

**Risk-based  
Disease Surveillance**  
**A Manual for Veterinarians**

**Angus Cameron**

Risk-based disease surveillance: a manual for veterinarians

Cameron, A.R.

© 2009

Published by:

The Food and Agriculture Organisation of the United Nations (FAO)

Viale delle Terme di Caracalla

00153 Rome, Italy

ISBN XX-X-XXXXXXX-X

### **Disclaimer**

This manual aims to give general advice about the use and analysis of risk-based surveillance systems for animal health management. Examples of diseases, surveillance and disease control strategies used in this book are intended for illustrative purposes only and are not to be interpreted as specific recommendations for disease control activities.

Neither the author nor publisher accept responsibility for any loss and/or damage, however caused (including through negligence), that you may directly or indirectly suffer in connection with your use of this manual, nor do they accept any responsibility for any such loss arising out of your use of or reliance on information contained in or referred to in this manual.

# Contents

---

<b>Preface</b>	<b>vii</b>
Introduction	vii
Purpose of this manual	viii
How to use this manual	viii
History and acknowledgements	ix
<b>Chapter 1 – Introduction to Animal Disease Surveillance</b>	<b>1</b>
Introduction	1
Terminology	2
Characteristics of a surveillance system	2
Origin of surveillance information	3
Disease focus	3
Purpose of the surveillance, and nature of the disease	4
Population Coverage	5
Representativeness	5
Type of data collected	7
Quality	10
Cost and practicality	12
Surveillance options	12
Passive disease reporting system	12
Abattoir	14
Sentinel herds	15
Surveys	16
Syndromic and indirect surveillance	16
Negative reporting (zero reporting)	18
Participatory disease surveillance	18
<b>Chapter 2 – Probability Theory</b>	<b>20</b>
Random variables	21
Notation and symbols	22
The rules of probability	22
Range	22
AND	22
OR	24
Sum of all possible outcomes	25
NOT	25
Conditional probabilities	26
General rules	27
General AND rule	28
General OR rule	28
Probability distributions	28
Bayes' Theorem	29
<b>Chapter 3 – Diagnostic Tests</b>	<b>31</b>
Sensitivity and Specificity	31
Combination of tests	33
<b>Chapter 4 – Concepts of Freedom from Disease</b>	<b>34</b>
Concepts and philosophy	34
Examples of sampling	36
Example 1: Disease free or high prevalence?	36
Example 2: Disease free or low prevalence?	37
Example 3: Imperfect sensitivity and specificity	38
Conclusion	38

Probabilities, confidence and freedom	39
Specificity of surveillance	39
Design prevalence	40
How to decide on an appropriate design prevalence	41
Integer design prevalence values	44
Design prevalence for early warning systems	44
Relative and absolute freedom	45
<b>Chapter 5 – Representative Surveys to Demonstrate Freedom from Disease</b>	<b>47</b>
Survey design	48
Calculation of sensitivity	48
Simple example	48
Imperfect sensitivity	50
Small populations	51
Imperfect specificity	52
Calculation of sample size with imperfect specificity	52
Two stage survey design	54
Clustering of infection	54
First stage calculations	55
Second stage calculations	55
Optimising the survey design	56
<b>Chapter 6 – Risk-based Surveillance</b>	<b>57</b>
Factors influencing sensitivity	59
Population variation	59
Risk-based surveillance	60
<b>Chapter 7 – Analysis of Complex Surveillance Systems</b>	<b>61</b>
Traditional approaches	61
Structured surveys	61
Expert panels	62
Ideal system	63
Overview – an analogy	63
Methodological requirements	64
Quantifying the sensitivity of complex surveillance	65
Combination of evidence from multiple surveillance components	65
Calculation of the probability of freedom from infection	65
Incorporating historical data	65
<b>Chapter 8 – Introduction to Scenario-Tree Modelling</b>	<b>66</b>
A simple example	66
Purpose of the scenario tree	68
Terminology	69
Branch probabilities	69
Node types	70
Infection node	70
Detection node	70
Category node	71
Building a scenario tree	73
Tree-building rules	76
Node order	77
<b>Chapter 9 – Incorporating Risk into a Scenario Tree</b>	<b>78</b>
Quantifying targeting in risk-based surveillance	78
Describing differences in risk	79

Describing targeting	81
Implementing risk in a scenario tree	82
What you need to know	82
Calculation of adjusted risk	82
The constraints	83
The solution	84
<b>Chapter 10 Calculating sensitivity with a scenario tree</b>	<b>87</b>
Calculation of unit sensitivity	87
Building the scenario tree	87
Organising the model parameters	88
Drawing the tree and adding parameters	88
Calculating the tree	89
Calculating the component unit sensitivity (CSeU)	90
Comparison with representative sampling	91
Calculation of component sensitivity (CSe)	91
What next?	91
<b>Chapter 11 – Probability Estimates in a Scenario Tree</b>	<b>93</b>
Summary of required values	93
Sources of estimates	95
Sensitivity	95
Proportions	99
Relative risk	101
Expert opinion	102
Gathering expert opinion	103
Combining expert opinion	104
Rinderpest example	106
<b>Chapter 12 – Incorporating Uncertainty</b>	<b>109</b>
Capturing uncertainty and variability in a model	109
Stochastic modelling	111
Describing distributions	112
Software for stochastic modelling	114
Palisade @Risk	114
PopTools	114
Freedom	114
Example exercises	115
Exercise 1: Combination of expert opinion	115
Exercise 2: Analysis of a simple scenario tree	119
PopTools reference	123
Installation	123
Random variable functions	123
Other useful functions	124
Selected Menus and Dialogs	125
<b>Chapter 13 – Clustering</b>	<b>128</b>
Clustering of disease and populations	128
Lack of independence between animals	128
Step-wise calculation of sensitivity	129
Herd level sensitivity calculation	130
Spreadsheet layout example	130
Herd-level sensitivity formulae	134
<b>Chapter 14 – Combining Multiple Surveillance Components</b>	<b>136</b>
Simple example	137

Overlapping surveillance components	138
Accounting for the overlap	138
Spreadsheet example	139
<b>Chapter 15 – Probability of Freedom</b>	<b>143</b>
Sensitivity versus freedom	143
Calculation of the probability of freedom from infection	144
Selecting a prior	145
<b>Chapter 16 – Incorporating Historical Surveillance Data</b>	<b>147</b>
Value of historical data	147
Risk of introduction	148
Calculation of posterior probability of freedom	149
Time period of analysis	149
Spreadsheet implementation	150
Examples	152
<b>Chapter 17 – Freedom Software</b>	<b>154</b>
Overview	154
Getting started	155
Log in and privacy	155
Building a scenario tree	156
Information required	156
Setting up the analysis	157
Building the node list	158
Specifying probabilities and distributions	161
Editing probability parameters	163
Editing the node list	163
Building the scenario tree	165
Displaying the tree structure	166
Refining conditional probabilities	166
Exporting the tree structure	169
Running the model	171
Interpreting the output	172
Modifying an existing scenario tree	173
Opening or cloning existing trees	174
Uploading data	175
Preparing the data for upload	175
Submitting the data	176
Identifying the columns	176
Classifying categories	177
Simulation parameters	177
Outputs for multiple time periods and clustered data	178
Combining multiple components	180
<b>Appendices</b>	<b>182</b>
Glossary	182
Abbreviations and symbols	186
<b>Index</b>	<b>187</b>

# Preface

“Science may be described as the art of systematic over-simplification.”

**Karl Popper (1902 – 1994)**

“We must plan for freedom, and not only for security, if for no other reason than that only freedom can make security secure”

**Karl Popper (1902 – 1994)**

## Introduction

---

From a disease point of view, the world is becoming a more dangerous place. Increasing global population and improvements in the standard of living mean that there is a rapidly increasing demand for animal protein. To meet this demand, animal production has intensified. The international movement of animals and animal products has been made cheaper and faster through improved transport infrastructure. Increasing human and livestock population has placed pressure on wildlife habitats, resulting in closer contact between wildlife, domestic animal populations and humans.

This complex mix of factors means that ‘traditional’ livestock diseases have the opportunity to spread and multiply much more quickly, and that ‘new’ diseases, arising from wildlife populations or genetic changes in existing pathogens, have a much greater chance to impact on animal and human populations.

Managing these disease threats poses enormous challenges and requires inputs from many disciplines. Good quality information is one essential

requirement – what diseases exist, where are they, what impact are they having, which populations are at risk, how can we prevent, control or eradicate these diseases. Animal disease surveillance plays a central role in providing this information.

Risk-based surveillance is not a particular technique, but describes a general approach to undertaking disease surveillance. The principle is simple and self-evident: the most efficient way to find disease is to look at those populations that are most likely to be affected. This is in contrast to the more traditional statistically-based approach of taking representative samples from a population.

While the idea of risk-based surveillance is simple, the implications are complex. The approach can be much more cost effective for some purposes, but if misused, can lead to serious errors. The analysis of data collected through risk-based surveillance has required the development of new analytical techniques.

## **Purpose of this manual**

---

This manual aims to present a comprehensive overview of the issues relating to risk-based surveillance. It is targeted at veterinarians interested in surveillance and the analysis of surveillance data. While a number of the concepts are necessarily complex and technical (particularly in relation to statistical data analysis and modelling), the manual assumes no prior knowledge of these areas and aims to introduce them in an easy-to-understand manner.

An attempt has been made to keep the language relatively simple, so that this manual may be accessible to those whose first language is not English. It also aims to be relevant to the animal health situation in developing countries, as well as more developed countries.

## **How to use this manual**

---

This manual contains a number of different types of material

- General background to disease surveillance
- Specific background to risk-based surveillance
- A description of techniques for the analysis of risk based surveillance
- Examples of the implementation of the analytical techniques using spreadsheet software
- A guide to using dedicated web-based software for the analysis of risk-based surveillance data, with practical examples.

Concepts are built up in a progressive manner. The later chapters make frequent reference to concepts introduced and explained in the earlier chapters.

The manual may be used in a number of ways: as a self learning tool, a reference book, or as a training course resource.

### **Self learning tool**

Interested and motivated individuals may use this book to teach themselves about the design and analysis of risk-based surveillance. In this case it is recommended that readers start at the beginning and work their way through each chapter. Any of the introductory material that the reader is already familiar with can be skipped.



### Reference book

Those that are already familiar with some aspects of risk-based surveillance and associated analytical methodologies may like to use the book as a reference, to dip into for specific information as required. The extensive examples may be helpful in this regard.

### Training course resources

The manual may also be used as a resource for structured training courses. The course presenter may use the manual when preparing the course syllabus, and course participants may be given copies to enable them to read more detailed descriptions of material touched on during the course.

## History and acknowledgements

---

While the text of this manual has been written by one person, many individuals and organisations have contributed to the development of the ideas and methodologies.

Professor Mo Salman deserves special mention as the grandfather of this approach. His thoughtful input into discussions on the topic got the ball rolling and he was instrumental in gathering a group of epidemiologists for a meeting on risk-based surveillance after the ISVEE meeting in Colorado in August 2000, to plan the way forward. The participants of this group were responsible for laying out the groundwork for the approach presented here.

However, further development would have been slow if it were not for the outstanding support of the Danish International EpiLab, who invited myself and Dr Tony Martin to use the extraordinary Danish animal health datasets to research and develop analytical methodologies for risk-based surveillance during 2002 and 2003. Tony Martin deserves credit as one of the key developers of the analytical methodology. Dr Matthias Greiner, the then head of the EpiLab and the rest of the team in Denmark contributed significantly to the work. The methodology was first unveiled at a training course held after the 2003 ISVEE meeting in November in Chile. Valuable and constructive feedback was received. Twenty training courses were to follow in the ensuing six years, in Europe, North America, Australia and Africa.

Further development was undertaken from 2004 to 2009, supported by the Australian Biosecurity Cooperative Research Centre for Emerging Infectious Diseases (ABCRC). The research team was composed of Tony Martin and myself along with strong contributions from Jenny Hutchison, Evan Sergeant and Nigel Perkins, all of AusVet Animal Health Services. The ABCRC supported the development of the web-based software, numerous case studies, and an initial users manual, as well as organising a number of training courses. The participants of these and other training courses have played an essential role in challenging and expanding the concepts presented.

The latest phase of development has been generously supported by the Food and Agriculture Organisation of the United Nations (FAO). The development of this manual was associated with a study on rinderpest in the Somali ecosystem. A training course and workshop held in Kenya in 2009, involving participants from Kenya, Ethiopia and Somalia, analysed a range of surveillance for rinderpest to estimate the probability of freedom from the disease in the Somali ecosystem, and, by implication, de facto global eradication. A wide range of veterinary and paraveterinary staff have been involved in rinderpest surveillance in the region over many years, often working in difficult and sometimes life-threatening

situations, and supported by a range of donors. The incredible work of all these people and organisations is acknowledged. The study demonstrated how the methodology could be applied in developing countries, and underlined the extremely high quality of surveillance that could be achieved by veterinary services with severely limited resources.

Finally, FAO deserves special thanks for their support in the development of this manual. It is hoped that, with FAO's assistance, this manual and the associated web-based software will provide veterinarians working in disease surveillance with the tools they need to implement and effectively analyse risk-based surveillance.

Angus Cameron, June 2009

# Chapter 1 – Introduction to Animal Disease Surveillance

To know that you do not know is the best.  
To pretend to know when you do not know is a  
disease.

Lao-tzu (604 BC - 531 BC)

## Introduction

---

The aim of this book is to assist those working in animal health to design and analyse appropriate disease surveillance systems, particularly for early warning, detection of disease or demonstration of freedom from disease. In order to design an effective surveillance system, two things are required:

- an understanding of available surveillance options, and
- an ability to compare and evaluate the different options so you can decide on the best combination

This chapter discusses the characteristics of animal disease surveillance systems that allow us to compare and evaluate their use for a variety of purposes, and introduces a range of different possible approaches to surveillance. Some of the material in this chapter and chapter 3 is based on information contained in another book by the same author<sup>1</sup>.

---

<sup>1</sup> Cameron, A.R. (2009) *Surveillance* in the 'Animal Health Management Essentials' series. Under publication by the OIE Regional Coordination Unit, Bangkok.

## Terminology

---

A range of different terms are used in this book with specific meanings. The appendix (page 186) lists the main abbreviations used, but the key terms and meanings are shown below.

Term	Meaning
Surveillance system	For a particular disease, this refers to the range of different activities that are able to produce data about the status of that disease in the population.
Component (SSC or surveillance system component)	A surveillance system may have one or many components. A component is a single activity that generates surveillance data. A component may be thought of as a single source of surveillance data
Disease	This book primarily deals with infectious diseases although may apply in places to other types of disease. Although formally, 'disease' refers to the clinical manifestations of an infection or other physiological abnormality, the term is often used more widely. In the context of 'freedom from disease', it is often used synonymously with 'infection'. For disease control purposes, it is the presence of the pathogen (infection) rather than clinical signs (disease) that is normally most important.
Infection	Formally, this means that a pathogenic agent has entered and is multiplying in an animal. Less formally, this can be generalised to mean that an animal has the characteristic of interest. For instance, when considering antibody tests, an animal may have antibodies that indicate previous exposure to an agent.
Country	Surveillance applies to a defined geographical region. For simplicity, this book has used the example of surveillance at the country level, but the techniques apply equally to a range of different levels. 'Country' can therefore be used interchangeably with terms such as zone, region, province, state, enterprise or compartment.

## Characteristics of a surveillance system

---

In order to design, evaluate and compare surveillance options, it is important to understand the different characteristics of a surveillance system. This section discusses a number of characteristics that can be used to describe surveillance.

### Active surveillance

Active surveillance describes an activity that is designed and initiated by the prime users of the data. The main purpose of the activity is disease surveillance and examples include:

- a serological survey to assess the prevalence of antibodies to brucellosis
- a farmer questionnaire to identify the level of mortality in their animals.

This is active because the users of the surveillance data (e.g. the veterinary authorities) are actively involved in generating the data.

One of the significant advantages of active surveillance is that the activity is designed by the users of the information. Therefore, it is possible to ensure that both the nature of the data collected and the quality of the data are adequate to meet the users' surveillance requirements.

### Passive surveillance

Passive surveillance describes a surveillance activity that uses data that has already been collected for some other purpose. In these cases, veterinary services do not initiate the data collection.

Examples of passive surveillance include:

- A farmer disease reporting system. In the process of seeking advice, diagnosis, or treatment for sick animals, the farmer 'reports' disease. The reason for the farmer making the report is not to help the surveillance system, but to seek veterinary assistance for the problem with their animals. The use of the data for surveillance is secondary.
- Abattoir meat inspection. The reason for the meat inspection is to ensure the quality of the meat sold to consumers. If the data were not used for surveillance, meat inspection would still be required.

The main advantage of passive surveillance systems is that they are cheap. As a result, they often can have much greater coverage of the animal population. However, the data may not fully meet the veterinary services' needs and there is little control over data quality. Data quality may be improved if farmers and veterinarians are provided with education or rewards to improve reporting for specific conditions.

### Targeted surveillance

Targeted surveillance describes surveillance that is focused on a specific disease or pathogen.

For example, a serological survey for brucellosis may use the rose bengal test (RBT). Blood from each sampled animal is tested and the result of the test is classified as RBT positive or RBT negative. An animal that has tuberculosis or foot-and-mouth disease, but not brucellosis, would be simply classified as RBT negative, as these other diseases are not of interest in the surveillance activity.

The term *targeted surveillance* can be used in two different senses. In this case, it is referring to surveillance targeted at a specific disease. Later in this book we will use the term in a different sense – surveillance targeted at a high risk portion of

the population. To differentiate between the two, it is preferable to refer to the second situation as *risk-based surveillance* rather than targeted surveillance.

### **General surveillance**

General surveillance is not focused on a particular disease, but can be used to detect any disease or pathogen. For example, the farmer disease reporting system is a general surveillance system, as any disease may be reported. However, not all diseases will be reported with the same reliability. Farmers are more likely to report diseases that show clear signs and have a significant impact (for example, many animals are affected or the disease results in death, such as haemorrhagic septicaemia), than they are to report diseases that display few signs or do not result in an immediate economic impact (for example, intestinal parasites).

Some laboratory tests, such as histopathology, allow detection of many different diseases, rather than just a single disease.

An important feature of general surveillance is that it is not only able to detect known diseases of interest, but may also be able to detect new, emerging, exotic or unknown endemic diseases. In other words, it is not necessary to be looking for a specific disease in order to find it.

The distinction between general and targeted surveillance depends on the disease detection system used. Targeted surveillance is based on the use of tests that are able to provide a yes/no answer for a specific disease. Examples include:

- polymerase chain reaction (PCR)
- enzyme-linked immunosorbent assay (ELISA)
- agar gel immunodiffusion (AGID).

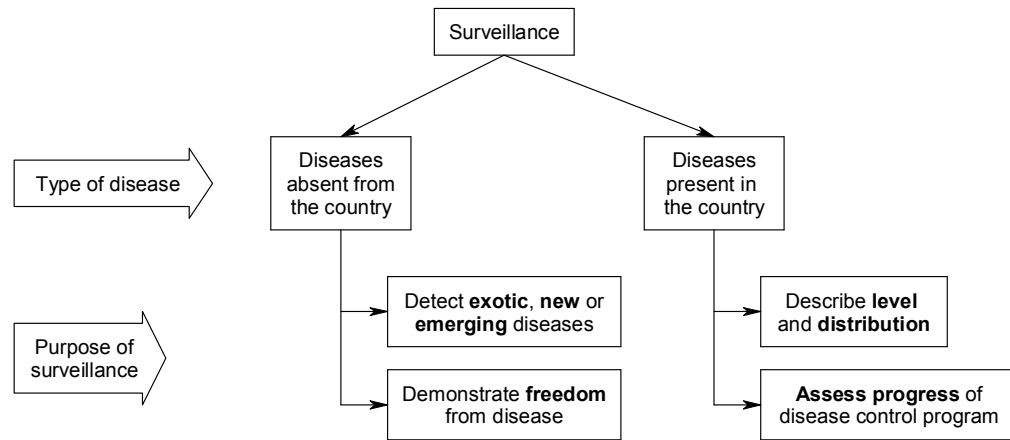
General surveillance is based on tests that are able to identify multiple diseases (in some cases, all diseases). These tests include

- clinical examination
- disease investigation
- *post mortem* investigation
- meat inspection
- histopathology
- various syndromic surveillance activities.

### **Purpose of the surveillance, and nature of the disease**

Although there may be some special cases, the purposes of most animal health surveillance can be divided into the following four categories:

- surveillance for diseases that are present
  - describing the level or distribution of disease (or a pathogen or risk factors for disease)
  - assessing the progress of disease control or eradication programs
- surveillance for diseases that are absent
  - detecting the incursion of new, emerging or exotic diseases (or pathogens or their risk factors)
  - demonstrating freedom from disease or pathogens.



