disease experience of the entire study population. The Kaplan–Meier survival curve (named after the people who developed it) is a graph that shows the proportion of farms or villages that have 'survived' (not had an outbreak) for a particular period of 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, is it possible to interpret a survival curve. However, some summary measures that describe the curve are easier to understand.

Single group analysis

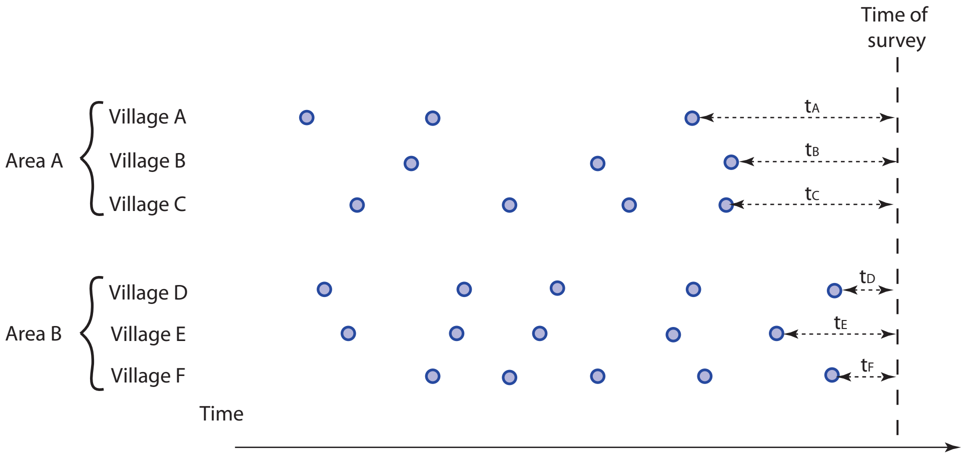
With single group analysis, the total number of observations and the total number of censored and uncensored observations are displayed. If weighted analysis is performed (because of the selection of farms or villages with random geographic coordinate sampling), these totals reflect the sum of the weights, 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 farms or villages have not had an outbreak and half have. An estimate of the median survival time is used for sample size calculations (page 214).

The *mean* or *average survival time* is the average interval 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), 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 below, each circle represents an outbreak. The mean survival time for the villages in area A is given by (t_A + tB + 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 only compare 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 that 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 are 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, and 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, nor 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 catfish farms 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 catfish farms 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. A hazard ratio produced by survival analysis can be interpreted in the same way as a risk ratio in other epidemiological studies.

> 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 farms that have suffered an outbreak in a fixed period can't be calculated. This section describes a simple approach, which takes advantage of data that already exists, to estimating village or farm-level incidence rate.

The fisheries and aquaculture services in some countries maintain good records of disease outbreaks, particularly for important or notifiable diseases. This passively acquired data (compared with active surveillance—see page 31) can provide some indication of the disease situation. However, reporting is almost always incomplete, so the population at risk is unknown and any incidence rate estimate based on this 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 the total number of disease outbreaks is.

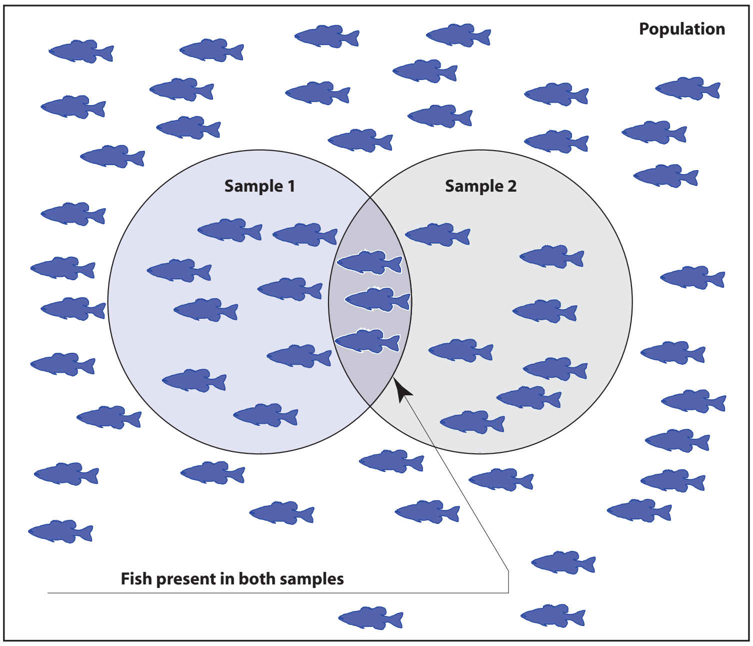
Background

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 she uses a different approach. 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 back 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 she catches a fish, she records 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:

$$\text{Total} \;=\; \frac{n_A \times n_B}{n_{AB}}$$

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 farm disease outbreaks in an area over a period of time, and this can be used to calculate the incidence rate. The population is now not fish, but all village or farm 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 disease outbreaks held by the fisheries authorities. 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 capture–recapture technique, there must be two sources of data on village or farm 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 survey. To be valid, both data sources have to refer to the same period of time, usually a period of one or two years. Only reports or specimens received in a clearly defined period should be analysed, and the survey should only record outbreaks occurring within that same period.

For the two data sources to be independent, the same people can't be responsible for collecting both. For instance, the district fisheries 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 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), this could also be used, saving the need for any field data collection.

Selecting the sample

If a survey is required, the villages or farms should be chosen by simple random sampling from a sampling frame. The **Random Village** program can be used (page 97), with 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 is similar to that used in the village outbreak surveys described above. The difference is that, instead of using only 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 past one or two years). Limiting the period to a relatively short time makes it easier for farmers to remember. However, if there has been more than one outbreak in the village or farm, farmers may find it difficult to reliably recall earlier outbreaks. The techniques described in Chapter 8 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) and the date of the outbreak. The date reported on a laboratory submission or a disease report will usually not match the date recalled by farmers during an interview. This is because there are often small errors in memory, and also because specimens or reports might 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 farm, 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 judgement 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

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, 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 the Calculate button.

Step 5: The results show the estimated total, and a 95% confidence interval.

The incidence rate is equal to the total number of outbreaks over the total number of villages, multiplied by the time period (page 58). 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 EUS outbreaks over a two-year period, 145 outbreak reports had been received by the fisheries authorities from a province containing 1293 villages. A survey was conducted of 85 villages, 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 outbreaks/(1293 villages × 2 years) = 0.073, or 7.3 outbreaks per 100 villages per year (with a 95% confidence interval of 6.3 to 8.2 outbreaks per 100 villages per year).

14

Surveys to demonstrate freedom from disease

Freedom from disease

Consider these four different situations:

- A shrimp farmer has had problems with white spot syndrome. In order to avoid significant losses, they want to purchase post-larvae that are not infected with white spot syndrome virus.
- One approach to disease control is the development of accreditation schemes amongst fry suppliers. These schemes involve testing each farm to provide a guarantee that the farm is free from disease. This means that other producers can buy from that farm without the risk of introducing disease. As a result, accredited farms are able to ask higher prices for their fry.
- After a recent outbreak of food poisoning, public health concerns require that oysters be shown to be free from potentially dangerous bacteria after harvesting.
- The development of export industries is an important way for developing countries to obtain foreign exchange and develop their economies. The export of live fish or aquatic animal products is one area where developing 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 diseases to the importing country.

In each of these four situations, it is necessary to demonstrate that a population (a single batch, farm, village, district, province or whole country) does not have a particular disease. This chapter describes survey techniques that are able to 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 survey 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 tank contains 342 fish, it is only free from disease if none of those fish is diseased. It is possible (though perhaps unlikely) that just a single fish is infected. If we take a sample from the tank and test the fish in the sample, we might, by chance, select the one infected fish, and be able to conclude that the tank is not free from disease. However, it is possible that we might not select that fish in the sample, and think that the tank is free from disease when it is not. The larger the sample size, the higher the chance that we will pick that one infected fish, but there is still a chance that we will miss it. The only way to be completely sure is to test every fish in the tank. If we are trying to show that a tank of 342 fish is free from disease, this may be expensive, but not too difficult. When trying to prove that a country with 8,000 million fish is free from disease, it is completely impossible to test every animal.

The second problem relates to laboratory tests. Chapter 2 discussed the ideas of sensitivity and specificity of tests (page 61). Very few laboratory tests are perfect, and most make a small number of mistakes, calling some diseased fish non-diseased, and vice versa. This means that if you test a large number of fish, it is difficult to interpret the results. If there is one positive test result from 342 fish, is this fish really infected, or is it just that the test has given the wrong result (a false positive)? Are all the fish that tested negative really disease free, or are some of them diseased and the negative test result wrong (a false negative)?

A typical test might have a sensitivity of 95% and a specificity of 99%. The sensitivity means that of every 100 diseased fish tested, 95 will give a positive test result, but 5 will give a false negative result. The specificity means that of every 100 disease-free fish, 1 will give false positive test results. This means that even if we test all the fish in a tank or all the fish 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 the test can produce false positives. If disease is present, we might also get some false negative test results, and miss the diseased fish.

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.

Even though we can't *prove* that a population is free from disease, if we test enough fish and take the performance of the test (its sensitivity and specificity) into account, we are able to show that it is *very unlikely* that the population has infected fish. 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 fish is smaller than some acceptable level (say, 5% or 1%).

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

Example

A outbreak of crayfish plague occurred in a previously disease-free area. One estuary was affected, and the entire crayfish population of that estuary was destroyed to eradicate the disease. All estuaries within 150 km were examined for clinical signs and other evidence of infection with the disease. The disease had not been present in the area before, so if it did get into one of the estuaries, it is likely that it would have spread quickly and affected a high proportion of the crayfish in the estuary, probably well over 50%. It is very unlikely that only 10% or fewer crayfish would be affected by such a disease. During the surveillance after the outbreak, surveys were carried out in the surrounding estuaries, to demonstrate that they were free from the disease. As it was impossible to prove that no crayfish at all had been infected, the surveys were designed to show that the chance that 10% or more of the crayfish 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 estuaries were declared free from disease, even though it was possible that some had as many as 10% of crayfish infected. This didn't pose any risk, because the highly contagious nature of crayfish plague meant that if the disease had entered an estuary, much more than 10% of the crayfish would have 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 is present in a group of animals. When conducting a survey, this is the lowest disease prevalence that we can expect the survey to reliably identify. If the disease is present in the population, but at a lower level than the prevalence specified, 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, crayfish plague may have a minimum expected prevalence of 50%, while other diseases with many causal factors (such as some parasitic diseases) may have a minimum expected prevalence of 5% or less.

Maximum acceptable prevalence

For diseases that 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, 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 can test every single animal, we can't prove that the prevalence is 0. Another way to think of this value is as the *minimum detectable prevalence*. This simply means that the design of the survey is not able to detect disease at a prevalence lower than this specified level.

As with any survey, it is easier to find diseased animals if the prevalence is higher. The measure of disease should therefore be the measure that gives the highest prevalence. Some measures of disease (e.g. clinical signs) may be relatively short-lived, and therefore give a low prevalence even when there is a lot of disease present. Others measures (e.g. poor weight gain) can be observed for a much longer period and therefore give a higher prevalence. This is why it may sometimes be easier to base a survey to demonstrate freedom from disease on the use of longer-lasting measures, rather than clinical examination of fish or identification of a pathogen. Clearly, the choice of approach is also influenced by the accuracy with which a diagnosis needs to be made.

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 farm, village, or tank. The second design uses a two-stage sampling scheme for surveys of larger areas (districts, provinces, states or countries).

Single-stage surveys

Single-stage surveys can be used when every member of the population can be listed on a sampling frame, or sampled using systematic sampling. Examples include all the fish or shrimp in a pond, all the fry in a batch, all the ponds in a village, all the fishers in a district, or all the tanks in a hatchery. 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 that determine how confident we are of the results.

Step 2: Calculate the sample size.

Step 3: Select the sample using simple random sampling (SRS).

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 findings.

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 determining the sample size. Performance is expressed in terms of sensitivity and specificity (page 61). If the test is not very reliable (either sensitivity or specificity, or both, are relatively low), the sample size will 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 65) for advice on ways to improve the specificity of a test.

Unfortunately, precise estimates for sensitivity and specificity are not 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 valid for the population being surveyed. If you don't know the sensitivity and specificity of tests, try the following steps:

Step 1: Ask your laboratory if it has conducted any studies in the local population to evaluate test performance.

Step 2: Ask if it knows of published figures based on other populations.

Step 3: Search the literature for published studies on the test. If you find more than one study, 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 study in the local population to measure test performance yourself.

If no reliable published figures are available, 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 (approaches to calculating sensitivity and specificity are described on page 63). 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. As the population size increases, the sample size reaches an upper limit. For instance, in one survey with a population of 20, the sample size required might be 16. If the population were increased to 25, the sample size would increase to 21. However, increasing the sample size from 100 to 1000 only changes the sample size from 27 to 30.

Minimum expected (maximum acceptable) prevalence

The choice of this figure (explained above) is based on a knowledge of the disease, or on practical limitations. The larger the prevalence chosen, the smaller the sample size and the easier the survey. With highly contagious diseases, which are likely to spread quickly, it is safe to chose quite a high prevalence. Other diseases, if present in a population, may have a very low prevalence. 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 α (alpha) level, is the probability that the results of the survey will indicate 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.

Type II error

The *type II* (β, beta) *error* 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.

Example

A hatchery is being tested as part of a farm accreditation scheme. If the hatchery is found to have disease, the farmer will not be allowed to sell fry to other farms. The producer 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 producer would want to set the type II error level very low, to minimise the chance of this mistake. On the other hand, a client of the farm that buys fry does not want to receive infected animals. The client would want to make sure, if the survey indicates that the farm is free from disease, that this is 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 spring viraemia in carp, 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 free, it is kept under quarantine. (If clinical cases are detected, the fish population of the village is destroyed.) 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 fish and enormous expense. The probability of a type I error should be kept very, very low. If a type II error is made, the village will be held under quarantine for a bit longer. This is inconvenient for the farmers in the village, but doesn't cause 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, FreeCalc. FreeCalc is a program both for sample size calculation and for 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, or the minimum detectable prevalence). This can either be entered as an estimate of the prevalence (a percentage) or as a direct measure of the number of diseased fish in the population. Click the radio button to choose if you want to enter the prevalence or the number of diseased fish, then type in the number. The equivalent value will be shown in the other box.

Step 5: Click the Options tab at the top of the window.

Step 6: You can usually leave the Formula for Calculation box, on the left, set to Modified Hypergeometric Exact. The different formulas are discussed on page 239.

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.

Step 8: Click the Sample Size tab again, and click 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 fish that can return positive test results, and still let us conclude that the population is free from disease. In other words, these are considered to be false positive test results. If we get fewer test-positive fish in the survey, we can still conclude that the population is free from disease; however, 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. Fish must be chosen using SRS. If a sampling frame already exists, you can use the manual technique described on page 73, or you can use the **Random Village** program on the accompanying CD (page 97). If no sampling frame exists, one must be made. If you are conducting a survey in a grouped population, use the technique described for building a sampling frame on page 99. In some situations, it may be possible to select a random systematic sample, as long as there is an appropriate management opportunity (page 77). See Chapter 5 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 above) and click 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 the Calculate button.

Step 7: A window is displayed with the results of the analysis.

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 were 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, 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, it is very unlikely that the population is free from disease. If it is large, this is consistent with there being no disease in the population. Small null and alternative probabilities suggest that the population is not free from disease, but that the prevalence is less than the minimum expected prevalence specified. The conclusion is given at the bottom of the window.

Two-stage surveys

Chapter 5 discussed the problem of drawing a simple random sample from a large population. For surveys at the district, state or national level, it is not possible to draw up a sampling frame that lists every animal or producer in the entire population. In such cases, two-stage sampling (page 82) is much more practical. At the first stage, we simply need a sampling frame that lists, for example, all the districts or villages. For those districts or villages that are chosen, we can then build a sampling frame of farmers, ponds or cages.

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 perkinosis in clams and abalone, most of the fish in most beds are not affected, and the overall prevalence in the population is very low. However, during an outbreak, a small number of beds might have a large number of clams affected, and the prevalence in those areas is very high.

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

Two-stage sampling for demonstrating freedom from disease can be used in a wide variety of situations. In this discussion, we will use the example of a farm using cage culture. The population is grouped at two levels—the cages on the farm and fish in each cage. 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 cages and amongst fish within a cage.

Step 2: Calculate the first-stage sample size (number of cages).

Step 3: Select the sample using SRS.

Step 4: Build a sampling frame of fish within the selected cages

Step 5: Calculate the sample size depending on the cage population.

Step 6: Select fish using SRS or random systematic sampling (RSS).

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 cage, and determine whether it is classified as diseased or non-diseased.

Step 12: When every cage in the sample has been classified, analyse these farm-level results to determine if the entire population is diseased or non-diseased.

Step 13: Report the results.

Sample size calculation

The sample size calculation takes place in two parts—first calculate the number of cages required at the first stage, and then the number of fish from each cage.

In addition to all the measures required for sample size calculation for small populations, there is another important measure to be considered. For each cage that is tested, it is necessary to decided whether the cage is to be classified as diseased or non-diseased. We analyse the results from fish within the cage to make this decision. However, this decision may 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 cage, the procedure (testing a sample of fish from the cage) can be thought of as a test. Just like any other test, its performance can be measured by sensitivity and specificity. The sensitivity of a cage-level test is the probability that a diseased cage will be classified as diseased. This is equal to 1 minus the type I error level. If, when testing fish within a cage, we set the type I error level to 0.05 or 5%, the sensitivity of the farm test is 0.95 or 95%. In the same way, the specificity of the cage test is equal to 1 minus the type II error level.

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

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

Step 1: Click the Sample Size tab.

Step 2: Under test sensitivity, enter the a value which is 1 minus the type I error used for selecting individual fish. For example, using a type I error level of 0.05 when selecting individual fish to test means that the farm test sensitivity is 95%.

Step 3: Enter the specificity in the same way. If the type II error level for selecting individual fish is 0.1, then the farm-level specificity is 90%.

Step 4: Enter the size of the total population. This is the total number of cages in the farm being studied (not the total number of fish).

Step 5: In the Prevalence box, enter either the prevalence or the total number of disease-positive cages, representing the maximum acceptable prevalence. Regardless of the disease in question, if the population is thought to be free from disease, the proportion of positive cages must be set to a relatively low value (usually less then 5%). This means that the number of cages that need to be tested will often be quite high.

Step 6: Click 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 232.

Step 7: Return to the Sample Size tab, and click Calculate.

The results will indicate how many cages need to be visited. The procedure for selecting fish from each of the selected cages is the same as that described above for surveys of small populations. 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 cage 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 fish in that cage. Because the sample size depends on the total population, and the population of every cage is likely to be different, the second-stage sample size should be calculated for each cage separately. If a portable computer is not available, the sample sizes for every possible cage size can be calculated before the field work, and written in a table for use in the field.

Sample size for minimum cost

When using two-stage sampling, a survey can produce results of the same accuracy using various combinations of first- and second-stage sample sizes. For instance, if we select a few cages and test many fish from each cage, it is possible to get the same accuracy as if we tested a few fish from each of many cages. By changing the type I and II error levels used for selecting the sample size for the second stage (testing fish within a cage), we are also changing the sensitivity and specificity of the cage test (used when selecting cages at the first stage). This enables us to produce a variety of 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 fish, and how much it costs to test a single cage. This was discussed in Chapter 11 (page 182). For prevalence surveys, there is a formula, used in the Prevalence program, to determine the cheapest combination. However, 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. These include the sensitivity and specificity of the laboratory test, the population of cages (first-stage population size), an estimate of the average fish population of the cages, the maximum acceptable prevalence of the disease amongst cages (first stage) and fish (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 fish, and the costs associated with sampling a single cage (see page 182).

Step 2: Pick starting values for the farm test sensitivity and specificity. The higher these values are, the fewer cages need to be tested, and the more fish need to be tested in each cage. If they are very high, there may not be enough fish in some cages to achieve this level. In general, try to make the specificity as high as possible.

Step 3: Calculate the number of cages needed, using the selected cage 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 minus the sensitivity, and the type II error to 1 minus the specificity. Change the sensitivity and specificity to those of the laboratory test, the population to the average cage size, and the prevalence to the maximum acceptable or minimum expected prevalence within the cage.

Step 5: Calculate the number of fish that need to be tested.

Step 6: Using the number of cages, and the number of fish, 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 these alternative combinations.

Step 8: Continue testing new values, until you find the one that gives the cheapest cost.

First- and second-stage sampling

At the first stage, cages must be selected using SRS from a sampling frame. Use a random number table to select from a written sampling frame or the **Random Village** program (described on page 97) to select from a sampling frame on computer disk. When using the program, set the sampling type to Simple Random, and select cages without replacement. The sample may be stratified, if this is convenient.

At the second stage, fish must again be chosen using SRS, or, if possible, RSS. For selecting from a grouped population, the technique described in Chapter 6 (page 100) can be used, either with the **Random Animal** program, or manually using a random number table.

Data analysis

The data is analysed in two stages. First, the data from each selected cage is analysed to provide a cage result, indicating that the cage is either diseased or non-diseased. Use the same approach described above for small population surveys (page 173), and record the status of each cage. When analysing the results from each cage, be sure to enter the correct population size for that farm, 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 cages have been analysed separately, the population of cages can be analysed. Use the **FreeCalc** program:

Step 1: Start FreeCalc and click the Analyse Results tab.

Step 2: Enter the cage test sensitivity and specificity.

Step 3: Enter the total number of cages 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 cages that were tested (the first-stage sample size).

Step 7: In the Number of Positive Reactors box, enter the total number of cages that were classified as diseased.

Step 8: Click the Calculate button.

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

FreeCalc options

Formulas

On the Options tab of the FreeCalc program, there is a choice of three different formulas 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. 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 is 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 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 formulas 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 formula and the infinite population formula, so the faster of the two is used. You can enter a population size above which the faster formula will be used.

15

Trainers' guide

Advice for trainers

It is often assumed that anybody who understands a subject well should be able to teach that subject to others. Unfortunately, this is not the case. There is a lot more to good teaching than just having an understanding of the subject. An understanding of the students and the way they learn is also necessary.

This chapter discusses who should be a trainer of the active surveillance techniques described in this book, and provides advice on techniques that may be used to help with the training.

Who should be a trainer

The two basic requirements for a trainer of active surveillance techniques are a good understanding of the subject and the ability to teach it to others. The most likely people to be involved in training are national or provincial level staff (epidemiologists) responsible for aquatic animal disease control. Epidemiologists or aquatic animal scientists working as development project staff may also be involved. Other people with different backgrounds may also successfully conduct training courses but, ideally, the trainer should have the following characteristics:

- A sound understanding of active surveillance, and the techniques described in this book. Experience with surveillance and sample surveys, and a knowledge of epidemiological principles, is important. However, in the absence of formal epidemiological training, a sound knowledge of all the concepts in this book and experience in conducting survey field work will provide a trainer with all the technical background necessary.
- Practical field experience. The trainer should be reasonably experienced in field techniques such as sampling and collecting specimens.
- Ability to use computers. Many of the technical calculations and analyses depend on the use of computer programs. The trainer should be familiar with computers and the programs being used (including a database program such as Epi Info), and be able to solve the types of computer problems that may arise.
- An ability to communicate easily with trainees. The trainer should be reasonably fluent in the trainees language, and understand the social and cultural issues that may impact on field work and training.
- A respect for the skills and experience of trainees, and the knowledge of producers.
- Enthusiasm for teaching and for active surveillance and survey field work. Enthusiasm is contagious.
- Experience with or an understanding of basic teaching techniques. These are discussed in more detail below.

Training skills

Every trainer has their own style. While you may wish to copy some good points from those who have taught you, there is no point trying to imitate them completely. Some people are more serious and strict, while others are casual and like to joke a lot. Both can make good trainers, as long as they are comfortable with the way they do things. Whatever your style, try to consider whether anything you do makes it more difficult for students to learn. If so, try to change it. Whatever your training style, it is always possible to learn new tricks and techniques, and to improve the effectiveness of your training.

Training is a process of communication, both from the trainer to the students, and from students to trainer. There are many ways to make this communication more effective. You are not simply transferring information to others, but you are trying to help the participants understand and solve problems, using tools they already have and new tools you give them. Some tips to encourage effective communication include:

- Keep eye contact with the participants. Don't talk with your back to them while you are working on the blackboard.
- Show interest in what you are saying, and make it more like a story. Don't speak in a droning voice. Speak clearly and loudly, but not too fast.
- Vary activities regularly, so participants don't become bored. Don't spend too much time in the classroom.
- Make sure that the training environment is comfortable and not too distracting.

Lesson planning

One of the keys to successful teaching is good organisation and planning. Regardless of how much technical knowledge the trainer has, if they are not organised, or are unsure what they are doing next, the students will find it difficult to learn.

A well-organised, carefully thought-out lesson plan will ensure that both the students and the trainer know exactly what is happening, and that effective learning can take place. Lesson plans have been prepared for the training courses suggested, and are presented in Chapter 16. These should be thought of only as a guide, as the specific needs of each training course will be different. You may use some of these lesson plans if they are appropriate, or develop your own. The structure that has been used for the lesson plans in this book is as follows:

- Title. The title of the lesson, so students know what to expect.
- Location. Where the lesson is to be conducted (classroom, village etc.)
- Duration. The expected time of the lesson. This can vary greatly depending on the level of knowledge and experience of the students.
- Objectives. These are the things that the student should be able to do at the end of the lesson.
- Key points. These are highlights from the lesson and things to keep in mind when teaching.
- Page references. The relevant pages from this book are listed for easy reference.
- Teaching methods. This is an outline of the activities during the class and the methods used to achieve the objectives.

Activities

Many of the items listed under teaching methods refer to activities. These may be games, discussions, role plays, field trips etc., as described in the next section. For each activity, an activity sheet has been included in Chapter 17. The activity sheets explain the purpose of the activity, how to run it, what equipment is needed, and suggested follow-up questions for discussion.

Teaching techniques

There are two main ways that people learn. The first is through being told something by somebody else. As we all know, it is easy to forget something that we are told. The second is to discover something on our own, either by doing something new, or using things that we already know to understand something in a new way. When we discover new knowledge on our own, it is much easier to remember. This is partly because it is fun, and gives us a feeling of satisfaction.

These two types of learning are described as 'teacher-centred learning', where all the knowledge comes from the teacher, and 'student-centred learning' where the knowledge is either discovered by the students or comes from a new understanding of things they already know. In many societies, teacher-centred learning is the most common way that people are expected to learn. Listening to lectures, taking notes from a blackboard and memorising lists of things has been used successfully for years. However, there are two problems with this approach. First, it is not fun, and because the facts being learnt are not connected to anything, they are easy to forget. The second problem is that things are not placed in context when they are taught. This means that it is harder for the students to use the facts in a real-world situation to solve problems.

Student-centred learning starts with the students' own experience of problems in the real world. Guided by the teacher, students are encouraged to come up with solutions to these problems, either using their own experience, sharing the experiences of their fellow students or making new connections with knowledge they already have. Naturally, the teacher is required to provide new information. However, if the new information provided by the teacher is given when students actually want or need it to solve a problem they are dealing with, the new information immediately has a useful purpose and is placed in context. These new facts will not be easily forgotten and will be able to used to solve other similar problems faced outside the classroom.

When either teacher or students are not used to student-centred learning techniques, the method can be quite difficult at the start. However, after trying for a while, both will realise that it makes teaching and learning more fun, and that the things being taught are useful. When running a training course, a quick check to see if you are using student-centred learning techniques is to listen for a while. If the students are doing most of the talking, then it's working the right way. If the teacher is doing most of the talking, something is wrong.

How, then, can a teacher encourage student-centred learning? The main technique is to use the experience that the students already have, and to present them with problems that they have already faced. To solve these problems they need to think, discuss with other students and discover new information. A skilful teacher is able to guide students so that they rarely need to be taught anything at all—most of the time, they discover things for themselves. It is surprising how often students already have a basic intuitive understanding of apparently complex concepts.

A range of different techniques and suggestions is presented below, to help trainers use student-centred learning effectively.

Learning landmarks

Effective learning has more to do with organising information than memorising new information. If a student understands the relationships between the different things they already know and the new things they learn, they are able to use that information to help with everyday tasks. Trainers need to help students organise the information. (How does the thing that is being taught now relate to other things that I already know and have previously learnt? How will I be able to use this knowledge?)

When people travel, landmarks help them navigate and know where they are. Learning landmarks are pointers for students that show where students are coming from, where they are now, and where they are heading. If students always know exactly where they are, it is much easier to organise the information. If they get lost, and don't know where they are going, or how this information is to be used, they don't know how the information relates to other things they know and they don't know how to properly organise it. Unless the connection is made, the information may be wasted.

There are three good ways to provide students with learning landmarks. The first is to give them a map of the lesson, so they can chart a course. At the beginning of every day, or every training session, give the students a brief outline of what is going to be covered. Make sure that it is clear how each topic is related the previous or the next topic, and why it is relevant.

The second technique is to regularly fix your position along the way. As each topic is dealt with, make sure the students know where they are up to. Introduce the topic and, better still, tick off the previous topics on the board. As each new concept is introduced, provide one or two examples of how this is relevant to the real world. The examples used throughout this book are there to help the readers understand how the topic being discussed is related to the real world.