## Population Coverage

---

Population coverage refers to the proportion of the population that is actually examined as part of the surveillance system. Two approaches can be used

### Sampling

When sampling, only some animals in the population are examined. For example, a sentinel herd system involves a relatively small number of herds; a small number of animals from these herds are tested or examined at regular intervals — animals that are not in the sentinel herds are not examined at all, so the herds are being used as a sample of the population.

A structured survey may involve randomly selecting a number of villages or farms, and then randomly selecting some animals from these villages or farms to test

### Comprehensive coverage (census)

In a census, all the animals in the population are examined. For example, if the population of interest is 'all farmed pigs in the country', a passive disease reporting system covers the entire population, as every single pig in the country is examined (even if only superficially) at more or less regular intervals; if a particular animal becomes diseased, there is a chance that that disease event will be captured by our surveillance system — the probability depends on many factors (for example, severity of the disease, relationships between farmers and veterinarians, whether a report is made), but each pig, if it becomes sick, has a chance of being recorded in the system.

## Representativeness

---

The representativeness of a surveillance system describes how well the information that is gathered describes the population of interest:

If the level of a characteristic of the animals in our surveillance system (for example, the percentage of animals with protective antibody titres) is approximately the same as the level in the source population, the system is representative of the population with respect to that characteristic. If there is a difference between the animals in the surveillance system and the animals in the source population — for instance, 90% of animals with protective antibodies,

compared with 60% in the source population — the surveillance system is not representative, but is biased.

Bias is the difference between the real value in the population and the value we measure through our surveillance. In many cases, bias due to a non-representative surveillance system can cause big problems.

### **Example**

Consider abattoir surveillance to assess the level of contagious bovine pleuropneumonia (CBPP) in a population. The system uses a sample of the population; the population of interest is ‘all farmed cattle’, but the surveillance examines only cattle that go through the abattoir. Animals infected with CBPP are likely to get sick or die on the farm, so an animal with the disease is much less likely to be sent to the abattoir than a healthy animal. As a result, the proportion of cattle with CBPP in the abattoir is likely to be much lower than the proportion on farms. Therefore, this type of surveillance is biased.

As the surveillance system is likely to detect a lower proportion of animals than the proportion that is truly infected, it is negatively biased.

The meat inspection system in some developing countries may be less developed than in other countries. This means that it is more common for sick animals to enter an abattoir than in countries where well-developed controls are in place. As a result, abattoir surveillance is more useful for detecting clinical disease in some less developed countries than in more developed countries.

Making animal health management policy decisions on the basis of biased information can be very dangerous. If this information were being used to monitor the progress of a control program, or to prioritise spending on future disease control programs, the wrong decisions might be made, which might have a negative effect on the health of the population. For example, the level of disease may seem to be low, so no action will be taken, but the true level of disease is high.

Surveillance systems that provide comprehensive coverage of a population are more likely to be representative. However, if the probability that some animals are recorded in the surveillance system is different from the probability that other animals are recorded, these systems can also be nonrepresentative.

### **Example**

A surveillance system for brucellosis may be based on farmer reporting of abortions or arthritis. If a control program is in place that involves modifying the management systems around calving to limit the spread of the disease, farms that adopt good management practices are less likely to have the disease. Farms that do not use good management practices may have higher levels of disease. However, farmers with poor management may also be less likely to report disease than farmers with good management.

That is, the disease rates may be higher, but the reporting rates may be lower from farms with poor management than from farms with good management. The outcome is that, even with a system in which every affected animal has a chance of being reported, differences in disease and reporting probabilities can result in a bias — in this case, making the total level of disease appear lower than it actually is.



Surveillance systems that aim to provide an accurate assessment of the level of disease typically produce results in terms of a proportion, such as the percentage of animals with CBPP, or the percentage of animals with protective antibody titres against FMD.

If you are making decisions (for example, evaluating the progress of a disease control program) based on data expressed in the form of a proportion or percentage, it is important that the surveillance system is set up to avoid bias.

## *Type of data collected*

---

### **Diagnoses**

Diagnoses refer specifically to disease, usually clinical disease. At the level of an individual animal, a diagnosis tells us what disease an animal has. In surveillance, a diagnosis is used to classify some animals as having a particular disease and other animals as not having that disease.

In order to make a diagnosis, the animal should be examined by a veterinarian. If necessary, specimens should be submitted for laboratory testing. However, as this is not always possible, some surveillance systems are designed to collect uninterpreted data, rather than the diagnoses that would result from their interpretation.

### **Classifications**

Often we are not interested solely in clinical disease, but also in some characteristic of the animal that is related to disease, as in the following examples:

- A serological survey to demonstrate freedom from foot and mouth disease (FMD) will classify animals as seropositive or seronegative. In this case, seropositive animals are unlikely to have the disease — we are simply using the serological status as an indicator of whether the animal has been exposed to the virus (or possibly a vaccine) at some time in the past.
- Surveillance to evaluate the progress of a vaccination program for foot-and-mouth disease can be done by estimating the proportion of animals that have protective antibodies. This is based on the antibody status of the animals rather than a diagnosis of disease.

Any measurable characteristic may be used to classify animals for the purposes of surveillance.

### **Analysis of specimens**

Both the diagnosis of disease and the classification of animals according to some characteristic (for example, antibody status) are usually achieved using some type of test. Some tests are laboratory based, such as

- enzyme-linked immunosorbent assay (ELISA) to measure antibody levels
- virus isolation
- polymerase chain reaction (PCR) to detect a pathogenic agent.

Other tests can be performed in the field, such as clinical diagnosis by a veterinarian or meat inspection in an abattoir.

When a laboratory test is used, the thing that is collected for surveillance is normally not the information but a specimen from the animal (blood, milk, a

tissue sample, etc). This specimen has one or more tests applied, to produce test results — the data required.

## Signs and syndromes

In the case of disease, the most commonly collected information is the diagnosis. In order to make a diagnosis, the animal should be examined by a veterinarian, and, if required, specimens submitted for laboratory testing. As this is not always possible, some surveillance systems are designed to collect uninterpreted data, rather than the diagnosis that would result from its interpretation.

To make a diagnosis, a veterinarian will observe the **signs** shown by a sick animal (for example, lameness, coughing, increased heart rate) and interpret them to decide on the disease causing the problem.

Many signs are easily observed by people without veterinary training. Although nonveterinarians may be unable to make a definitive diagnosis, people who work with livestock are often very good at identifying clinical signs in their animals. Village animal health workers are usually not veterinarians, but have been trained to recognise disease signs. However, there may be legal restrictions on who can make a diagnosis (for example, qualified veterinarians only).

Therefore, a surveillance system may collect data on the signs of disease observed. Changes in the patterns of signs observed in a population may indicate changes in the diseases that cause those signs. For instance, even if the diagnosis is not known, a sudden increase in the number of cases of coughing indicates the potential introduction and spread of a respiratory disease. This information can be used to initiate a detailed disease investigation to determine the cause of the coughing.

To make interpretation and reporting of this type of surveillance simpler, cases are often classified into **syndromes** according to the key sign or group of signs.

A syndrome is simply a defined collection of signs. In the above example, the syndrome may be ‘respiratory disease’ and include any case of disease that shows coughing, difficulty breathing and so on. Other syndromes include:

- acute febrile illness
- diarrhoea
- skin lesions
- sudden death
- lameness.

Both reporting of signs and reporting of syndromes are referred to as **syndromic surveillance**.

Syndromic surveillance is usually designed to help with the detection of changes in disease patterns or the early detection of new diseases. When a change is detected, it must be followed up by more detailed investigations to diagnose the disease causing the change.

Surveillance may collect data on the signs or the general syndrome associated with disease. The use of syndromes in data collection and reporting is more common than the use of signs because with syndromes, there is only one data item per case (for example, respiratory disease). With signs, a single case may have many different signs (for example, coughing, difficulty breathing, standing with neck extended, increased heart rate) — this makes reporting, collation and analysis of the data more complicated.

## Negative reporting

Negative reporting is a special case of disease reporting. The data item in this type of surveillance is the fact that an animal *does not* have a specified disease.

Negative reporting data may be used in two ways:

- To rule out a disease in a laboratory-based reporting system.
  - For instance, in a country seeking to demonstrate freedom from bovine spongiform encephalopathy (BSE), laboratory results may be collected from BSE tests on neurological cases. The results may all be negative. This does not provide any information on what neurological diseases are present, but it does provide evidence that BSE is not present.
- To rule out a disease in a clinical reporting system.
  - This is common for diseases that show clearly evident clinical signs and that spread quickly, such as FMD in a naive, susceptible population.

For example, a system may be established in which veterinarians complete a report after every farm or village visit, indicating that FMD was not present at the time of the visit. No special examination is necessary because, if FMD were present, it would normally be very easy to identify just by looking at the animals. The fact that the veterinarian visited the farm and did not see any evidence of disease provides information that the disease was absent. (There is a small chance that the veterinarian was wrong, but this is the case with any type of testing or surveillance). A surveillance system that collates large numbers of negative reports from a wide area can provide objective evidence that there are unlikely to be any animals with clinical signs of FMD.

Documentation of a clinical negative reporting system can provide valuable reassurance to trading partners about continued freedom from disease in a particular zone, compartment or country.

## Indirect indicators

Some surveillance systems do not collect data on the disease or health status of animals directly, but take a more indirect approach.

For instance, information provided by drug companies, distributors and feed supply stores on the sales of particular types of veterinary drugs and/or feeds can be used for indirect surveillance. Changes in the patterns of drug sales and commercial feed sales are likely to be good indicators that there is a change in the pattern of disease. However, this does not say what the disease is — any observed changes must be followed up by a detailed investigation to assess if there is really a change in disease incidence and, if so, what the disease is.

Surveillance for indirect indicators of disease is often described as being a part of syndromic surveillance. This approach is commonly used to assist with the early detection of disease. The ideal indicators are, therefore, those that change early in the disease process, as shown in the following examples.

The most common surveillance system used to detect disease is based on reporting by farmers to veterinarians when they have a disease problem. However, before the farmer calls the vet, they may try to treat the problem themselves. If a new widespread problem affects a population, it may be possible to detect the problem through the use of drug sales and/or commercial feed sales, rather than waiting for veterinary reports, which may come some time later.

In human disease surveillance, thermometer sales and business sick-leave records can be good early indicators of disease patterns in the population.

Indirect indicator surveillance is normally active surveillance. The veterinary authorities establish a relationship with the holders of the data (for example, drug suppliers), and ask that updates on sales be provided at regular (daily or weekly) intervals for analysis.

### **Risk factors**

Most surveillance, including indirect surveillance, collects information about disease or a disease-related state. Another approach to surveillance is to measure the risk factors that may be involved in causing the disease. This type of surveillance seeks to provide alerts before an outbreak of disease, so that preventative measures can be put in place.

Examples of risk factor surveillance are:

- Vector surveillance for vector-borne diseases. The vector for bluetongue is the *Culicoides* biting midge. Insect trapping sites provide surveillance information on the presence or absence of the disease vector.
- Surveillance for risk factors for the development of algal blooms, which may produce toxins that kill farmed aquatic animals or contaminate aquatic products, making them unsafe for humans to eat.
  - Surveillance systems can be established to monitor sunlight and water temperature, to assess the risk of the development of the blooms. This is risk-factor surveillance for the development of algal blooms
  - Surveillance may directly measure the amount of algae present and whether they are toxic or not. This is risk-factor surveillance for aquaculture or food safety.

External risk factors or factors not having a direct biological effect on the occurrence of disease in animals may be considered for surveillance activity. For example, in some regions, movement of animals during religious festivities from one area to another has resulted in an increase or resurgence of FMD outbreaks and other transboundary animal diseases. Data on prices and livestock movements may be used to predict times of increased risk and the location of potential new disease outbreaks.

### **Quality**

---

The way the quality of surveillance is measured depends on the type of surveillance.

#### **Surveillance to demonstrate disease freedom or detect disease**

When surveillance is undertaken to demonstrate freedom from disease, or for early detection of disease, the conclusion is either that disease has been detected and is therefore known to be present, or disease has not been detected, and is therefore believed not to be present.

With this 'yes/no' result it is possible to make two types of mistakes.

It is possible to falsely conclude that disease is present when it is not (a false alarm). False alarms may cause concern and expense, but do not ultimately endanger the disease status of the population (because no disease is present). A good surveillance system would be expected to generate a false alarm from time to time.

The second mistake is to falsely conclude that disease is not present when it truly is (surveillance failure). Missing a genuine case of disease can be a dangerous mistake, as it may spread undetected.

A surveillance system can be thought of as a type of diagnostic test on the entire population: the population has or does not have a disease and the surveillance is used to make a decision. The ability of a surveillance system to correctly identify a diseased population is analogous to the ability of a diagnostic test to identify a diseased animal. It is measured quantitatively by the sensitivity of the surveillance system.

Sensitivity is the key measure of the quality of a surveillance system that aims to detect disease or demonstrate freedom from disease. Sensitivity is discussed further in Chapter 3. The evaluation of the quality of the surveillance system therefore depends on an estimation of the sensitivity of the surveillance system.

### **Surveillance to measure the level or distribution of disease**

The key measure of a surveillance system to measure the level of disease is prevalence (the proportion of affected animals in a population). Various other measures may be used, such as incidence, but prevalence will serve as an example for this discussion.

Assessing the quality of a measure of prevalence involves assessing the two types of error that can occur: systematic error and random error.

#### **Systematic error**

Systematic error is the error produced by some systematic problem in the surveillance system. If the same surveillance is conducted repeatedly on the same population, the error will always be present, and the result would be the same.

Systematic error is measured by bias, which is defined as the difference between the true result and the expected result of the surveillance system (the expected result is the average of all results you would get if you repeated the same surveillance many times).

#### **Example**

Abattoir surveillance might be used to assess the prevalence of clinical paratuberculosis (Johne's disease) in cattle. This disease causes chronic diarrhoea and weight loss. Therefore, affected animals are less likely to be sent to an abattoir than healthy animals.

As a result, the prevalence of clinical cases of Johne's disease in an abattoir will always be lower than the prevalence in the general population. Abattoir surveillance for Johne's disease is therefore biased.

#### **Random error**

Random error is due to the fact that the result of our surveillance can vary randomly, according to the simple chance of selecting one animal or the next animal. With small sample sizes, the random error can be large.

#### **Example**

Consider a population of 1000 animals with a true prevalence of 10%. If only three animals are chosen at random, it is quite likely that all three would be healthy animals. This means our estimate of the prevalence from our sample would be 0%. The random error is 10%.

Consider the same population, but a sample of 300 animals instead of 3 animals. It would be much less likely to select all healthy animals for the whole sample of 300. It is more likely that the sample would have 10% of 300 (30) infected animals, but due to random sampling, the actual number of infected animals in the sample may be a little higher or a little lower. Selecting one or two infected animals more or less than the expected number is quite likely. It is much less likely that we would select a number of infected animals that is very different to the expected number (eg selecting only 4 or as many as 60 infected animals by chance).

The precision of an estimate describes how much random error there is. When calculating the results, the size of the random error is described by the confidence intervals around an estimate.

### *Cost and practicality*

---

An important characteristic of surveillance systems is their cost. The precision (when measuring disease) or sensitivity (when detecting disease) of a surveillance system increases as the number of animals examined increases, but so does the cost. A good surveillance system should be cost effective.

In addition to cost, the resources to undertake surveillance must be available. Practicality should always be considered.

## **Surveillance options**

---

There are a range of ways that surveillance can be carried out. This section describes a number of different approaches.

### *Passive disease reporting system*

---

Passive disease reporting systems describe the surveillance that is achieved when a farmer identifies that they have some sick animals and contacts a veterinarian for help.

Passive disease reporting systems are the most common and probably the most important form of surveillance in any country. They are a form of passive surveillance, as the reason the farmer contacts the veterinarian is not for surveillance, but in order to get help with the sick animals

They are also classified as general surveillance as they can be used to identify a wide range of diseases.

Passive disease reporting systems have a number of key advantages:

- The coverage of the animal population is usually very good because the person responsible for identifying the disease is the farmer. Most animals in the population are seen by their owners relatively frequently — this is in contrast to surveys, where only a very small proportion of the population is examined.
- The system is relatively inexpensive — farmers need to contact the veterinarian anyway, so the main extra cost is related to collecting the information for surveillance purposes.

Passive disease reporting systems are often the means by which new diseases — either incursions of exotic diseases or emerging diseases — are first discovered, because there is high coverage of the population, and they are capable of detecting any disease (as opposed to targeted surveillance).

Therefore, passive disease reporting systems play a very important role in any national surveillance system. These systems are far from perfect, however, due to the possibility of:

- farmers not observing their animals
- farmers not recognising signs of disease
- farmers being afraid to report because of the fear of negative consequences
- farmers being unable to report if they are in remote areas
- failure of the reporting system within the veterinary services to correctly register or diagnose the disease.

Efforts to address these limitations can significantly improve early detection of disease.

There are many variations in the detailed operation of farmer disease reporting systems, but a typical system may operate as described below:

- An animal gets sick, and is noticed by the farmer. The chance that the farmer notices a sick animal depends on the signs it is showing (more spectacular signs, such as sudden death, unusual neurological signs, or large, visible lesions, are easier for a farmer to notice) and the number of animals affected (if more than one animal is affected, a sick animal will be easier to notice).

Sometimes a farmer may experience problems that are not associated with clinical signs; for example, subclinical disease, nutritional deficiencies or mastitis at a herd level may cause production losses that are noticed by the farmer, prompting a call to the veterinarian.

- The farmer contacts somebody about a sick animal or animals. The simplest case is when the farmer contacts the local government veterinary officer directly. Alternatively, the farmer may contact a private veterinarian, who then contacts a government veterinarian.

There may be a number of other steps, such as contacting neighbours, the village head or the local animal health worker for assistance. Ultimately, if the official veterinary service knows about the case, the information can be used for surveillance purposes.

- Information about the case is recorded. Normally, this is done by the local government veterinarian, but it can happen at other stages. Information may be recorded in a number of ways — most often, a standard paper form is used.

- The written disease report is passed through a reporting hierarchy. If a report is filled out by the local village animal health worker, it will be passed to the district veterinary office. The information may then be passed from the district to the provincial office, then perhaps to a regional office, before it arrives at the national office.

At each stage, the information in the disease report may be analysed, summarised, or transformed into a different format. One common approach is for reports to be collated at the district level, with a summary report indicating the number of cases of different diseases sent to the provincial office each month. The provincial office combines all district reports into a single provincial summary of the number of cases, which is then sent to the national office. The national office then collates all the provincial reports.

Once surveillance data have been collected at the national level, they are available for use. Routine use of farmer reporting data often includes annual reports of the number of cases of different diseases reported each year and reports to meet international reporting obligations.

Diagnostic laboratories are often seen as alternative sources of surveillance data. However, the process by which samples arrive at the laboratory is basically the same as for the passive disease reporting system:

- The farmer notices that an animal is sick and seeks veterinary help.
- A diagnostic specimen may be collected and sent to the laboratory.
- Data from the laboratories are summarised and sent to the provincial or national offices for reporting, either linked to field reports or independent of them.

## *Abattoir*

---

Abattoir surveillance is commonly used as a form of passive surveillance. Its primary advantages are that:

- it is inexpensive — animals are processed and inspected for other purposes, so the costs are primarily related to data capture and any laboratory tests performed
- it can cover a very large number of animals
- it allows collection of diagnostic specimens, such as blood or tissue samples, for laboratory testing
- it provides a relatively constant supply of surveillance data
- it enables data to be collected from a relatively small number of abattoirs that slaughter animals from a large number of farms or villages (thereby decreasing the data collection costs).

Active, targeted surveillance can also be carried out at abattoirs, to take advantage of some of these benefits.

Abattoirs vary significantly from country to country and area to area. Highly industrialised commercial abattoirs are sophisticated factories with large workforces and tightly controlled food hygiene and safety requirements. Village abattoirs may operate outdoors and slaughter only a very small number of animals under poor hygiene conditions.

The types of surveillance information that can be collected from an abattoir include:

- routine meat inspection findings
- targeted specimens for laboratory analysis
- enhanced inspection findings.

### **Routine meat inspection findings**

In all but the smallest abattoirs, there is some form of meat inspection. Normally, a limited number of parts of the carcass and viscera are examined.

The aims of meat inspection are to ensure that the meat is fit for human consumption, or to detect or exclude a limited number of specified conditions. For instance, specific lymph nodes may be examined to detect granulomas, in order to be sure that the animal is not affected with tuberculosis.

If the findings of routine meat inspection are recorded and captured by the surveillance system, they may provide a useful source of surveillance data about diseases that can be detected.



In many abattoirs, animals are also examined before slaughter, and this information may be used to supplement the meat inspection findings. These examinations, which are rarely detailed, aim to detect obvious injuries or lesions, and signs that may indicate that an animal is clinically ill (such as signs of depression or fever).

### **Targeted specimens for laboratory analysis**

Abattoirs offer a valuable opportunity to collect specimens that cannot be collected easily from live animals. The simplest method is the collection of blood, but tissue specimens may also be collected. Large numbers of samples can be collected very rapidly at a busy abattoir, making this task simpler and cheaper than collecting similar specimens in the field.

The ability to collect specimens depends on the nature of the abattoir and the type of specimen required.

#### **Blood**

Blood is best collected as soon as the animal is killed and while it is being bled. In a busy commercial abattoir, this is one of the most dangerous and therefore strictly controlled areas of the plant, because it is the only place inside the abattoir where there are live animals, which pose a significant risk of injury to workers.

Even if there is plenty of blood available to be collected, it is necessary to consider carefully how it can be collected without danger or disrupting normal abattoir operations. Collecting blood at smaller, less busy abattoirs may be easier.

#### **Tissues**

Tissues can often be collected during or after removal of the viscera from the carcass. The ability to take tissue samples depends on the way in which tissues are used by the abattoir. If whole organs are going to be sold (such as livers), the abattoir may be reluctant to allow samples to be taken, and may require them to be purchased.

### **Enhanced inspection findings**

Routine inspection can detect only a limited number of conditions. It may be possible to do special inspections at the abattoir for a specific disease that can be detected at *post mortem* examination. This may be done by:

- external research
- surveillance staff
- existing meat inspectors, who have been trained to do more detailed examinations to detect disease.

These more detailed examinations may be further improved by the collection of specimens by the meat inspectors for laboratory confirmation.

---

### **Sentinel herds**

A sentinel is one who stands guard to warn when something happens. Sentinel herds act as indicators for the rest of the population to warn that disease is present.

A sentinel herd usually consists of a relatively small number of animals, kept together, that are visited on a regular basis and tested. Testing usually involves

blood testing to check for antibodies to specific diseases. It may also involve clinical examination or tests for a specific disease agent.

The typical operation of a sentinel surveillance system is as follows:

- A relatively small number of sentinel herds are established in areas considered at high risk of disease incursion.
- Where possible, animals are individually identified.
- When animals are first introduced into the sentinel group, they are tested to ensure that they are susceptible to the target disease (that is, they do not already have antibodies).
- At each subsequent test, the antibody status of the animals is assessed.
- If an animal is antibody positive, this indicates that the animal has been exposed to the disease in the time between the current test and the previous (negative) test.

Sentinel herds or flocks are therefore distinguished from other systems by being a relatively small group of identified animals, placed in a fixed strategic location, and monitored over time.

## **Surveys**

---

Surveys are often seen as the best way to do surveillance, but they can be costly and logistically challenging. They are a form of active surveillance, so the veterinary services have full control over the design of the survey and the data collected.

The key advantage of surveys is that the sampling strategy can be developed to exactly meet the needs of the veterinary services and decision makers. Many other forms of surveillance involve a compromise between the data needed to support decision making and the data that are available.

Surveys may be representative or risk-based (targeted to a subpopulation with a higher risk of having the disease).

Representative surveys are the most common form. With this approach, it is possible to confidently calculate measures of the level of disease, or probabilities of disease freedom, without the fear of error due to bias.

*Survey Toolbox* (Cameron 1999), Parts I and II (Chapters 2 to 9) deals with most aspects of livestock disease surveys, and Chapter 3 concentrates on techniques to ensure a representative sample.

Risk-based sampling is used to detect disease or to demonstrate freedom from disease. Animals are chosen from high-risk groups, so that if the disease is present, there is a better chance of detecting it than if purely representative sampling was used.

## **Syndromic and indirect surveillance**

---

Various forms of syndromic surveillance have been used for many years. However, recent interest from the field of human surveillance has led to a great deal of research in the area.

A syndrome is defined as a collection of signs that indicate the presence of a disease. Syndromic surveillance is therefore concerned with the signs and groups of signs that are associated with disease. The signs may be clinical signs, such as fever, lameness and diarrhoea, or indirect signs, such as a decrease in the feed consumption at the pen level in a piggery or an increase in antibiotic feed additive

sales from a supplier. When the signs do not relate to clinical signs, this type of surveillance is known as indirect surveillance.

Syndromic surveillance involves the identification of specific signs or groups of signs and analysis of the patterns of these signs in space and time.

The purpose is not to diagnose a specific disease, but to detect abnormal patterns of signs that may be due to one of a large number of diseases. When an abnormal pattern is detected, a disease investigation follows, in order to diagnose the actual cause of the disease.

Patterns of signs and syndromes are often much less clear than direct diagnoses of disease.

### **Example**

If diarrhoea is used as an indicator of the presence of classical swine fever (CSF), a syndromic surveillance system might collect farmer reports of diarrhoea in their pigs, or sales of treatments for diarrhoea.

Diarrhoea can have many causes, so there would be a constant stream of reports coming into the surveillance system. A single case of CSF would just be one more report among the many others. However, CSF usually occurs as significant outbreaks and can spread from farm to farm. When it enters the population as a new cause of diarrhoea, the normal pattern of reports of diarrhoea may change.

In order to detect these changes, large amounts of data are required to establish the normal patterns of the sign or syndrome being analysed. These patterns describe how much there is, seasonal variations, and normal random variations (in the absence of the target disease). An understanding of the normal patterns makes it possible to spot a change in these patterns when the new disease appears.

The source of data for syndromic surveillance systems should normally be fast, simple and cheap, and allow the routine collection of large amounts of data.

### **Example**

Commercial poultry farms expect a certain amount of mortality each day and routinely record the daily mortality in their sheds. Since death is a syndrome that can be used to detect disease, the data on mortality (if collected centrally for analysis) could easily be used to detect unusual patterns of mortality in the population and trigger a rapid investigation.

The above examples illustrate the three types of data that can be collected by a syndromic surveillance system:

- individual signs (diarrhoea, fever, lameness, agitation, etc) – farmers or veterinarians record the clinical signs that they observe, without making a diagnosis on the basis of these signs; patterns and combinations of the signs are analysed to determine what is normal and to detect what is abnormal
- syndromes (respiratory, gastrointestinal, neurological, death, etc) – cases are classified according to the dominant organ system involved; these classifications can be analysed to look for unusual patterns
- indirect signs (feed consumption, drug usage, etc) – signs that are not observed directly in sick animals, but are observed indirectly.

## *Negative reporting (zero reporting)*

---

A veterinary negative reporting system is a specialised surveillance system designed to provide evidence of freedom from disease. This system is a type of passive surveillance, which aims to document information that is being generated for other purposes.

Veterinary staff routinely visit farms, villages and other places where animals are kept for a range of reasons, such as examining and providing treatment to clinical cases, vaccination and other control activities or inspections and certifications. During the course of these visits, there is normally an opportunity to chat with the livestock owners and to see the other animals.

If the veterinary services are aiming to demonstrate that a country or zone is free from a disease that normally shows clear and obvious clinical signs, each visit by veterinary staff provides evidence. Even if specific examination of animals is not undertaken, it is very unlikely that a disease like FMD showing its normal manifestations in cattle or pigs could be present without the farmer asking the veterinarian about it, or the veterinarian noticing the disease in the animals. The fact that disease is not noticed at a routine visit can therefore be seen as evidence that the disease is not present.

After each visit, the veterinarian completes a brief report, which includes the location, the date, and confirmation that the target disease was not seen or reported during the visit.

The 'test' in this case (talking to the owner, and inspecting the animals from a distance) is not very sensitive and has very low sensitivity in early cases of disease. However, it is very inexpensive.

Information from the veterinary negative reporting system can be used in response to Carl Sagan's often quoted phrase: 'absence of evidence is not evidence of absence'. In other words, to provide evidence that the disease is absent, a simple absence of reports is not adequate. The veterinary negative reporting system generates documented evidence that the disease is not present. Over time, the number and coverage of these reports can provide significant evidence that the country or zone is free from the disease in question.

## *Participatory disease surveillance*

---

Participatory disease surveillance (or participatory disease searching, PDS) is a relatively new term to describe an approach to surveillance involving the engagement of farmers.

The method arose out of earlier work on participatory epidemiology and participatory rural appraisal. The common features of all these approaches are the use of trained teams to conduct semi-structured or unstructured interviews with farmers, and the use of a variety of tools to get an overall assessment of the problems and needs of the farmers. Typical tools include:

- participatory disease or risk mapping
- brainstorming
- participatory piling
- development of calendars
- prioritisation or ranking exercises
- open discussions

The prime objective of participatory approaches is still surveillance and a key output is quantitative data on the occurrence of disease. The participatory approaches from which PDS evolved are specifically designed to give investigators a general understanding of issues and problems from the point of view of the farmers and help address these problems without any preconceptions of what the most important issues might be.

PDS may be used in two ways:

- Targeted surveillance, investigating the occurrence of a single disease (for example, highly pathogenic avian influenza (HPAI) in Indonesia, or rinderpest in Pakistan or Africa).  
This application is at odds with the participatory philosophy, as the prime concern of investigators is finding out about the disease of interest — although they may be happy to learn about disease in general (or indeed other problems) from the farmers' point of view, they are not in a position to do anything about these more general problems.
- General surveillance, in which information about all diseases of importance to farmers can be collected and prioritised. The investigators are limited by their preconception that animal disease is a key problem, and the one that they are investigating.

Because PDS is a surveillance activity, rather than a component of a rural development activity, and its main reason is to collect data, it is better to separate it from the associated methods from which it evolved, and assess its value in terms of surveillance.

PDS is active surveillance (general or targeted). Trained teams visit villages and talk to farmers and the reason they do this is to generate surveillance data. The source of the information is the farmers and the way data are collected is through discussion with the farmers.

PDS may be thought of as an alternative approach to the passive disease reporting system, which overcomes some (but not all) of the problems of low farmer reporting rates.

The participatory tools used in PDS are not something special for this activity, but simply a documented approach to collecting good information from farmers. Aspects of this approach can and should be used (to the extent appropriate) whenever veterinary staff are discussing disease issues with farmers.

# Chapter 2 – Probability Theory

“If thus all events through all eternity could be repeated, by which we would go from probability to certainty, one would find that everything in the world happens from definite causes and according to definite rules, and that we would be forced to assume amongst the most apparently fortuitous things a certain necessity.”

**Jakob Bernoulli (1654-1705)**

Chance governs many of the events related to surveillance. When an infection enters a herd, not all animals become infected, but chance decides those that are and those that are not. When animals are selected in a survey, chance determines those that are selected and those that are not.

Understanding and analysing surveillance requires an understanding of these chance processes that govern a whole range of events. Probability theory provides us with a number of rules that help us understand and predict the outcome of chance events.

The early study of probability was concerned with games of chance, legal decisions and life insurance. The Bernoulli brothers, Jakob and Johann, were brilliant mathematicians and made significant contributions in these areas. However, Johann forced his son Daniel Bernoulli (1700-1782) to study medicine, on the grounds that there was no money in mathematics. Sharing his father's and uncle's mathematical talent, Daniel applied concepts of probability to the problem of smallpox vaccination. At that time, vaccination involved inoculation of a cut on the skin with live smallpox virus. A vaccinated person had a chance (about 1 in 200) of contracting the disease and dying because of the vaccination. On the other

hand, not being vaccinated meant that one had a chance of 1 in 7 of dying of smallpox in the longer term. Using principles developed in the analysis of lotteries, and applying them to estimates of life expectancy, Bernoulli concluded that smallpox vaccination was, despite the risks, the best course of action.

## Random variables

**Random variable: an unpredictable event that follows a long run pattern.**

In mathematics, a variable is something that varies, or that can take a number of different values. For example, the presence of the sun is variable – sometimes it is present (during the day) and sometimes it is not (during the night). In this case, the variable follows a clearly predictable pattern (night follows day regularly).

A *random* variable describes an unpredictable event. For example the toss of a die, the flip of a coin, or the gender of a baby all represent random, unpredictable events. We can never tell if a single toss of a coin will result in a head or a tail, nor if a particular natural conception will result in a male or a female.

However, as indicated in the quotation from Jakob Bernoulli above, while the outcome of individual random events may not be predictable, if they are repeated many times, a pattern becomes apparent. Probability theory provides rules through which we can understand and predict the outcome of repeated random events, or determine how likely different outcomes are.



### Example

When a six-sided die is thrown, it is not possible to predict what the result will be, whether it is a or a or any other outcome. However there are six possible results (, , , , , ) and with a fair die, each result is equally possible. The probability of getting a , for instance, can therefore be calculated as:

$$\frac{\text{The number of ways in which the required outcome can be achieved}}{\text{The total number of possible outcomes}} = \frac{1}{6}$$

Even if we know the probability of getting a is  $1/6$ , we still can't tell if we will get a or not for a single roll of the die. However, if we roll the die 60 times, we can calculate the *expected* number of times that we would get a . This is:

$$\text{The number of rolls} \times \text{the probability of 3 for each roll} = 60 \times \frac{1}{6} = 10$$

This means that if we throw the die 60 times, we would expect to get a ten times. But you don't always get what you expect. What this means is that you could also get a more than ten times or less than ten times. Getting 9 times or 12 times (for example) are both possible, but the *most likely* result is 10 times.

### Conclusion

The individual result is unpredictable, but probability allows us to predict the pattern of results for a random variable when an event is repeated many times.

## Notation and symbols

---

The probability of an event X is written as:  $P(X)$

Random variables or events are often assigned a short name or letter. For instance, the random event of tossing a coin where the possible result is either a *head* or a *tail*, getting a head could be called **H**. In probability formulae, the probability of a specified event is expressed as  $P()$ . Thus, the probability of getting a head when tossing a coin would be written as  $P(H)$ .

## The rules of probability

---

If the probability of an individual event is known, the rules of probability allow us to calculate the probability of various combinations of events. These rules can be summarised as:

- Range – the possible range of probability values
- AND – the probability of one event AND another event
- OR – the probability of one event OR another event
- SUM – the probability of all possible events
- NOT – the probability of an event NOT happening
- Conditional – the probability of an event, given that another event has already happened

### Range

---

Probabilities are always between 0 and 1

Probability values are proportions. As shown in the example above, when all outcomes are equally likely, a probability is calculated as:

$$\frac{\text{The number of ways in which the required outcome can be achieved}}{\text{The total number of possible outcomes}}$$

The number on the top of this equation (the numerator) is always a part of the number on the bottom (the denominator). This proportion can be thought of as *the proportion of all possible outcomes that are the required outcomes*.

As the numerator is always less than or equal to the denominator, proportions (and therefore probabilities) are always in the range from zero to one. If, in probability calculations, the result is less than zero or larger than one, it is a clear indicator that you have made a mistake.

An event with a probability of 1 is certain to occur. An event with a probability of 0 can never occur – it is an impossibility.

Probabilities are often expressed as percentages, ranging from 0% to 100%.

### AND

---

AND rule:  
 $P(A \text{ and } B) = P(A) \times P(B)$

Consider the example of flipping a coin. There are two outcomes: heads (H) and tails (T). The probability of each is  $\frac{1}{2}$ . Let us calculate the probability of first throwing a heads and then throwing a tails.

There are two ways to calculate this. In the first, for two throws of the coin, all possible outcomes can be listed:

- H, T
- H, H
- T, T
- T, H



There are now four possible outcomes, and only one of these matches our required result (H,T). Therefore the probability of throwing a heads and then a tails is one quarter:  $P(H,T) = 1/4$ .

The second way of calculating this is to consider the individual probabilities. Remember that all probabilities are between 1 and 0. If we are calculating the probability that first one event occurs, and then another event occurs, the probability of both occurring must be smaller than either event happening on their own (it is harder to throw a heads followed by a tails than it is to just throw a single heads, or just throw a single tails). With numbers between 0 and 1, they get smaller when you multiply them together. The AND rule therefore says that to calculate the probability of one event AND then another event, you **multiply** the probability of the first by the probability of the second.

For our coin flip, this means:

$$\begin{aligned} P(H,T) &= P(H) \times P(T) \\ &= \frac{1}{2} \times \frac{1}{2} \\ &= \frac{1}{4} \end{aligned}$$

#### Example

**Question:** The prevalence of disease in a herd is 15%. A pen-side test for the disease has a sensitivity (probability of giving a positive result in a diseased animal) of 85% and a specificity (probability of giving a negative result in a non-diseased animal) of 100%. If one animal is randomly selected from the herd and tested, what is the probability of getting a positive test result?

**Answer:** In order to get a positive test result, two events must occur. First an infected animal must be selected, and then that animal must give a positive result in the test.

$$P(\text{infected}) = \text{prevalence} = 15\%$$

$$P(\text{test positive}) = \text{sensitivity} = 85\%$$

$$P(\text{infected AND test positive}) = 15\% \times 85\% = 12.75\%$$

AND rule  
assumes events  
are  
independent.

It is important to know that this rule only holds if the two events are *independent*. Independence means that the probability of one event occurring is not influenced by whether the other occurs or not. When tossing a coin, the probability of getting H on the second throw is unrelated to whether you get H or T on the first throw.

Many other probabilities are not independent: consider the probability of the weather being windy  $P(\text{wind})$ , and the probability of rain  $P(\text{rain})$ . A meteorologist may tell us that, for a particular day:

- $P(\text{wind}) = 20\%$ , and
- $P(\text{rain}) = 40\%$

We may therefore calculate

$$P(\text{wind AND rain}) = 20\% \times 40\% = 8\%$$

**WRONG**

However, experience tells us that wind and rain often go together and that they are therefore not independent. In reality the probability of wind and rain

together is likely to be higher than 8%, but the probability cannot be calculated using just the individual probabilities.

## OR



**OR rule:**  
 $P(A \text{ or } B) = P(A) + P(B)$

Ten sided dice (or decimal dice) are often used to help with random sampling. These have ten faces numbered 0 to 9 and each side therefore has a probability of being selected of  $1/10$  or 10%.

For a single throw of the die, what is the probability of either throwing a 2 OR a 9? There are 10 possible outcomes, but now there are two that meet our requirements. The probability is therefore  $2/10$  or 20%.

When we say OR we mean that there are several different ways to achieve the outcome required, so the probability is greater than the probability of each individual outcome. The OR rule states that, to calculate the probability of either one outcome or another outcome we **add** the probabilities of each of the outcomes. This can be written as:

$$P(A \text{ OR } B) = P(A) + P(B)$$

### Example

**Question:** A village contains the following numbers of animals:

Cattle: 40

Goats: 30

Sheep: 20

Pigs: 10

If one animal is chosen at random, what is the probability that it will be either a sheep or a goat?

**Answer:** There are 100 animals. The probability of choosing a goat is  $30/100$  (30%) and the probability of choosing a sheep is  $20/100$  (20%).

$$P(\text{sheep or goat}) = P(\text{sheep}) + P(\text{goat}) = 20\% + 30\% = 50\%$$

**OR rule assumes that events are mutually exclusive**

The OR rule also has an important requirement. This rule only holds if the two events are *mutually exclusive*. In the case of our example, this means that it must not be possible to select an animal and for it to be both a sheep and a goat. In some situations, events are not mutually exclusive. If we again consider our example of the weather, it is possible to have both wind and rain together. We could calculate the probabilities of having either wind or rain:

$$P(\text{wind OR rain}) = P(\text{wind}) + P(\text{rain}) = 20\% + 40\% = 60\% \quad \text{WRONG}$$

This overestimates the probability of wind or rain because it does not take into account the occurrence of wind and rain, so the answer is not correct. This can be shown diagrammatically using a Venn diagram as show in Figure 1. The overlapping area represents the chance of having both wind and rain. To correctly calculate the probability of wind or rain you should use:

$$P(\text{wind OR rain}) = P(\text{wind}) + P(\text{rain}) - P(\text{wind AND rain})$$

This removes the overlap and prevents it from being ‘double counted’ in the probability calculation.