The third way to provide learning landmarks is to look back over the journey when it has ended. At the end of each lesson, run through all the topics covered, and highlight how they relate to each other.

Reinforcement and practice

Most subjects, including active surveillance for aquatic animal diseases, use knowledge that is built up layer by layer. It is only possible to learn the next level once the previous one is well understood. If new information is taught before the earlier information is properly learnt, the foundations become unsteady and students can become confused.

A good way to ensure that all the earlier information is well understood is to continually reinforce and practise it. Every time students are asked to remember something they have learnt, and to use it to solve a new problem, makes it harder to forget. The teacher should therefore take every opportunity to include previous concepts in new exercises and problems, to help students practise them.

Warmers

In student-centred learning, the students are expected to do most of the work, while the teacher guides them, providing new information when it is needed and giving them direction. At the start of a lesson, students are often not prepared to take this active role. They are not yet thinking about the problems that need considering, and they might feel shy or inhibited about speaking out in front of the group.

Warmers are exercises that are designed to 'warm up' the students, to start them thinking about the problems and topics to be dealt with, and to make them comfortable speaking aloud and discussing things with other students. Warmers should be relatively short exercises that involve a lot of student activity and, above all, are fun. It is best to use a warmer that deals with issues from the previous lesson, so students can practise what they have learnt while they prepare them for the upcoming topics.

Any of the following activities can be used as a warmer, but games and competitions are often the best. Warmers can also be useful during village interviews, to help farmers relax and feel comfortable about speaking out loud. A survey team that has had experience with warmers during its training will be much better able to use them during village interviews.

Questions and answers

The simplest way to get students to actively participate in the lesson is to ask direct questions. Questions may be asked of the group as a whole, or directed to individuals. Targeting individuals forces them to participate and avoids the problem of nobody being willing to speak first.

Questions can be used in two ways. First, questions are good for introducing a new topic. Pose a question on how to deal with a problem (for instance, how to select animals, how to collect blood from a fish, how to get participation from women during a village interview). This can then lead into a full discussion of the issue, maybe using some of the other techniques listed here.

The other way questions can be used is to check whether a topic has been properly understood. Using a new context or different example, ask students to use their new knowledge to solve a problem or explain one aspect of the topic. If a student is unable to do this, or makes mistakes, ask another student to comment or help. If several or all students show that they don't understand, the topic hasn't been adequately taught and you will have to think of a better way to explain or practise the concepts.

Using questions to check students' understanding is a quick and simple way to evaluate the effectiveness of your teaching. As a trainer, the only way to improve is to understand where there are weaknesses in your training, and think of new ways to overcome them.

Games or competitions

Much of the training involves serious or complex issues, but you can make it more enjoyable by using games or competitions. These allow students to relax and have fun, but still practise the ideas they have been learning, or learn new ones through the game.

Competitions can be effective warmers, such as the knowledge quiz competition (Activity 22). This provides students with an opportunity to recall and practise information they have learnt, as well as giving them a feeling of pride. You can also use the quiz competition to assess your students level of understanding of concepts.

As well as being used as warmers, games can be used to introduce or practise new concepts. The sampling jigsaw game (Activity 7) is an example of this. Through the game, students have an opportunity to see for themselves the effect of different sampling strategies, and have fun at the same time.

Group discussions

In group discussions, the class is divided up into a number of smaller groups. Ask each group to discuss a topic or consider some questions, and to record their ideas on a piece of paper as they go. At the end of an allotted time, one member of each group presents the findings to the rest of the class.

Group discussions are an opportunity for students to explore and discuss the issues with each other, to relate their own experiences and to share those of others. They are a good way for students to discover how much they already know about a subject.

Groups may be made up of just two people, but more usually have between 4 and 6. Before the discussion, you should make very clear the aim of the discussion and the topic or questions under consideration. During the discussion, you should move from group to group, monitoring the topic and checking that the discussion hasn't strayed onto something else.

While each group is reporting, their ideas should be recorded on the board for all to see. Finally, the trainer needs to summarise and organise the ideas, to give structure to the conclusion.

Brainstorming

Brainstorming is a technique in which class is given a topic and students are asked for the first ideas that come into their heads. It is used to collect a lot of ideas quickly and to encourage participation. Brainstorming exercises may be used as warmers, or to introduce a new topic.

To start a brainstorming session, present a question or idea to the class. Ask each student to respond quickly, using just one or two words, with their own ideas. Write down their answers as they go. Tell students that there is no such thing as a wrong answer in a brainstorming session—its purpose is just to collect lots of ideas.

To be successful, brainstorming should be done very quickly and with some excitement. You should choose the order of students in an unpredictable way, and jump to the next student quickly. Don't let any discussions or argument start at this stage, just collect the ideas.

When the list is complete and there are no more ideas, the list can be used as a basis for the next activity or session, depending on the objectives of the lesson.

Ranking

Ranking activities are used to set priorities or arrange things in order of importance. Ranking may be used during village interviews to identify priority livestock diseases, but it can also be a useful tool during training courses.

There are many ways to run a ranking activity, and these can be adjusted to the specific situation. Normally, the activity starts with the creation of a list, for instance a list of the common diseases affecting aquatic animals in the study area. Ask participants to identify which of the listed diseases are the most important, and which are less important. Make clear what 'important' means. You might choose to define it as 'most likely to cause death', 'causing the greatest financial loss', 'causing the most inconvenience', or 'most expensive to treat'. If you wished, you could conduct separate ranking exercises for each of these different criteria.

Each participant scores each disease on a piece of paper in order of importance. The most important disease gets a score of 1, the second most important disease gets a score of 2 and so on. When all participants have finished, add up the scores of all participants for each disease. The disease with the lowest score is the most important, through to the disease with the highest score, which is the least important.

One important use of ranking is to help participants identify their own preconceptions and biases. If a disease ranking exercise such as the one described is carried out, and then the same exercise is used with producers during a trial interview, the differences in the diseases and ranks between the participants and the producers can be highlighted. These differences represent differences in attitude or experience of the fisheries services compared to the producers.

Role playing

Role playing involves asking participants to act out some scene or situation. It allows participants to think about issues that are likely to be raised during field work, and to develop appropriate ways of dealing with them while still in a safe, non-threatening environment. Role playing is also a good way to get people participating and break down inhibitions.

A role play is like a very short play, acted by the participants. Give each of the players clear instructions about who their character is, and what position or attitude the character has. The players then act out the play, making up the dialogue as they go along. Usually, a role play involves some sort of conflict or disagreement that the actors need to resolve.

Activity 17 is an example of a role play. Participants use the play to explore a situation that is likely to arise during field work. Producers are reluctant to let the survey team collect specimens from their animals, and the survey team needs to explain why they should let them.

Field trips

The purpose of much of the training is to prepare participants for survey field work. By far the best way to do this is to actually do the work. Field trips give the participants an opportunity to practise the skills they have learnt, and to experience for themselves the problems and limitations of working in the field.

Field trips can provide valuable information for survey planning as well, as they can act as small pilot surveys. The activities of the survey staff and the responses of the producers can be assessed, and problems identified and corrected. However, a training field trip is necessarily very different from a real village visit during a survey. Participants are generally less confident with their new knowledge, and more importantly, there will usually be many more participants on the field trip than in a normal village survey visit.

Very good organisation is therefore necessary if the field visit is to be successful. The purpose of the trip and activities to be carried out should be carefully explained, and the roles and responsibilities of each of the participants must be clearly assigned.

While carrying out training activities in front of producers, the trainer should be aware of the sensitivities of both producers and participants. For instance, if inexperienced participants are practising handling animals or examining ponds, don't let too many participants use the same pond, or ponds belonging to a single farmer. Both animals and owner are likely to become stressed.

It is also important to try to maintain the participants' status while working in the village. If participants are seen by producers to be unskilled or ignorant, both the producers and the participants' confidence will be undermined. Teaching during field trips should therefore concentrate heavily on encouraging students to demonstrate and practise their knowledge and skills. Do this in a positive, supportive way, avoiding direct criticism.

After the field trip, there should always be a time set aside for discussion. A few questions should be used to stimulate the group to talk about their experience. In particular, it is important to identify any problems encountered and discuss how they could be addressed or avoided. Make a point of identifying the good things as well as the problems. Activity 18 describes a trial village visit.

Practical activities

In addition to field visits, training should include as many real-life, practical activities as possible. The idea is for participants to learn 'on-the-job', and to feel that the activities they are doing during the training are not just made-up exercises, but are actually contributing to the aim of the work.

One example is the tasks involved in survey planning. After the principles have been taught and practised with various exercises, they can be used, during the training, to prepare for the real survey. For instance, random selection of first-stage units (villages) in a two-stage prevalence survey can be done by the group during training. The participants work together to obtain and check the sampling frame, and then use the software to calculate sample size and select the required number of villages.

A similar approach can be used, after the survey, for data analysis. While the principles of analysis can be taught, analysing the actual data collected during field work will give the exercise much more meaning. There is also the advantage of having a larger group of people entering and analysing data, making it faster, and easier to use duplicate entry checking systems.

Field work

Although not formally part of the training course, learning continues during the actual field work of any survey. The field work should start as soon as possible after the training course, and be seen as a logical extension of it. Each survey team should be encouraged to hold a brief meeting at the end of each visit, to discuss problems that occurred, and how the procedures could be improved to address these problems.

The trainer should participate in the field work alongside the survey team as much as possible. At a minimum, you should accompany the team on a number of visits, especially during the early part of the survey. This is to help continue the training and refine the skills of the survey teams, and also to identify problems, errors, or poor practices that may have slipped into the work routine. If these are corrected at the start of the survey work, the quality of the survey will not be compromised.

16
Lesson plans

The lesson plans in this chapter are provided as a resource for trainers. They are divided into three separate courses.

Course 1, ‘Active surveillance, survey planning and sampling’, is designed for national staff, survey planners and coordinators, to prepare them for the task of coordinating surveys.
Course 2, ‘Field techniques for aquatic animal disease surveys’, is designed for field staff and the survey teams. This course provides training in all the practical data collection activities required. The design of these courses assumes that the participants of Course 1 will also participate in Course 2, and that the two courses will be followed almost immediately by the actual field work of a survey.

Course 3, ‘Computerised data management and analysis, and reporting’, is designed for national staff and coordinators. The aim is to run this course for the same participants as Course 1, soon after the completion of field work. The data collected can then be used as material for training.

After completing the training courses, and participating in the field work, trainees should be in a position to organise and conduct further aquatic animal disease surveys as required.

The lesson plans therefore provide a structured syllabus for teaching the techniques described in this book. However, they are not appropriate to every situation, and not every training course needs to cover all the material. While some trainers may wish to use the lesson plans much as they are presented, the plans can also be used simply to provide a guide and stimulate ideas for the running of similar courses. In particular, trainers should structure the lessons and use activities according to their preferences and the needs of the participants.

Course 1: Active surveillance, survey planning and sampling

Participants

National staff, survey planners and coordinators

Course structure

Lesson 1: Introduction to aquatic animal health information and surveillance
Lesson 2: Measures of disease
Lesson 3: Surveys and inference
Lesson 4: Sampling
Lesson 5: Sampling in practice
Lesson 6: Sampling aquatic animals
Lesson 7: Introduction to survey planning
Lesson 8: Trial survey
Lesson 9: Prevalence surveys
Lesson 10: Incidence rate surveys
Lesson 11: Surveys to demonstrate freedom from disease

Lesson 1: Introduction to aquatic animal health information and surveillance

Duration: 2 hours

Location: Classroom

Objectives

- Discuss the use and importance of information on animal diseases.
- Examine how information is collected in the current system (passive surveillance).
- Identify weaknesses in the collection of information.
- Introduce the concept of active surveillance to address these weaknesses.

Key points

- Many fisheries staff don't realise the importance of disease information to their jobs. Try to emphasise the relevance for participants' jobs of information and the need for good information.
- Passive surveillance systems suffer from under-reporting and bias.
- Active surveillance can overcome these problems.

Page references

Aquatic animal disease surveillance (page 27)

Teaching methods

- Introduce the course.
- Group discussion on the need for aquatic animal disease information. How do the participants use information in their own jobs? Who else needs information, and about what?
- Develop flowchart of collection of disease information. Have one participant draw the steps in information flow, while the others suggest different sources and paths.
- Use direct questions to investigate possible weak spots in the flow of information. Highlight the problem of under-reporting of diseases.
- Use group discussions to list possible reasons why a case of disease may not appear in national-level records.
- Explain the term passive surveillance.
- Use questions to find how it could be done better, and introduce the idea of active surveillance.

Lesson 2: Measures of disease

Duration: 3 hours

Location: Classroom

Objectives

- Understand prevalence and how it is measured.
- Understand incidence rate and how it is measured.
- Explore the relationship between incidence rate and prevalence.
- Consider examples of when to use incidence rate and when to use prevalence.
- Be able to interpret sensitivity and specificity as measures of a test's performance.
- Understand the difference between apparent prevalence and true prevalence.

Key points

- Prevalence is the number of cases of disease at one point in time.
- Incidence rate is the number of new cases of disease over a period of time.
- Diseases of long duration have a higher prevalence.
- Make sure students understand basic principles of immunity, antibodies and serological tests.
- Diagnostic tests usually make a small number of mistakes.
- Sensitivity is the proportion of true positives that a test detects; specificity is the proportion of true negatives.
- Sensitivity and specificity can be used to correct for the mistakes of a test, and calculate true prevalence.

Page references

Measures of disease (page 56)
Diagnostic tests (page 61)

Teaching methods

- Visual aids, examples and direct questions to introduce prevalence and incidence rate.
- Example calculations.
- Group discussion on which measures to use for two hypothetical situations.
- Activity 2: Sensitivity and specificity.
- Questions and example on apparent prevalence versus true prevalence.

Lesson 3: Surveys and inference

Duration: 3 hours

Location: Classroom

Objectives

- Explain the principle of a survey.
- Introduce the concepts of population and sample.
- Contrast surveys with complete counting of the population (censuses).
- Explain the process of inference.
- Define bias and explain the need for representative samples.
- Discuss estimation and precision.
- Identify the role of sample size in survey accuracy.

Key points

- Surveys examine only a small sample of the population.
- The sample is used to make inferences about the population.
- Inference can be wrong, giving a biased result.
- Representative samples ensure that inference is not wrong.

Page references

Disease surveys (page 48)

Teaching methods

- Activity 1: Classroom census and survey for average age.
- Questions on population and sample.
- Explain inference using visual aids, stressing that a survey estimate can be wrong.
- Explain bias, accuracy and precision, using visual aids.
- Activity 3: Biased sampling survey.
- Activity 4: Sample size effect and surveys.
- Group discussion: How are samples selected now? Are the samples representative? Ask participants to list the different ways they have selected samples in previous work. Have them consider potential for bias.
- Discuss techniques, highlighting potential for bias.

Lesson 4: Sampling

Duration: 3 hours

Location: Classroom

Objectives

- Understand the need for random sampling to reliably select a representative sample.
- Be able to distinguish probability from non-probability sampling techniques.
- Select random numbers using physical randomisation, random number tables, and a computer.
- Introduce the concepts of probability proportional to size sampling and stratified sampling.
- Understand the requirements of a good sampling frame.

Key points

- Random sampling is the only way to reliably select a representative sample.
- In simple random sampling, all elements have the same probability of being selected.
- Computers can simplify the task of selecting a random sample.
- Systematic sampling can sometimes be used to avoid the need for a sampling frame.
- Sampling frames should include every member of the population, once only.
- The sampling frame determines the level of inference.

Page references

The need for random sampling (page 70)
Random sampling techniques (page 72)
Sampling frames (page 81)

Teaching methods

- Assess the level of understanding of basic probability with questions.
- Introduce the concept of chance and probability.
- Examples of random outcome, using dice, cards, coins.
- Explain how we don't know the outcome of a single trial, but in the long run, we can predict what will happen over many trials.
- Activity 5: Random numbers. Predicting the outcome.
- Demonstration of using a random number table and computer-generated random numbers.
- Activity 6: Classroom age survey using random sampling.
- Discuss repercussions of an incomplete sampling frame or one with duplications.
- Activity 7: Sampling jigsaw game.

Lesson 5: Sampling in practice

Duration: 2 hours

Location: Classroom

Objectives

- Consider the problems of sampling from large populations.
- Understand the principles and advantages of two-stage sampling.
- Practise the actual selection of farms or villages from a sampling frame using a computer.
- Explain the meaning of replacement and without-replacement sampling.

Key points

- Building a sampling frame for large populations is usually too expensive or not possible.
- Two-stage sampling removes the need for a complete sampling frame, and makes field work easier.
- Computers can be used to select a sample from a sampling frame.

Page references

Sampling from a sampling frame (page 81)
Two-stage sampling (page 82)

Teaching methods

- Questions about problems with sampling from large populations.
- Explain the benefits of two-stage sampling.
- Activity 8: Selecting sample villages for survey (if using SRS or PPS).

Lesson 6: Sampling aquatic animals

Duration: 2 hours

Location: Classroom

Objectives

- Consider the problem of random sampling of individual aquatic animals.
- Identify the key principles we are trying to achieved in sampling.
- List management opportunities that may facilitate random sampling.
- Understand spatial sampling techniques
- Propose sampling strategies for a range of relevant production systems.

Key points

- Random sampling of individual aquatic animals is often very difficult.
- Systematic sampling may be used during some management activities.
- Spatial sampling may be used for static populations (e.g. molluscs).
- Compromises are often necessary, but efforts should be made to ensure a representative sample and avoid bias.

Page references

Sampling aquatic animals (page 105)

Teaching methods

- Group discussion: Sampling individual animals. Consider the problems and list the constraints to using traditional sampling approaches.
- Brainstorming: Identify situations where random or systematic sampling may be possible.
- Group activities: Design a sampling strategy for a range of specified situations.

Lesson 7: Introduction to survey planning

Duration: 2 hours

Location: Classroom

Objectives

- Consider the steps involved in running a survey.
- Understand the process of framing the question, and determining how to answer it.
- Understand the factors that influence sample size considerations.
- Understand the value of pilot surveys.
- Consider issues of analysis and reporting before the start of the survey.
- Plan a trial survey.

Key points

- The survey question must be able to be answered using a measurable value.
- Variance, precision and confidence are important factors in determining sample size.

Page references

Outline of survey procedures (page 168)

Teaching methods

- Group discussion: What are the major steps in running a survey?
- Organise and add any missed steps.
- Group discussion: The effect of variance. Present an example of two populations, such as one class in a school, and all the people in a village. Consider a survey to estimate the average age in each of the populations. How many people would be needed from the school class, and how many from the village?
- Group discussion: Things that need to be arranged before a survey.
- Compare with checklist.
- Explain the need for pilot survey. Examples of possible survey problems that would be avoided by doing a pilot survey.
- Questions to stimulate thought on the importance of reporting.

Lesson 8: Trial survey

Duration: 1 day

Location: Classroom and area of village/city in which training is taking place

Objectives

- Plan, implement and analyse a real survey.
- Develop an appropriate question and a way to answer it.
- Practise concepts of building a sampling frame.
- Carry out random selection from the frame.
- Practise interview skills and data collection.
- Perform simple data analysis.
- Practise oral reporting skills.

Key points

- This is the first real-world survey to be conducted during the course. Good organisation is important to maintain the confidence of the participants.
- Any appropriate unit of interest, disease, species or characteristic may be used, depending on the situation, although some changes in design may be necessary.

Page references

Outline of survey procedures (page 168)
Sampling (Chapter 5)
Prevalence (page 58)

Teaching methods

- Explain the activities carefully beforehand.
- Activity 9: Local survey.
- Presentation of results by different groups.
- Discussion of problems encountered during the survey.

Lesson 9: Prevalence surveys

Note: This lesson is necessary only if prevalence surveys are planned.

Duration: 3 hours

Location: Classroom

Objectives

- Understand the basic steps in carrying out a one- or two-stage prevalence survey.
- Be able to decide on the best survey design to use in a given situation.
- Calculate sample size, and understand the factors that are necessary.
- Decide on appropriate stratification variables.
- Select elements from a sampling frame.

Key points

- The design chosen depends on the sampling frame available.
- Variance and prevalence estimates are required for the sample size calculation. These usually need to come from previous surveys.
- The computer can be used to calculate sample size.
- Details of second-stage sampling are discussed in the next training course.

Page references

Prevalence surveys (Chapter 11)

Teaching methods

- Questions: Revise the need for two-stage sampling in large populations.
- Questions: Revise the concept of prevalence and when to use it.
- Present examples of different situations, and ask questions about how to carry out the survey (sampling frame available versus no sampling frame).
- Use the computer to calculate sample size. List the parameters necessary for the calculation, and have the group decide on appropriate parameters. Discuss their choices and your recommendations.
- Group discussion or questions on what an appropriate stratification variable would be.
- Highlight the problems of selecting elements from grouped populations, using examples. The solution to these problems will be discussed during the second training course (Course 2, Lesson 7).

Lesson 10: Incidence rate surveys

Note: This lesson is necessary only if incidence rate surveys are planned.

Duration: 3 hours

Location: Classroom

Objectives

- Appreciate the problems of collecting incidence rate measures.
- Appreciate difficulty of remembering events from many years ago.
- Know the limitations on which diseases may be studied.
- Understand the procedure for carrying out a retrospective disease outbreak survey.
- Use a computer to calculate sample sizes for a disease outbreak survey.
- Know the key pieces of data that must be collected.
- Understand the concept of two-sample analysis.
- Be able to identify appropriate data sources for two-sample analysis.

Key points

- Incidence rate can be measured at the animal level or pond/farm/village level. Village/farm-level incidence rate is easier to measure, and often more relevant to disease control programs.
- Surveys of past events are only reliable if the events are easily remembered.
- Sample size calculation for disease outbreak surveys depends on the difference in the average time since the last outbreak.
- Interview procedures for collecting the information will be covered in the next course (Course 2, Lesson 4).
- Two-sample analysis requires two different, independent sources of information on village or farm disease outbreaks.
- These can be used to estimate the total number of disease outbreaks.

Page references

Incidence rate surveys (Chapter 13)

Teaching methods

- Revise the meaning of incidence rate and the difference between incidence rate and prevalence.
- Use examples to distinguish between animal-level and village/farm-level incidence rate.
- Describe the survey procedure for village disease outbreak surveys.
- Use examples to explain the process of calculating sample size.
- Use a computer to practise calculating sample size.
- Use visual aids and the example of fish in a lake to explain the principle of two-sample analysis.
- Use questions and examples to determine what sort of data sources are independent.

Lesson 11: Surveys to demonstrate freedom from disease

Note: This lesson is only necessary if surveys to demonstrate freedom from disease are planned. The concepts covered are more complex than those in other lessons.

Duration: 3 hours

Location: Classroom

Objectives

- Understand the situations when it may be necessary to demonstrate freedom from disease.
- Understand the problems of using sampling and imperfect tests to demonstrate freedom from disease.
- Understand the concept of minimum expected (maximum acceptable) prevalence.
- Understand types I and II error and their importance in survey design.
- Understand the steps in a single-stage survey.
- Be able to calculate sample sizes using a computer.
- Understand how to use two-stage sampling in large populations.
- Understand the concept of clustering of disease.
- Be able to calculate optimal sample sizes for two-stage surveys (advanced groups only).

Key points

- It is not possible to prove that a population is free from disease, if using imperfect tests.
- Surveys are able to demonstrate that there is a low probability that, if the disease exists, the prevalence is greater than a specified level.
- Two-stage sampling can be used for large populations, and populations with disease clusters.

Page references

Surveys to demonstrate freedom from disease (Chapter 14)

Teaching methods

- Group discussion: Freedom from disease and when it may be necessary to be able to demonstrate it.
- Revise the concept of sensitivity and specificity.
- Activity 10: Classroom survey to demonstrate freedom. Use to stress that proof is impossible, and there must be a maximum acceptable prevalence.
- Use a computer to calculate sample size for single-stage surveys.
- Give examples of how disease clusters in a population.
- Questions on how to survey a large population (two-stage sampling).
- For advanced groups, demonstrate optimal two-stage sample size calculation.

Course 2: Field techniques for aquatic animal disease surveys

Participants

This course is designed for field staff, survey teams (including national staff, survey planners and coordinators from Course 1). The first three lessons cover much of the material presented in the first lessons of Course 1. At the end of the first course, split the participants into three groups, and have each group prepare one lesson. The members of the group can then share the responsibility of running the lesson and presenting the material. Be sure that they have a few days to prepare.

Course structure

Lesson 1: Introduction
Lesson 2: Surveys and inference
Lesson 3: Random sampling
Lesson 4: Village interviews
Lesson 5: Ranking and village outbreaks
Lesson 6: Trial village interview
Lesson 7: Selecting random elements from group populations
Lesson 8: Sample collection, processing and preservation
Lesson 9: Trial village visit (interview and specimen collection)
Lesson 10: Preparation for field activities

Lesson 1: Introduction

Note: Parts of this lesson may be presented by the participants from Course 1.

Duration: 2 hours

Location: Classroom

Objectives

- Discuss the use and importance of information on aquatic animal diseases.
- Examine how information is collected in the current system (passive surveillance).
- Identify weaknesses in the collection of information.
- Introduce the concept of active surveillance to address these weaknesses.
- Appreciate the balance between data quality and ease of collection.
- Be able to identify appropriate sources of data for different questions.
- Understand the advantages of village interviews for rapid, reliable data collection.

Key points

- The first part of this lesson is a summary of Lesson 1 in Course 1, explaining the need for information on animal diseases.
- Passive surveillance systems suffer from under-reporting and bias.
- Active surveillance can overcome these problems.
- Village interviews of aquatic animal producers can draw on the collective experience and memories of all producers, and get good quality information in a short time.

Page references

Aquatic animal health information (page 16)
Types of data and quality of data (page 120)

Teaching methods

- Introduce the course.
- If appropriate, invite participants from the first course to lead some of the sessions. Training is one of the best ways of learning. Monitor the performance closely and correct any mistakes, being careful not to undermine the confidence of the presenter.
- Group discussion: The need for aquatic animal disease information. Use direct questions to investigate possible weak spots in the flow of information. Highlight the problem of under-reporting of diseases.
- Explain the term passive surveillance.
- Use questions on how to improve surveillance; introduce active surveillance.
- Group discussion: Possible data sources for aquatic animal disease information. Nominate sources, and discuss aspects of reliability and difficulty of collection.
- Use examples of different types of information and ask questions to identify the best data source to use.

Lesson 2: Surveys and inference

Note: Parts of this lesson may be presented by the participants from Course 1.

Duration: 3 hours

Location: Classroom

Objectives

- Explain the principle of a survey.
- Introduce the concepts of population and sample.
- Contrast surveys with complete counting of the population (censuses).
- Explain the process of inference.
- Define bias and explain the need for representative samples.
- Discuss estimation and precision.
- Identify the role of sample size on survey accuracy.

Key points

- Surveys examine only a small sample of the population.
- The sample is used to make inferences about the population.
- Inference can be wrong, giving a biased result.
- Representative samples ensure that inference is not wrong.

Page references

Disease surveys (page 48)

Teaching methods

- If appropriate, invite participants from the first course to lead some of the sessions. Training is one of the best ways of learning. Monitor the performance closely and correct any mistakes, being careful not to undermine the confidence of the presenter.
- Activity 1: Classroom census and survey for average age.
- Questions on population and sample.
- Explain inference using visual aids, stressing that a survey estimate can be wrong.
- Explain bias, accuracy and precision, using visual aids.
- Activity 3: Biased sampling survey.
- Activity 4: Sample size effect and surveys.
- Group Discussion: How are samples selected now? Are the samples representative?
- Discuss techniques, highlighting potential for bias.

Lesson 3: Random sampling

Note: Parts of this lesson may be presented by the participants from Course 1.

Duration: 3 hours

Location: Classroom

Objectives

- Understand the need for random sampling to reliably select a representative sample.
- Be able to distinguish probability from non-probability sampling techniques.
- Select random numbers using physical randomisation, random number tables, and a computer.
- Introduce the concepts of probability proportional to size sampling and stratified sampling.
- Understand the requirements of a good sampling frame.

Key points

- Random sampling is the only way to reliably select a representative sample.
- In simple random sampling, all elements have the same probability of being selected.
- Computers can simplify the task of selecting a random sample.
- Systematic sampling may be used to avoid the need for a sampling frame.
- Sampling frames should include every member of the population, once only.
- The sampling frame determines the level of inference.

Page references

The need for random sampling (page 70)
Random sampling techniques (page 72)
Sampling frames (page 81)

Teaching methods

- If appropriate, invite participants from the first course to lead some of the sessions. Training is one of the best ways of learning. Monitor the performance closely and correct any mistakes, being careful not to undermine the confidence of the presenter.
- Assess the level of understanding of basic probability with questions.
- Introduce the concept of chance and probability.
- Examples of random outcomes, using dice, cards, coins.
- Explain how we don’t know the outcome of a single trial, but in the long run we can predict what will happen over many trials.
- Activity 5: Random numbers. Predicting the outcome.
- Demonstration of using a random number table and computer-generated random numbers
- Activity 6: Classroom age survey using random sampling.
- Discuss repercussions of an incomplete sampling frame or one with duplications.

Lesson 4: Village interviews

Duration: 3 hours

Location: Classroom

Objectives

- Be able to organise a village interview of producers.
- Identify people with the right skills to lead an interview.
- Be able to use techniques to get good information from village producers.
- Understand how to encourage all producers to participate in the interview.
- Know the appropriate order for conducting an interview.
- Be aware of potential problems that may arise, and how to address these problems during the introduction.
- Build a grouped population sampling frame during a village interview.