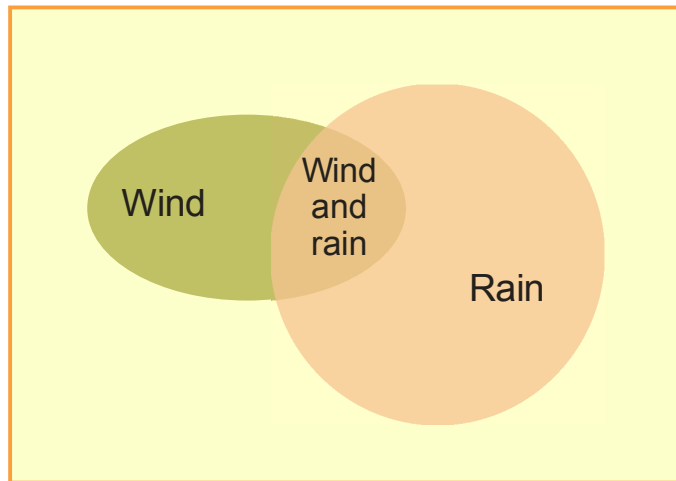


Figure 1: Venn diagram showing two non-mutually exclusive events

### Sum of all possible outcomes

**SUM rule:**  
sum of all possible outcomes equals 1

This simple rule says that the sum of all possible outcomes for a random event must add up to 1. For example, there are six possible outcomes when rolling a six-sided die. Each side has a probability of 1/6. The sum of the probabilities of the six sides is therefore 1. This can be written as:

$$\sum_{n=1}^N P(A_n) = 1$$

This is read as “the sum of the probability of all  $n$  outcomes ( $A_1, A_2, A_3 \dots$  up to  $A_n$ ) where  $n$  is from 1 to  $N$  (the total number of possible outcomes) is equal to 1”.

### NOT

**NOT rule:** P(NOT A) = 1 - P(A)

With our ten-sided die, what is the probability of not getting a 3? We can calculate this using our OR rule, as it is equivalent to getting a 0 or a 1 or a 2 or a 4 or a 5 or a 6 or a 7 or an 8 or a 9. This can be expressed as:

$$\begin{aligned} P(\text{NOT } 3) &= P(0) + P(1) + P(2) + P(4) + P(5) + P(6) + P(7) + P(8) + P(9) \\ &= \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} \\ &= \frac{9}{10} \end{aligned}$$

A simpler way of calculating this is to use the previous SUM rule. If the probabilities of all possible outcomes add up to 1, then the probabilities of all possible outcomes except for a single outcome A, must add up to 1 - P(A). So:

$$\begin{aligned} P(\text{NOT } 3) &= 1 - P(3) \\ &= 1 - \frac{1}{10} \\ &= \frac{9}{10} \end{aligned}$$

### Example

**Question:** In a surveillance system with the objective of detecting cases of disease, the probability that disease will be detected in a single animal is 10%. What is the probability of detecting disease in at least one animal, if a group of 8 animals are tested?

**Answer:** This problem needs to be addressed in several steps.

What is the probability that disease will *not* be detected in a single animal?

$$P(\text{detected}) = 10\%$$

$$\text{Therefore } P(\text{not detected}) = 1 - 10\% = 90\%$$

What is the probability that disease will not be detected in 8 animals?

This is an application of the AND rule. Restating the question, what is the probability that disease will not be found in the first animal, AND disease will not be found in the second animal AND... etc up to the eighth animal.

$$\begin{aligned} P(\text{not detected in 8 animals}) &= P(\text{not detected in animal 1}) \times \\ &P(\text{not detected in animal 2}) \times \\ &P(\text{not detected in animal 3}) \times \\ &P(\text{not detected in animal 4}) \times \\ &P(\text{not detected in animal 5}) \times \\ &P(\text{not detected in animal 6}) \times \\ &P(\text{not detected in animal 7}) \times \\ &P(\text{not detected in animal 8}) \times \\ &= P(\text{not detected})^8 \\ &= 0.9^8 = 0.43 \end{aligned}$$

What is the probability that disease *will* be detected in at least one of the 8 animals?

This is an application of the NOT rule, as it is simply one minus the probability that it wouldn't be detected in any of the eight animals. This makes the final calculation:

$$\begin{aligned} P(\text{detected in at least one of 8}) &= 1 - (1 - P(\text{detected in 1}))^8 \\ &= 1 - (1 - 0.1)^8 \\ &= 1 - 0.43 \\ &= 0.57 \end{aligned}$$

### Conditional probabilities

Consider a bowl containing balls that are selected at random. The bowl contains 5 yellow balls and 2 green balls. When a ball is selected it is not replaced in the bowl. If three balls are selected at random, what is the probability that all three will be yellow?

For the first ball selected, there are 5 yellow balls out of the total of 7 balls, so the probability of drawing a yellow is 5/7.

When the second ball is drawn, if the first ball was yellow, then there are only 4 yellow balls left, out of a total of six, so the probability is 4/6. However, if a

green ball was drawn in the first draw, there would still be five yellow balls and the probability would be 5/6.

In this example, the probability of drawing a yellow ball at the second draw depends on (or is *conditional* on) what ball was selected in the first draw. Conditional probability is expressed using the following notation:

$$P(A|B)$$

which is read as “the probability of A given B” or “the probability of A conditional on B”. In our example,

$$P(\text{ball 1 is yellow}) = 5/7$$

$$P(\text{ball 2 is yellow} | \text{ball 1 is yellow}) = 4/6$$

$$P(\text{ball 3 is yellow} | \text{ball 1 and 2 are yellow}) = 3/5$$

The probability that all three balls are yellow is therefore:

$$\begin{aligned} P(3 \text{ balls are yellow}) &= \frac{5}{7} \times \frac{4}{6} \times \frac{3}{5} \\ &= \frac{60}{210} \\ &= 0.286 \end{aligned}$$

**Sensitivity is a conditional probability**

The sensitivity of a diagnostic test is an example of a conditional probability. Sensitivity is the probability that a test will give a positive result if the animal is truly infected. This can be expressed as:

$$P(T+ | D+)$$

where T+ means that the test gives a positive result and D+ means that the animal is diseased.

Conditional probabilities refer to the situation where the probability of one event depends on whether another event has occurred. In the previous section on the AND (multiplication) rule (page 23), it was noted that the rule is only valid if the two events are independent. Independence is the opposite of conditional probabilities – this is when the probability of one event does not depend on whether another has occurred. Two events, A and B, are considered to be independent when:

$$P(A|B) = P(A)$$

or in words, the probability of A given that B has occurred is the same as the probability of A regardless of whether B has occurred or not.

## General rules

---

The AND (multiplication) and OR (addition) rules are only valid in certain cases. These rules can be extended so that they are valid in all cases.

### General AND rule

---

The probability of both A and B happening depends on whether A and B are independent. If B is conditional on A, then the AND rule can be rewritten as:

$$P(A \text{ and } B) = P(A) \times P(B|A)$$

For example, the probability that an animal is infected and that it gives a positive test result is equal to the probability that it is infected (prevalence of disease) multiplied by the probability that it will give a positive test result, given that it is infected (sensitivity of the test). This can be written as:

$$\begin{aligned} P(D+ \text{ and } T+) &= P(D+) \times P(T+ | D+) \\ &= \text{prevalence} \times \text{sensitivity} \end{aligned}$$

### General OR rule

---

The OR rule depends on the two events being mutually exclusive. If they are not, the rule can be expressed as:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

This was illustrated in Figure 1 on page 25.

## Probability distributions

---

The values for a random variable are unpredictable individually, but form some sort of pattern in the long run. We can use such patterns to make predictions about how likely various events are.

### Example

A population is infected with a prevalence of 20%. We take a random sample of 40 animals from the population. How many infected animals will be in our sample?

The expected number of infected animals is the probability of getting a infected animal multiplied by the number of animals selected, or  $20\% \times 40 = 8$  infected animals. However, as this is a random process, we will not always get exactly 8 infected animals after selecting 40. Instead, we may get a few more or a few less. However, if you repeat this experiment many times, a pattern starts to emerge.

Figure 2 shows what might actually happen in this example. In the first image ( $n=5$ ) the sampling has been repeated five times, and each time a different number of infected animals was selected (12, 9, 6 and twice 7). The expected number of 8 wasn't selected at all. The actual number of infected animals selected is a random variable, and this shows how it is not possible to predict individual outcomes, or even a small number of outcomes.

In the second image, sampling has been repeated 100 times. This time, 8 is in the middle of the values obtained, but the results are still rather irregular. The third shows the result of repeating the sampling 10,000 times, and reveals a distinct pattern.

The final image is based not on multiple repetitions but the theoretical probability of selecting each different possible outcome.

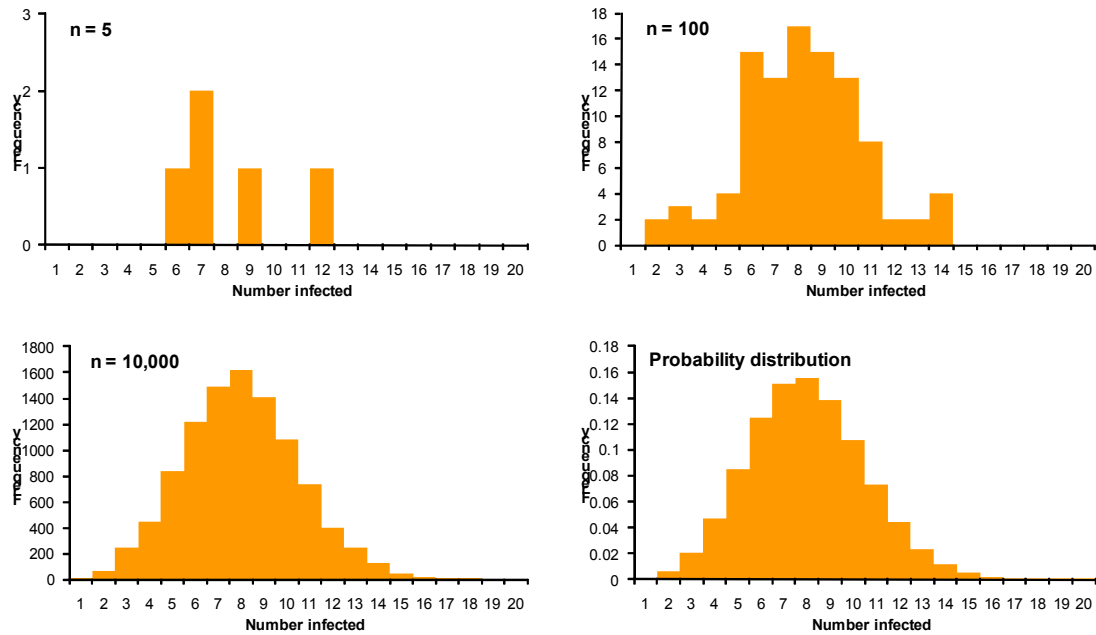


Figure 2: Number of infected animals selected based on different numbers of surveys randomly selecting 40 animals from a population with a prevalence of 20%.

## Bayes' Theorem

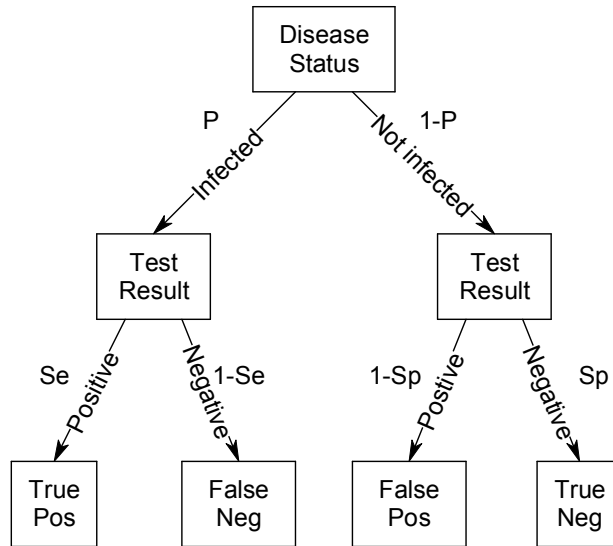
Conditional probabilities can be expressed in the form  $P(A|B)$ . Sometimes it is useful to be able to calculate the inverse probability, i.e.  $P(B|A)$ . This calculation is achieved by another probability rule, known as Bayes' Theorem.

### Example

An example of this situation arises when using tests for clinical diagnosis. As mentioned above, the sensitivity of a test is a conditional probability:  $P(T+|D+)$ . This is the probability of getting a positive test result, given that the animal is truly infected.

Knowing that diagnostic tests can sometimes make a mistake, it would be useful to be able to calculate  $P(D+|T+)$ , or the probability that the animal is truly infected, given that we have tested the animal and got a positive test result. This value is known as the positive predictive value.

There are two ways we can get a positive test result when testing an animal, as shown in diagram below.



A positive test result could be a true positive (the animal is infected, and gives the right test result) or a false positive (the animal is not infected but gives the wrong test result).

The probability that an animal is truly infected if we get a positive test result is the proportion of these positive outcomes that are true positives:

$$P(D+ | T+) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

The probabilities are indicated on the diagram above, so this can be calculated as:

$$P(D+ | T+) = \frac{P \times Se}{(P \times Se) + (1 - P) \times (1 - Sp)}$$

This simple result has surprisingly far-reaching implications. In this case, P stands for the disease prevalence in the population. However, it can be thought of as the probability that the animal is infected before any testing has been done. This is known as the *prior* probability that an animal is infected. The animal has been tested and we have a positive test result, which represents new information about the animal. Using the above equation, we can combine our prior knowledge with new information to produce a new estimate of the probability that the animal is infected (known as the *posterior*). The above formulae for the positive predictive value is an application of Bayes' theorem, which allows us to revise prior information with new information to give us an updated posterior probability.



# Chapter 3 – Diagnostic Tests

- A. If reproducibility may be a problem, conduct the test only once.
- B. If a straight line fit is required, obtain only two data points.

## Velilind's Laws of Experimentation

A test is broadly defined as any procedure that aims to divide a population into two groups: one with the characteristic of interest (disease, infection, presence of antibodies, etc), and one without.

All tests may produce results that make errors in this classification. To qualify as a test, the procedure should classify animals at least more accurately than a purely random procedure (such as tossing a coin).

The two types of **errors** that a test can make are:

- false positive — falsely identifying an animal that does not have the characteristic as having the characteristic
- false negative — falsely identifying an animal that does have the characteristic as not having it.

## Sensitivity and Specificity

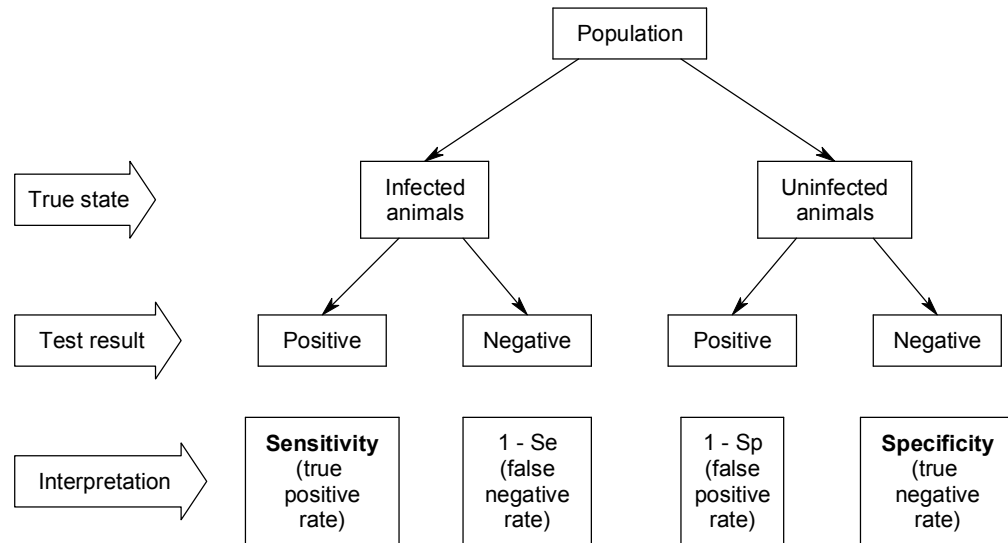
---

The **validity** of a test is the probability that it will get the classification correct. Validity is expressed in terms of **sensitivity** and **specificity**:

Sensitivity is the probability that a positive animal will be identified as positive by the test (1 – false negative rate) — this describes how well the test performs for truly positive (that is, infected) animals.

Specificity is the probability that a negative animal will be correctly identified as negative by the test ( $1 - \text{proportion of false positives}$ ) — this describes how well the test performs for truly negative (that is, healthy) animals.

These ideas are illustrated below.



Se = sensitivity; Sp = specificity

Sensitivity and specificity can be calculated using studies in which the test is applied to animals whose true status is known. The data are usually arranged in a two-by-two table as shown below.

		True status		Total
		Positive	Negative	
Test result	Positive	a	b	a + b
	Negative	c	d	c + d
Total		a + c	b + d	a + b + c + d

In this table, the sensitivity is the number of correct positive results (true positives),  $a$ , divided by the total number of truly positive animals,  $a + c$ . The specificity is the number of correct negative results (true negatives),  $d$ , divided by the total number of truly negative animals,  $b + d$ .

### Example

If a new test were applied to 100 animals, made up of 60 healthy animals and 40 infected animals, the results below might be obtained.

		True status		Total
		Positive	Negative	
Test result	Positive	36	10	46
	Negative	4	50	54
Total		40	60	100

In this example, the sensitivity of the test is  $36/40 = 90\%$  and the specificity of the test is  $50/60 = 83.3\%$ .

If the true status of animals is not known, this type of calculation cannot be used. New modelling techniques are available to estimate sensitivity and specificity when the true status of animals is not known. These techniques rely on the use of more than one test in a number of different populations. Detailed consideration of these techniques is beyond the scope of this book.

## Combination of tests

---

A country would not be considered infected with an exotic disease just because a farmer found a sick animal and reported it. The first test (examination of the animal by the farmer) is quickly followed by a series of other tests, for example:

- clinical examination by a veterinarian
- laboratory tests for antibodies
- confirmatory laboratory tests for the disease agent.

Combinations of multiple tests allow us to avoid certain mistakes. In this case, we want to be sure that we are not falsely identifying an exotic disease, so we are trying to increase the specificity and decrease the chance of a false positive. The animal would only be considered positive if all of the following occur:

- the farmer thought there was a problem
- the veterinarian also thought there was a problem
- the first (antibody) laboratory test gave a positive response
- the confirmatory (agent) laboratory test gave a positive response.

If the results of all these tests are positive, we can be very certain that the animal is truly infected.

There is always a trade-off when combining tests. In the above example:

- we increased specificity — with each extra test, the chance of making a false positive decreased
- we decreased sensitivity — the animal would be called negative if there was a negative result in any of the four tests, and, because each test has a chance of getting a false negative result, the chances of a false negative result increase with each extra test used.

In this case, in order to achieve high specificity, we need to sacrifice sensitivity. This is because with our interpretation of the results, the animal is only positive if it tests positive to all the tests.

Using a different interpretation would change the overall test characteristics. If we consider that the animal is only negative if it is negative to all tests, the result would be to increase the sensitivity, but decrease the specificity.

# Chapter 4 – Concepts of Freedom from Disease

No amount of experimentation can ever prove me right; a single experiment can prove me wrong.

Albert Einstein (1879 – 1955)

## Concepts and philosophy

---

When considering surveillance for diseases that are not known to be present, there may be two objectives:

- to demonstrate (or provide evidence) that the disease is not present, in order to support trade or stop unnecessary disease control activities, or
- to ensure that the disease would be able to be rapidly detected if it ever entered the country or region.

While it is common to talk about ‘disease’, this generally implies ‘infection’ and the objective is really to prove that the pathogen is absent from the country or region. In order to design appropriate surveillance, we must therefore first ask the question: How can you prove that infection is not present?

Let us consider some possible approaches in a stepwise fashion:

- A simple approach may be to visit a farm and to look at some animals, while asking yourself “Do any of these animals appear to be infected?” If the animals show no sign of the disease, then you may conclude that they are free from the infection.
  - This approach is quick and simple, but it does not prove that the country is free from infection, because:

**Freedom from disease implies freedom from infection**

- it is not representative (only one farm was examined), and
  - the test used (clinical observation) is not good – there could be subclinically infected animals present.
- To address these problems, a structured serological survey is often used to support claims of freedom from infection. Consider a large survey, in which a random sample of 10,000 animals is selected from all parts of the country. Blood is collected from each animal and is tested in the laboratory for antibodies that would indicate any previous exposure to the pathogen. Clearly, this is a much better than the first approach, however, if all the results from this survey are negative, does this prove that the country is free from infection?
  - If the population of susceptible species in the country is, for example 10 million, then our sample, even though very large, is still only a small part of the population. While we have tested 10,000 animals, there are still 9,990,000 animals in the population that we have not tested. It is certainly possible that one or more of those animals is infected, but we have missed them in our survey.

This example shows that examining a small number of animals cannot prove that we are free from infection. Examining a larger number gives us a much better chance of finding the infection if it is present, but it still cannot prove that we are free. The more negative animals we observe, the more evidence we have that we may be free, and our *confidence* that we are free increases. However, there is still a chance that the infection is present, so we don't have absolute proof that we are free.