Key points

- The aim is to have all village producers attend the village interview.
- Good organisation is needed to ensure that villagers know when the meeting is and that they attend.
- Obtaining good quality information requires skill and practice.
- Every effort should be made to ensure that producers are happy to participate in the survey and any future surveys.
- The introduction to the interview can be used to avoid problems, by explaining the purpose of the survey and addressing producers' concerns.
- Building a complete sampling frame requires careful and persistent questioning.

Page references

General guidelines (page 130)
Building a village sampling frame (page 99)

Teaching methods

- Use questions to revise the advantages of using a village interview to collect information.
- Group discussion: Problems with interview data and possible ways to overcome these problems.
- Group discussion: Problems with cooperation, during this interview and in the future, and ways to overcome them.
- Present the typical order of an interview.
- Activity 11: Role play introduction to village interview.
- Use questions to revise the idea of a sampling frame.
- Demonstrate data collection forms for building a sampling frame.
- Activity 12: Build a mock sampling frame in the class.

Lesson 5: Ranking and village outbreaks

Duration: 3 hours

Location: Classroom

Objectives

- Rank disease priorities or other information from the village.
- Understand techniques for determining the date of a disease outbreak in the past.
- Be able to use village histories and village calendars to help producers determine the date of an outbreak.
- Understand the need for censoring times when collecting village outbreak data, and how to collect them.

Key points

- Ranking can be used to determine the importance of different diseases.
- When collecting information about outbreaks, make sure that the producers clearly understand the disease being discussed.
- Village histories can help identify the year of an outbreak.
- Village calendars can help identify the month or season of an outbreak.
- Censoring times must be collected for villages that have had no outbreaks.

Page references

Collecting information from people (Chapter 8)

Teaching methods

- Activity 13: Disease ranking.
- Discussion: What criteria may be used to determine which diseases are important?
- Demonstration of data recording sheets for disease ranking.
- Example of asking for the usual months for different diseases.
- Activity 14: Retrospective questions.
- Examples of building village histories and calendars.
- Demonstrate and explain the use of the data recording sheet for village outbreaks.
- Discuss ways to determine censoring times for villages with no outbreaks.

Lesson 6: Trial village interview

Duration: Half day

Location: Village

Objectives

- Gain experience in conducting a village interview.
- Develop confidence addressing producers.
- Present an introduction that addresses producers' concerns.
- Build a sampling frame.
- Determine disease priorities in the village.
- Determine the normal time of occurrence for important diseases.
- Determine the date of the most recent outbreak of a particular disease.

Key points

- Careful organisation is important. The village must be notified beforehand, and the best time for a meeting decided. Unlike a normal interview, there will usually be a large number of participants involved. Each must clearly understand their role and responsibilities.

Page references

Collecting information from people (Chapter 8)

Teaching methods

- Preparation for visit, assigning responsibilities.
- Activity 15: Trial village interview.
- Discussion of successes, problems, and recommended solutions.

Lesson 7: Selecting random elements from grouped populations

Duration: 2 hours

Location: Classroom

Objectives

- Understand how to select random farmers, ponds, cages or animals, using both a random number table and a computer.
- Be able to identify individual elements that have been selected from the sampling frame.
- Be aware of the possible concerns of producers, and address these concerns convincingly.

Key points

- Random selection can take place during the village interview.
- To ensure a representative sample, it is important to avoid replacing animals when not necessary, and to follow the selection procedure carefully.

Page references

Sampling animals from grouped populations (page 98)

Teaching methods

- Use the sampling frame that was made during the village visit.
- Demonstration: Select random elements using a random number table.
- Demonstration: Select elements using the computer.
- Explain how to identify individual elements.
- Discuss the problem of bias if the survey staff count the elements.
- Activity 16: Selection of elements from a village sampling frame.
- Activity 17: Role play of producers’ concerns during selection.

Lesson 8: Sample collection, processing and preservation

Duration: Half day

Location: Classroom, field

Objectives

- Practise sample collection from all relevant species.
- Understand how to handle, process and transport specimens that have been collected.

Key points

- Be aware of the producers’ concerns about disturbing ponds or disrupting activities.
- Ensure specimens are properly handled—the effort in the field will be wasted if they are not suitable for laboratory examination

Page references

Specimen collection and processing (page 122)

Teaching methods

- Discuss and demonstrate sample collection equipment in the classroom.
- Have participants handle equipment, and practise its use.
- Practise techniques with real animals. The animals should preferably be research animals or purchased for the purpose.
- Demonstrate the technique once first, then have each participant practise.
- Ensure that animals are treated humanely, and demonstrate this through your own behaviour.
- Demonstrate sample processing and transport techniques, and have the participants practice for themselves.

Lesson 9: Trial village visit (interview and specimen collection)

Duration: 1 day

Location: Village

Objectives

- Implement all activities of a field visit.
- Improve interview skills through practice, and implement suggested changes arising out of previous interview.
- Select elements from a sampling frame.
- Invite farmers to submit animals for sample collection.
- Process specimens appropriately.

Key points

- The field trip should be as similar as possible to real survey field work.

Page references

Survey principles and specific techniques (Chapters 5 to 14)

Teaching methods

- Preparation, assigning roles.
- Activity 18: Trial village visit.
- Discussion of successes, problems and suggested improvements.

Lesson 10: Preparation for field activities

Duration: Half day or more

Location: Class and elsewhere

Objectives

- Ensure that all practical preparations for either pilot study or real field work have been completed.

Key points

- Stress the need for good planning and organisation.
- Completing and storing data recording sheets should get special attention.

Page references

Survey procedures (page 166)

Teaching methods

- Group discussions to develop checklist of all activities and preparations that need to take place before survey.
- Prepare list of tasks.
- Assign tasks.
- Complete preparations.

Course 3: Computerised data management and analysis, and reporting

Participants

National staff, coordinators

Course structure

Lesson 1: Introduction and review of field work
Lesson 2: Principles of analysis and introduction to computers
Lesson 3: Data processing procedures
Lesson 4: Simple data analysis, descriptive statistics
Lesson 5: Prevalence survey data analysis
Lesson 6: Incidence rate surveys—retrospective disease outbreak surveys
Lesson 7: Incidence rate surveys—analysis of two data sources
Lesson 8: Surveys to demonstrate freedom from disease
Lesson 9: Reporting

Requirements

This course uses computers in almost every lesson. Participants must have access to computers, preferably with no more than two people to one computer.

Lesson 1: Introduction and review of field work

Duration: 2 hours

Location: Classroom

Objectives

- Review fieldwork activities.
- Suggest improvements for future work.

Key points

- This lesson is an opportunity for the participants to use their knowledge and experience to improve future work.

Page references

None

Teaching methods

- Introduce the course and explain the content.
- Group discussion: Field work—strengths and weaknesses, suggestions for improvement. Ask groups to address each aspect of the field work in turn: training, preparation, village selection and so on.
- Be sure to record all suggestions, and act on them for future work.

Lesson 2: Principles of analysis and introduction to computers

Note: This lesson is necessary only for participants with no experience of using computers

Duration: 3 hours

Location: Classroom

Objectives

- Understand measures of central tendency (especially mean).
- Understand measures of spread (especially variance or standard deviation).
- Become familiar with computer hardware.
- Recognise the purpose of different types of major software.
- Understand the storage of data in a computer database table.
- Compile all data collection forms from the survey.

Key points

- Analysis of data converts a large amount of hard-to-understand data into a few easy-to-understand numbers, which can be used to understand the disease. Mean and standard deviation are two example measures.
- Computers use a set of instructions (programs) to process data.
- Data is stored in tables, made up of records (rows) and fields (columns).

Page references

Measures of disease (page 56)
Principles of data management and analysis (page 146)

Teaching methods

- Review measures of disease.
- Activity 19: Analysis of data on age of participants.
- Practical demonstration of the parts of a computer and the purposes of the different components. Open a computer and identify components, peripherals and different types of disks.
- Demonstrate different types of major software.
- Discuss different data types.
- Use questions to help participants identify the most appropriate way to store different types of data.
- Use visual aids to explain data storage (databases, tables, fields, records).
- Collect all data collection forms from the survey and check for completeness.

Lesson 3: Data processing procedures

Duration: 3 hours

Location: Classroom

Objectives

- Check data for completeness.
- Perform any manual coding necessary.
- Understand how to deal with missing data.
- Create a table.
- Enter data into a table.
- Check data for errors.
- Manipulate data.

Key points

- If participants are familiar with a database program, use that program. If they are not, use any program that is available and familiar to the trainer. Epi Info is the recommended choice, as it can be made freely available to all participants.
- Don't go into detail about how to check data for errors after data entry. The techniques will be taught in the next lesson, and are the same as those used for simple data analysis.

Page references

Data processing procedures (page 150)

Teaching methods

- Activity 20: Data checking and data entry.
- Most of this lesson should be taken up with individual or paired work on computers, with brief breaks to explain the procedures required for different operations.
- Training will be much easier if several tutors experienced in the use of computers are available to help answer participants' questions.

Lesson 4: Simple data analysis, descriptive statistics

Duration: 3 hours

Location: Classroom

Objectives

- Calculate means, standard deviation and confidence intervals for survey data.
- Calculate proportions and confidence intervals.
- Generate frequency tables and cross-tabulations.
- Analyse subsets of the data.
- Create graphs of the results of analysis.

Key points

- Simple descriptive statistics can be generated quickly using the computer.
- Analysing and graphing data in different ways gives a more complete understanding of the data.

Page references

Epi Info manual (recommended)

Teaching methods

- Perform simple data analysis.
- Most of this lesson should be taken up with individual or paired work on computers, with brief breaks to explain the procedures required for different operations.
- Training will be much easier if several tutors experienced in the use of computers are available to help answer participants' questions.

Lesson 5: Prevalence survey data analysis

Duration: 3 hours or more (depending on data entry time)

Location: Classroom

Objectives

- Estimate the prevalence and calculate confidence interval for the prevalence estimate based on one- or two-stage sampling data.
- Understand how the data requirements and analysis differ according to the survey design.
- Calculate true prevalence from apparent prevalence.
- Compare the prevalence estimates from two different surveys.

Key points

- Different data is required, depending on survey design.
- With poor tests or low prevalence levels, apparent prevalence may be quite different from true prevalence.

Page references

Review of survey designs (Chapter 10)
Prevalence data analysis (page 187)

Teaching methods

- Enter, check and recode data from prevalence survey.
- Analysis of one-stage prevalence survey data (using Epi Info or other software).
- Demonstrate the use of the Prevalence program for analysis.
- Use questions and examples to explain the interpretation of the program output.
- Use the True Prevalence program to convert to true prevalence based on test performance.

Lesson 6: Incidence rate surveys—retrospective disease outbreak surveys

Duration: 2 hours

Location: Classroom

Objectives

- Understand the difference between traditional incidence rate measures and the survival curve measure of incidence rate.
- Generate a survival curve describing village outbreak experience.
- Interpret the summary measures of the survival curve.
- Compare the results of two surveys.
- Understand the interpretation of the hazard ratio.
- Understand the need to adjust for seasonal patterns.
- Be familiar with options for more complex analysis.

Key points

- Incidence rate uses a single number to summarise disease occurrence. Village outbreak surveys use a curve (the survival curve) to summarise disease occurrence.
- Just as incidence rates can be compared and the difference measured, so can survival curves.
- The hazard ratio measures the risk of disease outbreaks in one group compared to another.
- If two surveys are conducted at different times of the year, and the disease shows a seasonal pattern, you need to adjust for this to avoid bias.

Page references

Review of survey design (Chapter 10)
Data management (page 146)
Data analysis (page 212)

Teaching methods

- Enter, check and recode data from village outbreak survey.
- Analyse data using Survival program.
- Analyse different data sets to explain the need to adjust for seasonal patterns.
- Use questions to clarify the interpretation of survival curves.

Lesson 7: Incidence rate surveys—analysis of two data sources

Duration: 2 hours

Location: Classroom

Objectives

- Match outbreaks from two sources.
- Analyse data from two data sources to estimate the total number of disease outbreaks.
- Use this estimate to calculate the incidence rate.

Key points

- Clear rules have to be established for matching outbreaks between two sources.
- If the number of outbreaks appearing in both sources is small, the estimate will have very wide confidence intervals.
- Incidence rate requires a knowledge of the size of the total population (the total number of villages in the study area).

Page references

Review of survey design (Chapter 10)
Data analysis (page 212)

Teaching methods

- In small groups, match outbreaks between the two sources and calculate totals.
- Use the Capture Recapture program for analysis of the results
- Calculate incidence rate and the confidence interval.

Lesson 8: Surveys to demonstrate freedom from disease

Duration: 2 hours

Location: Classroom

Objectives

- Analyse data to calculate the probability that an area is free from disease.

Key points

- Analysis requires a knowledge of test performance (sensitivity and specificity), as well as maximum acceptable prevalence and type I and II errors.

Page references

Farm or village survey data analysis (page 234)
Two-stage survey data analysis (page 238)

Teaching methods

- Enter, check and recode survey data.
- Use the FreeCalc program to analyse the data.
- Use questions to clarify the interpretation of the results, and the meaning of the null and alternative hypotheses.

Lesson 9: Reporting

Duration: 1 hour, plus homework

Location: Classroom

Objectives

- Understand the need for reporting at different levels.
- Consider the best way to communicate the results at these different levels.
- Understand techniques for clear communication of results.
- Prepare reports of the survey results.

Key points

- Survey results should be reported back to everybody who has participated and everybody who may need the results.
- Reports should be simple, clear and easy to understand, and targeted at the user.
- Written reports may not be appropriate for all users.

Page references

Survey procedures (page 168)

Teaching methods

- Group discussion: Who might need the information from the survey? Highlight why different groups might need the results (including producers and the international community).
- Group discussion: How best to communicate with the different users of the information.
- Activity 21: Report preparation.

17
Activity sheets

This chapter contains a set of activity sheets, which provide a guide to running different types of activities during training courses. The activities are directed at the specific training courses described in the previous chapter, but can be modified or used as they are for different types of training courses.

Each activity is followed by several questions that may be used for discussion, to focus the participants' attention on what they have been doing.

Activity 1: Introductory classroom census and survey

Location: Classroom	**Duration:** 20 minutes
Reference: Course 1, Lesson 3 Course 2, Lesson 2	**Software:** None
Objectives: Differentiate between a complete count (census) and a sample (survey).	**Concepts practised:** Sampling; Estimation; Inference
Equipment and materials: Board for recording ages	

Description of activity

Explain that the aim is to determine the average age of all the participants in the room, and that this will be done in two ways. First, ask a sample of a small number of people. Record their ages in a column on the board. When finished, have the participants calculate the average of the ages.

Next, ask everybody in the class in turn, and write their ages on the board. Calculate the average and compare to the first result.

Questions for discussion

Which way took longer?

If there were 3000 people in the room, which way would be best?

Was the answer from the survey right?

Was the answer from the census right?

Activity 2: Sensitivity and specificity game

Location: Classroom	**Duration:** 30 minutes
Reference: Course 1, Lesson 2	**Software:** None
Objectives: Introduce an understanding of sensitivity and specificity	**Concepts practised:** Sensitivity and specificity; Probability
Equipment and materials: Board for recording responses	

Description of activity

The aim is to demonstrate how participants intuitively use the concepts of sensitivity and specificity all the time. It is demonstrated using a non-diagnostic example.

Invite one participant to the front of the room. Ask them to make a decision about some unknown characteristic of each of the other members of the group, such as whether they originally come from the capital city, belong to some ethnic or cultural group, or come from some broad region (the north or the south).

Whatever characteristic is chosen should be culturally appropriate (not cause discomfort), and several people in the room should representing each option (some from the north, and some from the south). It should also be some characteristic that is not immediately obvious, but about which the observer will be correct for most people most of the time.

Draw a two-by-two table on the board, and record the responses. For each person, the participant must first choose (e.g. north). The person in question says whether the choice is right or wrong.

At the end of the exercise, ask how well the group thought they went. Ask if the participant was better at picking people from the north than people from the south.

Use the figures to calculate sensitivity and specificity, and show how these can be interpreted. If possible, repeat the activity using a different participant to decide, but using the same characteristic. The second participant should not have heard the answers the first time. Compare the performance of the two people.

Questions for discussion

How does this relate to diagnostic tests?

How could we improve the proportion of correct decisions?

Activity 3: Biased sampling

Location: Classroom	**Duration:** 20 minutes
Reference: Course 1, Lesson 3 Course 2, Lesson 2	Software: None
Objectives: Understand how non-representative samples produce biased estimates	**Concepts practised:** Sampling; Populations; Bias
Equipment and materials: Board for recording responses	

Description of activity

Conduct a survey of the participants to calculate their average age from a small sample. Before selecting the sample, chose a group that will be clearly biased. For example, if younger participants mostly speak English, and the older participants don't, explain how this survey is going to be conducted in English. Select several people and ask their age in English, recording the answers only of those who can respond in English. Another biased sample can be obtained by only asking those in the group with the higher positions or ranks, as a sign of respect or in recognition of their greater experience. Other approaches can be used to select a biased group, with ages either greater or lower than the overall average.

Select the group, record the ages, and calculate the average. Compare this average with the result of the census used in Activity 1.

Questions for discussion

Why are the results different?

If we were selecting animals, how could a similar thing happen?

Activity 4: Survey sample size

Location: Classroom	**Duration:** 20 minutes
Reference: Course 1, Lesson 3 Course 2, Lesson 2	Software: None
Objectives: Recognise the effect of different sample sizes on the reliability of a survey	**Concepts practised:** Surveys, sampling; Sample size; Variance
Equipment and materials: Board for recording results	

Description of activity

Conduct two classroom surveys to calculate the average age of participants. In the first, use a small sample size, say 4 people. In the second, use a large sample size, almost all the people in the room.

Record the results, and calculate the average age from both. Contrast the results with the true average from a census (Activity 1).

If desired, calculate confidence intervals around the estimates.

Questions for discussion

Which survey produced the most accurate result?

Which was fastest and easiest?

Was either survey correct?

Do the confidence intervals contain the real average?

Activity 5: Random numbers

Location: Classroom	**Duration:** 40 minutes
Reference: Course 1, Lesson 4 Course 2, Lesson 3	**Software:** None
Objectives: Become familiar with physical randomisation procedures. Introduce probability and the prediction of random outcomes.	**Concepts practised:** Random numbers
Equipment and materials: Pack of playing cards, dice, coin	

Description of activity

First, demonstrate the concept of a single random outcome in several ways. Shuffle the cards, and ask one participant to select one at random. Ask another to say if the suit is black or red. Have another roll the dice. Ask another to say what the result is without looking. Have another flip a coin, while another predicts if it will be heads or tails.

Stress that the outcome of all of these things is random, and that the individual result cannot be predicted.

Next demonstrate how the average of many outcomes can be approximately predicted.

Split the group into three. In the first group, have one person flip the coin 100 times while the rest record the results. In the second, have one person shuffle the cards, and another draw one at random, and record if it is black or red. Replace the card and repeat 50 times, recording the results. In the third group, roll one die 120 times, recording the results in a table.

Explain the activities first, and before doing them ask each group to predict the outcome (how many heads, how many tails etc). Record the predictions on the board. Then conduct the trials, and record the final results on the board.

Questions for discussion

Why can't you predict whether a single flip will be heads or tails?

When there were many trials, how well were you able to predict the average outcome?

Was it exactly as you predicted, or just close?

How could this be used for survey sampling?

Activity 6: Random sampling using a random number table and computer

Location: Classroom	**Duration:** 30 minutes
Reference: Course 1, Lesson 4 Course 2, Lesson 3	**Software:** Random Village, or Random number generator (e.g. Epi Info)
Objectives: Select random samples using random numbers	**Concepts practised:** Sampling; Random numbers
Equipment and materials: Board; Random number table; Computer	

Description of activity

Conduct a classroom survey to determine the average age of participants. Have one participant build a sampling frame on the board, by asking each person's name, and writing a number next to it from 1 to the total number of participants.

Divide the group in two. Have one half use a random number table to select a random sample (with a reasonably large sample size), and the other half use the computer to generate random numbers for the sample.

Conduct two surveys, using the sets of random numbers, and calculate the average ages.

Questions for discussion

Were the results the same?

Were the results correct?

Are they closer to the real value than those from other techniques?

Are they representative?

Activity 7: Sampling jigsaw game

Location: Classroom	**Duration:** 1 hour
Reference: Course 1, Lesson 4	**Software:** Random Village
Objectives: Implement four different sampling schemes. Examine how non-random sampling causes problems with inference.	**Concepts practised:** Convenience sampling; Haphazard sampling: Random sampling; Systematic sampling; Inference to the population
Equipment and materials: Jigsaw puzzle, with 40 to 80 pieces, showing a scene with lots of variation. Each piece should be numbered on the back, sequentially from 1 to the total. Four sets of pen and paper for four groups, or an overhead projector and four sets of overhead transparencies and pens for four groups.	

Description of activity

Divide the group into four. Explain that the class is going to use four different sampling techniques to select a sample from a population, and that the sample will be used to estimate what the population is like.

Use the Sampling Techniques figure to introduce or revise convenience, haphazard, systematic and random sampling.

In another part of the room, assemble the jigsaw puzzle and turn it over so only the backs of the numbered pieces are showing. The population is all the pieces. The sampling frame is a list of numbers from 1 to the total population.

Each group will choose the same number of pieces. For a puzzle of about 50 pieces, each group chooses 5 pieces. For larger puzzles, select more pieces.

Have each group select the pieces to sample, using four different sampling techniques:

- Group 1 uses convenience sampling, and selects pieces 1, 2, 3, 4 and 5 (all from the top corner of the puzzle).
- Group 2 uses haphazard sampling. Ask each member of the group to pick a number between 1 and the total. Use the first 5 numbers chosen.
- Group 3 uses systematic random sampling. If there are 40 pieces, the sampling interval is 40 / 5 = 8. Have the group use ether a random number table or the Random Village program to pick one number at random, between 1 and 8, as the starting number. Then pick every eighth number.
- Group 4 uses simple random sampling. Use a random number table or the Random Village program to select 5 numbers (without replacement) between 1 and the total.

One by one, have each group come to the puzzle, and examine the 5 selected pieces. Do not let the other groups see the front of the pieces selected. When the pieces have been examined, have each group draw a picture of what they think the jigsaw puzzle is showing. This is an example of inference, guessing what the population is like, based on a sample. Give each group 5 minutes to draw its picture. They may want to write on the picture to explain what certain things are.

When all groups are finished, show each of the pictures (preferably on an overhead projector), and note which sampling technique was used. Then show the complete jigsaw puzzle, so the class knows the true state of the population. Finally, have the class vote on which group's picture was most like the real population.

Questions for discussion

Talk about how well each sampling technique was able to provide a representative sample.

Usually, systematic sampling and random sampling provide a good picture. Haphazard may sometimes give a good picture, but convenience never does. If haphazard is better than simple random, discuss how simple random sampling does not *always* provide a representative sample, but is the only way to give a representative sample most of the time.

Activity 8: Selecting elements from a disk-based sampling frame.

Location: Classroom	**Duration:** 1 hour
Reference: Course 1, Lesson 5	**Software:** Random Village
Objectives: Select a sample of farms or villages from computerised sampling frame	**Concepts practised:** Use of the computer for random sampling
Equipment and materials: Computer; Village or farm sampling frame (with or without population figures), in either Paradox or dBASE format	

Description of activity

This is part of the preparation for the real survey. Discuss the source of the sampling frame, what information is included (population figures or not), how well the elements (farms, villages etc) are identified, and how complete the list is likely to be.

Ask the group to consider options for stratification. The sample size should have already been calculated.

Demonstrate the use of the Random Village program, using a dummy data set.

Have several members of the group use the program to select the sample. Save and print the results, and have staff with local experience examine the list for any villages or farms that are impossible to survey. Use the program to replace these with new villages/farms.

Questions for discussion

What is the effect of replacing inaccessible farms?

How do the results of the survey suffer?

Activity 9: Local survey

Location: Urban area	**Duration:** 1 day
Reference: Course 1, Lesson 8	**Software:** Random Village or Epi Info random number generator
Objectives: Implement field survey	**Concepts practised:** Sampling and sampling frames; Data collection and analysis
Equipment and materials: Transport to survey site; Data recording sheets	

Description of activity

The aim of this activity is to implement all the newly learned skills through a survey to estimate some characteristic of a local urban area. The characteristic may be related to aquaculture (the proportion of householders that ate fish the day before the survey; the proportion that fish regularly), or may be completely unrelated (a survey of shops, estimating what proportion sell food). The survey may include other questions as well (to estimate a proportion or an average, or generate a categorical list).

In the classroom, explain the aim of the survey. Have participants frame the question (e.g. How important is fish in the local diet?) and refine it to a measurable quantity (What proportion of people at fish yesterday?). Define the target population (all people in the local urban area), and the study population (probably a much smaller area that can be surveyed in a few hours).

Discuss the options for sampling frames and sampling strategies. Assuming that no list of people is available (use it if it is), a household frame can be used instead. If no household sampling frame exists, one must be generated.

Select a small area nearby, consisting of one or two streets, or a small number of blocks, with a total of several hundred houses. Draw a sketch map of the area. Divide the group into pairs and assign each pair to a particular part of the study area. Travel to the area, and have each pair prepare a map of their section showing all households (preferably identified by house number).

Return to the classroom, and compile these maps into a map of the overall study area. Number each household consecutively, starting at 1.

Have the group select random numbers using either a random number table or computer, to select a sample of households. When selecting, stratify by section of the study area. The sample size should be as large as possible. The group will divide into pairs to do the survey work, and each pair should have about 10 to 20 households to survey. If there are 10 participants, there are 5 groups, and the sample size can be about 50 to 100.

Prepare data recording sheets. The questions should be kept to a minimum. Discuss the brief questionnaire and the need for a short explanation to householders of what the survey is about.

Return to the study area, split into groups and collect the data. When finished, return to the classroom for data analysis. On the board, tally the total number of people and the total number who ate fish, and calculate the proportion. Use Epi Info to calculate a confidence interval for the proportion. Have participants prepare a half-page report of the survey, its findings and conclusions, suitable for submission to the local authorities.

Questions for discussion

Is fish an important part of the diet?

What further information may be necessary to answer this question?

Can we use inference to estimate the situation in the whole town/city?

What population do the results relate to?

What problems did you have?

How could the survey be improved?

Activity 10: Survey to demonstrate freedom from disease

Location: Classroom	**Duration:** 30 minutes
Reference: Course 1, Lesson 11	Software: FreeCalc
Objectives: Understand the difficulty of proving freedom from disease with imperfect tests	**Concepts practised:** Sensitivity and specificity; Freedom from disease
Equipment and materials: Computer	

Description of activity

Begin this activity with a revision of sensitivity and specificity. Then, using a characteristic identical or similar to the one used in Activity 2 (Sensitivity and specificity), have one member of the group make a decision about a sample of the others. Ideally, the characteristic should be rare, but quite possible. One possibility is whether or not the person was born abroad.

The assessment of the one person is recorded on the board, but the true status is not yet revealed. When the decisions have been made, the class is asked whether the results indicate that there is nobody who was born abroad. Explain how a sample is unable to prove this, because not everybody has been checked.

Repeat the survey using everybody in the room, and record the decision.

You can extend this exercise to analyse the results with the FreeCalc program.

Questions for discussion

How confident can you be of the final conclusion?

How could you improve your confidence?

Activity 11: Introductions for interviews

Location: Classroom	**Duration:** 30 minutes
Reference: Course 2, Lesson 4	**Software:** None
Objectives: Practise explaining the purpose of the interview, and address problems which are likely to be of concern.	**Concepts practised:** Addressing groups; Awareness of the concerns of others; Avoiding problems
Equipment and materials: None	

Description of activity

Select one person to act as the leader of the interview. The rest of the group takes the role of the village producers. Ask the leader of the interview to spend 5 minutes preparing the main points to be covered during the introduction to the interview (for example, why we are here, what we intend to do). While they are preparing, ask the rest of the group to imagine that they are farmers and fishers. Have them think of various objections to participating in the interview or giving information. Ask them to try to make things difficult for the leader.

Start the role play by asking the leader of the interview to introduce the survey and to try to explain things in order to avoid as many problems as possible. When the leader is finished, the farmers should ask questions or raise concerns about those issues that the leader has not explained.

Repeat the exercise with a different leader (perhaps choosing the most vocal of the producers to play the role).

Questions for discussion

How well did the leaders explain the survey?

Were there any major points that were missed?

Are there any potential producer concerns that can't be addressed during this introduction?

Activity 12: Building a grouped sampling frame

Location: Classroom	**Duration:** 30 minutes
Reference: Course 2, Lesson 4	**Software:** None
Objectives: Develop a sampling frame by asking producers about their aquatic animals	**Concepts practised:** Interview skills
Equipment and materials: Data record sheets for the sampling frame	

Description of activity

Select one member of the group to be the leader of the interview. Have that person leave the room, and ask the rest of the group to play the role of farmers, and decide how many ponds they each have. Then ask different farmers to adopt personalities that may be difficult during an interview. Ask one person to act as if they know it all, and to try to answer for others or correct their mistakes. Ask another pair to be bored and want to chat all the time. Ask some to pretend to be deaf, some to be unsure of what is wanted, and some to be suspicious. Ask some to try to give the wrong information, without lying.

Ask the leader to return, and collect information about the number of ponds.

Questions for discussion

How well did the leader do the job?

Is all the information correct?

Could they have done it any better?

Activity 13: Disease ranking

Location: Classroom	**Duration:** 40 minutes
Reference: Course 2, Lesson 5	**Software:** None
Objectives: Rank diseases in order of importance	**Concepts practised:** Ranking; Interview skills
Equipment and materials: Board; Paper and pens	

Description of activity

Divide the class into several small groups, and ask each group to list on paper all the important aquatic animal diseases that occur in the local area, in the species of interest. When finished, record all these diseases on the board. Next, have the groups discuss how the importance of these different diseases may be assessed, and to write a list of 3 or 4 different criteria for ranking.

Write the criteria on the board, and agree on four to use to rank the diseases. Ask each person to rank the diseases according to each of the criteria, scoring the most important disease with 1, down to the least important.

Add up everybody's scores for each disease, and each criterion, and assign overall ranks.

Questions for discussion

Will producers rank the diseases in the same way?

Is a disease with a rank of 10 twice as important as a disease with a rank of 20? (No)

How could ranking be done in a more quantifiable manner?

Activity 14: Retrospective questions

Location: Classroom	**Duration:** 30 minutes
Reference: Course 2, Lesson 5	**Software:** None
Objectives: Emphasise the difficulty in remembering past events, and understand how it can be made easier	**Concepts practised:** Group memory; Dating landmarks
Equipment and materials: Paper to record responses	

Description of activity

Break the class into small groups. Identify an event that occurred some years ago and that every participant knows about. Examples include elections, natural disasters, major news items etc. Ask each group to try to remember the month and the year of the event.

Record the different responses from each group.

Compare the results from the different groups, and discuss how they remembered.

Questions for discussion

Are the answers the same? If not, what sort of errors were made (wrong year, wrong month)?

How did participants remember?

How can remembering be made easier?

Activity 15: Village interview

Location: Village	**Duration:** Half day
Reference: Course 2, Lesson 6	**Software:** None
Objectives: Practise a village interview as it will be done during the real survey	**Concepts practised:** Public address; Explanations; Ensuring cooperation; Encouraging participation
Equipment and materials: Transport; Data recording sheets, pens, paper	

Description of activity

Make sure that the village interview has been organised beforehand, and that the producers know when it is. Prepare the group well, so they know their responsibilities and roles.

Try to use a village with a large number of producers. Divide the village into a number of smaller groups, so that as many of the participants as possible have the chance to lead the interview. There should be two or three participants running each interview, with a minimum of 5 or 6 producers. The exercise is easier if there are a number of tutors available, experienced in village interviews, to supervise each group.

Have the group run through the complete interview, as it will be run during the survey. For example, the following sections may be included:

- Explain the purpose of the interview
- Build a grouped sampling frame
- Rank disease problems
- Ask about the usual dates of disease problems
- Ask about disease outbreak history
- Invite questions on disease problems and offer advice

Questions for discussion

What information did each group collect?

Did each group get the same answer to the same questions?

What problems were encountered? How could these be addressed?

Activity 16: Random selection of grouped elements

Location: Classroom	**Duration:** 40 minutes
Reference: Course 2, Lesson 7	**Software:** Random Animal
Objectives: Use a sampling frame to practise selecting grouped elements	**Concepts practised:** Random selection
Equipment and materials: Sampling frames from village visit (1 copy for each participant); Random number table; Computer; Pen and paper for results	

Description of activity

Copy the sampling frame generated during a village visit (e.g. Activity 15). Give a copy to each participant, with a random number table. Have each participant select a group of 10 animals at random. While they are doing this, have each student in turn use a computer and the Random Animal program to enter the sampling frame and select 10 animals. When the samples have been selected, use a role play to practise identifying the individual animals. Have one participant play the role of the survey team leader, another the farmer. Select several ponds, then move to a new farmer, and new ponds.

Questions for discussion

Which technique was simplest for selecting elements, the random number table or the computer?

Which is most appropriate for use in the village?

How well did the team leader select the elements?

Are there any suggested improvements?

Activity 17: Specimen-collection role play

Location: Classroom	**Duration:** 30 mins
Reference: Course 2, Lesson 7	**Software:** None
Objectives: To understand aquatic animal producers' concerns about specimen collection. To practise addressing these concerns and encouraging cooperation.	**Concepts practised:** Communicating with producers
Equipment and materials: Role cards—one for each actor explaining their attitude.	

Description of activity

Select four students to participate in the role play. One person plays the role of the member of the survey team, and three play village producers. Explain the scene: The member of the survey team has come to collect specimens. They visit the producers, and want to collect specimens from one animal belonging to each owner.

Give each of the actors a role card to explain the position they are to take during the play.

The member of the survey team needs to collect specimens from the selected animals. They are not allowed to change the animals for others, and must get specimens.

The first farmer is happy for them to try to collect specimens, but doesn't believe that they will be able to catch the fish.

The second farmer doesn't want a scraping to be collected from the selected fish, because she is brood stock.

The third farmer doesn't want their fish to be used because they are very suspicious about what the information will be used for. They are afraid that the team is trying to prove that they are not looking after their fish properly.

Conduct the play, telling the actors that it should last no more than 10 or 15 minutes. Each of the participants should stick to their role, and try to argue their position as strongly as possible.

The role play can be repeated with different actors, or with the same actors in different roles.

Questions for discussion

How well did the survey team member perform?

Are there any important points that they missed in their explanation?

What should you do if the farmers completely refuse to cooperate?

If you were the farmer, would you be happy to be involved in the survey if it were to be conducted again next year?

How can these problems be avoided?

Activity 18: Village interview and specimen collection

Location: Village	**Duration:** 1 day
Reference: Course 2, Lesson 9	**Software:** Random Animal
Objectives: Implement a complete village visit	**Concepts practised:** Interviews; Restraint; Blood collection
Equipment and materials: Transport; Data recording sheets; Restraint and specimen-collection equipment	

Description of activity

Organise the visit carefully beforehand. Make sure everybody knows their role and responsibilities.

Divide the class into several groups and conduct the village interviews as described in Activity 15.

During the interview, have one or two people select elements from the grouped sampling frame (e.g. ponds).

Collect specimens from the selected elements. Initially, have all participants as one group while identifying and collecting specimens from the first few ponds. Then, if tutors are available to supervise, split the class into smaller groups so that each person has an opportunity to practise every role in collecting specimens or making observations (e.g. capturing animals, euthanasing and preserving, taking pH measurements from the water, testing transparency etc).

Process the specimens appropriately.

Questions for discussion

What problems arose?

How could they be addressed?

Activity 19: Analysis of data

Location: Classroom	**Duration:** 30 minutes
Reference: Course 3 Lesson 2	**Software:** Epi Info or other database, or statistical software
Objectives: Understand basic principles of data entry and analysis	**Concepts practised:** Summarising data
Equipment and materials: Computers (preferably one each or one between two participants)	

Description of activity

Collect data on the age of each participant, and record it on the board. Using a table that has already been created, have each participant enter the data. Use the Epi Info Analysis program to calculate the mean, minimum, maximum, variance, standard deviation, and confidence interval around the mean.

Questions for discussion

How is each of these values interpreted?

What population does the data relate to?

Activity 20: Data management

Location: Classroom	**Duration:** 3 hours
Reference: Course 3 Lesson 3	**Software:** Epi Info
Objectives: Become familiar with data management procedures	**Concepts practised:** Table creation; Data entry; Data checking
Equipment and materials: Computers	

Description of activity

Using the raw data collected during surveys, set the task of calculating basic village-level descriptive statistics on population, proportion of villages suffering outbreaks, and most important diseases.

Lead participants through each of the operations step by step, and then let them carry out the operations themselves and explore the procedures:

- Check data
- Create appropriate tables and data entry forms
- Set up checks during data entry
- Check data after data entry
- Recode data
- Export data
- Analyse data.

Questions for discussion

How should the results be interpreted?

What do the numbers tell us about the disease situation?

Activity 21: Report writing

Location: Classroom	**Duration:** 2 hours
Reference: Course 3 Lesson 9	**Software:** None
Objectives: Practise writing appropriate reports	**Concepts practised:** Reporting
Equipment and materials: Paper, pens, possibly a word processor	

Description of activity

After discussing the levels of reporting and appropriate types of report presentation for different audiences, assign several participants to work together to produce a short report for each of identified groups.

Give them two hours to work together in small groups to collect the data and generate tables and graphs. Set them the job of preparing the report as homework, over two or three nights or a weekend.

Have each group present its report to the class. Some reports should be presented orally (e.g. reports for illiterate producers; briefings for busy decision makers, such as the Minister). Other reports should be written, and the contents and presentation can be explained.

Questions for discussion

Did each report contain the information that was needed?

Was it easy to understand?

How could they be improved?

Activity 22: Knowledge quiz competition

Location: Classroom	**Duration:** 30 minutes
Reference: None	**Software:** None
Objectives: Encourage participation; Fun break; Warmer	**Concepts practised:** Any concept from the previous topics covered during the training
Equipment and materials: Scoreboard (blackboard / whiteboard / paper); paper for writing questions	

Description of activity

Divide the group into halves. Give each group 10 minutes to think of a list of 15 questions (with the right answers), based on the topics covered during the previous day, or throughout the training. Choose a spokesperson for each group. Group 1 asks the first question and Group 2 has 30 seconds to answer (the trainer keeps time), during which they may discuss the answer amongst themselves. Group 1 says if the answer is correct or not. If the answer is correct, Group 2 gets a point. If the answer is wrong or they are unable to answer in the time, Group 1 gets a point. If the answer is either right or wrong, but Group 1 judges it incorrectly, they lose a point.

It is then the second group's turn to ask a question. This continues until all questions have been asked. The group with the highest score wins.

Questions for discussion

Ask individuals to provide the correct answer for any questions that were answered incorrectly. If nobody (including the asking group) is able to, set aside time to cover the area again.

Appendix A

Introduction to epidemiology for aquatic animal scientists

This appendix contains a set of notes on aquatic animal epidemiology, prepared by Chris Baldock, AusVet Animal Health Services. It is intended to give the interested reader a broader overview of epidemiology and an understanding of some of the issues involved in better understanding disease.

1 Overview of epidemiology

The purpose of this chapter is to provide an overview of epidemiology with an emphasis on key concepts and definitions.

Objectives

Basic epidemiological concepts and definitions are introduced in this chapter. At the completion of this section, readers should be able to:

- understand the role of epidemiology in studying disease;
- give a definition of epidemiology;
- define some basic epidemiological concepts; and
- consider disease problems from an epidemiological perspective.

Epidemiology and where it fits

As animal production systems have intensified, the interaction of disease agents with other factors such as the physical environment, nutrition and genetics has become more complex. This complex interplay among a variety of factors sits in delicate balance while the goal of increasingly efficient production is sought. In such a system, even small changes in some factors can provide enough stress to cause the expression of disease. Resultant morbidity and mortality translate to lost production and reduced profitability.

The traditional approach to the emergence of new disease entities is to seek interventions which will prevent or cure disease at the individual animal level. This traditional perspective requires developing an understanding of disease processes at the individual animal, organ, tissue, cellular and molecular level. Such an 'inside the animal' approach largely ignores the complex interplay, which occurs among individual animals when aggregated in populations, particularly when these populations often exist in a less than ideal environment for good health and optimal production. A population of animals has attributes which go beyond the mere summation of its constituent animal units in the same way that the individual animal is more than just the sum of its individual organ systems. In addition, epidemiology looks at higher levels of populations. For example, the all ponds on a particular farm may be regarded as a population, as could all the farms in an area such as a province or country. These relationships are shown in Figure 1.1.

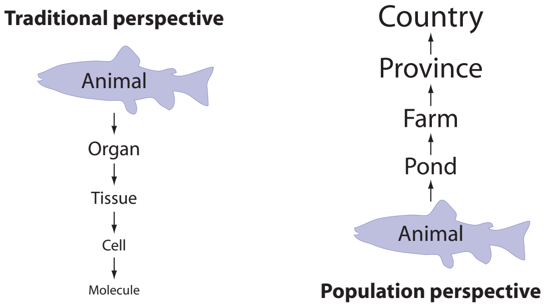


Figure 1.1 Representation of the relationship between the traditional perspective of investigating disease and a population perspective

For the epidemiologist, the population is the patient. The science of epidemiology is that of identifying the patterns of disease in populations and, through analysing those patterns, gaining insight into the cause of disease with the view to improved control. Quantitative epidemiological approaches to the investigation of production loss and disease occurrence can help unravel the complex interplay among the many factors which result in the expression of disease.

Epidemiology is the study of the patterns and causes of disease in populations.

Epidemiological studies provide insight not only into those factors which are unique to the population *per se* but can also raise hypotheses worth exploring further at the individual animal, organ, cellular and genetic level.

Thus, the understanding of disease processes operating at the population level require both a 'downward' and 'upward' approach to investigation. By using such a bi-directional approach, fresh insights into the mechanisms and control of disease can be obtained.

Epidemiology is an integrating science with close links to clinical and laboratory medicine as well as biostatistics and health economics. In addition, it is the basic science, which underpins state veterinary medicine, preventive veterinary medicine and herd health programs. Epidemiologists usually use the word 'disease' in its broadest sense to include any health related condition or event of interest, in addition to clinical illness.

Epidemiology is concerned with:

- Detecting the existence of a disease or other production problem.
- Identifying the causes of disease.
- Estimating the risk of becoming diseased
- Obtaining information on the ecology and natural history of the disease.
- Defining and quantifying the impact and extent of the problem.
- Planning and evaluating possible disease control strategies.
- Monitoring and surveillance to prevent further disease episodes.
- Assessing the economic effects of disease and control programs.

Introducing epidemiological concepts

We all see problems in different ways depending on our prior experience and knowledge. Epidemiologists are interested in diseases and their production impacts, but take a different approach to say the pathologist or microbiologist. Some of the underlying principles of epidemiology are introduced in this section.

Cause of disease

The epidemiologist interprets causality in quite a wide sense. This is somewhat different to the more traditional view of the role of cause being restricted to aetiological agents with all other contributions being designated as 'contributing' or 'predisposing' factors. An epidemiological definition of a cause of a disease is 'an event, condition or characteristic that plays an essential role in producing an occurrence of the disease'.

A **cause** is an event, condition or characteristic that plays an essential role in producing an occurrence of the disease in question.

Under such a definition, the presence of say a fungus in a pond of fish may not of itself be a sufficient cause of disease. It may require a stress trigger to cause an outbreak of disease in the group. Under this concept, the fungus is a *necessary* cause (no disease would occur if the fungus was not present) but not a *sufficient* cause of the particular syndrome, whereas the stress is neither necessary nor sufficient but can be a component of a sufficient cause. In fact, for any particular expression of a particular disease, there may be a range of possible sufficient cause complexes.

The challenge for the epidemiologist is to help identify some of the more important components of sufficient causes for a particular disease with the view to devising cost-effective intervention strategies at critical points to either prevent disease expression or reduce the production effects.

One of the problems for the epidemiologist is that it is often difficult to unequivocally prove a causal relationship from a single study. One only has to witness the amount of epidemiological research undertaken to link smoking and lung cancer to appreciate this point. In an epidemiological sense, factor 'X' causes disease 'Y' if and only if:

- X occurs prior to Y in time,
- a change in the frequency of X is associated with a change in the frequency of Y and
- the association is not due to X and Y each being correlated with some other factor.

Risk factors for disease

Risk factors are those characteristics of some individuals which, on the basis of epidemiological evidence, are associated with increased risk of disease. Risk factors may be either causal or non-causal. Non-causal risk factors are sometimes called risk markers.

For example introduction of wild crustacean carriers of *Lagenidium* spp. is a causal risk factor for larval mycosis in farmed crustaceans but when its is in the larval stage of development it is a non-causal risk factor for the disease (a risk marker). Being nearer the ocean may be a risk factor for shrimp farms experiencing a white spot disease outbreak, but it may or may not be a cause.

A knowledge of risk factors is useful in determining how to control disease in populations. An important objective of many epidemiological studies is to identify risk factors and quantify the magnitude of their effect on disease frequency. For example, we may find in a research study that shrimp farms using recirculating water systems are at five times the risk of having a white spot disease outbreak during the grow-out period when compared with those using closed systems. We might expect from this finding that a farmer who changed from a recirculating system to a closed system would dramatically reduce the risk of having an outbreak.

Epidemiological measurements

The basic units of measurement in quantitative epidemiology are those of counting. These basic units are translated to measures of *frequency* and measures of *association*.

Measures of association are obtained by comparing disease frequencies between different groups of animals, ponds, farms etc. Both of these types of measures are explained below.

There are two basic epidemiological measures of disease frequency: *incidence* and *prevalence*.

Incidence is the proportion of individuals within the population at risk who convert from a non-diseased to diseased state during a specified time period.

Prevalence is the proportion of individuals within the population at risk who have the disease at a particular point of time or during a particular period.

Thus, incidence reflects the number of *new* cases within a given period of time, while prevalence reflects the number of *existing* cases at a point in time. An understanding of the previous incidence of disease is sometimes used to estimate the risk of disease in the future. For example, if the experience in an area is that white spot outbreaks usually occur in one of every four (25%) crops of shrimp cultured, then a farmer may think that if he purchases a batch of post-larvae he has a 25% risk of experiencing an outbreak of white spot during grow-out.

Risk is the probability (likelihood) of experiencing disease in a defined future period of time and can be estimated from previous incidence of disease.

Measures of association are used to compare the frequency of disease or other characteristics among different groups of animals in a population. A commonly used measure of association is *relative risk*.

For example, say a farmer has two ponds in which he rears fish. If 20% of fish in pond A developed epizootic ulcerative syndrome (EUS) compared with 10% in pond B, then we would say the relative risk of EUS in pond A compared with pond B is 2.

The calculation of measures of frequency and association appears deceptively simple in theory but can be extremely difficult in practice. The challenge in the above example is to identify whether the difference in incidence between pond A and pond B is a real difference or is just due to chance. If the difference is not due to chance, the next challenge is to identify what is different about pond A (what are the risk factors) that leads to the increased risk of disease.

Random error, bias, confounding and effect modification (interaction)

When any attempt is made to collect information about a particular disease in a population and the factors that might be associated with its occurrence, the data that is collected, recorded, analysed and reported is likely to differ to some degree from the true values. These inaccuracies may be either non-systematic or systematic.

For example, the true weight of a fish selected for measurement in a study may be 567.3 gm. However, when we read the scales, we may read 567 gm, slightly lower than the true weight. The next fish we weigh may have a true weight of 543.8 gm, but we may read 544 gm, slightly higher than the true weight. These types of errors are known as *non-systematic*, *chance*, or *random* errors and should average out to approximately zero when we weigh many fish. Other random errors such as with data entry where we may enter 5667 gm into the computer rather than 567 gm can be allowed for to some extent during statistical analysis of data. It is vitally important that they be minimised, but their effect on an epidemiological study is usually less significant than errors due to *bias.* Bias occurs in an epidemiological study when the observations do not reflect the true situation because of some *systematic* error. The problem of bias and how to avoid the different types are of central importance to the validity of all epidemiological studies.

In the example above, there are two opportunities for bias. First, if the sample of fish chosen from the pond and weighed, is not *representative* of all the fish in the pond, then the average weight estimated for the whole pond of fish may be biased. If we chose only the first few fish that came to eat feed thrown on top of the water, the selected fish may well be the larger fish in the pond. This is a simple example of what is called *selection bias*. Second, the scales used to weigh the fish may be adjusted incorrectly and always weigh 3 gm too heavy so that, in addition to random error in the weights, we also have a systematic error or bias of 3 gm for each fish. Our estimated average weight will also be 3 gm too heavy in this instance. This is a simple example of a *measurement bias*.

Bias is any systematic error in the design, conduct or analysis of a study which results in estimates that depart systematically from the true value.

The above definition of bias, in contrast to conventional usage, does not imply prejudice, such as the observer's desire for a particular result. Many types of bias have been defined which may impact on the results of an epidemiological study. However, there are essentially two broad categories: selection bias and measurement bias.

Selection bias is due to systematic differences in characteristics between those individuals selected for study and those who are not.

Measurement bias is due to faulty measuring equipment or systematically dissimilar methods of measurement used on individuals in different groups being compared.

Measurement bias is sometimes called observation or information bias. For example, we may be interested in comparing the average weight of fish in pond A with those in pond B based on the weights of 30 fish selected from each pond. Say we select and weigh 30 fish from pond A immediately before feeding time but select and weigh 30 fish from pond B immediately after feeding. We then compare the average weights and find that the fish in pond B are, on average, slightly heavier than the fish in pond A. The difference we found may be entirely due to the feeding — a measurement bias — the reality may be that, on average, the fish in the two ponds are the same weight.

In addition to chance and bias, an epidemiological study may be affected by *confounding*. Confounding arises when the purpose of the study is to identify risk factors for disease.

Confounding occurs when two risk factors are interrelated and it is incorrectly concluded that one of the factors is causally related to the disease in question.

For example, it might be observed that shrimp in ponds with cloudy water do not grow as well as those in clearer water. We might conclude from this that light penetration of the water is important for normal growth. However, it may be that the cloudiness is due to the presence of particular algal species in the water which inhibit growth of the shrimp through toxin production. In this theoretical example, confounding has meant that we have incorrectly concluded that light penetration is associated with poor shrimp growth when the true cause was the presence of toxic algae. The relationships are represented in Figure 1.2.

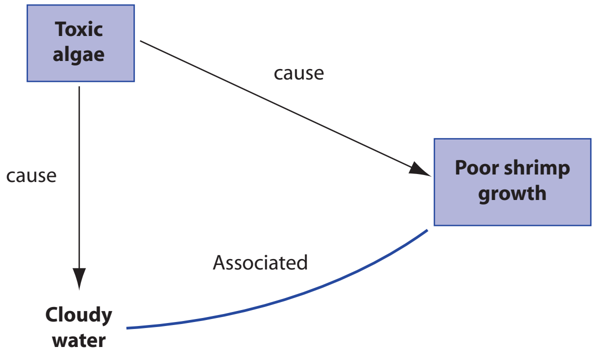


Figure 1.2 Theoretical example of relationships resulting in confounding leading to incorrect conclusions on the cause of poor shrimp growth

Confounding is one of the critical problems to watch for when undertaking an epidemiological study. It would probably be better to name the problem 'confusing' as it occurs when the effects of two or more factors are mixed and it is difficult to determine which factors are truly 'causal' in an epidemiological sense.

Where two or more risk factors play a role in the causation of a disease, the possibility exists for *effect modification* (also known as *interaction*) to occur between two or more of the factors. Interaction is different from confounding.

Effect modification (interaction) is said to occur when the incidence of disease in the presence of two or more risk factors differs from the incidence expected to result from their individual effects.

When effect modification occurs, the effect can be greater than what we expect (positive interaction or *synergism*) or less than what we would expect (negative interaction or *antagonism*). The problem when evaluating effect modification is to ascertain what we would expect to result from the individual impacts of the different risk factors.

For example, say we find that the incidence of EUS is 5% in fish in ponds with acidic water but with a smooth lining, 2% in ponds with neutral water but with a rough lining and 15% in ponds which have both acidic water and a rough lining. The 15% incidence is a lot higher than we would expect if the two factors of acidity and rough pond lining operated independently to increase the risk of EUS. We would therefore suspect synergy between these two risk factors and would need to investigate further.

In complex epidemiological studies, information is often collected on a wide variety of factors to identify the important risk factors for the disease of interest. To achieve this, statistical analysis is required to sort out the effects of confounding and effect modification among the potential causal factors for the disease.

Types of epidemiological study

There are many types of quantitative epidemiological study but they can be broadly grouped into *observational*, *intervention* and *theoretical* studies as shown in Figure 1.3. The underlying principles for all types of study are similar. In the process of finding causes of disease, factors which are statistically linked with the disease of interest and suspected to be causal for the disease (known as *risk factors*) are identified. A risk factor is not necessarily a causal factor. Two general concerns of epidemiological studies are *internal* and *external validity*. Internal validity is the likelihood that any observed differences between groups reflect the true state of nature and are not due to chance, bias or confounding. External validity is the relevance of results to the wider population beyond the limited study groups.

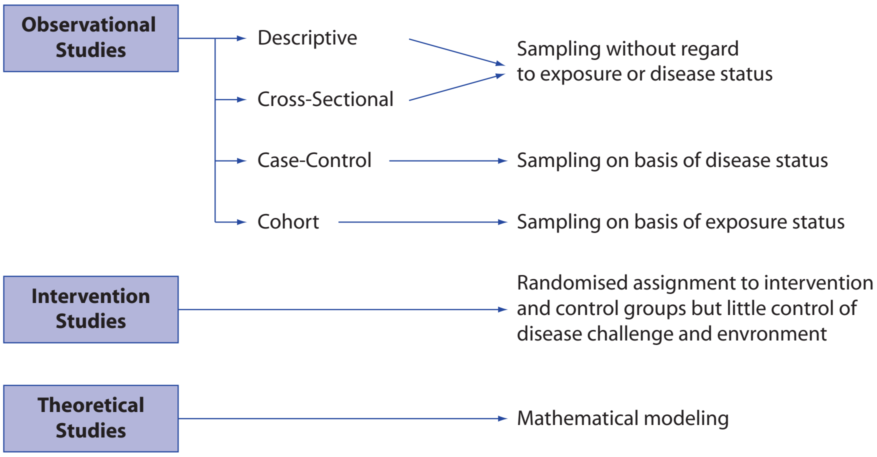

Observational Studies
Descriptive
Cross-Sectional
Case-Control
Cohort
Sampling without regard to exposure or disease status
Sampling on basis of disease status
Sampling on basis of exposure status
Intervention Studies
Randomised assignment to intervention and control groups but little control of disease challenge and envronment
Theoretical Studies
Mathematical modeling

Figure 1.3 Classification of quantitative epidemiological study types

Observational studies

In observational studies nature is allowed to take its course, while differences or changes in the characteristics of the population are studied, without intervention from the investigator. There are basically four types of observational study: *descriptive*, *cross-sectional*, *case-control* and *cohort*.

In *descriptive* studies, the focus is on describing the distribution and frequency of a disease in a population in terms of animal, place and time. The described patterns of disease may lead to hypotheses which can then be further elucidated. Surveys are an example of descriptive studies.

In *cross-sectional* studies, prevalence of the disease in question is measured and compared among those with and those without the risk factor(s) of interest. A weakness of cross-sectional studies is that evidence for causation is only realistically produced for ‘permanent’ (sometimes called ‘fixed’) factors such as species and sex.

In *case-control* studies, the investigator selects fish, ponds or some other units with the disease of interest as the 'cases' and units without the disease as the 'controls'. The frequencies of suspected risk factors are then measured for the two groups and compared. Case-control studies are well suited to rare diseases and many suspected risk factors can be compared at the same time. They are relatively quick and inexpensive to perform but are susceptible to many biases and do not yield estimates of the frequencies of disease in the exposed and unexposed populations.

In *cohort* studies, representative fish (or ponds, farms etc.) exposed to a suspected risk factor are selected along with a comparison group which is not exposed. The investigator does not assign the factor of interest, but merely observes the course of natural events. After a suitable period of observation, the frequency of the disease of interest is compared between the groups. Cohort studies can provide a complete description of the development of disease and true incidence rates in exposed and unexposed groups. They are particularly suited to say examining the differences between species of fish (where these can be reared together) because of the rapid turnover in populations which means that periods of observation are not too long.

Intervention studies

Intervention studies are in reality epidemiological experiments imposed at the population level. These are sometimes called field trials. This is in contrast to laboratory or pen experiments, which are conducted under much more rigidly controlled conditions. Intervention studies include field and clinical trials. The purpose is to evaluate the effects of some preventive or treatment (intervention) strategy. We commonly think of such studies as pertaining only to testing vaccines or drugs. However, the same methodology is applicable to other interventions such as changes in management or pond design. Basically, eligible experimental units are allocated randomly to two or more groups, the treatments applied and the outcomes measured.

Theoretical studies

Theoretical studies are based on mathematical modelling using a computer and are designed to answer 'what-if' type questions in an attempt to extend the limits of existing knowledge. They are particularly useful in examining the behaviour and impact of infectious diseases as well as the possible effects of a range of interventions. The results from such studies need to be confirmed with follow-up observational or intervention studies wherever possible.

For example, if the infectious behaviour of the EUS agent in a pond of fish could be modelled, then the required attributes of an intervention could be approximately specified in terms of the required efficacy and the effects of timing of application.

Conclusion

Study of the behaviour of disease in populations is a complex matter and an epidemiologist has specialist skills to help identify key risk factors and possible causes for disease, as well as evaluate and monitor possible interventions. Just as good disease investigation in the individual animal includes contributions from the clinician, microbiologist, pathologist and other disciplines, so the investigation of disease in populations must be made from an epidemiologically sound perspective.

Exercises

Exercise 1.1

List three diseases of aquatic animals which occur in your region where an epidemiological approach may be useful.

Exercise 1.2

List the possible causes (from an epidemiological perspective) of one of these diseases.

Exercise 1.3

Describe the impact at the farm level of the disease you chose in Exercise 1.2.

Exercise 1.4

List the main options for control at the farm level of the disease you chose in Exercise 1.2.

2 Patterns of disease

The purpose of this chapter is to explore some basic disease principles which result in the patterns that are seen in populations.