How then can we get this proof? We could try testing every single animal in the entire population. If all animals were negative, would this prove that we were free from disease? The problem here is that virtually all laboratory tests can make mistakes. When the sensitivity of the test is less than 100%, it means that there is a risk that a truly infected animal may give a negative test result. Even testing every single animal can't prove that we are free from infection. However, even if there is not absolute proof, we would be very very confident that the infection was extremely unlikely to be present.

If we had a perfect test that never made mistakes, by the time we had finished testing the animals, there is still a chance that those animals that were tested first had become infected.

The simple conclusion is that it is not possible to prove that a population is free from infection. This problem has been examined extensively by philosophers concerned with science and knowledge. In this example, we have developed a theory – that the population is free from infection. Each time we observe an animal that is not infected, it lends support to this theory, and as we see more and more uninfected animals, we become more and more confident that our theory is likely to be correct. However, no matter how many uninfected animals we see, we cannot prove that the theory is correct.

On the other hand, it is very simple to prove that theory is *not* correct. All that is needed is to find a single infected animal, and we have disproved the theory that the population is free from infection. Karl Popper established this principle of falsifiability as a critical foundation for science.

So, from a philosophical point of view, we cannot prove that a country is free from infection, no matter how many animals we test, but we *can* prove that a

**It is impossible to prove that a population is free from infection**

**A single infected animal can prove that a population is not free**

Surveillance for disease freedom provides evidence that allows practical decisions to be made

country is infected by finding a single infected animal. If our surveillance objective is to demonstrate that a country is free from infection, what can be done?

If absolute proof is not possible, then we must work with that which is possible. We cannot prove that we are free, but we can describe the level of confidence that we have, based on repeated observations of many non-infected animals. As in many areas of epidemiology where there is often uncertainty, instead of trying to achieve absolutes we are forced to work with probabilities.

The reason for seeking to demonstrate freedom from infection is to support decision-making. For example:

- A trading partner may need to decide if it is safe to import animals from another country.
- Veterinary authorities need to decide if the vaccination program can be stopped.

In order to support these decisions, absolute certainty of the disease status is not necessary. Having a high level of confidence about the disease status, so that the risk of being wrong is acceptably low, is normally adequate. The key requirement is that enough evidence is available to provide the confidence which allows practical decisions to be made. In order to do this, we have to work with probabilities.

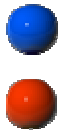
## Examples of sampling

---

The easiest way to understand the factors that influence our confidence in freedom from disease is to do some practical examples. The examples are based on classroom exercises that illustrate some of the concepts. The aim is to understand our intuitive feelings about how confident we are that a population may be infected or may be free from infection.

### Example 1: Disease free or high prevalence?

---



Consider a bag containing 1000 coloured balls. Blue balls represent uninfected animals, and red balls represent infected animals. A number of bags have been prepared, some containing all blue balls (a disease free population), and others have 200 red balls (a population that is infected with a prevalence of 20%). Without looking inside the bags, the task is to decide if the bag is one of the disease free bags, or one of the infected bags.

→ One ball is drawn from the bag and it is blue.

**Q** Are you able to guess if the population is the uninfected population, or the infected population?

If the ball had been red, then we would have proven that the population was infected, but as the ball is blue, we are not sure whether the population is infected or not, and a single ball does not provide much confidence.

→ Four more balls are drawn from the bag and they are all blue.

**Q** Do you feel confident that the bag represents a disease-free population?

Your level of confidence has increased, as there is now more evidence available, but you can't be sure.

**Q** If the population is infected and the prevalence of disease (proportion of red balls) is 20%, how many red balls would you expect to have seen after drawing five balls from the bag?

The expected number of red balls is equal to the probability that each ball is red (20%) multiplied by the number of balls chosen (5) which equals 1. So if the population is infected, you would expect to have seen 1 red ball after having drawn five balls.

**Q** You have seen no red balls after drawing five balls. Does this mean that the population must be free?

Of course, the population could still be infected. The *expected* number of red balls is 1, but this doesn't necessarily mean that you would get a red ball after five draws. The process is random, so even from an infected population with a prevalence of 20%, you may draw a red ball as the first ball, or you might not find a red ball until quite a few blue balls have been chosen. On *average*, you would expect to have one red ball after having drawn five balls.

→ Five more balls are drawn at random, and they are all blue.

**Q** How confident do you now feel that the population is free from disease?

After having drawn 10 balls, you would expect to see 2 red balls if the population were infected. We have seen no red balls. By now, most people would feel pretty confident that the bag represented a disease-free population.

→ Ten more balls are drawn, and they are all blue.

There are now 20 balls, and we would have expected 4 red ones if the population were infected. Most people would be ready to conclude that the population is most likely free from disease. Note that it is possible, by chance, to draw 20 blue balls in a row from a population with 20% red balls, but it is very unlikely (the probability of this happening is only 1.15%).

### ***Example 2: Disease free or low prevalence?***

---

Let us repeat this exercise, but this time the choice is different. Again, a number of bags have been prepared, but those that are infected have only 20 red balls out of 1000, or a prevalence of 2%.

→ Five balls are drawn at random and they are all blue.

**Q** How confident are you that this population is free from disease?

This time, if the bag is infected there are only a few red balls in the bag. It will therefore be much harder to find them. After drawing five balls, there is still a good chance that we wouldn't have found one of the infected balls, so our level of confidence is very low.

→ Fifteen more balls are drawn and they are all blue.

**Q** How does your level of confidence this time compare with your level of confidence after drawing 20 balls in the previous example?

In the previous example, the expected number of red balls from an infected population after having selected 20 balls was four. In this example, the expected number is  $2\% \times 20$ , or 0.4 balls. Even if the population is infected, it is still likely that we would not have detected an infected animal (red ball) yet. Our confidence is much lower this time, compared to the same number of samples in the previous example.

→ Thirty more balls are drawn and they are all blue.

Now fifty balls have been drawn and the expected number of red balls, if the population is infected is 1. After drawing fifty balls, our level of confidence is about the same as it was when we had drawn only five balls in the previous example – we are still unable to make a reasonable guess as to whether the population is infected or not.

### Example 3: Imperfect sensitivity and specificity

---

In the previous two examples, the colour of the ball has indicated the true disease status of the animals – if a ball is blue, the animal is not infected and if the ball is red, the animal is infected. The test (using our eyes to detect the colour) can be considered to be perfect – we never call a red ball blue, nor a blue ball red.

Consider an example where the test is not perfect. All the balls are the same colour, and they have to be tested by a machine to determine whether they are infected or not. However the machine makes mistakes – the sensitivity is 95% and the specificity is 90%.

Using this machine, you could never be sure if a ball that gave a negative test result was truly negative, or maybe a false negative. A positive test result could indicate a true positive or a false positive.

Let us repeat the exercise in Example 1 where we were trying to distinguish between a population that is free and a population that had 20% infection.

→ After drawing five balls, they all test negative.

**Q** What is our level of confidence after testing five balls, compared to the same stage in Example 1?

Although each ball tests negative, the sensitivity of the test is 95%. This means that an infected ball has a 5% chance of giving a false negative result. We may have already found an infected ball, but our test gave the wrong result. We are therefore a little less confident this time than we were in Example 1.

→ Five more balls are drawn and one tests positive.

**Q** Is the population infected or not?

We now have ten balls with one positive test result. If the population is infected at 20% we would expect to have 2 infected balls by now, although it is still quite likely that we would have only found 1. However, the specificity of our test is 90% which means that it would, on average, produce a false positive 10% of the time. If the population was not infected, we would expect to see one positive result after drawing 10 balls. We are now not sure if the positive is a true positive or a false positive, or if the negatives are correct either. Our confidence level about the status of the population is lower than in Example 1, and it is very difficult to make any useful guess about the status of the population.

### Conclusion

---

**Factors influencing confidence: sample size, assumed prevalence, sensitivity, specificity**

These examples illustrate some important factors that influence our confidence as to whether a population is free from disease or not:

- 1) Our confidence increases as the number of samples increases.
- 2) Our confidence depends on the assumptions about the level of disease in the population. When we are trying to decide if the population is infected or not, if we assume that disease would be common in an infected population, it would be easier to detect, and our confidence grows more quickly when we fail to detect it. On the other hand, if the disease is assumed to be rare in an infected population, more sampling is required to achieve the same levels of confidence.
- 3) Our confidence depends on the sensitivity and specificity of the test we use.

These relationships can be expressed mathematically as follows:

Confidence  $\propto$  sample size, assumed prevalence, sensitivity, specificity



This can be read as: “confidence is proportional to sample size, assumed prevalence, sensitivity and specificity”. It means that if any of these factors are increased, the confidence increases, and if any are decreased, the confidence decreases.

## Probabilities, confidence and freedom

---

The term ‘confidence’ has been used here in its general English meaning, to give an indication about how confident we feel about the surveillance and its ability to detect disease if it is present. When dealing in probabilities, it is important that the exact technical meaning of the various terms used is clearly understood.

When analysing surveillance, the aim is to determine the probability that the surveillance system would find at least one diseased animal based on the assumption that the population is infected at a specified prevalence. This may be written using probability notation:

$$\text{Confidence in surveillance} = P(T+ | D+)$$

Where:

- T+ means getting a positive result from our surveillance. Here surveillance is considered as a type of test of the entire population.
- D+ means that the population is infected (at the specified prevalence).

This is exactly the same concept as the sensitivity of a diagnostic test (the probability of getting a positive test result, given that the animal is infected):

$$\text{Sensitivity} = P(T+ | D+)$$

**Sensitivity is a measure of the confidence in a surveillance system**

Our measure of ‘confidence’ in our surveillance system is therefore a measure of the sensitivity of the surveillance system. The result of our analysis of a surveillance system is normally expressed in terms of sensitivity, but usually requires more detailed explanation. For instance:

“The sensitivity of the surveillance system is  $x\%$ , which means that the probability of finding at least one infected animal, assuming that the population is infected at a prevalence of  $P$ , is  $x\%$ .”

Sensitivity is a useful measure of the quality of a surveillance system and its ability to detect disease. However, for decision makers, a more intuitive measure is the probability that the country is free from disease. In probability notation, this may be expressed as:

$$\text{Probability of freedom} = P(D- | T-)$$

Or in words, the probability that the country is free from disease (D-), given that our surveillance did not detect any infected animals (T-). Calculation of the probability of freedom from disease, based on the sensitivity of the surveillance system is discussed in Chapter 15.

## Specificity of surveillance

---

The performance of diagnostic tests on individual animals is described by the sensitivity and the specificity. The specificity is the probability that the test will

give a negative result in an uninfected animal (the true negative rate). If we can also talk about the sensitivity of a surveillance system, then there must also be a specificity for a surveillance system – the probability that, if the country is free from disease, the surveillance system will give negative results.

When the purpose of surveillance is to demonstrate freedom from disease, imperfect specificity means that there is a possibility of false positives. A false positive means that we will conclude that the country is infected, when it is truly uninfected. This is a major mistake as it may result in the implementation of costly emergency control activities and the loss of trade opportunities. For these reasons, steps are normally taken to ensure that the specificity of any diagnostic system in such surveillance is as good as possible. Normally there are a series of confirmatory tests, and an animal is only considered positive if it gives a positive result to each of the confirmatory tests. This makes the specificity very high (but decreases the sensitivity).

Even with multiple tests, there is still a theoretical possibility that an animal in a surveillance system could give a false positive result. However, the specificity of the surveillance system is based not on the individual test results, but on the conclusions that are made about them.

If there is a positive test result that has been followed up with confirmatory tests and it is still positive, the conclusion will be that it is a true positive and that the country is infected. Once this conclusion has been reached (even if it is occasionally incorrect), the question of freedom no longer arises – the country is deemed to be infected. If an animal that initially tested positive later tests negative on confirmatory tests, then it is assumed to have been a false positive, and the conclusion is that it is truly negative.

Based on this logic, the specificity of a surveillance system to detect or demonstrate freedom from infection is normally assumed to be 100%.

## Design prevalence

---

In the examples above, you were asked to distinguish between two options: a bag representing a disease-free population, and a bag representing an infected population. The assumed level of disease (prevalence) in the infected bag influenced our confidence in the decision.

When analysing surveillance to demonstrate freedom from disease, this assumed prevalence value is important. If the value is high, the ability of the surveillance to detect disease (at that level) will be high. If the value is low, the ability to detect disease will be low.

The difficulty with this value is that it is not a real prevalence. We are dealing with a population that is free from disease, and therefore the real prevalence is zero. Instead, the value represents a hypothetical prevalence that is used to set the standard for our surveillance. To distinguish this value from a true prevalence, it is given the name *design prevalence* as it is used to establish the design of our surveillance. In equations, prevalence is usually denoted by the letter P, but design prevalence is represented with the symbol P\*.

If there is no disease, then it is not possible that the surveillance system would be able to detect disease. When we analyse surveillance, we are calculating the probability that the surveillance undertaken would be able to detect disease, *if disease were present at a specified level*. The design prevalence specifies the hypothetical level of disease that is used to measure the quality of our surveillance.

In order to account for disease clustering, it is often necessary to specify two levels of design prevalence: the proportion of infected herds in the population

**Design prevalence: a hypothetical prevalence that sets the standard for surveillance**

( $P_H^*$ ), and the proportion of infected animals in those infected herds ( $P_A^*$ ). Clustering is discussed in detail in Chapter 13.

### ***How to decide on an appropriate design prevalence***

The design prevalence sets the standard of proof for the surveillance. There is no right or wrong design prevalence. It is simply a value that has to be fixed in order to evaluate the surveillance. The main requirement of the design prevalence is that it is acceptable to those that need to make decisions on the basis of the surveillance.

#### **Example**

Surveillance has been undertaken in country A to demonstrate freedom from disease in order to support animal exports. In analysing the surveillance, the exporting country uses a design prevalence of 20%. This results in an estimate of the sensitivity of the surveillance which is very high (99.5%).

The country that wishes to import animals (country B) examines the analysis of the surveillance. They point out that this simply means that the country A has a 99.5% chance of finding the disease if 20% or more of the population is infected. Failing to detect infection simply means that the population could be infected at anything less than 20%. Country B suggests instead that a design prevalence of 1% should be used.

Country A objects, because if a design prevalence of 1% is used, the estimated sensitivity of the surveillance decreases to 64%.

For a given surveillance system, increasing the design prevalence will increase the sensitivity and vice versa. The requirement in this example is that both countries agree on a design prevalence value, and assess country A's surveillance against this single fixed value.

Unfortunately, the process of agreeing on an appropriate design prevalence is not simple. There are a number of possible approaches, and these are listed below, in order of preference.

#### **Global standards**

The OIE Terrestrial Animal Health Code and Aquatic Animal Health Code contain recommendations and standards for surveillance. For a small number of diseases, these standards include information on the required design prevalence, although these values may be expressed in a number of different ways. Where such standards exist, these should be used. Examples include:

##### **Bovine tuberculosis**

“Regular and periodic testing of all cattle, water buffalo, and wood bison herds did not detect *M. bovis* infection in at least 99.8% of the herds and 99.9% of the animals in the country or zone for 3 consecutive years”

**Terrestrial Animal Health Code, 2008, Article 11.7.2, section 3**

##### **Rinderpest**

“Annual sample sizes shall be sufficient to provide 95% probability of detecting evidence of rinderpest if present at a prevalence of 1%”

of herds or other sampling units and 5% within herds or other sampling units.”

**Terrestrial Animal Health Code, 2008, Article 8.13.22, section 5:a:ii**

#### **Bovine spongiform encephalopathy**

The application of Type A surveillance will allow the detection of BSE around a design prevalence of at least one case per 100,000 in the adult cattle population in the country, zone or compartment of concern, at a confidence level of 95%.

**Terrestrial Animal Health Code, 2008, Article 11.6.22, section h:1**

#### **Regional standards**

When the OIE code doesn't specify appropriate figures, regional standards may. The Council Directives of the European Economic Community (EU regulations) provide an example of regional standards.

#### **Trading partner requirements**

Where no standards exist, and the purpose of the surveillance is to support international trade, the requirements of the importing country should be used. This is appropriate when a country establishes clear standards for the import of animals, but this is not always the case.

#### **Acceptable level of protection (ALOP)**

A theoretical approach to determining the appropriate design prevalence is to calculate it based on the importing country's acceptable level of protection. The exposure assessment of import risk analysis normally starts with the prevalence of disease in the exporting country, and ends by calculating a probability of introduction of the disease. This may be compared to a national standard of acceptable risk (the ALOP) – if the risk is higher, imports are not permitted or risk mitigation strategies are required. If the risk is lower, trade is permitted. If a quantitative risk analysis is performed, and the ALOP is specified quantitatively, it is possible to do a risk analysis in reverse. This means that the risk of introduction is determined from the ALOP, and the prevalence in the country of origin that would result in this exact risk is calculated. This prevalence can then be used as the design prevalence, as if the level of disease is lower, the risk for importations will be acceptable.

In most cases, this approach is only theoretical, because:

- Virtually no countries have an explicit, quantitative ALOP. This is a concept that is embedded in the WTO's Agreement on the Application of Sanitary and Phytosanitary Measures (SPS agreement) but is almost never translated into reality.
- Conducting a quantitative import risk analysis is complex and time consuming.

#### **Biology**

The most common way to determine suitable design prevalence values (as the previous options are frequently not possible) is to base the value on an understanding of the biology of the disease.

### Example

In a naïve population, foot and mouth disease (FMD) normally spreads rapidly and infects a high proportion of the exposed animals. Typically, 60% to 80% of a herd would become infected and seroconvert. If the disease had been introduced to a susceptible herd, it would be extremely unusual for less than, say, 50% of the herd to be seropositive within a month or two of the infection.

In this case, if a design prevalence of 50% is used, the surveillance would be able to conclude, if no infection is found, that the disease, if present, has infected fewer than 50% of the animals. Normally, this would not be considered adequate proof, but for a highly contagious disease like FMD, it is biologically implausible that the disease could be established and infect fewer than 50% of the animals. Demonstrating that the disease, if present, is present in less than 50% is logically equivalent to demonstrating that the disease is not present at all.

Normally, when this approach is used, the design prevalence is decreased somewhat to account for unusual circumstances where disease spread is slower. Even if it is biologically extremely unlikely that less than 50% of the herd would be infected, a design prevalence of 20% or even 10% is often used.

This approach is appropriate for highly contagious diseases. For less contagious and slowly developing diseases, it is often biologically plausible for an extremely small proportion of the herd to have been infected, without significant further spread. In these cases, the biology of the disease does not help guide the decision of design prevalence.

A particular problem arises in vaccinated populations, even when dealing with highly contagious diseases. For example, surveillance to demonstrate freedom from FMD infection is sometimes conducted in vaccinated populations using non-structural protein (NSP) ELISA tests that can distinguish between antibodies derived from vaccination and those derived from natural infection. Normally, for FMD, a reasonable design prevalence would be 10% or 20%. However, in a vaccinated population, the disease can no longer be considered highly contagious. It is biologically feasible for a small number of (non-immune) animals in a herd to have been exposed and seroconvert, without the rest of the herd being affected. The choice of the design prevalence therefore depends not just on the disease, but the characteristics of the population being studied.

Vaccinated populations require a lower design prevalence.

### Practical considerations

For slow moving diseases or diseases that are difficult to transmit (such as tuberculosis or Bovine Spongiform Encephalopathy (BSE)), the final choice of design prevalence is often dictated by practical considerations. As these diseases may affect a small proportion of the population, the design prevalence should be as low as possible. However extremely low design prevalence values mean that very large sample sizes are required to achieve an acceptable level of sensitivity. In practice, the design prevalence is made as small as possible, while still being able to conduct affordable surveillance.

The judgement of what is practical depends on the nature of the disease, the resources of the countries involved, and the consequences of infection. Typically, the lowest design prevalence values that are used are 0.1% as this is judged to be the lowest for which surveillance can be affordably run. Tuberculosis is one example of a disease for which a design prevalence of 0.1% has been used. The only disease for which a lower design prevalence has been used is BSE (0.001%), and this is due to the perceived high consequences of human exposure.

## Arbitrary choice based on commonly used values

Finally, if none of the above considerations help define an appropriate value for design prevalence, the choice becomes arbitrary. It is more important to have a fixed agreed design prevalence value than to worry too much about getting the 'right' value. The most commonly used values are 1% at the herd level and 1%, 5% or 10% at the animal level.

### Integer design prevalence values

Design prevalence, as with real prevalence values, is a proportion and is often expressed as a percentage. It describes the proportion of animals in a herd, or the proportion of herds in the population, that may be infected.

Consider a herd with 15 animals. If the design prevalence in this herd is 1%, it means that  $1\% \times 15 = 0.15$  animals are infected. It is not possible to have a fraction of an animal infected – the whole animal is either infected or not infected. The possible prevalence values for this herd are therefore 0%, 6.7%, 13.3%, 20%, 26.7% and so on. It is not possible to have an infected herd with a prevalence lower than 6.7%.

When the number of animals in a herd is small and the design prevalence is also small, the effective design prevalence is determined by rounding up the target design prevalence to the nearest value that is possible based on a whole number (integer) of infected animals. Herds of different sizes will have different effective design prevalence values, and may have different numbers of animals that are assumed infected.

Design prevalence may be expressed as the integer number of infected animals per herd, or integer number of infected herds in the population

One approach to simplify this situation is to express the design prevalence in terms of an integer number of infected animals, instead of as a proportion. For instance, a design prevalence that is sometimes used is one animal per herd or one herd in the population. For herds of different sizes, this represents a varying proportion, but it is still an acceptable and unambiguous definition of the design prevalence. A single infected animal per herd is the most commonly used integer design prevalence, but larger numbers may also be used.

One interesting side-effect of using a design prevalence of one animal per herd is the ability to assess confidence in absolute freedom, rather than freedom relative to a specified level of disease. The sensitivity measures the probability of detecting disease at the specified design prevalence, and if no disease is detected, we can conclude that, if present, the disease prevalence is lower than the design prevalence. When the design prevalence is one single infected animal, it is not possible to have a lower disease prevalence, so failing to find disease at this prevalence means that disease is not present at all.

### Design prevalence for early warning systems

Thus far, the discussion has focused on surveillance to demonstrate freedom from disease or infection, mainly for the purpose of supporting international trade. Another purpose of surveillance is to ensure that if disease enters the country, it can be detected as quickly as possible, so that an emergency response can be launched.

Consider a country with 100,000 herds. A herd-level design prevalence of 1% means that the surveillance has a good chance of detecting disease if at least 1000 herds are infected. This represents a very large number of infected herds, and while it may be adequate for trade purposes, it is not adequate for early detection and emergency response. Surveillance systems for this purpose have to be able to

detect disease at much lower levels, preferably before the disease starts spreading and when only the first or second herd is infected. In this example, it would mean a design prevalence of, say 2/100,000 or 0.002%.

Normally, we set the design prevalence, and then calculate the sensitivity of the surveillance system. For early warning systems, it is also possible to do it the other way around. Instead of asking “What is the sensitivity of my surveillance if the design prevalence is 1%?”, we could ask the question “How many herds would have to be infected before my surveillance system could detect them with a sensitivity of 95%?”

## Relative and absolute freedom

The use of the design prevalence means that the sensitivity of our surveillance system is being measured against an agreed standard. If our surveillance has failed to detect disease and the sensitivity of the surveillance is good, we may conclude that the disease is not present at a level equal to or higher than the specified design prevalence. However, this still means that the disease could be present at a level lower than the design prevalence.

This raises the question of what we actually mean when we talk about demonstrating freedom from disease and specify a certain design prevalence.

### Example

A surveillance program is in place for bovine tuberculosis, following an eradication program. Tuberculosis is a disease that spreads slowly, so it is biological feasible that a very small proportion of a herd could be infected and a small number of herds in the country could be infected. The herd level design prevalence ( $P^*_h$ ) set by OIE for demonstration of freedom from tuberculosis is 0.2%

The surveillance program detects a small number of infected herds, representing 0.1% of the population.

**Q** Is the population free from tuberculosis?

There are two possible answers to this question. The first states that, as the prevalence detected is less than the design prevalence, the population can be considered as ‘officially free’. This reflects the concept of ‘relative freedom’. Freedom is defined as a prevalence of disease less than the specified design prevalence.

The second approach is to recognise that any infected animals in the population mean that the population is not free from infection. The best that can be claimed is that the prevalence of disease is very low. This reflects the concept of ‘absolute freedom’.

### Example

A second country is also completing an eradication program. Their surveillance (designed using a design prevalence of 0.2%) has failed to find any infected herds. Their conclusion is that they are free from disease.

Both countries have conducted detailed surveillance and have demonstrated that, if the disease is present, the prevalence is lower than 0.2%. The difference between the two countries is that the first *knows* that there is still disease present,

but in the second, they may be free, or disease may be present but remains undetected. Should the second country be considered to have a better disease status than the first?

This is a question that is difficult to answer. Many prefer to use a definition of freedom based on what we know. If we know that infection is present in a country, then the probability that the country is free is zero (it is known to be infected). If surveillance has failed to find infection in the country (even though it is possible that some infected animals remain undetected), then the probability of freedom is greater than zero and may be calculated.



# Chapter 5– Representative Surveys to Demonstrate Freedom from Disease

Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.

John von Neumann (1903 – 1957)

In the past, representative surveys were considered the best way to gather evidence to demonstrate freedom from disease. Representative surveys are based on random sampling and have two major advantages:

- All animals in the population are represented. This avoids bias and gives you confidence that you have not missed part of the population.
- Analysis of surveys based on random sampling is relatively simple.

The disadvantages with representative surveys are that they are often very expensive and inefficient. More recently, approaches to the analysis of risk-based surveillance (that is, non-representative surveillance) have been developed. While the purpose of this book is to give the reader the skills required to design and analyse risk-based surveillance, it is important to first understand how representative surveys for freedom from disease are designed and analysed.

## Survey design

---

The simplest form of a representative survey to demonstrate freedom from disease involves single-stage simple random sampling.

### Example

A flock has 2000 birds. The objective is to demonstrate that the flock is not infected with avian influenza.

A simple random sample of 50 birds is selected, meaning that each bird in the flock has the same probability of being selected as every other bird (50/2000 or 2.5%).

In this type of survey, we assume no knowledge about the individual birds. Some birds may be older or younger; some birds stronger or weaker; birds in different parts of the house may have a different risk of exposure to wild birds carrying the disease. In representative surveys using random sampling, none of these factors are taken into account. The beauty of simple random sampling is that it ensures that the sample selected will be as representative of the population as possible. If 10% of the population is made up of older birds, the sample will have approximately 10% older birds. If 5% of the population are exposed to a higher risk of infection than the rest, the sample will have approximately 5% at higher risk.

More complex representative designs are possible, including multi-stage surveys and those using different approaches to random sampling. These are discussed in more detail on page 54.

## Calculation of sensitivity

---

The sensitivity of a survey is the probability that, if the population is infected (it is disease positive, D+) at a given design prevalence ( $P^*$ ), at least one infected animal will be detected by the survey (the survey, as a test of the population, would have a positive test result, T+). In probability notation:

$$\text{Survey sensitivity} = P(T+ | D+, P^*)$$

In representative surveys where animals are chosen by simple random selection, the probability that each animal is infected is equal to  $P^*$ , the design prevalence. This makes the calculation of sensitivity relatively simple.

### Simple example

---

Consider the flock of 2000 birds. Let us use a design prevalence ( $P^*$ ) of 5%, meaning that our survey is aiming to detect disease if at least 5% of the population is infected. Our aim is to calculate the sensitivity of our survey or the probability that we would successfully detect disease if it were present.

The method of calculating survey sensitivity is based on simple application of probability rules in a step-by-step manner.

**Q** If a single animal is chosen at random, what is the probability that it would be infected?

If the population is infected, our survey design uses the design prevalence to specify the level of infection that would be present. In this case, our design

prevalence is equal to 5%. If the prevalence of infection is 5% then the probability that any animal chosen at random would be infected would also be 5%.

$$P(\text{infected}) = P^* = 5\%$$

**Q** What is the probability that a single animal chosen at random would *not* be infected?

This is an application of the NOT probability rule:

$$\begin{aligned} P(\text{not infected}) &= 1 - P(\text{infected}) \\ &= 1 - P^* \\ &= 95\% \end{aligned}$$

**Q** What is the probability that two animals chosen at random would not be infected?

This could be rephrased – what is the probability that the first animal chosen is not infected *and* the second animal chosen is not infected. Using the AND probability rule:

$$\begin{aligned} P(2 \text{ animals not infected}) &= P(1\text{st not infected}) \times P(2\text{nd not infected}) \\ &= (1 - P^*) \times (1 - P^*) \\ &= (1 - P^*)^2 \\ &= 95\%^2 \\ &= 90.25\% \end{aligned}$$

**Q** What is the probability that 50 animals chosen at random would not be infected?

This is a simple extension of the previous example. The probabilities of each animal not being infected are multiplied together. If we use  $n$  to represent the sample size, then:

$$\begin{aligned} P(50 \text{ animals not infected}) &= (1 - P^*)^n \\ &= 95\%^{50} \\ &= 7.69\% \end{aligned}$$

**Q** What is the probability that at least one animal out of those 50 *is* infected?

This again is an example of the NOT probability rule. If one or more animals is infected it means that all fifty animals are not uninfected:

$$\begin{aligned} P(\text{at least one infected}) &= 1 - P(\text{all uninfected}) \\ &= 1 - (1 - P^*)^n \\ &= 1 - 95\%^{50} \\ &= 92.3\% \end{aligned}$$

We have calculated the probability that we would get one or more positive test results using a perfect test (sensitivity and specificity both equal to 100%) if we sampled 50 animals from the population, with a design prevalence of 5%. This is therefore the sensitivity of our survey.

$$\begin{aligned}\text{Survey sensitivity} &= \Pr(T+ | D+, P^*) \\ &= 1 - (1 - P^*)^n\end{aligned}$$

### Imperfect sensitivity

The previous example gave the formula for selecting at least one infected animal in our survey. The problem is that once an infected animal has been selected, we need to test the animal to determine if it is infected or not. Unfortunately, our diagnostic tests are virtually never perfect.

Let us assume that the sensitivity (Se) of our test is 90%, but the specificity (Sp) is 100%. This assumption of perfect specificity was discussed on page 39. We are now interested not in the probability of selecting an infected animal, but the probability of getting a positive test result. If we have perfect specificity, we cannot have a false positive, so any positive result indicates that the population is truly infected.

**Q** What is the probability that a bird chosen at random from the population will give a positive test result?

In order to give a positive test result, the bird must first be infected (D+), and then it has to test positive (T+) to our diagnostic test, given it is infected (D+). These two events mean that we need to use the AND probability rule.

$$\begin{aligned}\text{P(animal tests positive)} &= \text{P}(D+) \times \text{P}(T+ | D+) \\ &= P^* \times \text{Se} \\ &= 5\% \times 90\% \\ &= 4.5\%\end{aligned}$$

**Q** What is the sensitivity of the survey, taking imperfect sensitivity into account?

The same logic from the previous example can be used, but this time,  $(P^* \times \text{Se})$  replaces  $P^*$ . The final formula is:

$$\text{Survey sensitivity} = 1 - (1 - (P^* \times \text{Se}))^n \quad (1)$$

This is an important result, as it is used as the basis for the analysis and design of both representative and risk-based surveillance. It is therefore worth writing it larger, so it is easier to remember.

$$\text{Survey sensitivity} = 1 - \left(1 - (P^* \times \text{Se})\right)^n$$

## Small populations

The formula for survey sensitivity presented above is adequate for most situations, but it is based on two assumptions to simplify matters. This section and the one following discuss these assumptions. These are slightly more advanced topics and are not essential to the understanding of the fundamentals of risk-based surveillance. The calculations involved in these sections are normally left to computer software to implement.

The first assumption is that the probability of selecting an infected animal is independent of the result of other selections, that is, it is not affected by whether an infected animal was selected previously or not.

### Example

Consider a small herd of 20 animals, with a design prevalence  $P^*$  of 20%. This means that, if the herd is infected,  $20\% \times 20$ , or 4 animals would be infected.

The probability of selecting an infected animal when the first animal is chosen would be  $4/20$  or 20%. However, if an uninfected animal were selected first, the probability of selecting an infected animal at the second draw would be  $4/19$  or 21%. On the other hand, if the first animal selected were infected, then the probability that the second animal chosen would be infected is  $3/19$  or 15.8%. The probabilities of selecting a positive or negative at each step for the first three animals sampled are shown in Figure 3.

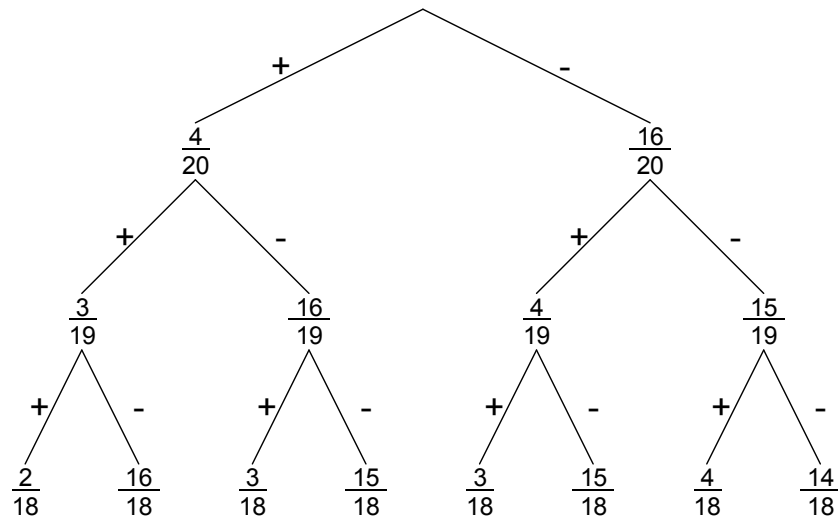


Figure 3: Probabilities of selecting infected (+) or uninfected (-) animals from a population of 20 with a design prevalence of 20%

With each animal sampled, the probability of selecting an infected animal changes. The formula for survey sensitivity shown above is no longer valid, as that formula assumes a constant probability of selecting infected animals ( $P^*$ ). There is a more complex formula for survey sensitivity that takes this effect into account, based on the hypergeometric distribution. This formula is implemented in the computer software discussed in Chapter 17.

When the sample size is small relative to the population size, the effect of changing probabilities is very small. For instance, in the above example with a population of 20 animals, the change in the probability of selecting an infected animal between the first and second selections (if an infected animal were chosen first) is from 4/20 (20%) to 3/19 (15.8%) or a decrease of 4.2%. If the population size was 2000, the probabilities would be 400/2000 (20%) to 399/1999 (19.96%) or a decrease of 0.04%. In large populations, it is common to assume that the change in probabilities is so small that it can be ignored.

The other approach that is sometimes used is to carry out ‘sampling with replacement’. This means that, whenever an animal is chosen from the population, it is sampled and then returned to the population (which means that it has a chance of being chosen a second time). This means that the probabilities do not change as more animals are selected and the formula is valid.

Because computer software is able to make the complex calculations to take changes in probability with each selection into account, there is normally no need to make these assumptions.

### ***Imperfect specificity***

---

If a test has imperfect specificity, it is possible for it to give a positive result when the animal is truly negative (a false positive). This means that when testing an animal, it is possible to get a positive result because the animal is truly infected and the test gives a true positive OR because the animal is uninfected and the test gives a false positive. The chance of selecting an uninfected animal is 1 minus the chance of selecting an infected animal ( $1 - P^*$ ) and the chance of the test giving a false positive is 1 minus the chance of it giving a true negative ( $1 - \text{specificity}$ ). This can be expressed as:

$$\begin{aligned} \text{Pr}(\text{animal tests positive}) &= \text{Pr}(\text{true positive}) + \text{Pr}(\text{false positive}) \\ &= (P^* \times \text{Se}) + ((1 - P^*) \times (1 - \text{Sp})) \end{aligned}$$

Using this approach, the formula for survey sensitivity becomes:

$$\text{Survey sensitivity} = 1 - [1 - ((P^* \times \text{Se}) + ((1 - P^*) \times (1 - \text{Sp})))]^n$$

This is the probability of getting at least one animal with a positive test result. With imperfect specificity, the animals with positive test results could well be false positives, so don’t necessarily mean that there are infected animals in the population. To overcome this problem, the survey design is often modified when using tests with imperfect specificity, so that a certain number of positive (assumed false positive) results are permitted before classifying the population as infected. Determining the acceptable number of positives results is discussed in the next section.

The formula is rapidly becoming more complex, so where we need to take imperfect specificity into account, it is better to leave the calculations to computer software.

### **Calculation of sample size with imperfect specificity**

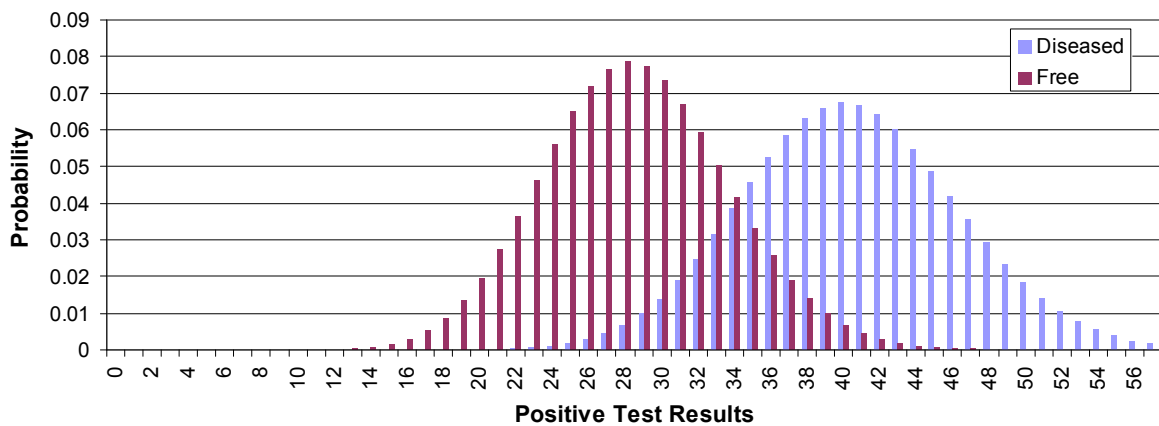
---

Most of the surveillance techniques in this book are based on the assumption of perfect specificity as discussed on page 39. This section looks at a special case,

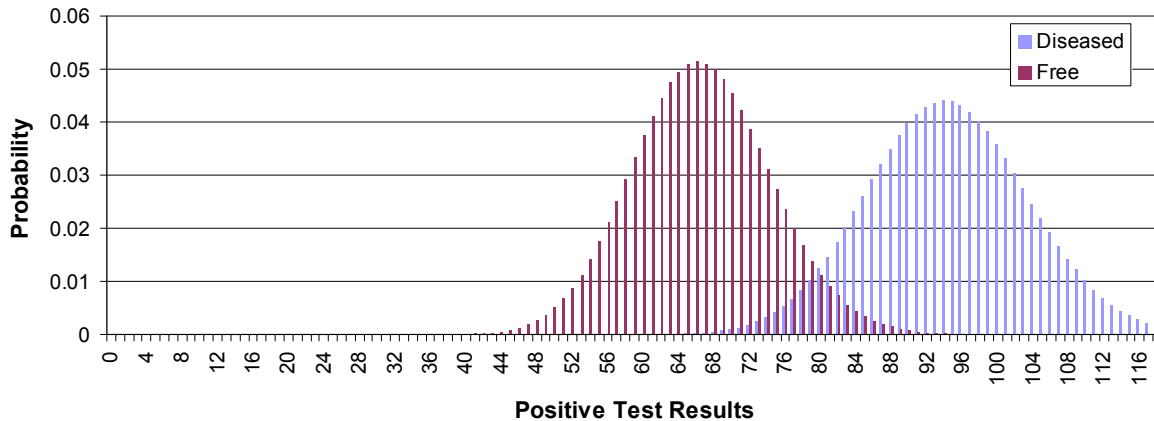
which is the use of representative surveys with *imperfect* specificity. The calculation of sample size is based on a comparison of two sampling distributions: the distribution of the likely number of positive results if the population is free from infection (i.e. false positives), and the distribution of the number of positive results if the population is infected at the design prevalence. When these distributions overlap significantly, it means that it is quite possible to get the same number of positive results from either a free or infected population. Increasing the sample size spreads out the two distributions and decreases the overlap. The size of the overlap determines the probability of making an error. If we set the level of the acceptable error to 5% (equivalent to a confidence level of 95%) we can determine the sample size that achieves an overlap just less than 5% between the two curves.

**Example**

The probability distributions below are based on a design prevalence of 5%, a test sensitivity of 80%, a specificity of 90% and a sample size of 300. The overlap between the free and the positive distribution is large, so the sample size is not large enough.



In the second example, the sample size has been increased to 700. Now, if we use a cut-off of 80 animals to define the population as positive or negative, the proportion of the free curve greater than 80 is less than 5%, and the proportion of the infected curve less than 80 is less than 5%. The sample size in this case needs to be 700 or more.