Objectives

At the completion of this section, participants should be able to:

- describe the natural history of selected diseases;
- understand different mechanisms of disease transmission and spread;
- describe mechanisms for the maintenance of disease;
- describe the ecology of selected diseases; and
- describe disease occurrence in terms of animal, time and spatial patterns.

Introduction

A basic premise of epidemiology is that, in a population of animals, ponds or farms, disease does **not** occur randomly in animal groups, over time or in space. Although the transmission of disease among individuals involves chance events, the resultant effect at a population level is to produce distinct patterns which can be described and analysed by the epidemiologist to gain insight into the cause and behaviour of disease with a view to prevention or control.

> Although disease transmission has chance elements at the individual animal, pond and farm level, the resultant effect at the population level is to produce distinct patterns which can be described and analysed to better understand the causes and behaviour of disease and how it can be prevented and controlled.

To begin to understand why we see patterns at the population level, we need to understand the behaviour of disease in the individual animal and how disease agents move from animal to animal and farm to farm. In later chapters we will learn how to more formally analyse disease patterns.

Unit of study

We can examine patterns of disease by looking at individual animals or some other unit of study which is an aggregation of animals (an animal group), sometimes assumed to be randomly mixing for the purposes of disease transmission. Examples are tank, cage, pond, farm, village, district, province, state etc.

Before talking about patterns of disease, it is therefore important to understand the concept of *unit of study*. In medical epidemiology and with many livestock diseases, the *unit of study* is often the individual person or animal. Thus, a medical epidemiologist may be interested in identifying factors which make some people more susceptible to influenza than others. In aquaculture, the unit of study may also be the individual animal. For example, it may be observed that in a farmer's pond, a particular disease seems to be more common in male fish than in females. Here the unit of study is the individual fish and the characteristic which seems to be associated with disease is the sex of the animal. However, the unit of study can also be an aggregation of individuals such as a pond or a farm in the case of aquatic animals. For example, in a goldfish breeding farm it may be observed that some tanks of fish have a greater problem with a particular disease than others. In this case, the epidemiologist would be interested in identifying factors associated with the tanks that lead to greater problems in some than others. Some factors might be size, location on the farm, water quality.

When describing patterns of disease and relating those patterns to characteristics (or factors) of the unit of study, it is important that the chosen characteristics are relevant to the chosen unit of study. For example, the characteristics of species, sex and age are relevant where the unit of study is the individual animal but are not relevant where the unit of study is the pond or farm. Examples of different units of study and characteristics appropriate to each are shown in Table 2.1 in hierarchical order. By this we mean that a number of animals are contained in a pond (or cage) and then a number of ponds make up a farm, a number of farms make up a village and so on.

Table 2.1 Some possible units of study in hierarchical order with examples of relevant characteristics applying to the particular unit of study

Unit of study	Examples of relevant characteristics
animal	species, sex, age
pond, cage	size, shape, location, stocking density, stage of production
farm	location, number of ponds, source of stock, production method
village	location, number of farms, farming practices, income level
district	location, number of villages, government services,

A characteristic which is relevant to a certain level in the hierarchical order is assumed to apply equally to all units lower in the hierarchical level. For example, if the unit of study is the pond and if we wish to know if the size of the pond affects disease occurrence, then it is assumed that the size of a particular pond affects all fish equally in that pond.

Progress of disease in individuals and populations

The progress of disease in an individual animal over time without intervention as it occurs in the natural rather than a controlled situation such as in a laboratory or tank experiment is known as the natural history of disease. The natural history begins with exposure of the host to the disease agent and progresses through to either recovery or death. The epidemiologist is interested in using population based methods to identify the important factors affecting this natural history with the intention of identifying possible methods of prevention and control.

In the simplest sense, at the outset, the host will be in a *susceptible* state (stage). Following exposure to an infecting dose of an agent (pathogen), pathological processes begin and the animal moves to the *incubatory* or *pre-clinical* stage of the disease. As the pathological process progresses and the host's various response mechanisms come into play, *clinical* signs are manifested and it is usually during this stage that a *diagnosis* is made. Finally there will be *recovery* from the disease or *disability* and in some cases *death* if the disease is severe enough. In some instances, an infected animal may not show detectable signs of disease, in which case it is said to be in a *sub-clinical* state. These different stages of disease through which the host progresses are collectively referred to as the *spectrum of disease* and are illustrated in Figure 2.1.

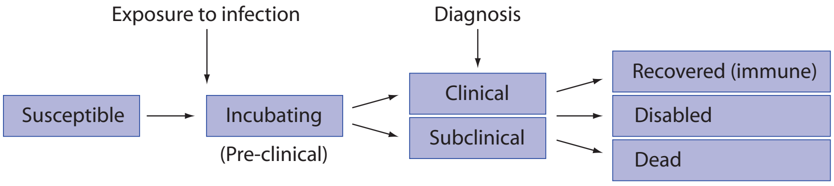


Figure 2.1 Spectrum of a disease simplified into a number of discrete states or stages through which an individual progresses with time

With infectious diseases, we refer to the *incubation period* which is the time period from exposure to infection through to when clinical signs are first manifested. For physical agents such as toxins we usually refer to *induction* or *latency* periods to mean a similar thing. For gastrointestinal parasitic diseases, the term *prepatent period* is used to mean the time from initial infection to when parasite eggs are passed in the host's faeces. Readers are referred to the glossary for details of definitions.

When an infectious disease agent is first introduced into a *susceptible population*, there will be very few animals in the clinical and subsequent states. As the epidemic progresses, the number of animals with clinical disease will increase and then slowly decrease while the number of susceptible animals will decrease and the number of recovered animals will increase (assuming no mortality). This phenomenon is shown in Figure 2.2.

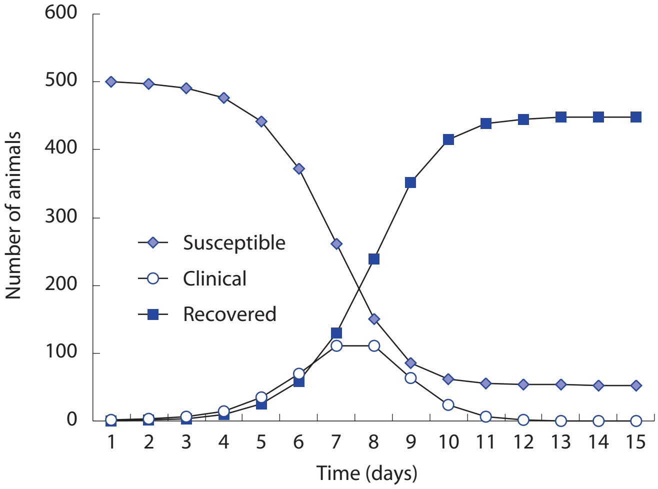


Figure 2.2 Epidemic pattern in a population of 500 susceptible individuals following the introduction of a directly transmitted infectious agent which begins by affecting a single animal (Reed–Frost model with p = 0.005)

It is very important to remember these basic concepts when using diagnostic tests and estimating their usefulness in populations. Often a particular test is useful to detect an infected animal at one stage of disease but not another. Its overall usefulness on a population basis will therefore depend on the proportion of individuals in the population of interest in the various stages of disease at the time that samples were taken. For example, from Figure 2.2, we can see that if we were to use a test which detected recovered animals (e.g. an antibody detection test) at Day 4 of the epidemic, we would find that only 5% of the animals had recovered. However, by Day 12, almost 90% of the animals have recovered.

The terms infectivity, pathogenicity and virulence also need to be understood when considering the progress of an infectious disease in a population. The definitions for each of these terms as they are used in epidemiology are shown below.

Infectivity — the percentage (or proportion) of susceptible individuals exposed to a particular agent who become infected

Pathogenicity — the percentage of infected individuals who develop clinical disease due to the particular agent

Virulence — the percentage of individuals with clinical disease who become seriously ill or die

Transmission, spread and maintenance of infection

To understand how disease patterns are created, we must understand how disease agents (organisms) move around in the population — from animal to animal, pond to pond, farm to farm etc. We also need to know how agents can persist in a population and not be easily detected.

Transmission and spread

The *chain of infection* is the series of mechanisms by which an infectious agent passes from an infected to a susceptible host. To move around in a population, a disease agent must escape from infected hosts and find new susceptible hosts. This is summarised in Figure 2.3.

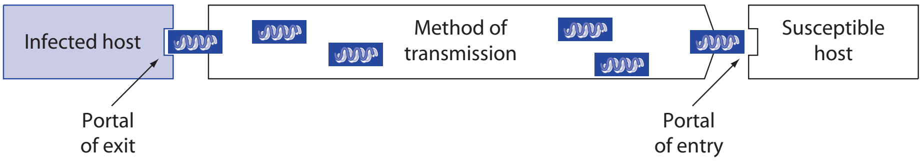


Figure 2.3 *Chain of infection* for infectious disease agents

The terms transmission of disease and spread of disease have related meanings but are used for different purposes in this text although these terms are often used synonymously.

Transmission refers to the movement of infection from an infected animal to a susceptible animal within an infected population.

Spread refers to the movement of infection from an infected population or sub-population to a susceptible population or sub-population

Interest in how a particular disease agent moves around will focus on different mechanisms depending on the unit of interest of the epidemiological investigation.

For example, the most fundamental level of interest is transmission from animal to animal. However, within a particular farm there may be interest in methods of spread from pond to pond. At a higher level again, the interest will be in methods of spread from farm to farm. Finally, quarantine authorities are interested in mechanisms of spread from country to country.

A number of ways of looking at different methods of transmission have been devised. These are summarised in the Table 2.2.

Table 2.2 Methods of transmission for infectious diseases in aquaculture settings

Direct transmission or spread			
	Horizontal –	Direct contact	
		Contact with discharges	(vomitus, faeces)
		Cannibalism	
	Vertical–	Spawning	
Indirect transmission or spread			
	Airborne –	Droplet nuclei	(< 5 microns)
		Dust	(> 5 microns)
	Vector –	Mechanical	(bird, other aquatic species)
		Biological	(other aquatic species)
	Vehicle –	Fomites	(vehicles, personnel, equipment)

Maintenance of infection

When active in a population, an infectious agent must be able to survive in host animals and the external environment or vectors and reservoirs. In most instances, within the host animal, defence mechanisms will either eventually terminate the infection or the host will die. However, in some cases, infection will persist and the host will appear relatively normal. Such animals are said to be *carriers*.

> A carrier is an animal which is capable of transmitting infection but shows no clinical signs

A carrier can be *incubatory, convalescent,* or *chronic*. Carriers are very important in the maintenance of infectious diseases in populations. Some helminths and protozoa form protective cysts within the host and can survive for long periods of time.

Some infectious agents such as white spot syndrome virus can infect more than one crustacean host species. In such instances, persistence of infection in a particular area is facilitated by the presence of a range of host species of varying susceptibility to disease. A particular host species is said to be a *reservoir host* when it is the host species in which the disease agent normally lives and persists in a population and from which it can spill over to other species of hosts and cause disease. More generally, a *disease reservoir* is any animal, plant or environment or combination of these in which an infectious agent normally lives and multiplies and upon which it depends as a species for survival in nature. A disease reservoir can be a source of infection for susceptible hosts of different species. An outbreak of disease in a susceptible population may occur when circumstances permit effective contact to be made with the reservoir of infection. Some infectious agents can also persist in a population by surviving for long periods of time within *vectors*.

A disease reservoir is any animal, plant or environment or combination of these in which an infectious agent normally lives and multiplies and upon which it depends as a species for survival in nature.

In the external aquatic environment, infectious agents are exposed to variations in temperature, concentrations of various chemicals (e.g. oxygen, salinity) and ultraviolet light. Once out of an aquatic environment, infectious agents are also susceptible to drying out (desiccation). The period of survival of an agent in the environment will depend on the particular set of conditions existing at the time. Some agents have the capability to produce resistant forms in response to harsh environments and thus persist for longer periods of time. For example, some species of fungi form spores which are quite resistant to harsh environments.

Ecology of disease

To investigate disease in natural populations, we need to understand the relationships among the hosts, agents and natural environments. These relationships determine the eventual observed pattern of disease both in time and space. For example, climate has a large impact on the geographical distribution of animal species, disease agents and potential disease vectors. The study of the relationship among animals, plants and their environment in nature is known as *ecology*. *Ecology of disease* extends this basic concept to include pathogens (agents of disease).

Ecology of disease is the relationship among animals, pathogens and their environment in a natural situation without intervention.

When humans intervene in natural ecological relationships such as by removing mangrove stands and intensively farming shrimp, organisms that may have always been present in the area but not causing a problem may cause disease. Similarly when application of fertilizer to farmland close to rivers may result in alterations to the local ecosystem with disease occurring such as with Epizootic Ulcerative Syndrome in wild fish.

This relationship is often expressed as a Venn diagram as shown in Figure 2.4. What this diagram says is that it is not until a particular set of conditions relating to the agent, host and environment come together will disease occur.

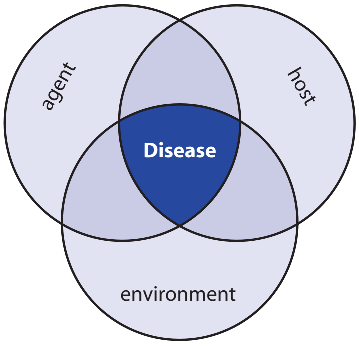


Figure 2.4 Venn diagram expressing relationship between agent, host and environment resulting in disease occurrence

Some of the components or factors belonging to each of these are shown below:

Agent	Host	Environment
infectivity	species	climate/weather
pathogenicity	genotype	water system
virulence	phenotype	water quality
immunogenicity	age	food
antigenic variation	sex	geology
survival	nutritional status	
	physiological status	
	pathological status	

Patterns of disease

The epidemiologist uses methods which document the patterns of disease in populations and by analysing these patterns, a better understanding of the cause of disease can be obtained. Although disease patterns are traditionally thought of as occurring in time (temporal patterns) and place (spatial patterns), we can extend this concept to include the identification and analysis of patterns as they occur in animals ponds, farms etc.).

Patterns by animal or other unit of study

Some species, sex or age class of animal can be more affected by disease than others even though they share a similar environment. For example, in Vietnam, Red Spot Disease (not the same as Epizootic Ulcerative Syndrome) is more common in grass carp than other carp species even when they are grown in the same pond. Many disease agents cause more severe disease in immature animals than adults, though this is not always the case. If the reasons for these differences can be understood, then it may be possible to design prevention and control strategies.

For example, say a farmer has a pond in which he rears both grass carp and Chinese carp. An outbreak of red spot disease occurs in the pond and on investigation it is found that 30% of a sample of grass carp have the disease while only 10% of Chinese carp are affected. Moreover, when affected animals are carefully examined, the disease in the grass carp appears to be generally more severe with larger lesions and more moribund animals. We have now described the pattern by animal and can proceed to analyse our findings. Through further investigation we might uncover why the disease was worse in grass carp and this may lead to a solution which in this case may be to reduce the proportion of grass carp in the pond in future.

If the unit of study was, say, a cage of fish, it may be observed that for a particular farm, cages which have a higher stocking density have a higher overall mortality per batch on average than those with a lower stocking density. This can be expressed in a number of ways and an example is shown in Figure 2.5.

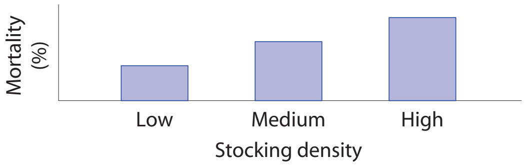


Figure 2.5 Histogram showing relationship between mortality and stocking density in caged fish batches on a single farm

Patterns by time

The timing of onset of cases tends to follow one of 4 patterns.

1. Cases may occur *sporadically*, that is they do not seem to be associated with any other identifiable factor, nor with each other, e.g. injuries in fish
2. Cases may occur regularly at a fairly constant level. The disease is often referred to as being *endemic*. It virtually always occurs, often at low levels, e.g. cut tails in eels
3. Cases may occur in time clusters, a pattern typical of outbreaks or epidemics e.g. yellow head or white spot syndrome in prawns,
4. If an epidemic takes international proportions and affects a large proportion of the population, it is termed a pandemic. Typical pandemics are AIDS and influenza in humans and parvovirus in dogs. White spot syndrome in prawns could possibly be described as pandemic in parts of Asia

Epidemic curves

A useful means to represent the temporal (time) pattern of disease events is to construct an epidemic curve, which can illustrate both the magnitude of the problem, (i.e. the number of new cases occurring, and the rapidity with which the epidemic progresses).

The epidemic curve represents in a graphic form the onset of cases of the disease, either as a histogram, a bar graph, or a frequency polygon. The frequency of new cases (or outbreaks) is plotted on the ordinate (y-axis) over a time scale on the abscissa (x-axis). A typical epidemic curve may be conceived of as having 4 and occasionally 5 segments as displayed in Figure 2.6.

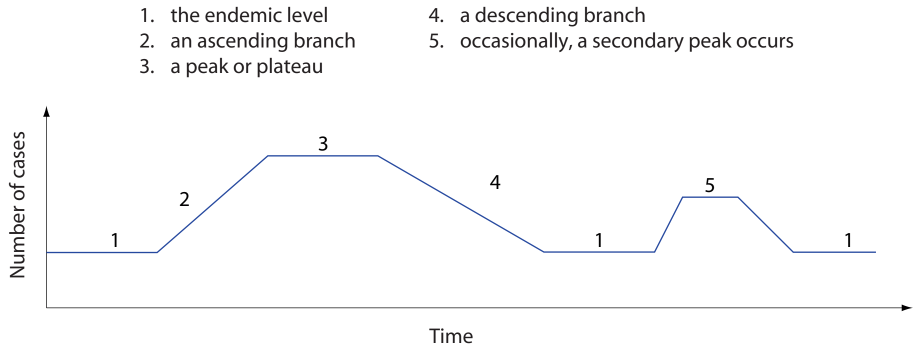


Figure 2.6 Five stages of an epidemic curve

An *epidemic* is said to occur when the frequency of cases (or outbreaks) in a population clearly exceeds the normally expected level for a given area and season. The slope of the ascending branch of the epidemic curve can reveal something about the type of exposure or about the mode of transmission of the disease agent. If transmission is fast and effective the slope of the ascending branch is likely to be steeper than if transmission is slow or if the incubation period is long. Exposure of a large number of animals to an agent at once or within a short period of time, e.g. through exposure to a common source, results in a *point epidemic*, typically a feed or waterborne disease. Most often toxins are associated with this type of outbreak but it is also possible for food or water spread of infectious agents if a large proportion of the population is exposed at once. The ascending branch of the corresponding curve would be almost vertical before reaching its peak. When the disease agent is transmitted via contact or vectors the ascending branch is more gradual and the resulting curve is typical of a *propagating epidemic*. The slope of the curve also depends on some agent characteristics such as its ability to survive outside the host; on some host factors such as contact rates, population density, etc.

The extent of the plateau and the slope of the descending branch are mainly a function of the availability of susceptible animals in the population. This is turn may be a function of such things as the composition of the population-at-risk with respect to their immune status — a concept referred to as herd immunity — or some intervention, such as vaccination or treatment. In addition, the contact rate among animals will also exert a major impact on the rate of spread. In a population experiencing an outbreak of an infectious disease, individuals may be classed as:

1. Susceptible
2. Resistant
3. Immune
4. Incubating ± infective
5. Diseased ± infective
6. Dead
7. Convalescent ± infective ± immune

During the course of an epidemic, individuals may move through a number of these states. Consequently, the numbers of individuals in any one category will not remain constant. The secondary peak in an epidemic curve is usually due to:

1. introduction of susceptible animals into the previously epidemic area, or
2. movement of infected animals from the epidemic area and contact with susceptible animals.

The main peak of the curve is at times preceded by a smaller peak which could represent the *index case*(s) (the first case to occur in the epidemic). The interval between this first peak and the beginning of the next or main peak could indicate the incubation period. Identifying the index case can be important in identifying the source of an outbreak.

In a closed population the pattern of disease may be easily appreciated, but when the population structure changes the pattern often becomes far more complex. For example, in most livestock populations there are births and introductions which often increase the number of susceptible animals; and deaths, culls and harvesting which decrease the number of immune animals. Furthermore, intervention by methods such as quarantine, treatment, vaccination or removal of a toxic source will potentially change the shape of the epidemic curve.

Identifying trends in the temporal distribution of disease

If the epidemic curve extends over a relatively long period of time and is based on frequent observations at short intervals, it may be examined for such patterns as cyclic fluctuations, seasonal variations, or long term (secular) trends as opposed to erratic or random fluctuations which have no recognisable pattern. Examples of these trends are shown in Figure 2.7.

Cyclical fluctuations exist when the variations occur at rather regular intervals; these intervals are usually longer than seasons.

Seasonal variation exists when the ups and downs occur at periodic intervals, coinciding with ‘seasons’, (where seasons be as short as a week or as long as a year, depending on what biological phenomena one is measuring).

Long term (secular) trends are long-term changes where, in addition to short term ups and downs, the curve either climbs or declines more or less steadily over an extended period of time, usually years.

Erratic variations occur in a totally unpredictable fashion.

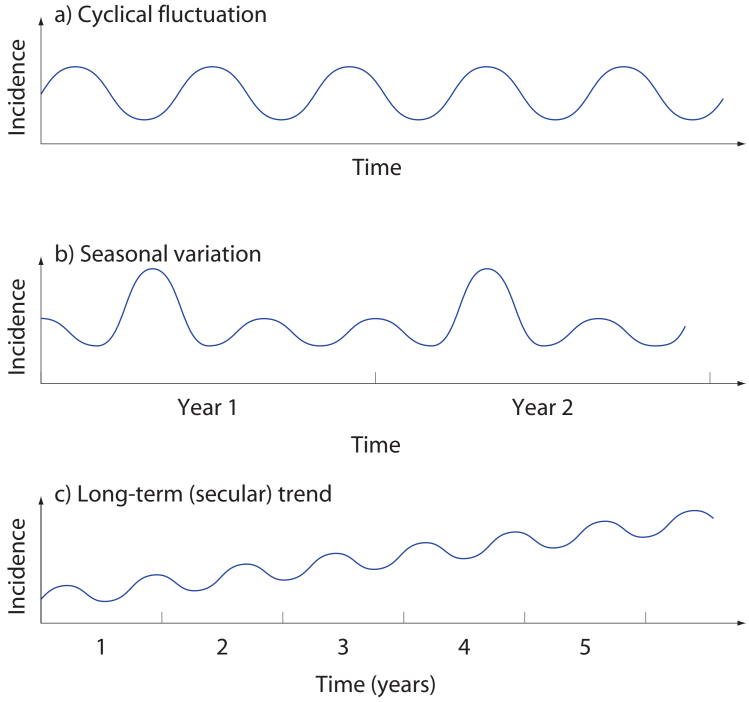


Figure 2.7 Temporal patterns in the distribution of disease

The types of time variations shown in Figure 2.7 may not always be obvious from the curve in its raw form. Cyclical and seasonal fluctuations can sometimes be identified by plotting moving averages of the raw data. The long term (secular) trend can be represented by a straight line which can be obtained using least squares regression. Time series analysis is a set of statistical methods, used to formally detect whether any of these types of variations exist and to determine the effect of each.

Why do epidemics occur?

Some of the reasons for the occurrence of outbreaks or epidemics due to infectious disease agents are listed below. There are probably many others.

- Recent introduction of the agent into a susceptible population.
- Recent introduction of a susceptible group of animals into an infected area.
- Recent increase in virulence or amount of the agent.

- Change in the mode of transmission of the agent.
- Change in host susceptibility or response to the agent.
- Factors causing increased host exposure or involving new portals of entry.

Patterns by place

Disease can also be characterised by spatial patterns. On a broad scale (such as across a country) such patterns are typically influenced by differences in environment and farming practices, but even at the local level (such as within a farm) spatial patterns may exist and if they can be identified and analysed, insight into disease cause and its prevention can be obtained. Computerised mapping and statistical methods for spatial analysis permit formal analysis of spatial patterns where large amounts of data are involved. However, simple hand drawings of the spatial distribution of disease can often reveal interesting patterns providing insight into the behaviour of disease. Figure 2.8 shows a theoretical pattern of disease in a series of fish culture cages where those further downstream tend to have a worse problem.

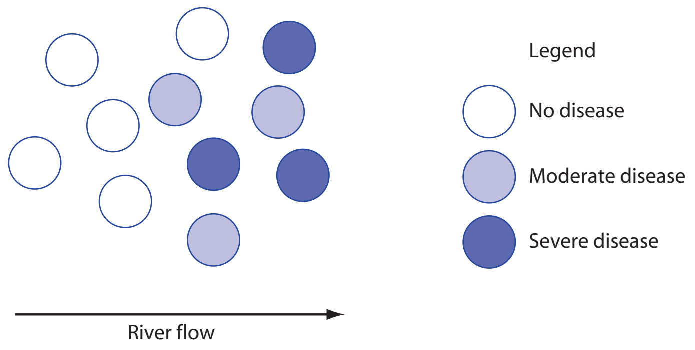


Figure 2.8 Theoretical spatial pattern of disease in fish culture cages in a river

Exercises

Exercise 2.1

a) Name three diseases of interest to you.

b) For the diseases you chose in part a), describe how they behave with regard to timing of disease events i.e. sporadic, endemic, outbreak, epidemic, pandemic. If you think that a particular disease may behave differently depending on the circumstances, describe the different circumstances which you believe cause it to behave differently.

c) For the diseases you chose in part a), select the unit of interest (individual animal, pond, farm etc.) you believe would be appropriate for an epidemiological investigation

Exercise 2.2

a) Describe the natural history of one of the diseases you chose in Exercise 2.1 a).

b) Describe the chain of infection for this disease.

c) List the possible means of transmission from animal to animal, and spread farm to farm and country to country for this disease.

Exercise 2.3

List three diseases where a carrier state and / or vectors exist or are suspected to exist.

Exercise 2.4

The following table shows stocking densities and estimated batch mortality rates for a fish farm. Describe the pattern.

Cage/ batch	Stocking density (fish/m^3)	Mortality for batch (%)	Cage/batch	Stocking density (fish/m^3)	Mortality for batch (%)
1/1	16	15	1/2	27	17
2/1	28	18	2/2	37	22
3/1	19	12	3/2	26	20
4/1	35	21	4/2	48	27
5/1	40	30	5/2	43	27
6/1	21	10	6/2	17	10
7/1	45	22	7/2	21	13
8/1	32	16	8/2	18	15
9/1	33	19	9/2	44	23
10/2	50	23	10/1	36	18

3 Cause of disease

The purpose of this section is to provide a greater insight into the meaning of cause from an epidemiological perspective and the process used to determine if a particular factor is causal for a particular disease or not. Risk factors, bias and effect modification (interaction) are also discussed in this section as they are central to the understanding of cause.

Objectives

At the completion of this section, participants should be able to:

- List possible causal factors for selected diseases.
- Identify which are necessary and which are sufficient causal factors for selected diseases.
- List sources of potential bias when studying selected diseases.
- Describe the process followed in determining whether a causal association exists

Cause of disease

Traditional view of causality

In the late 19th century, the traditional view of causality was deterministic, i.e. agent 'X' produced effect 'Y'. Specificity of both cause and effect was implied. The development of the Henle–Koch postulates reinforced this view and was helpful in formulating the link between microorganisms and disease:

1. The organism must be present in every case of disease.
2. The organism must not be present in other diseases.
3. The organism must be isolated from tissues in pure culture.
4. The organism must be capable of inducing disease in experiments.

In light of development of current knowledge, Koch's postulates were considered too restrictive in thinking about causality for a number of reasons:

1.	Multiple aetiologic factors	*e.g. epizootic ulcerative syndrome*
2.	Multiple effects of single causes	*e.g. pond bottom deterioration*
3.	Carrier state	*e.g. white spot in some crustaceans*
4.	Quantitative causal factors	*e.g. level of pH in pond water*
5.	Non-agent factors	*e.g. age, species, sex.*

Epidemiological view of causality

The epidemiologist interprets causality in quite a wide sense. This is somewhat different to the more traditional Henle–Koch view of the role of cause being restricted to aetiological agents with all other contributions being designated as 'contributing' or 'predisposing' factors. An epidemiological definition of a cause of a disease is 'an event, condition or characteristic that plays an essential role in producing an occurrence of the disease'.

A cause is an event, condition or characteristic that plays an essential role in producing an occurrence of the disease in question.

From an epidemiological perspective, factor 'X' causes disease 'Y' if and only if:

- X occurs prior to Y in time,
- a change in the frequency of X is associated with a change in the frequency of Y, and
- the association is not due to X and Y each being correlated with some other factor.

The presence of a fungus such as *Aphanomyces invadans* in a pond of fish may not of itself be a sufficient cause of disease. It may require a stress trigger to cause an outbreak of epizootic ulcerative syndrome in the group. Under this concept, the fungus is a *necessary* cause (no disease would occur if the fungus was not present) but not a *sufficient* cause (other component causes are required for disease to occur) of the particular disease, whereas the stress is neither necessary nor sufficient but can be a component of a sufficient cause. In fact, for any particular expression of a particular disease, there may be a range of possible sufficient cause complexes. This is shown theoretically in Figure 3.1.

A **sufficient cause** is a set of minimal conditions that inevitably produces disease.

A sufficient cause comprises a group of component causes.

A **necessary cause** is a component of every sufficient cause.

The completion of a sufficient cause is equivalent to the onset of disease.

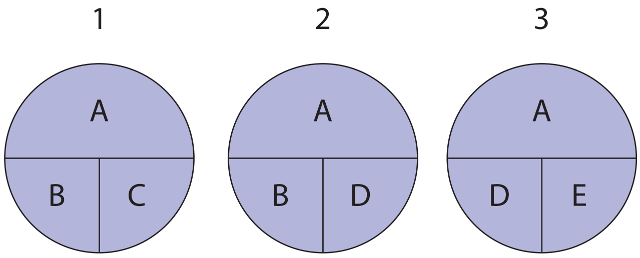


Figure 3.1 Three sufficient causes of a disease where A is a necessary but not sufficient cause while B, C, D and E are component causes but each is neither necessary nor sufficient

Examples of necessary causes are *Pfiesteria piscicida* for *Pfiesteria* toxicity, lead for lead poisoning and *Aphanomyces invadans* for epizootic ulcerative syndrome.

Some large fish kills in Maryland, USA appear to have been due to excessive nutrients in river systems, warm weather, river flow characteristics and exposure to *Pfiesteria piscicida*. These factors constitute a *sufficient cause* for *Pfiesteria* toxicity as illustrated in Figure 3.2.

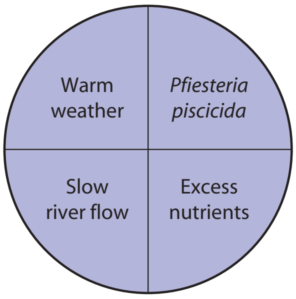


Figure 3.2 A *sufficient cause* for mass fish kill due to *Pfiesteria* toxicity in Maryland, USA

The relationships among causal factors can also be represented as a *web of causation*, which shows the direct and indirect associations among different causal factors. This is a valuable way of portraying relationships when the temporal sequence of events is known e.g. epizootic ulcerative syndrome in fish illustrated in Figure 3.3.

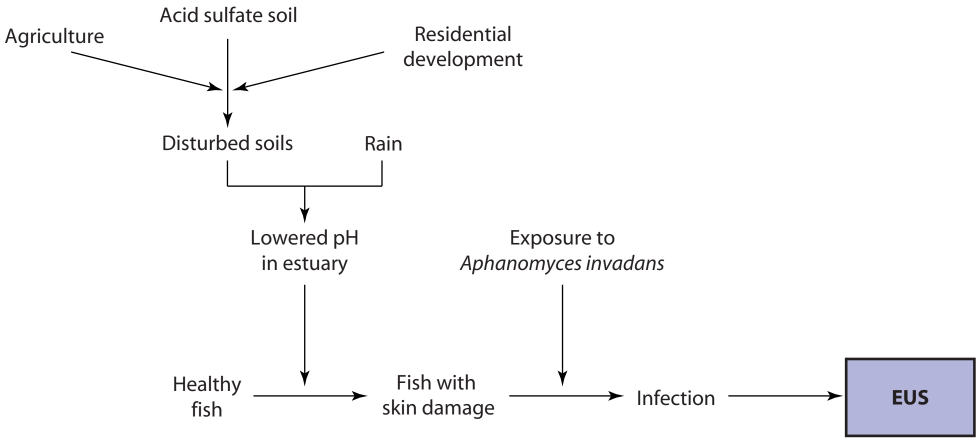


Figure 3.3 Some factors influencing the occurrence of epizootic ulcerative syndrome in Australian coastal estuaries

The challenge for the epidemiologist is to help identify some of the more important components of sufficient causes for a particular disease with the view to devising cost-effective intervention strategies at critical points to either prevent disease expression or reduce the production effects. For example, in the case of EUS, reduction in the disturbance of acid-sulphate soils should lead to a reduction in the occurrence of EUS.

Causal criteria

To prove causality is usually very difficult particularly for chronic diseases or for infectious diseases with long incubation periods. For example, the change in disease frequency may be too subtle to reliably detect or the appropriate measuring tool may not exist. Therefore, we end up having to make judgements about whether or not a particular factor causes a particular disease. A variety of criteria to assist in making judgements have been proposed and these include:

1. Time sequence
2. Strength of association
3. Dose–response relationship
4. Consistency of the association on replication
5. Biological plausibility
6. Specificity of the association
7. Coherence
8. Experimental evidence
9. Analogy
10. Intervention

Each of these is discussed below.

1. *Time sequence*

A cause must occur before its effect. This is the only absolute criterion on the list of ten which must be met. However, just having the correct time sequence is extremely weak evidence for causation. Time sequence is hard to establish in 6 instances:

- i. retrospective data
- ii. intermittent exposure
- iii. transient exposure
- iv. long latent period
- v. variable latent period
- vi. indefinite onset of disease

2. *Strength of Association*

The proportion of study units with the disease should be higher in those exposed to the proposed causal factor than those not exposed. The stronger the statistical association between the frequency of occurrence of disease and the risk factor, the more likely the association is a causal one, although there are exceptions.

3. *Dose–response relationship*

A demonstrable dose–response relationship between the risk factor is, like strength of association, strong evidence for causation if present, but only indeterminate evidence if absent. Absence of a dose–response relationship, though, isn't worrisome. Some risk factors (e.g. sex) don't necessarily have reasonable dose levels. Also, there can be a threshold phenomena (where a certain minimum amount is needed before there can be any response) and a plateau phenomena (where the response has peaked and it won't matter how much more of the risk factor is added).

4. *Consistency of the association upon replication*

This means that studies by different researchers in different locations have found similar results. The idea is that several independent studies won't all have committed the same errors in selection, information, and confounding. Further, it's unreasonable to think that the net effect of all the biases in each individual study would be similar to the net effect in most of the other studies. Consistency is moderate-to-strong evidence for causation.

5. *Biological plausibility*

It's easier to 'believe' an association if we can understand an underlying mechanism, a process through which 'X' could affect 'Y'. However, this is only moderate evidence at best. Also, absence of an 'explanatory mechanism' may simply reflect ignorance.

6. *Specificity of the association*

Specificity refers to the extent of '1-to-1' correspondence between the cause and the effect. Perfect specificity would imply that the risk factor has no effect other than the one being studied. Also, perfect specificity would imply that the risk factor was both a necessary cause (the disease can't happen without the risk factor) and a sufficient (the disease always occurs if the risk factor occurs). High specificity of the association is fairly strong evidence for causation. However, in general, we think of the specificity criterion as being weak evidence because we believe in 'multiple causation' and a 'web of causation'. We don't expect to see high or even moderate specificity, even in truly important causal associations.

7. *Coherence*

This implies that a cause-and-effect interpretation of an association does not conflict with what is known about the natural history and ecology of the disease.

8. *Experimental evidence*

Findings from laboratory and tank experiments should be consistent with the findings from epidemiological studies as is the case with EUS.

9. *Analogy*

It's easier to believe in the causal nature of an association if the situation is analogous to another one that we already know to be causal. However, this only is weak supportive evidence.

10. *Intervention*

If elimination of the putative causal factor results in a lower incidence of disease, this would provide additional evidence for causation.

A summary of the relative importance of the 10 criteria in assessing causation is provided in Table 3.1

Table 3.1 Relative importance of different criteria for assessing causation.

Criteria	Strength of evidence provided for causation if criterion present	Strength of evidence against causation if criteria absent
1. Time sequence	Very strong	Very strong
2. Strength of association	Strong	Minimal
3. Dose response	Strong	Indeterminate
4. Consistency on replication	Moderate to strong	Moderate
5. Biological plausibility	Moderate	Moderate
6. Specificity of association	Strong	Minimal
7. Coherence	Moderate	Weak
8. Experimental evidence	Strong	Weak to moderate
9. Analogy	Weak	Weak
10. Intervention	Strong	Moderate

Evaluating associations in a search for cause

An a*ssociation* is a relationship, a linkage in occurrence, or a dependency between 2 variables.

> An **epidemiological association** is a relationship between the frequency of occurrence of disease and the frequency of occurrence of factors, which may be causal for the disease in question.

Note that 'correlation' is not a synonym for 'association'. 'Correlation' has a precise statistical meaning among scientists. A correlation is a measure of a linear relationship between two ordinal or continuous variables. All significant correlations are associations, but you should not think of the reverse as being true because not all associations can be measured as correlations.

To evaluate an association that we suspect might be a causal association, we look at several pieces of information:

1. the chance that the observed association occurred just because of random variation (the P value);
2. the possibility that the so-called cause and effect are related intrinsically in some non-causal fashion ('night-and-day go together');
3. the chance that there was bias (systematic or non-random, error) in the study; and
4. the 10 causal criteria described above.

These four considerations are described in detail below.

1. *Random variation*

Random variation results from within- and among-animal variation, and measurement errors that occur due to imprecision (lack of perfect repeatability) in the measuring equipment or people using the equipment.

Because there are many usual sources of ordinary, anticipated random variation, we don't expect that all samples from the same population will have exactly the same mean or that factors will have the same frequency of occurrence. Rather, we anticipate and accept that there will be 'some' variation, even if the null hypothesis (H_o) is true. What we end up doing in statistical hypothesis testing is to predict how much random variation we'd expect to see if the null hypothesis is true. Then, we relate the observed difference between the groups to that predicted or expected variation. If the observed difference is relatively 'small', then we call the groups 'similar' or 'not significantly different'. (We say that the variables are 'not associated' or are 'independent'). We summarise the relative 'smallness' in the P-value, but it is an inverse relationship (big P value = small relative difference). These concepts are discussed further in Chapter 8.

2. *Intrinsic non-causal relationships*

Not all statistically significant associations are causal. We just said that the significance could be due to an 'unlikely but true' very large amount of random variation. Another explanation is that some associations simply are 'intrinsic' (for want of a better word). Night and day are associated in a regular repeating pattern that has no random variation. Most people who have a left hand also have a right hand. Neither hand 'causes' the other, nor is it useful to worry about the presence of the hands being linked by some useful-to-analyse ‘common cause’. The association is simply intrinsic — we don’t have a useful point for intervention. Another example of a non-causal association is the use of suntan lotion and drowning.

3. Bias

The concept of bias is introduced here but described more fully in a later section. Bias is a systematic (non-random) error in the data resulting from inadequacy in the study design or measuring instruments. It is not a matter of random variation or imprecision. The term, *validity* refers to a lack of bias.

While you can work hard to protect against bias, you can never rule bias out completely as an explanation. But, if you’re satisfied that bias seems reasonably unlikely, you can ‘go on’ to consider the information provided by the causal criteria. Actually, the criteria incorporate judgements about bias, so in fact many of these pieces of information are considered simultaneously.

4. Causal criteria

After ruling out chance, intrinsic relationships and bias as a possible explanation of an association, it is usual to consider the possible causal factors as ‘risk factors’ for disease.

Risk factors are those characteristics of some individual study units which, on the basis of epidemiological evidence, are associated with increased risk of disease. Risk factors may be either causal or non-causal, depending on the outcome when the causal criteria are applied. Non-causal risk factors are sometimes called risk markers. For example being nearer the ocean may be a risk factor for shrimp farms experiencing a white spot disease outbreak, but it may or may not be a cause. The process of considering the 10 causal criteria and deciding if a risk factor is causal or not is subjective and requires impartial judgement on behalf of the investigator.

> **Risk factors** are those characteristics of some individual study units which, on the basis of epidemiological evidence, are associated with increased risk of disease.
>
> Risk factors may be either causal or non-causal.

This whole process of evaluating associations to identify causes is summarised in Figure 3.4.

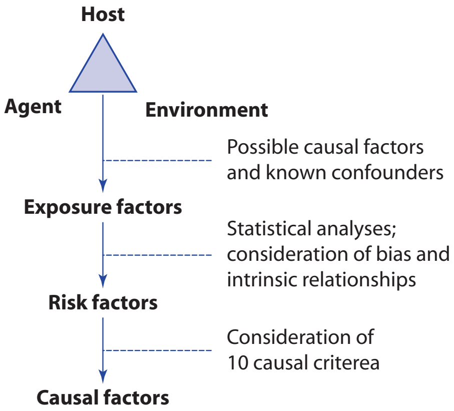

Host
Agent
Environment
Possible causal factors and known confounders
Exposure factors
Statistical analyses; consideration of bias and intrinsic relationships
Risk factors
Consideration of 10 causal criterea
Causal factors

Figure 3.4 Evaluating associations to identify causal factors

Developing hypotheses in the search for causes

A primary objective of many epidemiological studies is a search for causes of disease, establishment of the relationships among multiple causal factors, and quantification of the relative magnitudes of these factors. A vital step in this process is the development of hypotheses.

In a relatively informal way we generate hypotheses by essentially four methods:

1. ***Method of difference*** — if disease is more commonly associated with the factor than with the absence of the factor; e.g. epizootic ulcerative syndrome occurs more frequently in situations of low pH than in situations of normal pH.
2. ***Method of agreement*** — if the factor is common to cases or outbreaks which differ in other characteristics; e.g. with 'Mad Cow Disease' in England, all affected herds had cattle that were fed diets with animal protein but other factors such as exposure to vaccines, hereditary factors, potential exposure to toxins varied among herds.
3. ***Method of concomitant variation*** — if the intensity or frequency of the disease varies with the frequency or intensity of the factor; e.g. the prevalence and severity of epizootic ulcerative syndrome in fish may increase with reduced levels of lime application to ponds especially during preparation.
4. ***Method of analogy*** — if the pattern of the disease or its association with other factors is similar to patterns or causal associations for similar diseases; e.g. citrullinaemia in dairy cattle was thought to be a heritable condition because it is heritable in humans.

Bias in epidemiological studies

There is some variation in all biological systems. Within an animal, there are changes almost from moment-to-moment due to excitement, meals, circadian rhythms, season, pregnancy status, etc. When you examine any cross-section of a population, animal-to-animal variation (due to genetics, personality, owner-determined use, athletic training, etc.) adds even more variation on top of the within-animal variation. In addition, in the process of conducting studies there are random measurement errors that occur due to imprecision (lack of perfect repeatability) in the measuring equipment and in the people using that equipment.

Probably the biggest challenge when undertaking population based studies is the avoidance of bias. Bias occurs in an epidemiological study when the observations do not reflect the true situation because of some systematic error. For example, if a sample of fish is chosen from a pond and weighed, then the sample would need to be representative if the value was to be used to estimate the average weight for the whole pond of fish. If the sample was not representative then the weight estimate would be a biased estimate. This is a simple example of what is called selection or sampling bias.

> **Bias** is any effect at any stage of an investigation tending to produce results that depart systematically from the true values i.e. a systematic error (lack of validity) rather than a random error (lack of precision).

Although there are many different types of bias, they can be broadly classified into three general categories: *selection*, *measurement* and *confounding*. The differences between these categories are not always clear-cut and the strategies for preventing bias are not always exclusive to a single type. Many of the other terms you will read which describe a type of bias will in fact be specific examples of these three fairly broad categories of bias.

Selection bias

> **Selection bias** is a systematic error in the way that the samples of study units were drawn from their underlying populations, or in the way that study units were assigned to interventions

The potential for selection bias is very prominent in the selection/assignment procedure of an epidemiological study. For example, if in a cross sectional sample of prawns from a pond, the easy-to-catch prawns in the shallow water near the edge of the pond were caught the sample may not be representative of all prawns in the pond and a selection bias would result.

If, in an intervention study, an investigator can't describe a formal decision rule that he or she used to assign treatments, then you need to worry about selection bias because of the extreme potential for subjectivity in the assignment 'decision'. An important source of almost unavoidable selection bias is that volunteers are different from non-volunteers. Similarly, responders are different from non-responders. Another important source of selection bias arises from differences in access to extension activities and technical advice. For example, prawn farms that have regular input from trained specialists are very unlikely to be representative of all prawn farms.

There are many strategies for protecting against selection bias. These include having formal, clear criteria for study units to be acceptable for inclusion and exclusion from the study. Also, the comparison group should be truly appropriate. (For example, a comparison group that, even with the risk factor under study, probably couldn't develop the outcome would be inappropriate.) *Randomisation* is a selection or assignment method which involves a formal decision rule that makes use of chance. A truly random assignment procedure offers important protection against selection bias. Randomisation avoids any challenge to the subjectivity of the investigator.

Measurement bias

> **Measurement bias** is a systematic error in the way that data were gathered or measured.

We understand that all data potentially are subject to imprecision (random variation). We understand that there even may be a systematic error due to an inaccuracy in the measuring equipment used when compared to the best possible equipment. What especially concerns us here, though, are the systematic errors that are applied differentially between groups. For example, if you're looking for evidence of previous exposure to a chemical that is suspected as a causal factor for a disease in fish, farms with and without the disease may be investigated regarding chemical use. In such a case, all farms should be questioned with equal vigour, and with equal adherence to non-leading questions to avoid triggering 'recall bias' as much as possible.

When investigating causal associations equal effort should be expended in searching for old records for the diseased and the non-diseased groups. If you're following farms or ponds which are exposed and not exposed to a suspect causal factor you must guard against checking the exposed groups twice as frequently or using more sensitive methods of disease detection in the exposed group. This is to avoid 'diagnostic work-up bias'.

There are lots of research design features that help to protect against measurement bias. Some of the more obvious include:

1. Blind the measurer/data collector.
2. Get better measuring equipment or tests.
3. Standardise the protocol for data collection.
4. Use prospective rather than retrospective data.
5. Use objective rather than subjective measurement criteria.

Confounding

In addition to chance and bias, an epidemiological study may be affected by confounding (which is also a form of bias, although some epidemiologists like to characterise it separately). Confounding arises when the purpose of the study is to identify risk factors for disease. It occurs when two risk factors are interrelated and it is incorrectly concluded that one of the factors is causally related to the disease in question. For example, it might be observed that shrimp in ponds with cloudy water do not grow as well as those in clearer water. We might conclude from this that light penetration of the water is important for normal growth. However, it may be that the cloudiness is due to the presence of particular algal species in the water which inhibit growth of the shrimp through toxin production. In this theoretical example, confounding has meant that we have incorrectly concluded that light penetration is associated with poor shrimp growth when the true cause was the presence of toxic algae. The relationships are represented in Figure 3.5 below.

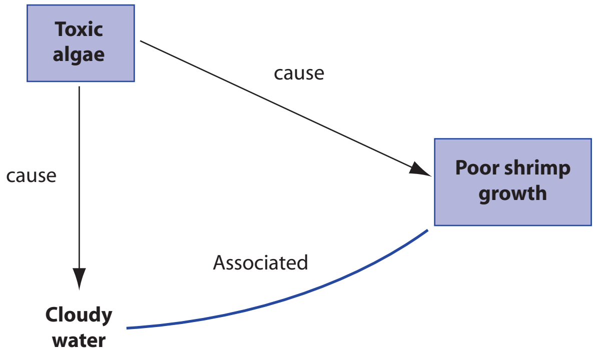


Figure 3.5 Theoretical example of relationships resulting in confounding leading to incorrect conclusions on the cause of poor shrimp growth

Confounding is one of the critical problems to watch for when undertaking an epidemiological study. It would probably be better to name the problem 'confusing' as it occurs when the effects of two or more factors are mixed and it is difficult to determine which factors are truly 'causal' in an epidemiological sense.

Confounding is a systematic error that results from unaccounted-for differential distributions of particular covariates. Confounding is therefore a form of bias.

These covariates are 'third' variables (i.e. not either of the two of primary interest: the so-called cause and it's effect). These third variables are called 'confounders' because they mask (alter the appearance of) the true relationship under study. In order to confound, the third variable must have a true association with both the risk factor and the disease. In Figure 3.6 a), the true cause is the confounding factor which leads to the finding of a 'spurious' association between a suspected cause and the disease because of the non-causal association between the confounder (true cause) and the suspected cause.

To be a **confounder**, an exposure factor must:

1. be a risk factor for the disease in question;
2. be associated with the exposure factor under study in the source population; and
3. **not** be affected by the exposure factor or the disease. In particular, it can not be an intermediate step in the causal path between the exposure and the disease.

Although it may not be appropriate for all examples of confounding, an easy way to visualise confounding is to think of *common-cause associations*. A *common-cause association* is an apparent but non-causal association between two variables in which each of the two variables is the result/effect of the same previous, third variable as illustrated in Figure 3.6 b).

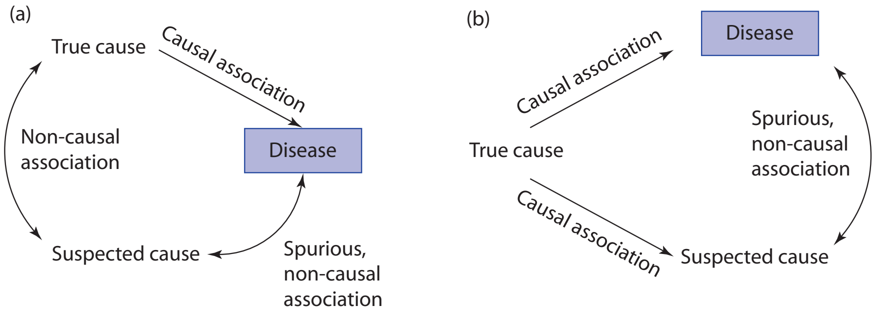


Figure 3.6 Two models of causal and non-causal associations

Confounding is situation-specific, and you have to know something about the biology and logic of the situation to guess at things that should be explored as confounders. In general, in most studies you should at least think about the following kinds of variables: species, age, breed, season, sex and physiological status (e.g. spawning, nursing and growing), level of production.

Whatever else is pertinent to the specific risk factor and disease process, one of the best protections against confounding bias is randomisation. Randomisation assures that, on average, most confounders will be distributed roughly evenly between treatment groups or sub-samples. It is not a guarantee that all will be exactly equal, so even if randomisation is used (and especially with small group sizes), it is still important to look at the baseline descriptions of the randomised groups. However, you should not re-randomise if an important variable is unevenly distributed, rather, you should adjust in the analysis. The reason randomisation is so important, though (and one of the reasons you don't 're-randomise'), is that randomisation is the only available method for controlling confounding due to unknown or unmeasured variables. All other methods to control confounding assume that you know enough to have a measurement of the potential confounder. These other methods for controlling confounding include:

1. Restriction of entry into the study
2. Stratification (and its extreme: matching) in the design
3. Standardisation of rates
4. Stratification in the analysis
5. Adjustment using multivariable statistical methods in the analysis

Figure 3.7 illustrates the stages of a study where each of the 3 main types of bias must be addressed. Bias at any stage of the study will mean that any generalisations back to the target population of interest are invalid.

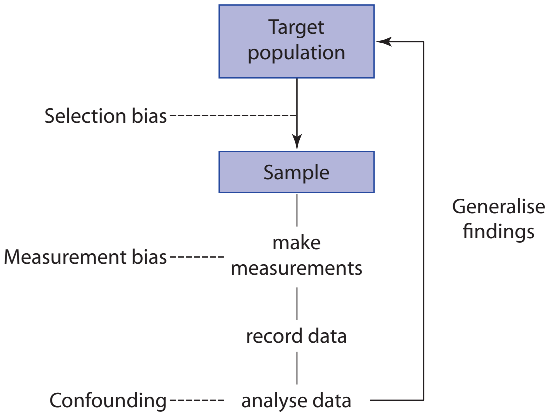


Figure 3.7 Occurrence of bias in an epidemiological study. Elements of the study are shown inside the clear, rectangular box

Effect modification (interaction)

Where two or more risk factors play a role in the causation of a disease, the possibility exists for effect modification (also known as interaction) to occur between two or more of the factors. Interaction is different from confounding.

> **Effect modification (interaction)** is said to occur when the incidence of disease in the presence of two or more risk factors differs from the incidence expected to result from their individual effects.

When effect modification occurs, the effect can be greater than what we expect (positive interaction or synergism) or less than what we would expect (negative interaction or antagonism). The problem when evaluating effect modification is to ascertain what we would expect to result from the individual impacts of the different risk factors.

For example, say we find that the incidence of EUS is 5% in fish in ponds with acidic water but with a smooth lining, 2% in ponds with neutral water but with a rough lining and 15% in ponds which have both acidic water and a rough lining. The 15% incidence is a lot higher than we would expect if the two factors of acidity and rough pond lining operated independently to increase the risk of EUS. We would therefore suspect synergy between these two risk factors and would need to investigate further.

In complex epidemiological studies, information is often collected on a wide variety of factors to identify the important risk factors for the disease of interest. To achieve this, statistical analysis is required to sort out the effects of confounding and effect modification among the potential causal factors for the disease.

Exercises

Exercise 3.1

a) For the diseases you chose in Exercise 2.1 a), list the factors you would like to evaluate to see if they are associated with the particular disease.
b) Group each of these factors according to the level at which they operate i.e. at the individual animal level (e.g. age), at the group level (e.g. pH level of the pond), at the farm level (e.g. water source), at the regional level (e.g. soil type).
c) Identify which of these factors are potentially necessary causes for the disease in question

Exercise 3.2

What the major limitations of Koch's postulates in investigating the cause of a disease?

Exercise 3.3

Give examples of how hypotheses could be formulated concerning factors associated with disease using:

a) The method of difference
b) The method of agreement
c) The method of concomitant variation.

Exercise 3.4

List some methods which are appropriate for reducing the chance of the

a) selection bias
b) information bias
c) confounding bias

occurring when studying disease in aquaculture environments.

4 Diagnosis and screening

The purpose of this chapter is to briefly consider methods of diagnosing disease and the important characteristics of laboratory tests from an epidemiological perspective.

Objectives

At the completion of this section, participants should be able to:

- List different methods to diagnose disease
- Understand the difference between diagnosis and screening
- Evaluate precision, sensitivity and specificity
- Interpret test results for disease diagnosis and screening

Diagnosing disease

Clinicians and pathologists devote substantial time in arriving at the 'correct' diagnosis when investigating disease. The diagnosis is reached through a process of application of various tests and comparing the results from each. Competent investigators use good judgement, a thorough knowledge of the literature, past experience, diagnostic tests and intuition to organise their observations to reach a diagnosis.

A *test procedure* is usually taken to mean a test performed on a specimen in a laboratory. However, the principles discussed in this chapter also apply to information obtained from the clinical history, physical examination, gross pathology etc.

Table 4.1 lists a number of ways that a disease might be diagnosed. These methods may be used alone or in combination to arrive at a final diagnosis. However, all of these methods are subject to error which will be discussed in some detail later in this chapter.

Table 4.1 Some methods used to diagnose disease

• history	• microbiology	• economics
• behaviour	• serology	• biochemistry
• clinical signs	• epidemiology	• physiology
• physical examination	• response to therapy	• imaging
• autopsy	• production	• transmission tests

Three levels of diagnosis have been defined to assist in the surveillance and control of aquatic animal disease in Asia. These are summarised in Table 4.2 (from Reantaso et al., 2000). Level I diagnosis can be made on the farm without any laboratory confirmation. Level 2 diagnosis requires some laboratory support, while Level 3 requires the use of advanced laboratory techniques.

Table 4.2 Three levels of diagnosis for surveillance purposes

Level – Activities	Skills and Equipment	Responsibility	Requirements
Level I – Activities Observation of animal and the environment; Clinical examination; Gross pathology	Knowledge of normal (feeding, behaviour, growth of stock, etc.); Frequent/regular observation of stock; Regular, consistent record-keeping and maintenance of records – including fundamental environmental information; Knowledge contacts for health diagnostic assistance; Ability to submit and/or preserve representative specimens.	Farm Workers/ Managers; Fisheries Extension Officers; Field Veterinarians; Local Fisheries Biologists.	Field keys; Farm record keeping formats; Equipment list; Model clinical data sheets; Pond-side check list; Protocols for preservation/ transport of samples.
Level II – Activities Parasitology Bacteriology Mycology Histopatholo gy	Laboratories with basic equipment and personnel trained/ experienced in aquatic animal pathology; keep and maintain accurate diagnostic records; Preserve and store specimens; knowledge of/contact with different areas of specialisation within Level II Knowledge of who to contact for Level III diagnostic assistance.	Fish biologists; Aquatic Animal Veterinarians; Parasitologists; Mycologists; Bacteriologist; Histopathologists; Technicians.	Model laboratory record-keeping system; Protocols for preservation/transport of samples to Level III; Model laboratory requirements/equipment/ consumable lists; Contact information for accessing Level II and Level III specialist expertise; Asia Region Aquatic Animal Disease Manual; OIE Diagnostic Manual for aquatic Animal Diseases; Regional General Diagnostic Manuals.
Level III – Activities Virology Electron Microscopy Molecular Biology Immunology	Highly equipped laboratory with highly specialised and trained personnel; keep and maintain accurate diagnostic records; Preserve and store specimens; Maintenance of contact with people responsible for sample submission.	Virologists; Ultrastructural histopathologists; Molecular biologists; Technicians.	Model Laboratory requirements, equipment, consumable lists; Model job descriptions skills for requirements; Contact information for reference laboratories; Protocols for preservation of samples for consultation/validation; OIE Diagnostic Manual for Aquatic animal diseases; General molecular and microbiology diagnostic references; Asian Aquatic Animal Health Guide.

It is important when investigating disease on a population basis that consistency of diagnosis is maintained within the particular study, regardless of the method(s) of diagnosis used. This involves developing a *case definition,* which is applied uniformly to all groups in the study. Failure to do so will lead to bias in the study.

A **case definition** is a set of standard criteria for deciding whether an individual study unit of interest has a particular disease or other outcome of interest. The unit may be an individual animal or a group of animals such as a pond of fish, a pen of pigs, a village or a farm.