The calculations are complicated, so it is always better to use software tools to calculate sample sizes. There are a series of web-based tools for such calculations available at the EpiTools web site: <http://epitools.ausvet.com.au>.

## Two stage survey design

Surveys to demonstrate freedom from infection in small populations are relatively simple to design. However, when the population is large, there are new complications.

### Clustering of infection

Infection is rarely evenly distributed through a population. Normally it forms clusters.

#### Example

Consider Foot and Mouth Disease (FMD). When the disease is present in a country, it is not spread evenly across the whole country. Instead, at any one time (even if the disease is endemic), a small proportion of herds are infected, but most herds are not infected. However, in those infected herds, a high proportion of animals may be infected.

The result is small patches of a high prevalence of infection, while most of the other herds have zero prevalence. This 'patchy' distribution of infection is known as clustering.

Disease may cluster due to a range of factors, but population groupings are the most common – for example, herds, villages with shared grazing, fish with a common water source and so on.

Where clustering occurs (which is almost always with large populations), the use of a single value for the design prevalence to describe the level of infection is not enough. The overall prevalence of FMD in the population may be only 0.1%, but in infected herds, 60% of animals may be infected. In these cases, the level of infection is described by using two different values for the design prevalence – one at the animal level, and one at the group (herd) level.

For FMD, the animal-level design prevalence in an infected herd (known as  $P^*_A$ ) may be 20%, but the herd-level design prevalence (proportion of infected



herds in the entire population,  $P^*_H$ ) may be 1%. For surveys of large populations, it is normal to define these two levels of design prevalence.

Where the grouping structure of a population is more complex there may be even more design prevalence levels. For instance, pigs may be grouped into pens or sheds and then into farms. This may require three levels of design prevalence to describe accurately. Increasing the number of design prevalence levels complicates calculations considerably so should be avoided wherever possible. This explanation will be limited to two levels, but can be extended to more.

### ***First stage calculations***

Sample size calculations can be thought of as doing surveys at different levels. First we do a survey of each herd to find out if it is infected or not. Our herd-level survey has a risk of error, so has a defined sensitivity and specificity (these are the probabilities that we set to define the cut-off values for the overlap of the free and positive probability distributions). Each herd has a result (positive or negative) with a certain sensitivity and specificity. We can treat these herd-level results as if they were just another test, and analyse the population of all the herds we have tested to determine if the whole population is free.

The first stage calculations are therefore the same as those illustrated above. We use software to determine the sample size required, based on:

- The individual animal test sensitivity and specificity
- The animal-level design prevalence
- The error levels that we wish to set

The error levels can be a bit confusing so some terminology is explained in the table below.

<b>Error type</b>	<b>Probability value</b>	<b>Error description</b>	<b>Herd-level equivalent</b>
Type I	Alpha	False positive rate	1 - Specificity
Type II	Beta	False negative rate	1 - Sensitivity

#### **Example**

If we choose a type I error level (alpha) of 5% and a type II error level (beta) of 1% it means that our herd-level survey will have a sensitivity of 99% and a specificity of 95%.

Changing the error levels will change the sample size and the herd-level sensitivity and specificity.

### ***Second stage calculations***

Calculations at the second stage are very similar. The difference here is that we are calculating the number of herds that need to be sampled. The design prevalence is the herd-level design prevalence, and the sensitivity and specificity are not the animal-level diagnostic test values, but are instead those values that we defined by our type I and type II error rates at the first stage.

The error levels at the second stage determine the overall survey sensitivity and specificity.

### *Optimising the survey design*

---

The relationship between error levels and sensitivity and specificity means that, in the survey design, we can set different first stage error levels, but still achieve the same overall survey sensitivity and specificity. This simply means either testing more animals per herd and fewer herds, or fewer animals per herd and more herds.

This flexibility gives us the opportunity to optimise the survey design based on cost. These calculations assume that there is a per-herd cost and a per-animal cost. By varying the parameters, the least cost combination of animals per herd and total herds can be calculated.

These calculations are also available on the EpiTools web site:

<http://epitools.ausvet.com.au/content.php?page=2StageFreedomSS>

# Chapter 6– Risk-based Surveillance

The obscure we see eventually. The completely obvious, it seems, takes longer.

**Edward R. Murrow (1908 - 1965)**

The more original a discovery, the more obvious it seems afterwards.

**Arthur Koestler (1905 - 1983)**

Up until this point this book has discussed different aspects of surveillance, but has largely limited itself to trying to develop a good understanding of representative surveillance based on random sampling. Representative sampling is good when we want to:

- Measure the level of disease in a population and avoid bias
- Detect changes in the level of disease over time
- Describe the distribution of disease

This type of surveillance asks the questions “How much disease or infection is there and where is it?” These questions are answered with measures of the level of disease, such as prevalence.

However, the main focus of this book is surveillance to demonstrate freedom from disease, or for the early detection of disease. The question being asked is “Is disease or infection present?” The answer takes the form of a probability – the probability that our surveillance would have detected disease if it were present

(the sensitivity of our surveillance). For this type of surveillance, representative sampling is often not the best approach.

### Example

You arrive in a new city that you have never visited before. You start to feel sick and realise that you need to visit a doctor. What do you do?

- A) Select shops and houses at random, knock on the door and ask if there is a doctor available?
- B) Visit a cinema, a car repair shop, a computer shop and a bus station to see if you can find a doctor?
- C) Go to the hospital and ask to see a doctor?

Option A can be thought of as a representative survey. It is possible that, if the sample size was big enough, you'd eventually find a doctor by visiting randomly selected houses and shops. However the proportion of people in the population who are doctors is relatively small (there is a small design prevalence), so it could take a long time to find one.

Option B is a form of targeted surveillance. You decide to concentrate your search for a doctor in a number of different locations. However, the choice of locations is not good. It is possible that, through chance, you could find a doctor in a car repair shop, a computer shop or a bus station, but the probability may be even lower than if you had used random sampling.

Option C represents risk-based surveillance. In order to find a doctor, we go to a place where we know that doctors are most likely to be. There is a very small chance that there won't be any doctors (the hospital is closed, or the doctors are on strike), but by choosing a hospital, we give ourselves the very best chance of finding a doctor quickly.

Risk-based surveillance involves looking for something where we think it is mostly likely to be. The main thing that distinguishes risk-based surveillance from representative surveillance is our knowledge about the disease and the risk factors associated with the disease. In our example, we know that doctors often work in hospitals, so there is a higher probability of finding a doctor in a hospital than in most other locations.

**Risk-based surveillance means looking for something where it is mostly likely to be**

### Example

Consider another example. You have arrived in the city and feel perfectly healthy. A friend from home told you that you should meet somebody that he once knew, called Ahmed, but didn't know where Ahmed lives or works or even if he is still in the city.

In this example, we know almost nothing about Ahmed except his name. Looking in a hospital, a bus station, a car repair shop or a computer shop would all have some chance of finding him, but so would randomly selecting shops and houses. If we know some risk factors, we would be able to search more efficiently (perhaps if we had know that he likes Chinese food, we could search in the Chinese restaurants), but if we don't know any risk factors, random sampling is a reasonable way to search (but not very efficient, as the prevalence of 'Ahmed' may be very low – just one person in the whole city).

**To use risk-based sampling you must know some risk factors**

We could try to guess at some risk factors. For instance, your friend who told you to meet Ahmed likes swimming and goes to the pool often. You could assume that because they are friends, Ahmed likes to go to the pool as well, so you should search at swimming pools. The value of this approach depends on whether your assumption is right or wrong. Searching in swimming pools if Ahmed doesn't like swimming (and therefore never goes to the pool) could be a much worse way to try to find him than using random sampling.

In conclusion:

- Risk-based surveillance involves using knowledge of risk factors to improve the probability that we will find disease or infection
- Risk-based surveillance is more efficient at finding disease or infection than representative (random) sampling
- If we don't know about the disease or any suitable risk factors, it is not possible to use risk-based surveillance
- Surveillance that is based on some factor that is *not* a risk factor for the disease may be *less* efficient than representative sampling

## Factors influencing sensitivity

---

In Chapter 4 we presented an example of sampling from different bags to try to work out if the bag was 'infected' (had red balls present) or not. In the discussion about this example on page 38, the factors that influence our confidence about surveillance (or the sensitivity of the surveillance) were listed:

- The design prevalence ( $P^*$ )
- The sensitivity (Se) and specificity (Sp) of the test used
- The number of animals included in the surveillance ( $n$ )

Equation (1) on page 50 showed how these factors are related (when specificity is assumed to be 100%):

$$\text{Surveillance sensitivity} = 1 - (1 - (P^* \times \text{Se}))^n$$

## Population variation

---

The equation above makes an important assumption. Remember that the middle term  $(1 - (P^* \times \text{Se}))$  represents the probability that an animal will *not* provide a true positive test result. This is raised to the power of  $n$  animals in the surveillance, which implies that all those animals are assumed to have the same values for  $P^*$  and Se, i.e. that all animals have the same probability of being infected, and that they all have the same probability of being detected.

This is clearly not true. If disease is present in a population, some animals are at a greater risk of becoming infected than others, depending on the nature of the disease. Some diseases affect young animals more than old, while some affect females and not males. There are many possible risk factors that describe differences in the risk of infection for different parts of the population. Similarly, infection may be easier to detect in some animals compared to others. For example, while both cattle and sheep can become infected with Foot and Mouth Disease (FMD), cattle often show clear clinical signs, while in sheep these signs may be absent or very subtle. Species is therefore a factor that influences the probability of clinical detection of FMD.

When random sampling is used to ensure a representative sample, the average probability of infection,  $P^*$ , of the animals sampled will be the same as the average

Animals vary in their probability of infection and the probability of detection.

for the population (that's the purpose of representative sampling). Similarly the average sensitivity,  $Se$ , will be the same. Even though there may be significant variation in the probability of infection and sensitivity between individual animals, the average is the same as in the population, so the equation above can be used.

## Risk-based surveillance

When the selection of animals for surveillance is not representative, Equation (1) can no longer be used. This is because the average  $P^*$  and  $Se$  in the sample is no longer necessarily the same as the average for the entire population. While this may complicate things, it also provides an important opportunity for increasing the efficiency of surveillance.

Risk-based surveillance aims to take into account the differences in risk for animals in the population. By selecting animals with a higher probability of being infected ( $P^*$ ), or a higher probability of being detected if they are infected ( $Se$ ), the sensitivity of the surveillance can be increased without increasing the total number of animals being tested.

Risk-based surveillance also aims to account for differences in  $P^*$  and  $Se$  in different parts of the population. It does this by dividing the population into separate risk-groups.

### Example

In a population, the  $P^*$  for disease X is set at 5%, and the average sensitivity of the test being used is 90%. If we sampled 20 animals using random (representative) sampling, the sensitivity would be:

$$\begin{aligned}\text{Surveillance Sensitivity} &= 1 - (1 - (P^* \times Se))^n \\ &= 1 - (1 - (0.05 \times 0.9))^20 \\ &= 60.2\%\end{aligned}$$

However, if disease X is present in the population, it is three times more likely to affect young animals than older animals. The average probability of being infected is 5%, but as only 20% of animals are young, the probability in those young animals is 10.7% while the probability in older animals is 3.6% (the way these figures are determined will be explained later). If we use risk-based surveillance, we would concentrate on the part of the population with the higher risk. By sampling only young animals, the sensitivity for a sample of 20 would be:

$$\begin{aligned}\text{Surveillance Sensitivity} &= 1 - (1 - (P^* \times Se))^n \\ &= 1 - (1 - (0.107 \times 0.9))^20 \\ &= 86.8\%\end{aligned}$$

By focusing our surveillance on the group at the higher risk, we were able to increase the sensitivity by about 26% without testing any more animals.

# Chapter 7 – Analysis of Complex Surveillance Systems

“Things that are complex are not useful. Things that are useful are simple.”

Mikhail Kalashnikov (1919 – )

## Traditional approaches

---

Data collected by surveillance systems can be analysed and used for a variety of purposes. This chapter considers the situation where surveillance is used to provide evidence that a zone, country or region is free from disease, in order to support trade in animals or animal products. In the past, two distinct approaches have been available, sometimes used in combination.

## Structured surveys

---

The most common approach to demonstrating freedom from disease is to design and conduct a structured survey. The survey is designed in such a way as to achieve a specified sensitivity (e.g. 95%). This approach may be used at various levels, including a single herd, farm or pond, up to an entire country or group of countries.

The advantages of this system are that it is:

- Quantitative: as the survey is designed by those conducting the surveillance, it can be designed so that the results can be easily analysed using traditional probability theory, as discussed in Chapter 5. This usually means that random sampling is used to ensure a representative sample.
- Transparent: The method used to collect the sample and analyse the data can be easily documented, so that anybody using the results of

**Structured surveys ignore other available surveillance data**

the analysis (an importing country, for instance), can see exactly what was done.

- Repeatable: The results of the analysis are likely to be very similar (except for any random error present), no matter who does it.
- Objective: Once the methodology is documented, the surveillance and analysis is completely objective. As the result is quantitative, there is no element of personal judgement involved. Normally, a target or standard is established, and if the surveillance meets this standard, it is considered to be adequate, otherwise it is not.

The disadvantage of structured surveys is that they are often very expensive and wasteful. Before conducting a structured survey, it is important to be reasonably confident that the area is already disease-free, otherwise the expense of the survey is wasted. This means that there is likely to be a great deal of surveillance data that has already been collected, and that provides a reasonably high level of confidence that the disease is not present. This prior evidence usually comes from complex surveillance activities that are hard to analyse, but nevertheless provides valuable evidence.

When a structured survey is used to provide evidence of freedom from disease, this normally implies that all previous surveillance is ignored, and a single survey is used to provide new evidence.

### *Expert panels*

---

An alternative approach has been used in some cases, either in bilateral trade negotiations, or for a small number of diseases for which OIE grants official disease-free status. This involves the use of a small panel of experts who visit the area of interest, examine all the available surveillance data, laboratories, veterinary services and other infrastructure, and based on the overall picture, provide a judgement on whether the evidence is adequate to support a claim of freedom from disease.

This approach has some important advantages:

- It is able to take into account all the available surveillance data, including not only structured surveys, but also complex surveillance systems (such as farmer disease reporting systems), and historical surveillance data
- It is able to use information on the quality of the veterinary services to assess the reliability of the surveillance.

The use of an expert panel therefore is less wasteful, as it takes into account a whole range of complex factors and weighs all available evidence. The problem is that this process goes on in the heads of the expert panel members, and the results are therefore:

- Qualitative. It is very difficult to decide whether the available evidence meets a specified standard or not.
- Subjective. The decision depends very much on the personal opinions, experiences and biases of the experts involved. For instance, a prominent virologist may feel that the diagnostic test that they have developed is the best to use, and be unreasonably biased against a country that is using a different test that was developed by a competing laboratory.
- Non-reproducible. Different expert panels may come up with different conclusions about the same situation.



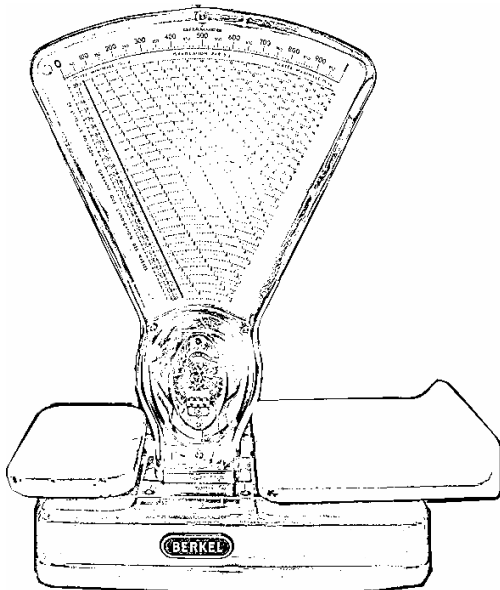
## Ideal system

The ideal system for analysing surveillance to demonstrate freedom from disease would provide a combination of the advantages of these two approaches. Specifically, it should:

- Be able to incorporate all available evidence, including both structured surveys and complex non-structured surveillance activities such as farmer disease reporting systems, abattoir surveillance and so on, as well as current and historical surveillance data.
- Be able to capture information about the quality of surveillance and the quality of the veterinary services.
- Be objective. The results of the analysis should not depend on who is doing the analysis.
- Be repeatable. Repeated analysis, either by the same or different people, should provide the same result (allowing for random error).
- Provide a quantitative outcome that allows simple evaluation of whether the evidence meets the required standard, and also allows comparison of the strength of evidence provided by different surveillance activities or by different countries.
- Be easily communicable. The principles, methods, assumptions and results should be able to be clearly documented and relatively easily understood by those with an interest in the analysis.

## Overview – an analogy

This book presents a collection of analytical techniques that meet most of the requirements of the ideal system. The approach used may be easiest to grasp with the use of an analogy.



Consider a set of old grocer's scales as shown on the left. Imagine that these scales are designed to help weigh evidence of freedom from infection. The numbers at the top indicate the probability that the country is free from infection, starting at 0 on the left (infected) to 1 on the right (definitely free). Evidence (in the form of weights) is placed on the tray on the right of the scale and makes the needle rise showing an increasing probability of freedom, depending on the strength of the evidence (size of the weight).



In this analogy, the weights represent the evidence gathered from surveillance, and the size of the weight represents the sensitivity of that surveillance. It is possible to achieve a high probability of freedom by putting a single heavy weight on the tray (i.e. by doing surveillance that has very high sensitivity). However, it is also possible to achieve the same probability of freedom by putting several smaller weights on the tray (i.e. by combining the evidence from several different surveillance activities). This

shows how different types of surveillance, even surveillance with relatively low sensitivity, can be combined with others to produce a high probability of freedom from infection.

The weights don't have to be put on the tray all at the same time. It is possible to put relatively small weights (use surveillance with poor sensitivity), but keep putting more and more small weights on the tray over a period of time. Eventually, given enough time, it is possible to accumulate enough small weights to provide a high probability of freedom. This is how surveillance evidence can accumulate over time.

However, there is also a tray on the left of the scales. When weights are put on this tray, the needle is pushed back towards the left, decreasing the probability that the country is free from disease. The left tray represents the risk that new infection may be introduced, because of poor biosecurity. If the biosecurity is perfect, and there is no chance of introducing new disease, then no weights will be put on the left tray. It is relatively easy to build up enough surveillance evidence on the right tray to give a high probability that disease is not present. But if the level of biosecurity is poor, then there is an ongoing risk that new infection will be introduced. This means that some weights are steadily being put on the left tray. If there is no new surveillance, then the needle will gradually move to the left, decreasing the probability of freedom.

The probability of freedom from infection is therefore a balance between two factors – on one side, the evidence of freedom from infection based on surveillance (which may be made up of multiple different types of surveillance with different sensitivities), and on the other, the probability that infection may be introduced.

To complete the analogy, the designer of the scales unfortunately decided to put a spring in the mechanism that tries to pull the needle back towards zero. As the needle moves towards 1, it gets harder and harder to move it further. This means that the first weight that is put on the right tray gives us a relatively high probability of freedom, but if an equal weight is added again, it results in a smaller increase in the probability of freedom than the first weight. Evidence of freedom is therefore not additive.

## Methodological requirements

---

In order to understand and measure the balance between new surveillance evidence from different surveillance activities and the risk of introducing infection (and therefore to describe the probability that the country is free from infection) a number of distinct methodological tools are required.

- A method to quantify the sensitivity of a component of a surveillance system. In our analogy, this is a method for determining the weight of evidence that a surveillance activity contributes. Analytical methods for structured random surveillance already exist, but a new method is needed for complex surveillance.
- A method to combine the evidence provided by different surveillance components. This is the equivalent of placing multiple different weights on the right tray of the scales.
- A method that allows us to calculate the probability of freedom from infection, based on the combined sensitivity of the surveillance. This represents the internal mechanism of the scale that determines where the needle points to.

- A method to account for the balance between the risk of introduction of new disease, and the progressive accumulation of evidence from surveillance over time. This will allow us to determine the true value of historical surveillance data.

### ***Quantifying the sensitivity of complex surveillance***

---

Chapter 5 discussed the analysis of structured surveys to demonstrate freedom from disease. Some relatively simple formulae were developed to analyse surveillance data and calculate the sensitivity of the surveillance.

Complex surveillance systems can't be analysed in the same way, because there are a whole range of different biases which mean that some animals are more likely to be infected than others and some animals are more likely to be selected than others.

The methodology that is used to estimate the sensitivity of complex surveillance systems is known as scenario-tree modelling, described in detail from Chapter 8 to Chapter 13. This method uses a tree structure to describe the population and surveillance structures, and to explicitly capture the probability that any given animal might be infected with the disease or that it might be detected. Scenario trees are the tool we use for quantifying risk-based surveillance as they provide a quantitative measure of the sensitivity of surveillance components.