For example, the investigator may be interested in comparing the occurrence of a particular disease in farmed fish in two different countries. Great care would need to be exercised if in one country the disease was diagnosed based on microbiological findings whereas in the other country a pathological basis was used.

Some examples of case definitions which might be used when investigating white spot in shrimp are given in Table 4.3. The choice of case definition will depend on the objectives and methods used in the investigation.

Table 4.3 Examples of case definitions for white spot in shrimp

Study unit	Case definition
Animal	A shrimp with one or more visible, discrete white patches on the inside of the carapace.
Animal	A shrimp which returns a positive PCR result for white spot syndrome virus.
Pond	A pond where one or more shrimp have one or more visible, discrete white patches on the inside of the carapace.
Pond	A pond where one or more shrimp return a positive PCR result for white spot syndrome virus.
Pond	A pond subject to emergency harvest because, in the opinion of the manager, there is a risk of mass mortality from white spot syndrome.

It is often useful to have definitions for a *suspect case* based on field observations (history, clinical signs, gross pathology etc.) and a *confirmed case* based on laboratory findings especially where it may take some time to confirm cases. Where a previously unrecognised and potentially serious syndrome is being investigated, it is advisable to initially use a very broad case definition to minimise the risk of missing any cases. In this instance, a revised classification can be applied later when time permits.

A distinction is usually also made between the use of tests in *screening* and *diagnosis*. Screening begins with apparently healthy individuals whereas diagnostic testing begins with animals showing signs consistent with the disease in question. Screening tests are used for the presumptive identification of unrecognised disease in apparently healthy populations. A screening test should have both high sensitivity and precision, be easy to perform and of low cost if a large number of individuals are to be tested. A screening test is not intended to be diagnostic: individuals which return a positive result in a screening test should be subject to a more thorough investigation to establish a diagnosis. On the other hand diagnostic tests are used to distinguish between animals that have the disease in question and those which have other diseases on the differential list.

Screening –	**Diagnosis –**
· applied to healthy population	· applied to sick individual
· seeks unrecognised disease	· differentiates among likely diseases
· high sensitivity	· high specificity
· large numbers tested	· small numbers tested
· low cost important	· cost not so important

Accuracy of test procedures

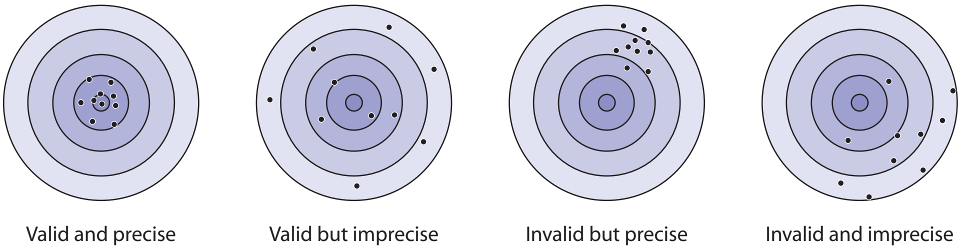
An accurate test is both precise and valid. In other words the result is repeatable (a measure of precision) and gives a true measure of the value being measured (sensitive and specific – measures of validity). More formally, precision is defined as a lack of random error while validity is a lack of systematic error or bias.

A **precise** test has a low level of random error i.e. a high level of repeatability.

A **valid** test has a low level of systematic error (bias) i.e. high **sensitivity** and **specificity**.

A test can be precise without being valid and vice versa. A good test is both precise and valid.

The concepts of precision and validity are most easily understood by thinking of shooting at a target as shown in Figure 4.1.



Valid and precise

Valid but imprecise

Invalid but precise

Invalid and imprecise

Figure 4.1 Validity and precision in test procedures

Precision

Tests performed on presumably identical material under apparently similar conditions do not, in general, yield identical results. This variation is attributed to unavoidable random error inherent in every test procedure because factors that may influence the result of a test can not all be completely controlled. When interpreting test results, this variability must be taken into account. There are many different factors which contribute to the variability of a test procedure, including:

1. uniformity of test material;
2. transport and storage of test material;
3. reagents;
4. equipment and its calibration;
5. operator;
6. environmental conditions such as temperature, humidity, light, air pollution.

Precision is a general term used to describe the variability between repeated tests on apparently identical material. A test with a high level of precision has low variability and vice versa.

The British Standards Institution (BSI) defines two complementary measures required to assess the precision of test methods: repeatability and reproducibility. Repeatability refers to the tests performed under conditions that are as constant as possible in the one laboratory by one operator using the same equipment over a short period of time. An example is replicate wells on the one plate in an ELISA procedure. Reproducibility refers to tests performed under widely varying conditions in different laboratories at different times by different operators. Thus, repeatability and reproducibility are two extremes, the first measuring the minimum and the second the maximum variability in results due to random error. The BSI Standard goes on to provide methods for the determination of repeatability and reproducibility of test procedures based on specified experimental designs.

Other measures of precision can be used for measurements made on a continuous scale. These include the error standard deviation determined from duplicate measurements for each specimen, coefficient of variation (error standard deviation as a percentage of the mean), line of identity (where the fitted regression line for the duplicate measurements is compared to the line of identity which has a slope of 1 and passes through the origin).

Validity

The validity of a test procedure is measured by its *sensitivity* and *specificity*.

Sensitivity from a population perspective is the proportion of animals with the disease (or infection) of interest which test positive (i.e. proportion of true positives). This equates to the laboratory definition where it means the ability of an analytical method to detect very small amounts of the material (such as an antibody or antigen). Thus a test which is highly 'sensitive' from a laboratory perspective is also likely to be 'sensitive' from a population perspective.

> **Sensitivity** (True Positive Rate): The proportion of animals with the disease of interest who test positive. Sensitivity is also defined as the conditional probability that a test will correctly identify those animals that are infected (Pr T+ | D+).

Factors affecting sensitivity in antibody assay estimates include:

- number of animals in study
- method used to determine disease or infection status
- stage of disease
- cut-off point selected
- anti-species conjugate type
- non-specific inhibitors
- incomplete antibody
- suppression of immunoglobulin production

Specificity from a population perspective is the proportion of animals without the disease of interest which test negative (i.e. proportion of true negatives). This equates to the laboratory definition where it means the ability of the test to react only when the particular material is present and not react to the presence of other compounds. Thus a test which is highly 'specific' from a laboratory perspective is also likely to be 'specific' from a population perspective.

Specificity (True Negative Rate): The proportion of animals without the disease of interest who test negative. Specificity is also defined as the conditional probability that a test will correctly identify those animals that are not infected (Pr T– | D–).

Factors affecting specificity in antibody assay estimates include:

- number of animals in study
- method used to determine disease or infection status
- cut-off point selected
- anti-species conjugate type
- non-specific inhibitors
- group cross-reactions
- non-specific agglutinins

In order to estimate the test attributes of sensitivity and specificity, one must conduct the test on specimens from a number of animals for which the status of infection or disease is known. Often, experimental infections are used to determine these parameters, although field samples collected from known diseased and non-diseased animals (if a 'gold' standard is available) are much better. Results from such a study can then be tabulated in a two-by-two table from which the sensitivity (Se) and specificity (Sp) can be calculated as shown in Table 4.4.

Sensitivity and specificity are generally considered 'fixed' values, although for some diseases the amount of the agent and stage of disease may affect test sensitivity. Specificity may vary with geographical region because of differences in cross-reacting agents.

Confidence intervals for test characteristics can be easily calculated for sensitivity and specificity estimates using computer software such as EpiInfo. The precision of the estimates will improve as the sample size increases. Ideally, there should be several hundred infected and non-infected animals for these calculations. For some diseases, sensitivity and specificity may be difficult or costly to determine but such characteristics should be known for any test that is used in a national eradication program or for official health status certification.

Table 4.4 Classification of individual animals according to test results (T+, T–) and disease status (D+, D–):

	State of nature		
Test result	Diseased (D+)	Not diseased (D–)	Total
Positive (T+)	a	b	a + b = T+
Negative (T–)	c	d	c + d = T–
Total	a + c = D+	b + d = D–	n = a + b + c + d

a = true positives; b = false positives; c = false negatives; d = true negatives

Test characteristics Sensitivity, Se = a/(a + c)
Specificity, Sp = d/(b + d)

Prior probabilities Probability of having disease, Prevalence = (a + c)/n
Probability of not having disease, (1–Prevalence) = (b + d)/n

Example with numbers for a theoretical test procedure:

	True disease status		
Test result	D+	D–	Total
Positive (T+)	35	0	35
Negative (T–)	2	43	45
Total	37	43	80

Se = 35/37 = 94.6 (95% Confidence Interval: 80.5–99.1)
Sp = 43/43 = 100.0 (95% Confidence Interval: 89.8–100)

There is an inverse relationship between Se and Sp for a particular test, as shown in Figure 4.2 below for a procedure with a continuous reading such as an ELISA. The frequency distribution of test results in a healthy population (D–) usually overlaps with the frequency distribution of test results in the diseased population (D+). Animals to the right of the cut-point (C–C) are classified as reactors (diseased or infected) and animals to the left are classified as negative (non-infected). If fewer false positives are required, C–C is moved to the right; specificity increases and sensitivity decreases. However, if fewer false negatives are required, C–C is moved to the left: sensitivity increases and specificity decreases.

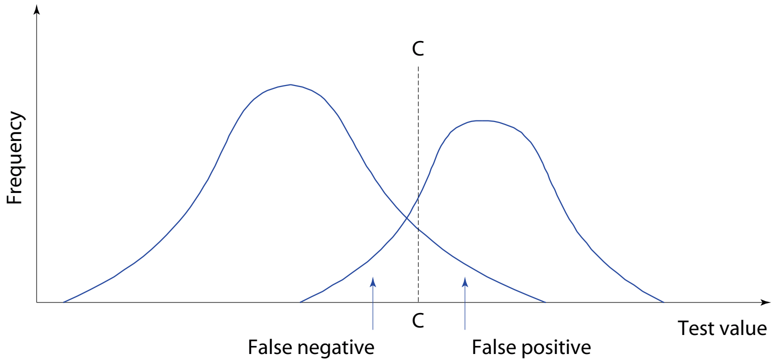


Figure 4.2 Frequency distribution of test results measured on a continuous scale for healthy and diseased groups of animals with a theoretical cut-off point (C–C) separating reactors from non-reactors

Selection of the appropriate cutoff value will depend on a number of issues including the relative cost of false positives and false negatives, the stage of an eradication program, if any, and the availability of other tests. An important consequence of imperfect specificity (i.e. < 100%) is that if a large number of animals are tested from a population free of the disease in question, there is a significant chance of abnormal results. For example, if 10 independent tests, each of specificity 90% were performed, the probability of at least 1 positive test is 65%.

Test interpretation at the individual animal level

Predictive values are conditional probabilities which answer the two related questions: What proportion of the test positive animals are really infected, and what proportion of the test negative animals are truly not infected? Using probability notation, the predictive value of positive test results (PPV) is the P(D+ | T+); for negative test results (NPV) it is the P(D– | T–). Formulae for calculating predictive values are based on Bayes' theorem of conditional probability and are as follows:

$$\text{Positive predictive value} = \frac{a}{a+b} = \frac{\text{Prev} \times \text{Se}}{\text{Prev} \times \text{Se} + (1-\text{Prev}) \times (1-\text{Sp})}$$

$$\text{Negative predictive value} = \frac{d}{c+d} = \frac{(1-\text{Prev}) \times \text{Sp}}{(1-\text{Prev}) \times \text{Sp} + \text{Prev} \times (1-\text{Se})}$$

Se: sensitivity; Sp: specificity; Prev: pre-test probability of disease (or true prevalence).

Predictive values are functions of prevalence and the test characteristics of sensitivity and specificity. As prevalence declines so does positive predictive value. The converse is true for negative predictive value (Table 4.5 and Figure 4.3). If the sensitivity and specificity of a diagnostic test are known for a particular target population, then predictive value graphs can be drawn for the range of all possible pre-test probabilities of disease from 0 to 1 (100%).

Table 4.5 Effect of prevalence on positive predictive value (PPV) with a hypothetical test procedure (Se and Sp = 0.95)

Prevalence (%)	0.1	1	2	5	10	50	90	100
PPV	1.9	16.1	27.9	50.0	67.8	95.0	99.4	100

The important point that this table indicates is that despite using a good test (Se and Sp = 0.95) most reactors are non-infected (false positives) when the disease is present in the population or region at low prevalence. Of the 2 test properties, it can be shown that specificity exerts a greater influence on PPV than does sensitivity. On the other hand, sensitivity exerts a greater influence on negative predictive value (NPV).

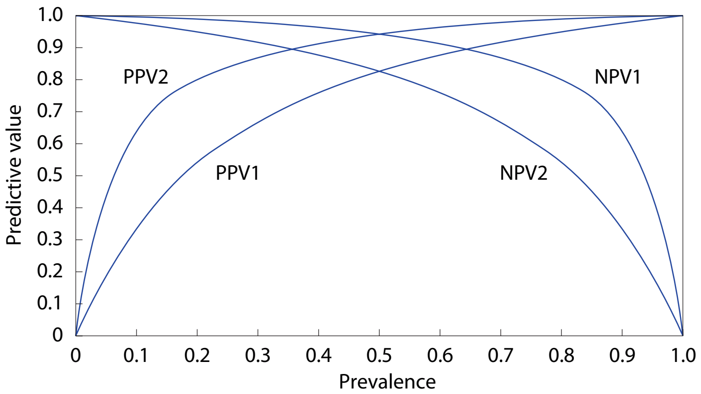


Figure 4.3 Relationship between prevalence (or pre-test probability of disease) and predictive values for two tests. Test 1 has Se = 95% and Sp = 80% while test 2 has Se = 80% and Sp =95%. Note that positive predictive values are better for test 1 (PPV1) because of its higher specificity while negative predictive values are better for test 2 (NPV2)because of its higher sensitivity.

Some strategies that can be used to improve the predictive value of a positive test result include:

1. Test ‘high risk’ groups — those with clinical signs rather than normal animals.
2. For the same test use a cutoff with to provide higher specificity or use another test with a higher specificity than the original test.
3. Use multiple tests for interpretation of results (see below).

Some rules of thumb

With an understanding of the principles of predictive values, the following rules of thumb for using tests in the diagnostic process at the individual animal level can be recommended (modified from Baldock, 1988):

1. Decide on the pre-test probability of disease or non-disease after clinical work up, but before performing any diagnostic tests. Revise probability estimates in the light of new information.
2. Consider what action will be taken for both a positive and negative test outcome. If the planned actions are the same regardless of test outcome, then performance of the test may not be justified.
3. If the objective is *to confirm a likely diagnosis* (the ‘rule-in’ situation), then choose a test has *high specificity* (>95%) and at least moderate sensitivity (>75%). If a positive result is returned, then it is highly likely the individual has the disease in question (PPVs are high for tests with high specificity). If a negative result is returned, then further diagnostic work up is required.
4. If the objective is to *confirm that an individual is free from a particular disease* (the ‘rule-out’ situation), then choose a test with high sensitivity (>95%) and at least moderate specificity (>75%). If a negative result is returned, then it is highly likely the individual is free from the disease in question. If a positive result is returned, then further testing is required with more specific tests to ascertain whether or not the result was a false positive result or not.

Multiple testing

Two or more tests can be used either sequentially or simultaneously and results interpreted in series or parallel. In *parallel interpretation*, an animal is considered positive if it reacts positively to either or both tests — this increases sensitivity but tends to decrease the specificity of the combined tests. In *series interpretation*, an animal must be positive on both tests to be considered positive — this increases specificity at the expense of sensitivity.

Example: Test 1: Se = 50%; Sp = 98.7% Test 2: Se = 60%; Sp = 98.6%

Test 1	Test 2	Infected	Uninfected
+	–	30	70
–	+	50	80
+	+	70	30
_	–	50	7620
	Total	200	7800

	Sensitivity	Specificity
Test in parallel	150/200 = 75%	7620/7800 = 97.7%
Test in series	70/200 = 35%	7700/7800 = 99.6%

In general, the greater the number of tests involved, the greater the increase in sensitivity or specificity depending on the method of interpretation that is used. To minimise the overall misclassification rates with many tests requires relatively elaborate statistical techniques such as discriminant analysis. Note that in practice, the results may not be this good, because if the 2 tests are biologically correlated, they will tend to give similar results on samples from the same animal.

Test agreement

For many new tests, the true state of nature is not known and an investigator may only be able to measure how well a newly developed test agrees with an existing test. Unfortunately, sensitivity and specificity are often not available for the original test either! For the same specimens, the investigator records the appropriate frequency data into the 4 cells, a (both tests positive), b (test 1 positive and test 2 negative), c (test 1 negative and test 2 positive), and d (both tests negative). The value Kappa (K), a measure of relative agreement beyond chance, can be calculated.

Kappa has many similarities to a correlation coefficient and is interpreted along similar lines. It can have values between –1 and +1. Recommended criteria for evaluating agreement are:

Kappa value	Evaluation
>0.8–1	Excellent agreement
>0.6–0.8	Substantial agreement
>0.4–0.6	Moderate agreement
>0.2–0.4	Fair agreement
>0–0.2	Slight agreement
0	Poor agreement
–1–<0	Disagreement

If two tests agree well, they could be equally good or equally bad!. It may be possible to justify use of a newly developed test, if it agrees well with a standard test and if it is cheaper to run in the laboratory. If two tests disagree, one test is likely to be better than the other although there may no way to tell which is better!

Example:

	Test 2 results		
Test 1 results	+	–	Total
+	121	36	157
–	34	931	965
Total	155	967	1122

Kappa = 0.74. Good agreement but no indication of which test is more sensitive or specific.

Estimation of true prevalence from apparent prevalence

Application of a test of known performance characteristics in a population often results in some positive results. An important question that the epidemiologist must answer is whether the positives are true or false, i.e. is infection present in the population? True prevalence can be estimated from apparent prevalence (AP) by the formula:

$$\text{True prevalence} = \frac{\text{AP} + \text{Sp} - 1}{\text{Se} + \text{Sp} - 1}$$

which has a solution for situations other than when Se + Sp = 1. All values are expressed as proportions (between 0 and 1) rather than percentages for these calculations. When the prevalence is 0, the AP = 1 – Sp, the false positive test rate. This calculation assumes that test characteristics are known which is often not the case. In these situations, rational interpretation of results is difficult unless those animals that test positive can be evaluated by some 'gold standard' method.

Example: Say we have conducted a survey with a test whose sensitivity is 90% (0.9) and specificity is 95% (0.95) and we find a reactor rate (apparent prevalence) of 15% (0.15). By using the formula, we can calculate the true prevalence to be approximately 11.8% (0.118).

Group (aggregate) level test interpretation

The previous discussion described the testing of individual animals. However, in epidemiological investigations, the study unit can often comprise a group of animals such as a batch of shrimp postlarvae or a cage of fish. For example, it is common practice in Asia to test batches of postlarvae for viral diseases such as monodon baculovirus and white spot syndrome virus. This is done by examining a small number of animals from the batch and then deciding the infection status of the entire batch based on the findings from the sample.

What is not commonly realised is that testing for disease at the group or aggregate level incorporates a number of factors additional to those relevant to testing at the individual animal level. Thus, techniques such as PCR which may be highly sensitive and specific at the individual animal level can still result in misclassification of a high proportion of groups where only a small number of animals in each group are tested.

At the individual animal level, diagnostic test performance is determined by its sensitivity and specificity. Additional factors which come into play when a group of animals is to be classified are the number of animals tested, the prevalence of disease in the group, the level of statistical confidence required and the number of individual animal positive results (1, 2, 3 etc.) used to class the group as positive. Just as we do for individuals, we want high sensitivity and high specificity in our group level interpretation.

The only way to be 100% confident that no animals comprising a particular group are infected with a particular agent is to test every animal in the group with a diagnostic test which has perfect sensitivity and specificity. However, if only a low proportion of individual animals in the group are infected and only a small number are tested there can be quite a high chance that infected groups will be misclassified as uninfected. Table 4.6 shows the number of infected animals which can be in a group of 100,000 despite a sample testing negative using a test with perfect sensitivity and specificity at the individual animal level.

Table 4.6 Number of diseased or infected animals which could remain in a group of 100,000 after a small number are tested and found to be negative using a test which has perfect sensitivity and specificity at the individual animal level.

No. of animals in sample tested from group of 100,000 and found negative	No. of infected animals which could be in the group despite the sample testing negative (two levels of confidence shown)	
	95%	99%
100	2,950	4,499
500	596	915
1,000	298	458
10,000	29	44

The situation is further complicated where the test procedure being used has imperfect sensitivity and specificity which is the case for many of those in use for aquatic animals.

When testing a group of animals for the presence of disease, there are a number of important points to keep in mind:

1. Individual and group level test characteristics (sensitivity and specificity) are not equivalent.
2. The number of animals to be tested in the group (sample size) is relatively independent of group size except for small groups of approximately 200 animals and fewer.
3. The number of animals to be tested in the group depends much more on individual animal specificity than it does on sensitivity.
4. The number of animals to be tested in the group is linearly and inversely related to the expected prevalence of infected animals in the group.
5. As the required level of statistical confidence increases, so the required sample size increases. The usual level is 95%. If this is increased to 99%, there is an approximate increase of 50% in the required sample size. For a reduction from 95% to 90% confidence, there is a decrease in sample size by 25%.
6. As the sample size increases, group level sensitivity increases.
7. As the number of animals used to classify the group as positive is increased, there is a corresponding increase in specificity.
8. As group level sensitivity increases, group level specificity decreases.
9. When specificity = 100% at the individual animal level, all uninfected groups are correctly classified i.e. group level specificity also equals 100%.

Exercises

Exercise 4.1

Suppose there is a disease in a population which affects one individual in every 1,000. Say an individual is chosen at random from the population and tested for disease 'X' using a procedure which has a sensitivity of 100% and specificity of 95%. Assume no other information (e.g. clinical signs) about the individual is available. If the test result is positive, what is the likelihood (probability) that the individual has the disease in question? Alternatively, if the test result is negative, what is the likelihood that the individual does not have the disease in question?

Exercise 4.2

Say a new test for QX disease in oysters has been developed and evaluated with the following results:

	D+	D–
T+	123	2
T–	8	540

a) What are the estimates for sensitivity and specificity of this test?

b) Which estimate is more precise, sensitivity or specificity?

Exercise 4.3

Using the test mentioned in Exercise 4.2, a survey is undertaken with 23 test positive oysters found in 863 sampled. What is the likely true prevalence of QX disease in this population? If there were no test positives, how confident would you be that the population was in fact free of QX disease?

Exercise 4.4

We are interested in introducing 500 fish from area A with a low prevalence (0.1%) of a sub-clinical disease X into area B which is free of disease.

a) What is the probability of introducing disease to area B if 500 fish are randomly selected from area A?

b) If we have a test of sensitivity = 90% and specificity = 100% and introduce only fish which test negative, what is the probability that disease will be introduced because of false negative tests?

c) How could the risk be further reduced and/or eliminated?

Note:

1. You would need to examine a random sample of approximately 3325 fish from area A to be 95% confident of detecting at least one infected fish with a test of sensitivity 90% and specificity 100%.
2. This problem demonstrates the difficulty of preventing the spread of disease based on health certification.

5 Disease surveillance

Introduction

This section gives an overview of aspects of disease surveillance relevant to reducing the risk of international spread of aquatic animal diseases as well as control of important epidemic diseases. These issues have always been the focus of international aquatic animal health authorities. However, they have become even more important since the formation of the World Trade Organisation and subsequent implementation of the various multilateral agreements on trade. Consequent increased international trade in aquatic animal commodities has resulted in increasing scrutiny of the risk of international spread of disease. As a result there has been a growing interest in developing better systems for investigating and reporting of animal diseases. Reliable evidence for freedom from particular diseases is also becoming an issue of major interest. For this reason, the emphasis in this paper is on those aspects disease surveillance that provide reliable information both to support trade and meet international reporting requirements.

Disease surveillance should be an integral and key component of all government aquatic animal health services. This is important for early warning of diseases, planning and monitoring of disease control programs; provision of sound aquatic animal health advice to farmers; certification of exports; international reporting and verification of freedom from diseases. It is particularly vital for animal disease emergency preparedness.

In the OIE *International Aquatic Animal Health Code*, disease is defined as *clinical or non-clinical infection with one or more of the aetiological agents of the diseases listed in this Code* while disease surveillance is defined as *a systematic series of investigations of a given population of aquatic animals to detect the occurrence of disease for control purposes, and which may involve testing samples of a population*. The Code does not include a definition of monitoring. However, monitoring can be defined as *a systematic series of investigations of a given population of aquatic animals to detect changes in the prevalence and geographical distribution of disease, and which may involve testing samples of a population*. Thus, surveillance is concerned with detection of new or exotic diseases, while monitoring is concerned with understanding changes in endemic disease levels and distribution. The term *surveillance program* is often used in a wider sense to incorporate both surveillance and monitoring activities.

Figure 5.1 provides a conceptual summary of the relationships among the broad components of a surveillance program. This figure incorporates the OIE *Code* concepts of providing an effective surveillance infrastructure as well as including a description of host population and environmental characteristics. The concepts of passive and active surveillance, aquatic animal health information systems and national and international disease reporting are discussed below.

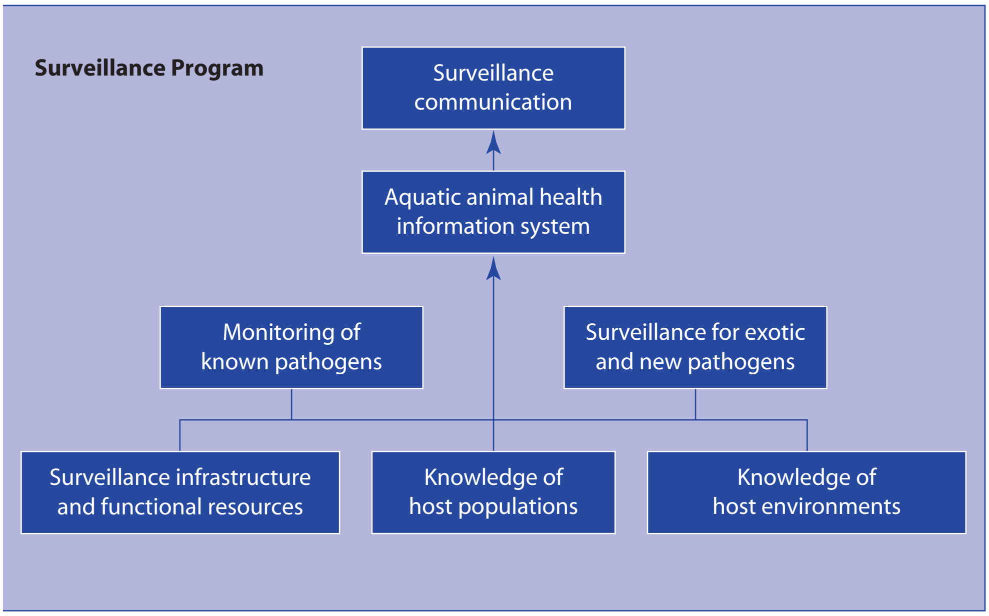

Surveillance Program
Surveillance communication
Aquatic animal health information system
Monitoring of known pathogens
Surveillance for exotic and new pathogens
Surveillance infrastructure and functional resources
Knowledge of host populations
Knowledge of host environments

Figure 5.1 Relationships among different components of a surveillance program incorporating OIE *Code* concepts.

Purpose and objectives of surveillance

The primary purpose of aquatic animal disease surveillance is to provide cost-effective information for assessing and managing risks associated with trade in aquatic animals and products, animal production efficiency and public health.

This statement of purpose is consistent with the OIE *Code* and international perceptions of what surveillance is meant to achieve.

The disease focus of a surveillance program should be based on the OIE listed diseases, any national list of notifiable diseases and other diseases of special concern to the particular country. The recommended statements to precisely articulate objectives which define the boundaries of surveillance are:

1. *Rapidly detect new and exotic infectious diseases in aquatic animals.*
2. *Provide evidence of freedom from diseases relevant to domestic and international movement of aquatic animals and products.*
3. *Describe the distribution and occurrence of diseases relevant to disease control and domestic and international movement of aquatic animals and products.*
4. *Assess progress in control or eradication of selected diseases and pathogens.*

As written, the above objectives are unambiguous and clearly set boundaries on what surveillance is meant to achieve, whether the activity be undertaking a survey to describe the distribution and prevalence of an important disease, collecting information to ensure that disease zones are maintained etc.

Types of surveillance

A variety of names have been used to describe different types of surveillance programs. There are many ways that different surveillance activities can be described. Terms such as *passive surveillance* and *active surveillance* are commonly used, but it is not always clear what they mean. A brief explanation is given below, since a comprehensive surveillance program can be seen as comprising both active and passive surveillance components.

Passive surveillance

Passive surveillance is the secondary use of routinely collected data that was generated for some other purpose. This involves the routine gathering of information on disease incidents such as requests for assistance from farmers, reports from field officers, submission of diagnostic specimens to laboratories and results of laboratory investigations.

Passive surveillance can be used to create a national disease reporting system based on the day-to-day disease investigation activities of field officers and laboratory network. Such a reporting system should include feedback loops. A theoretical example is shown in Figure 5.2. At the first stage of investigation, sufficient data is collected to assist the farmer with his/her problem. It is necessary to report only a small portion of this data to the next administrative level. However, an audit trail should be maintained of all records generated at each stage of reporting. Thus, although the national system may contain only a very brief summary of each investigation, the full information can be accessed if required.

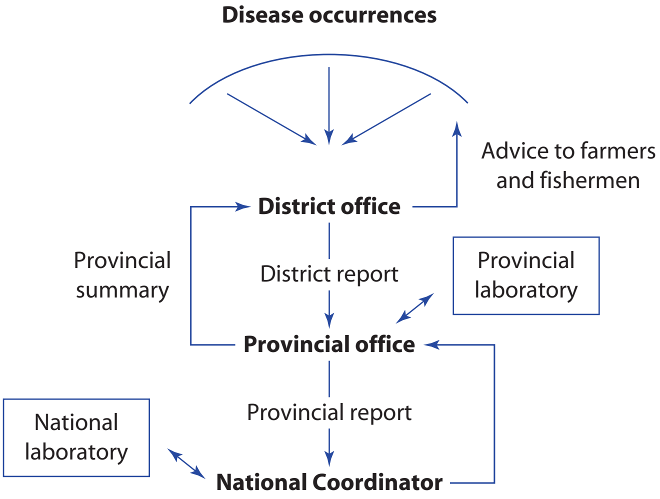


Figure 5.2 Example of information flows in a national disease reporting system

It is important that passive surveillance systems are strengthened and that the disease information that they yield be effectively captured and analysed. However, it should be recognised that complete reliance on passive surveillance usually leads to very significant under-reporting of diseases. It is essential that passive surveillance be supplemented by a strong system of active disease surveillance, particularly for emergency diseases.

Active surveillance

In contrast to passive surveillance, active surveillance involves the active collection of accurate and representative field data on the health of the livestock population. Active disease surveillance includes deliberate and comprehensive 'searching' for evidence of disease in animal populations and, in some instances, verification that specified populations are free of specific diseases. Active disease surveillance programs may be of a 'catch all' nature to detect any significant disease occurrences, or may be purposeful to target specific high-threat diseases or to monitor the progress of individual disease control or eradication campaigns. In order to maximise the value of active surveillance it should be based on survey techniques which provide representative samples of the population of interest. Bias is avoided by the use of probability sampling techniques, and appropriate analysis provides valid measures of the precision of estimates.

The movement of infected animals very frequently spreads epidemic diseases. Emphasis in active disease surveillance for such diseases must be given to situations where animals are on the move, especially where they are then brought together from a number of different sources. This includes markets, trading routes, border areas and floods.

Surveillance in populations of wild aquatic animals should also not be overlooked. Wild animal populations may provide a reservoir of infection for some diseases, but may also act as a sensitive indicator of diseases that are not very clinically apparent in adjacent farmed populations.

Requirements for a surveillance program

Diagnosis can be at different levels and in the Regional Guidelines, three levels are proposed with Level 1 essentially being the diagnosis made by the investigating field officer. However, at each level, a *case definition* is required for each disease to be reported. Case definitions may be quite general or very specific, depending on the perceived value of the resultant information.

Each investigation which is undertaken as part of the surveillance program will result in a diagnosis at some level of certainty with respect to the specified problems/diseases/pathogens of interest. In some instances, the investigation may not even result in a diagnosis, but merely describe the incident (e.g. in terms of morbidity, mortality, duration of the problem, clinical signs, appearance of gross lesions). The level of diagnostic certainty will be largely determined by the investigating officer's ability to recognise the characteristics of specific diseases as well as whether or not the report is followed up with a more detailed investigation by people with greater expertise. In most instances, the highest level of diagnostic certainty will be achieved when the investigation includes positive results with an internationally approved laboratory examination. Thus, it will be necessary to include an assessment on the diagnostic certainty with each record of a disease investigation.

Investigations of suspected disease occurrences which eventually result in meaningful surveillance require:

- appropriately trained and motivated personnel;
- standardised field and laboratory methods supported by quality control;
- access to manuals and training opportunities.

Thus, the basis of all good surveillance programs are observant and skilled people with appropriate resources who understand what is normal, are alert to changes and can describe the abnormalities they see. The design and structure of a surveillance program depend on its purpose. However, all surveillance programs have some basic common features, including:

- a clearly stated and valued purpose;
- a list of problems/diseases/pathogens of interest;
- the capability and capacity to undertake investigations to the required level of diagnostic certainty;
- specifications for the information to be collected;
- a system to collect, record and collate data as well as report findings.

Aquatic animal health information systems

To provide access to surveillance findings, some form of information repository or warehouse is required from which various communications can be produced in a variety of formats. This is usually given the name *Aquatic Animal Health Information System* (AAHIS) in the case of a repository for information on the health of aquatic animals.

An aquatic animal health information system is a system for the collection, storage, analysis and reporting of information related to the health of aquatic animals. As such, virtually every organised society that keeps aquatic animals has some form of aquatic animal health information system. This may range from the system used in a single village in a developing country (in which information is gathered by owners, passed by word of mouth, stored in the memory, analysed mentally, and further reported by word of mouth) to a national system such as that used in developed countries (involving a network of government officers, laboratory diagnostic resources, complex sampling strategies, high powered computerised data management and analysis systems, and extensive procedures for distributing and acting upon the information gathered).

The word system implies a collection of many different components working together for a particular purpose. All too often, the expression information system gets mixed up with concepts of information technology, and is understood to refer to a computer system. Computers certainly play a role in most modern aquatic animal health information systems, but they are merely one component, a tool for handling the information. Instead, system here refers more to a set of operational and administrative procedures for the collection of data from a range of different sources, the processing of that data to produce useful information, and the application of that information to protect the health of aquatic animals and improve the well-being of their owners.

International disease reporting

A comprehensive surveillance program with data and reports collected in a national aquatic animal health information system can provide the basis for international disease reporting. Most countries report disease occurrence in some way. There are various international levels of formal reporting, the most important of which is through the Office International des Epizooties (OIE). However, there may be a number of other levels of reporting in a region. Examples are briefly described here.

Office International des Epizooties (OIE)

OIE has obligatory disease reporting requirements for member countries. This should be factored into the national disease surveillance system. A staff member in the national office of the relevant authority should be allocated the responsibility of preparing draft international disease reports, for OIE and elsewhere, for the approval of the Responsible Officer. The head of the national epidemiological unit would generally be the most appropriate person to carry out this function.

In brief, countries should notify OIE within 24 hours of any of the following events:

- for *diseases notifiable to the OIE*, the first occurrence or re-occurrence of a *disease*, if the country or zone of a country was previously considered to be free of that particular disease;
- for *diseases notifiable to the OIE*, important new findings which are of epidemiological significance to other countries;
- for *diseases notifiable to the OIE*, a provisional diagnosis of a *disease* if this represents important new information of epidemiological significance to other countries;
- for *diseases* not notifiable to the OIE, if there are new findings of exceptional significance to other countries.

Thereafter, monthly reports are sent to OIE to provide further information on the evolution of the disease incident until the disease has been eradicated or the situation has stabilised.

Quarterly and annual reports are sent on the absence or presence and evolution of *diseases notifiable to the OIE* and findings of epidemiological importance to other countries with respect to *diseases* that are not listed.

In addition, there are requirements to report on changes to the status of *infected zones*.

Regional organisations

Regional organisations may be established whose mandate may include the fostering of international cooperation on aquatic animal health issues and facilitation of safe international trade in aquatic animals and products. These organisations may also have requirements of their member countries for reporting and sharing of information on diseases.

The NACA/OIE/FAO Quarterly Disease Reporting System is an example of such cooperation in the Asian region. The NACA/OIE/FAO list includes all diseases listed by OIE, the notifiable ones as well as the other significant diseases. This list, however, more specifically reflects the Asian situation. Additional diseases are listed which occur in parts of the Asia–Pacific Region and thus are of concern because they may spread further within the region.

Special arrangements with neighbouring countries and trading partners

Many epidemic aquatic animal diseases do not respect borders and can spread very rapidly from country to country. Neighbouring countries should therefore cooperate closely in the control of these diseases. Unless this is done the disease control efforts of individual countries will be continually frustrated. Part of this cooperation should be the rapid sharing of information on new disease occurrences and the spread of existing epidemic diseases to new areas, particularly near shared borders. Arrangements should not only be made for this information to flow between the respective Responsible Officers, but also at a local level between contiguous District, Provincial or Regional offices along borders.

Likewise, arrangements should be made for the rapid flow of disease information between the Responsible Officers of major trading partner countries for aquatic animals and their products.

An example in Asia is the group of countries influenced by the Mekong River.

Conclusion

Aquaculture production is expanding throughout the world during a period of rapid change in international trading arrangements. Acquiring, analysing and reporting information on the health of aquatic animals will become increasingly important to aid decision makers in developing sound policy not only for disease control but also for quarantine and health certification to permit the safe movement of aquatic animal commodities both within countries and internationally.

Appendix B

Glossary of epidemiological terms

(Courtesy of Dr Ian Gardner, University of California, Davis)

Accuracy – the degree to which a measurement, or an estimate based on measurements, represents the true value of the attribute that is being measured.

Agent – a factor, such as a microorganism or chemical substance, whose presence or excessive presence is necessary for the occurrence of a disease.

Analytical study – a hypothesis testing method of investigating the association between a given disease, health state or other outcome variable, and possible causative factors.

Benefit–cost ratio – the ratio of the net present values (usually monetary values) of measurable benefits to costs. Used to determine the economic feasibility or probability of success of a time-bounded program.

Bias – any effect at any stage of an investigation tending to produce results that depart systematically from the true values (i.e. a systematic error).

Bias (response bias) – a systematic error due to differences in characteristics between those who volunteer to participate in a study and those who do not.

Bias (selection bias) – error due to systematic differences in characteristics between those animals or farms that are selected for study and those that are not.

Categorical data – qualitative data that can be allocated to specific groups. May be nominal (i.e. named) or ordinal (i.e. ordered) or dichotomous (e.g. presence/absence).

Chi-square test – a method of testing to determine whether two or more series of proportions or frequencies are significantly different from one another or whether a single series of proportions differs significantly from an expected distribution. Pearson's Chi-square is used for unmatched data and McNemar's Chi-square for matched data. See definition of association for further explanation.

Clustering – a closely grouped series of events or cases of a disease in relation to time or place or both. The term is normally used to describe aggregation of relatively uncommon events or diseases.

Confidence limits – an interval whose end points can be calculated from observational data and which has a specified probability of containing the parameter of interest.

Confounding – a situation in which the effects of two factors are not separated. The distortion of the apparent effect of an exposure or risk factor brought about by association with other factors that can influence the outcome.

Confounding factor – a confounding factor or variable is one that is distributed non-randomly with respect to the independent (exposure) variable and is associated with the dependent (outcome) variable being studied. The association with the dependent variable is usually established from results of previous studies.

Contingency table – a tabular cross-classification of data such that subcategories of one characteristic are indicated horizontally (in rows) and subcategories of another characteristic are indicated vertically (in columns), and the number of units in each cell is indicated. The simplest contingency table is the fourfold or 2 x 2 table, but a contingency table may include several dimensions of classification.

Continuous data – quantitative data with a potentially infinite number of possible values along a continuum.

Cost–benefit analysis – method of identifying, in monetary terms, the losses and gains incurred by society as a whole from the effects of a disease.

Cross-sectional study – (syn: prevalence study) – a study, carried out on a representative sample of a population, that examines the relationship between a disease or other health-related characteristic and other variables of interest as they exist in a defined population at one particular time.

Crude rate – a rate that applies to a total population irrespective of the attributes of that population (cf. specific rate).

Data – facts of any kind.

Database – a systemised collection of information, commonly on electronic media, about a specific subject such as animal disease.

Denominator – the population at risk in the calculation of a rate or ratio. See also Numerator

Dependent variable – (syn: outcome / response variable) a variable or factor, the value of which depends on or is hypothesised to depend on the effect of other (causal) variable(s) in the study.

Endemic disease – the constant presence of a disease or infectious agent within a given geographic area or population group. It also implies a prevalence that is usual in the area or in the population.

Epidemic – the occurrence in a population or region of cases of disease clearly in excess of normal expectancy – this is frequently taken as more than two standard deviations greater than the mean occurrence.

Epidemic curve – a histogram in which the X-axis represents the time of occurrence of disease cases and the Y-axis represents the frequency of disease cases. It is a useful tool to determine the epidemiology of disease occurrence in an outbreak investigation.

Epidemic, propagating – an outbreak or series of outbreaks resulting from animal-to-animal spread.

Epidemiology – the study of the distribution and determinants of health-related states and events in populations. It is a term now in common usage for studies in animal populations, although epizootiology is still occasionally used.

Epidemiology, descriptive – study of the occurrence of disease or other health-related characteristics in populations. Implies general observation rather than analysis.

Error, sampling – after testing a sample from a large population, the mean or any other statistic calculated from the sample will have a different value from the true value if the whole population was measured. The difference between the value for the whole population and its estimate calculated from the sample is called the sampling error.

Error, systematic – an error due to factors other than chance, such as faulty measuring instruments.

False negative – when the result of an individual test is negative but the disease or condition is present.

False positive – when the result of an individual test is positive but the disease or condition is not present.

Frequency – a count, or number of occurrences, of an event in a specified population and time period.

Frequency distribution – any arrangement of numerical data obtained by measuring a parameter in a population.

Histogram – frequency distribution plotted in the form of rectangles whose bases are equal to the class width and whose areas are proportional to the absolute or relative frequencies.

Hypothesis – a proposition that can be tested by facts that are known or can be obtained. The assertion that an association between two or more variables, or a difference between two or more groups, exists in the larger population of interest.

Incidence – the number of new cases of disease or other condition that occur in a specified population during a given period. Mathematically, two types of incidence rate can be distinguished. These are incidence density rate and cumulative incidence.

Incubation period – the interval of time between invasion by an infectious agent or contact with a chemical and the appearance of symptoms of the disease or condition in question.

Independent variable – the characteristic being observed or measured that is hypothesised to influence an event. An independent variable is not influenced by the event or manifestation but may cause it or contribute to its variation.

Index case – the first diagnosed case of an outbreak in a farm or other defined group.

Infectivity – the ability of an agent to enter, survive and multiply in the host. Epidemiologically, it is measured as the percentage of the individuals exposed to an agent who become infected.

Inference – the process of passing from observations to generalisations.

Latent infection – persistence of an infectious agent within the host without symptoms of disease.

Linear regression – statistical method used to study the relationship between independent and dependent variables when the dependent variable consists of continuous data.

Longitudinal study – a study, conducted over a defined period of time, which may be either retrospective or prospective.

Mean, arithmetic – a measure of central tendency computed by adding all the individual values together and dividing by the number in the group.

Median – the middle value of a set of observations arranged in order of magnitude.

Mode – the most frequently occurring value in a set of observations. A given set of observations can have more than one mode.

Monitoring – the performance and analysis of routine measurements aimed at the early detection of changes in the prevalence or incidence of disease, health, or alteration in a production parameter.

Multistage sampling – a term applied to the selection of a sample in two or more stages (e.g. selecting a sample of farms and then a sample of animals within those farms).

Nominal data – a type of data in which there are limited categories but no order, such as breed and eye colour.

Normal – within the usual range of variation in a given population or population group; or frequently occurring in a given population or group.

Normal distribution – a continuous symmetrical frequency distribution where both tails extend to infinity, and the arithmetic mean, mode and median are identical. Graphically, it is a bell-shaped curve and its steepness or shape is completely determined by the mean and variance.

Null hypothesis – the hypothesis that two variables have no association at all, or two or more population distributions do not differ from each other.

Numerator – the upper portion of a fraction used to calculate a rate or ratio.

Observational study – an epidemiological study where nature is allowed to take its course while changes or differences in one characteristic are studied in relation to changes or differences in other(s) without intervention of the investigator (e.g. descriptive, cross-sectional case-control, cohort).

Occurrence – a statement indicating the presence of disease without signifying the frequency. This definition describes the use of the word in international animal disease reports.

Ordinal data – a type of data in which there are limited categories with an inherent ranking from lowest to highest (such as severity of disease).

Outbreak – the occurrence of disease in a farm or any other identifiable group of animals. For practical purposes, the term is synonymous with epidemic.

Outliers – observations differing so widely from the rest of the data as to lead one to suspect that a gross error in recording may have been made, or suggesting that these values came from a different population.

Pandemic – an epidemic occurring over a very wide area, involving many countries and usually affecting a large proportion of the population.

Parameter – a summary descriptive characteristic of a population (cf statistic – which is a sample-based measure).

Pathogenicity – the ability of an organism to produce disease. Epidemiologically, it is measured as the percentage of infected individuals who develop clinical disease.

Power – probability of finding a difference between two or more groups, given that a difference exists. Power = 1 – Beta = 1 – Probability of a type II error.

Precision – the quality of being sharply defined or stated. Refers to the ability of a test or measuring device to give consistent results when applied repeatedly. Sometimes also called repeatability.

Predictive value – in screening or diagnostic tests, the predictive value of a positive test is the proportion of test-positive animals that have the disease. The predictive value of a negative test is the probability that an animal with a negative test does not have the disease. The predictive value of a test is determined by the sensitivity and specificity of the test, and by the prevalence of the condition at the time the test is used.

Prevalence – the proportion of cases of a disease or other condition present in a population without any distinction between old and new cases. When used without qualification, the term usually refers to the number of cases as a proportion of the population at risk at a specified point in time (point prevalence).

$$\text{Prevalence} = \frac{\text{Number of cases at specific point in time}}{\text{Population at risk at same point in time}}$$

Prevalence study – see Cross-sectional study

Primary case – the individual that introduces disease into a farm, pond or other group under study. Not necessarily the first diagnosed case in that group. See Index case.

Proportion – a fraction where the numerator is a subset of the denominator.

Prospective study – a study that collects data as events occur.

Qualitative data – data denoting specific qualities, such as breed, gender, or colour. See *Nominal data*.

Random – governed by chance.

Randomisation – allocation of individuals to groups by chance. Within the limits of chance variation, randomisation should make control and experimental groups similar at the start of an investigation and ensure that personal judgement and prejudices of the investigator do not influence allocation. Note that random allocation follows a predetermined plan often devised with the aid of a table of random numbers or an electronic random number generator.

Random sample – a sample of a population assembled so that each member of the population has an equal and non-zero opportunity to be selected.

Random sampling – procedure for selecting individuals from a population so that each has an equal chance of being selected in the sample.

Rate – an expression of the change in one quantity per unit time. It is a ratio whose essential characteristic is that time is an element of the denominator and in which there is a distinct relationship between numerator and denominator. See also Ratio and Proportion.

Ratio – the expression of the relationship between a numerator and denominator where the two are separate and distinct quantities (i.e the numerator is not included in the denominator).

Relative risk – the ratio of the disease incidence in individuals exposed to a hypothesised factor, to the incidence in individuals not exposed; a measure of association commonly used in cohort studies.

	Diseased	Not diseased
Exposed	a	b
Unexposed	c	d

The relative risk is [a/(a+b)] ¸ [c/(c+d)]

Repeatability – the ability of a test to give consistent results in repeated tests. See Precision.

Response rate – the number of completed or returned survey instruments (questionnaires. interview etc.) divided by the total number of individuals selected for study.

Retrospective study – a study that collects and utilises historical data. A case-control study is retrospective because it looks back from the point of known effects to determine causative factors.

Robust – a statistical test is robust if the inferences hold true even when assumptions inherent in the test are violated.

Sampling – the process of selecting a number of representative subjects from all the subjects in a particular group. Conclusions based on sample results may be attributed only to the population sampled. See also Random sample.

Screening – implies subjecting a population or sample of a population to a diagnostic test or procedure, with the objective of detecting disease. Tests used for this purpose are usually cheap, easily performed and sensitive, but often not very specific.

Sensitivity – is the proportion of truly diseased animals, in the screened population, that are identified as diseased by the test. It is a measure of the probability that a diseased individual will be correctly identified by the test.

Sentinel farms – farms that are reasonably representative of the population as a whole and which are tested at regular intervals for infectious disease to determine whether and to what extent the diseases are occurring in the population.

Seroepidemiology – epidemiological studies based on an examination of sera taken from the population or a sample of the population.

Significance, level of – also known as alpha (α) or type I error rate. The probability of saying a difference exists when none does.

Spatial distribution – the relationship of disease events to location of individual animals or clusters of animals.

Specificity – the proportion of truly non-diseased animals correctly identified by a test. Like sensitivity, specificity is a conditional probability.

Specific rate – expresses the frequency of a characteristic per unit size of a specific population.

Sporadic – a disease occurring irregularly and generally infrequently and without any apparent underlying pattern.

Standard deviation – a measure of dispersion or variation. Equal to the positive square root of the variance. The mean indicates where the values for a group are centred. The standard deviation is a measure of how widely values are dispersed around the mean in the population.

Standard error – measure of the variability of a sample statistic that specifically relates an observed mean to the true mean of the population.

Statistic – a summary value calculated from a sample of observations, usually to estimate a population parameter.

Statistical significance – statistical methods allow an estimate to be made of the probability of the observed degree of association between independent and dependent variables being exceeded under a null hypothesis. From this estimate the statistical 'significance' of a result can be stated. Usually, the level of statistical significance is stated by the 'P value' or probability value. See also Significance, level of.

Statistics – the science and art of dealing with variation in data through collection, classification, and appropriate analysis.

Stratified sample – involves dividing the population into distinct subgroups according to some important characteristic (e.g. pond size) and selecting a random sample out of each subgroup.

Surveillance – a system or measurement technique to gain knowledge about a population by collection, analysis, and interpretation of data, with a view to the early detection of cases of disease or changes in the health status of the population. The goal of surveillance is directed action in the treatment or prevention of the condition.

Survey – an investigation in which information is systematically collected.

Systematic sample – the procedure of selecting according to some simple systematic rule, such as every fifth fish in the tank as they are transferred to another tank.

Temporal distribution – the relationship of disease events to time.

Trend – a long-time movement in an ordered series (e.g. a time series). An essential feature is that the movement, whilst possibly irregular in the short term, shows movement consistently in the same direction over a long term.

Type I error – an error which occurs when using data from a sample that demonstrates a statistically significant association when no such association is present in the population. Equals the level of significance, or Alpha.

Type II error – an error that occurs from failure to demonstrate a statistically significant association when one exists in a population. Equals beta. The power of a study equals 1 minus Beta.

Validity – the extent to which a study or test measures what it sets out to measure.

Variable – see Dependent variable, Independent variable.

Variance – the variance of a set of observations is the sum of squares of the deviation of each observation from the arithmetic mean of the observations, divided by one less than the number of observations.

Virulence – the degree of pathogenicity, indicating the potential severity of the disease produced by an agent in a given host. Epidemiologically, it is measured as the percentage of individuals with disease who become seriously ill or die. Sometimes, the case-fatality rate is considered an indicator for the virulence of disease.

Appendix C

Sample data collection forms and questionnaires

Grouped population sampling frame

No.	Group identifier	Total units	Cumulative total	Selected units
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				

Random number table

9537	7654	2531	7467	2873	5885	5154	6419	9346	9458	2281	4520	1241	6730	4263
3014	7669	2948	7241	0139	3841	1369	1123	8300	7790	3632	9154	4698	3874	2423
2682	4082	3359	0932	6215	9668	4282	7428	2833	7014	0217	2737	6768	4218	3007
5531	1283	5400	7610	3466	2697	0649	2159	4803	7655	3325	7537	5885	1465	4746
4534	4703	1566	8974	8989	3953	5752	4976	1253	1041	2678	0067	1001	1802	8224
4202	9222	0395	0882	0406	5696	4204	7995	0571	0744	6751	8284	7202	2610	2531
4783	0798	7713	5203	3246	9008	1017	6802	5738	9416	0092	3831	4662	7819	5152
1515	3328	4102	2777	3867	8974	0632	1175	6051	8063	2795	5037	2319	6941	0285
7824	5298	1243	0754	4284	9480	4027	6284	1251	7275	9796	9015	0199	7321	3200
3894	3231	2288	0103	7834	2159	6589	7655	4435	2457	0141	3600	6792	1631	0840
4495	1477	3933	1570	7080	6521	1885	5664	2691	7577	8866	2425	0383	5134	1282
2495	0365	0326	0856	7851	0801	9001	7861	6828	4483	6681	8913	5735	9767	7244
6941	7266	1482	6315	5838	5539	3608	9895	4136	7294	5075	7471	0057	4551	1275
7136	7584	1352	4940	4637	4448	5390	8329	0559	3921	7029	2652	4622	4366	2786
6602	5200	3213	4913	6662	9579	7025	1113	1206	9229	5973	9585	0994	1648	9597
3346	4427	2525	5519	0821	0334	2335	4005	0598	6894	8161	1447	3213	7990	9132
5327	7977	9909	7696	3362	8331	3798	3732	6549	9457	6097	2249	9890	5228	6924
2541	7991	9425	0987	0809	2695	2051	1145	4111	8633	3193	5735	2601	8008	2604
9611	9655	9767	5203	6374	2752	2562	0175	8457	0393	2300	3658	9471	2385	6007
5322	9436	8575	7562	3770	7711	7100	0856	8138	1847	3270	9227	5393	7474	8566
7959	2467	2482	8581	4816	5323	0199	7210	2602	9477	7211	4004	2738	9695	7642
7906	6113	8081	2517	9752	4073	3221	3255	0388	0730	7586	9013	9009	1631	3952
1374	9257	1451	0624	1662	5929	1230	2935	6900	3504	0815	3387	5632	0377	4424
1676	9319	6404	8020	8916	9174	0284	2252	3169	0590	1531	6276	1788	3408	6972
6970	1559	4110	7432	2041	3362	5336	4365	9501	8548	0159	0352	4491	4694	4804
5850	7679	9254	5612	3905	0924	1378	0962	0437	3103	2957	7646	5019	2527	1399
4712	3274	0387	0697	4663	2449	3002	5661	9899	5543	7188	1043	6954	0520	5805
3291	6142	4611	1300	5324	5192	0015	7741	7972	7192	6577	7169	8827	3935	9888
7277	9996	9284	0611	6375	6807	9284	6975	3175	1465	4700	8996	3251	8478	7923
9425	0618	5866	1284	0362	8875	5458	2846	6681	5532	6480	8909	7075	4222	1831
3045	3952	3590	9404	9828	7222	5711	3926	7353	6153	0426	5545	9608	9806	7823
4299	8225	3096	8302	4524	8587	6188	5714	9020	6674	6780	0167	8418	4586	2754
2207	4564	2702	5504	4287	5653	0294	7690	3897	4751	9238	0857	4756	8867	0935
7750	3178	2451	8603	6500	4976	1476	2884	8548	2806	0380	5326	3127	4905	4731
6009	4643	3594	8319	9547	4857	5677	5734	1317	5770	3484	5591	2051	3796	4675
7711	8280	3680	9546	6147	2663	1095	6521	2602	3125	5871	0333	5523	5951	7422
9115	2208	9888	3651	2995	3651	5409	3153	1912	4784	1442	3188	7233	5272	2297
1634	2060	5774	7820	5607	5813	3150	3583	8092	2846	2552	7785	2049	9719	9730
5092	7923	9073	9726	9775	7783	8331	4648	1630	3745	3901	2776	1808	3408	7362
1041	1523	0736	8295	3543	9323	0040	5601	0440	7831	3570	2664	4956	7887	2088
5022	8169	7826	3863	6097	6440	1104	7124	3058	5921	8873	2708	2044	2776	8838
9198	0531	5469	3493	2502	5640	2531	9095	5617	4837	7192	8672	1628	8392	9365
9246	3728	5474	9748	5657	4377	8841	2910	1538	6470	4421	4721	3605	5547	6820
1925	3806	1808	3684	9405	7201	1973	6606	5327	7402	6204	5216	9511	0145	4407
2225	4105	5575	5354	9190	9667	3896	3610	4398	8622	9613	8722	7660	8141	8922
1507	6559	4651	7610	9162	4502	0623	8353	5306	7346	5421	4992	6490	0868	7323
0525	7467	5629	1470	7150	7088	2736	4571	3323	5504	3615	8199	0720	6842	1583
3757	9743	8240	3837	1403	9785	0110	4526	6744	1897	7339	2223	2982	0299	4867
3934	6211	4903	0863	5501	7117	0980	9984	9837	7574	2885	6252	6631	9876	7689
8185	0935	0549	2719	0349	6359	8011	8187	0842	6450	5905	1492	0645	8788	4341
9698	8154	0394	8064	4653	0565	6530	8610	3923	5696	6513	2257	8723	5929	5121

Retrospective disease outbreak questionnaire

(using epizootic ulcerative syndrome (EUS) as an example)

Farm / Village Name		Date
District Name		ID No.

Question 1	
Has there ever been an outbreak of EUS in this village?	
Yes G	**No G**

If **Yes**	If **No**

Question 2		Question 2	
When was the last outbreak of EUS in the village? (Month and year that the first fish showed signs)		What is the earliest date since which you are sure there has been no EUS in the village?	
Date		Date	

Question 3		Question 3	
At the time of the last outbreak, how ponds were there in the village?		At that time, how ponds were there in the village?	
Ponds		Ponds	

Disease ranking and seasonal patterns

Disease or problem	Description	Species	Rank	Usual month											
				1	2	3	4	5	6	7	8	9	10	11	12

Appendix D

CD contents

- Survey Toolbox software
- Other epidemiological software
- Electronic version of the text
- Mapping software and data
- Aquatic animal health resources