### ***Combination of evidence from multiple surveillance components***

---

Chapter 14 looks at possible approaches to combine evidence from different surveillance components. The basic technique of calculating the combined sensitivity of two or more components is very simple, but becomes much more complex when there is overlap between the coverage of the surveillance components, as this means that we need to avoid 'double-counting' the same evidence.

### ***Calculation of the probability of freedom from infection***

---

The quality of surveillance is most commonly described in terms of sensitivity. However, the probability of freedom from infection is a more intuitive and often more useful measure of the quality of surveillance. Chapter 15 discusses how this is calculated, based on the combined sensitivity of all the available surveillance.

### ***Incorporating historical data***

---

As illustrated with the analogy of the scales, evidence can accumulate over time, but its value is decreased if there is an ongoing risk of introduction of new disease. Chapter 16 describes how to incorporate historical surveillance data and quantitatively measure the balance between evidence and risk of introduction of infection.

# Chapter 8 – Introduction to Scenario-Tree Modelling

Absence of proof is not proof of absence.

William Cowper (1731 – 1800)

## A simple example

---

In Chapter 5, we derived the basic formula that is used to analyse representative surveys to demonstrate freedom from infection. In that formula, the probability that an animal gives a positive test result is given by  $P^* \times Se$ , or, in words, the probability that the animal is infected times the probability that it gives a positive test result, given that it is infected. This formula was based on the assumption that the specificity of the test was perfect.

This can be represented by the simple tree shown in Figure 4. This diagram shows that there are two ways to get a positive test result (T+): an infected animal with a true positive test result, or an uninfected animal with a false positive test result. To calculate the probability of getting a positive test result, we use our two probability rules: the AND rule to multiply probabilities down the branches of the tree, and the OR rule to add the resultant probabilities. Thus the result is:

$$\begin{aligned} P(\text{Infected AND true positive}) &= P^* \times Se \\ P(\text{Uninfected AND false positive}) &= (1 - P^*) \times (1 - Sp) \\ P(\text{either one OR other of the above}) &= (P^* \times Se) + [(1 - P^*) \times (1 - Sp)] \end{aligned}$$

If specificity is equal to 1, this simplifies to  $P^* \times Se$ .

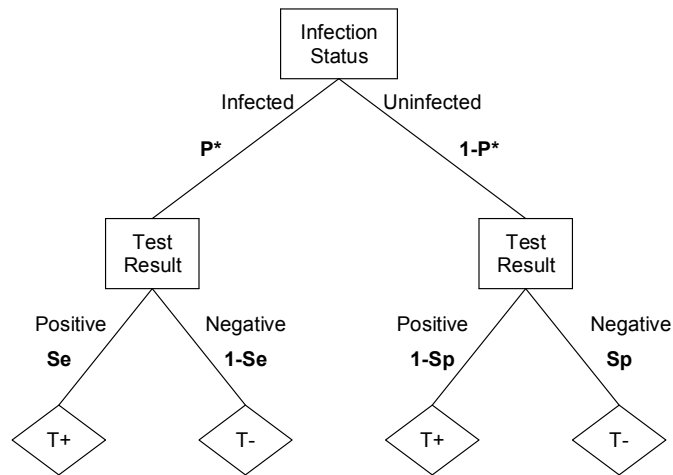


Figure 4: Simple scenario tree

This example assumes that all animals have the same probability of being infected, and the same sensitivity. If all animals are not the same, we can add some new nodes in the tree to describe the differences between animals. The tree shown in Figure 5 looks much more complicated, but is really still very simple. Two new nodes have been added: **AGE** and **SPECIES**. For this example disease, age influences the probability that an animal will be infected (young animals are at higher risk of becoming infected than older animals). Species influences the probability that an infected animal will be detected (sensitivity). This tree may represent a clinical surveillance system – cattle show typical clinical signs, while sheep often show only mild clinical signs.

The first tree in Figure 4 divided the population into four groups, based on infection status and test result. The tree in Figure 5 has divided the population into 16 different groups. Within each group, animals are similar but each group is different with respect to the factors included in the tree.

Including different factors in the tree allows us to assign different probabilities. If young animals have a greater risk of being infected, we can use a different probability of infection for young animals (on the left side of the tree) compared to old animals. If infected cattle are easier to detect than infected sheep, the sensitivity (probability of a positive test result in infected animals) will be different for cattle than sheep.

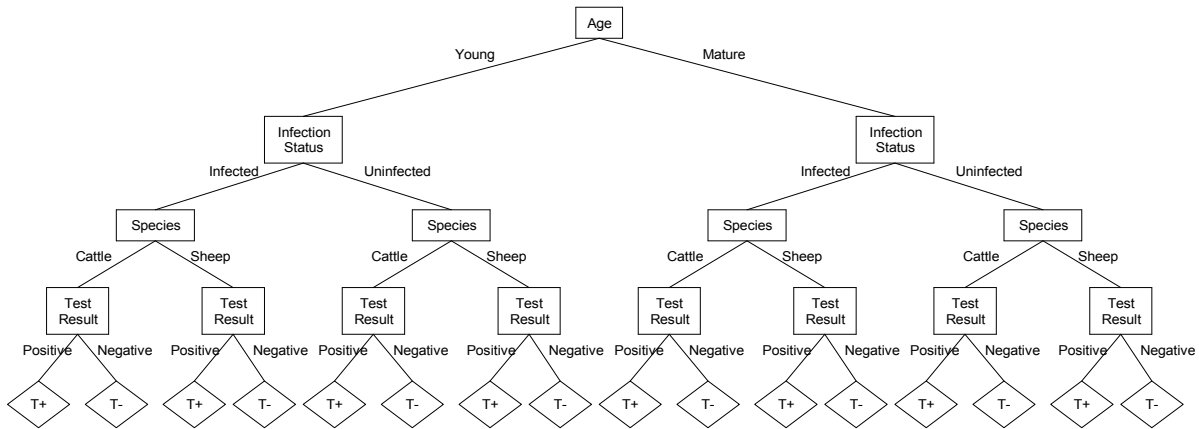


Figure 5: Scenario tree with one factor affecting probability of infection (age) and one factor affecting the probability of detection (species).

## Purpose of the scenario tree

A scenario tree is a tool to assist in the calculation of the sensitivity of a component of a surveillance system. In contrast to the simple analysis of representative surveys, the purpose of a scenario tree is to take into account the fact that not all animals in the population:

- have the same probability of being infected (some are at greater risk than others);
- nor do they have the same probability of being detected (the sensitivity of detection is greater in some animals than others).

Remember our formula for surveillance sensitivity:

$$\text{Surveillance sensitivity} = 1 - (1 - (P^* \times Se))^n$$

In this formula, all  $n$  animals in the population are assumed to have the same values for  $P^*$  (probability of being infected) and  $Se$  (probability of being detected). The scenario tree divides the population into smaller sub-populations, based on risk factors and detection probabilities.

The reason this is so valuable is because it allows us to analyse non-representative surveillance. If surveillance is targeted towards a group of animals that are at higher risk of being infected, a scenario tree allows us to calculate the sensitivity that we achieve for that particular group.

A scenario tree represents a series of different limbs, or paths from the beginning to the end and each limb defines a sub-population. For instance, in Figure 5, the first of the 16 sub-populations is made up of those animals that are young, infected, cattle and test positive, while the last is made up of animals that are mature, uninfected, sheep and test negative. The probability that a randomly selected animal falls into one of these groups can be calculated by multiplying the probabilities at each step down the limb (using our AND probability rule). For instance, the probability that an animal is in the last group is the probability that an animal is mature multiplied by the probability that it is uninfected multiplied by the probability that it is a sheep multiplied by the probability that it gives a negative test result.

**Scenario trees divide the population into homogenous subpopulations. Animals in a subpopulation have the same risk of infection and probability of detection.**

## Terminology

---

There are a number of terms that are commonly used when talking about scenario trees, and they need to be defined.

- NODE** A node represents a factor that is used to divide the population into a number of groups. In Figure 5, nodes are shown as square boxes. A good way to think about nodes is that they are asking a question about the animal or group of animals. For example the *AGE* node is asking “What is the age of the animals?” and the *SPECIES* node is asking “What species are the animals?”
- BRANCH** A branch represents the answer to the question, dividing the population into different groups on the basis of the node from which the branches come. Branches are shown in Figure 5 as lines coming out of the bottom of each node box. The branches from the *AGE* node answer the question: “Is the animal *YOUNG* or *OLD*?”, thereby dividing the population into two groups. Branches have probabilities associated with them, indicating the probability that an animal or group of animals will belong to that branch.
- Outcome** The outcome (sometimes called a ‘leaf’) is at the end of the last branches of the scenario tree. In Figure 5, these are shown as diamonds. The outcome represents the final conclusion about animals in that group. The two possible outcomes are normally *TEST POSITIVE* or *TEST NEGATIVE*.
- Limb** This describes the path through a particular series of nodes and branches, starting at the beginning and continuing to the end, that results in a single outcome.

## Branch probabilities

---

Branch probabilities are often proportions

Each branch is associated with the probability that an animal or group of animals will fall into that branch. In Figure 5, the *AGE* node has two branches: *YOUNG* and *MATURE*. The probability that an animal falls into the *YOUNG* branch is given by the proportion of young animals in the population. Branch probabilities are therefore often proportions.

Branch probabilities can also be sensitivities or specificities

Another example is the *TEST RESULT* node. Here, the proportion of animals that have a positive test result depends on whether the animals are infected. If they are infected, the probability is the individual animal test sensitivity. If the animal is not infected, it is one minus the test specificity.

In the example of *TEST RESULT*, the probability depends on the previous branch – is the animal infected or not. In fact, probabilities in a scenario tree are always conditional, which means that all probabilities depend on all the previous branches in the tree.

Probabilities are conditional on all previous branches in the tree

For instance, the *SPECIES* node has two branches – *CATTLE* and *SHEEP*. However, there are four different *CATTLE* branches. The probability for the first *CATTLE* branch is based on the sub-population of *YOUNG* and *INFECTED* animals. Some young infected animals are cattle, and some are sheep. The proportion of young infected animals that are cattle is the correct branch probability to use.

The second *CATTLE* branch is for *YOUNG UNINFECTED* animals, the third is for *MATURE INFECTED* animals and the last is for *MATURE UNINFECTED* animals. Each of the

probabilities for the four cattle branches may therefore be different, and depends on the previous branches.

Changing the order of nodes in a tree changes the conditional probabilities. In the previous example, it was necessary to estimate the proportion of young animals that are cattle. This is a non-intuitive value and may be difficult to estimate. Changing the node order so that *SPECIES* comes first, then *AGE* would mean that the probability to estimate is the proportion of *CATTLE* that are *YOUNG*.

## Node types

---

There are three main types of nodes in a scenario tree, each of which serves a different purpose. When developing a scenario tree model, it is important to make sure you are clear what type each node is. The three main types of nodes are infection nodes, detection nodes and category nodes.

### Infection node

---

**The probability for the infected branch of an infection node is always the design prevalence ( $P^*$ )**

An infection node represents the question “Is the animal or group of animals infected?” Infection nodes always have two branches: *INFECTED* and *NOT INFECTED*. Remember that analysis of surveillance for freedom from infection is based on estimating the surveillance system sensitivity, which is the probability of detecting disease if the disease is present at a defined level. An infection node defines the level of disease that is present. The probability associated with the *INFECTED* branch of an infection node is the design prevalence ( $P^*$ ).

Scenario trees must always have at least one infection node. Often, for surveillance in large populations, there will be two (or more) levels of design prevalence specified to take clustering of infection into account (see Clustering of infection on page 54). In this case there will be more than one infection node (one for each level of clustering).

The probability for an infection node at a given level never changes. For example, in Figure 5, there are two infection nodes (one for *YOUNG* and one for *OLD* animals). The value for the infected branch for both is the same – the design prevalence.

If a tree describes national surveillance for a disease that clusters, there are likely to be two infection nodes, one at the herd level and one at the animal level. The *INFECTED* branches for all the herd-level nodes will all have the same probability,  $P^*_{H}$ , the herd-level design prevalence (which, for example, may be 1%). The value for the branches for the animal-level nodes will be the animal-level design prevalence,  $P^*_{A}$  (which, for example, could be 20% for a highly infectious disease).

However, like other nodes, infection nodes are conditional on the previous branches in the limb. The probability that an animal is infected if the herd is not infected is equal to zero. Where there are more than one infection nodes, the probability for the *INFECTED* branch will be zero if any of the previous infection node branches were *NOT INFECTED*.

### Detection node

---

A detection node describes the probability that an animal will be detected as being infected. Each scenario tree must have at least one detection node, but some have many. The last node in a tree is always a detection node. Detection



**Laboratory surveillance detection systems**

nodes always have two branches: *DETECTED* (the ‘yes’ answer to the question), or *NOT DETECTED* (the ‘no’ answer).

The sensitivity of a surveillance system is the probability that infection will be detected if it is present at a defined level. If there are no detection nodes, there is no way to describe how the disease can be detected.

Typically, scenario trees can have two different detection structures based on laboratory or clinical surveillance.

For surveillance based on laboratory testing (e.g. a structured survey where sampled animals are all tested with a specified test), the detection node corresponds to the test used.

**Example**

For instance, if an ELISA is used to detect antibodies, then:

- the detection node would be **ELISA RESULT**,
- the question would be “Is the sample positive to the ELISA test or not?”
- the branches would be “Yes (*POSITIVE*)” and “No (*NEGATIVE*)”
- for animals that are infected, the *POSITIVE* branch probability would be the sensitivity of the ELISA test
- for animals that are not infected, the *NEGATIVE* branch probability would be the specificity of the ELISA test

**Clinical surveillance**

Where follow-up tests are used, these can be added as a series of further detection nodes, one for each test.

For clinical surveillance, the detection nodes describe steps in the detection process. A typical example of the steps that have to occur for clinical detection of an infected animal are:

- Infected animal shows clinical signs
- Owner notices clinical signs
- Owner contacts veterinarian
- Veterinarian examines animal
- Veterinarian takes appropriate samples
- Samples tested for the disease

Each of these steps is represented by a separate detection node, and has a probability of occurring. This is then normally followed by one or more laboratory tests to detect the infection and possibly confirm the diagnosis.

**Category node**

**Category nodes are used to describe factors influencing infection or detection.**

Infection and detection nodes describe the probability of being infected and of being detected. However, on their own, they are not able to describe differences between sub-populations. Category nodes are used to divide the population into groups according to relevant factors. Category nodes allow us to explicitly account for the impact of a wide range of factors in our surveillance.

There are three types of category nodes: risk category nodes (which influence the risk of infection), detection category nodes (which influence the probability of detection), and group category nodes (which are used to describe the coverage of the surveillance).

While every scenario tree must contain at least one infection node and one detection node, category nodes are optional. However, without at least one category node, the tree is, in effect, assuming that all animals in the population

**Category nodes have two or more branches**

have equal probability of being infected and detected. The results of analysis of the scenario tree will therefore be identical to a simple analysis assuming representative sampling. The aim of scenario trees is to take into account differences in sub-populations, and category nodes are the way to achieve that.

Category nodes must have at least two branches, but they may have more than two, depending on the nature of categories being considered. For instance, the risk of infection may vary geographically. If there are seven geographical regions, each with a different risk, there would be seven branches to **REGION** category node.

The probability associated with the branches of a category node represents the proportion in each of the groups. For instance, if 20% of animals are young, then the *YOUNG* branch of the **AGE** category node would have a probability of 20%, and the *OLD* branch would have 80%.

Two different proportions are used, referring to two different populations. The first is the population proportion (*PrP*) that uses the entire population as the reference (conditional on earlier nodes). The second is the surveillance proportion (*PrSSC*) that only uses those animals included in the surveillance system component as the reference. The use of population and surveillance system component proportions is explained in Chapter 9.

A simple way to remember the difference between the node types is to consider the values used for probabilities with the node branches:

- Infection nodes: design prevalence
- Detection nodes: sensitivity
- Category nodes: proportions

### **Risk category nodes**

**Risk category nodes: factors influencing risk of infection**

A risk category node is used to describe the effect of a risk factor for infection. In Figure 5, **AGE** is a risk category node, as it describes a risk factor that influences the probability of infection. Young animals are more susceptible to infection than older animals.

Risk category nodes are more complex than the other types of nodes as they need to describe the difference in risk of infection between the two categories. The way in which this is achieved is discussed in detail in Chapter 9.

### **Detection category nodes**

**Detection category nodes: factors influencing probability of detection**

The sensitivity of a test or other component of a detection system may vary. For instance, some tests are more sensitive in earlier stages of infection than later stages; the probability that an animal will show clinical signs may depend on the serotype infecting the animal. These factors may not influence the probability of infection, but they can influence the probability of detection. Detection nodes allow these factors to be included in the scenario tree.

### **Group category nodes**

**Group category nodes: factors to describe coverage**

Group category nodes describe factors that have no direct impact on the probability of infection or the probability of detection. They therefore do not influence the results of the scenario tree. The only reason for including these nodes is to allow the scenario tree to be analysed separately for different populations of interest.

For example, there may be no geographical difference in the risk of infection. However, it may be useful to include a group category node for region. This will

allow the surveillance system sensitivity to be analysed region by region (instead of just a single summary measure for the whole country), allowing the quality of surveillance to be compared between regions.

## Building a scenario tree

---

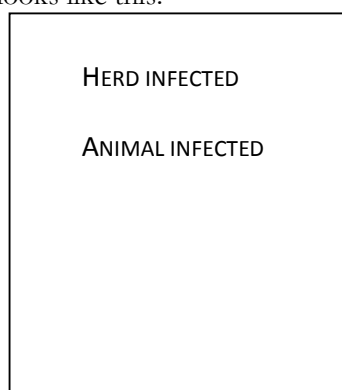
A scenario tree describes:

- The risk that an animal or group of animals might be infected (based on the risk factors for infection and the structure of the population); and
- The way in which an infected animal may be detected, based on the structure of the surveillance system.

The best way to build a scenario tree is to consider these two areas separately. Let us use as an example a simple scenario tree to describe surveillance for bovine brucellosis. Rather than drawing a tree as shown in Figure 4 and Figure 5, it is usually much simpler to just write a list of nodes.

### Step 1: infection nodes

Start with the infection nodes, as every tree must have at least one infection node. Surveillance systems for diseases that cluster normally have two. Because brucellosis clusters at the herd level we will start with two infection nodes, so our list looks like this:



Note that we start from the largest unit (herd) and progress down to the smallest (animal).

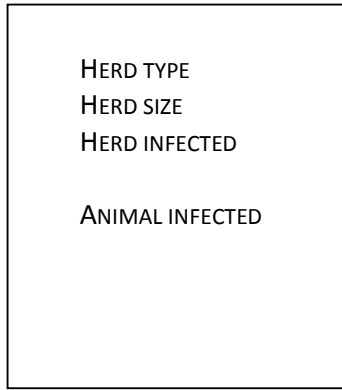
### Step 2: risk factors at the herd level

The next step is to identify all the possible risk factors (factors that influence the probability of infection). This can be done in two parts: factors operating at the herd level, and factors operating at the animal level.

If brucellosis is present in the population, there may be factors that mean that one herd is more likely than another to become infected. For instance, brucellosis is often more common in dairy herds than beef herds, so type of herd should be included as a risk factor. Small herds buy in fewer animals, so may have a low risk of being infected than large herds, so herd size could be included as a risk factor.

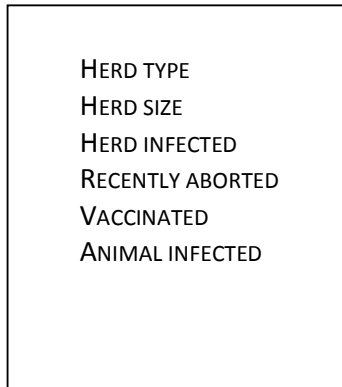
Risk factors are added to our list *before* the infection node that they influence. These two risk factors are therefore added *above* the herd infection node.

Our list now looks like:



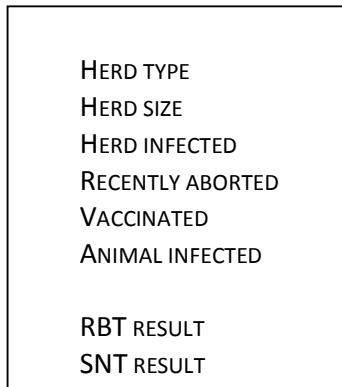
**Step 3:  
Animal-level risk  
factors**

The next step is to list those risk factors operating at the animal-level. This could include recently aborted cows. Vaccination is sometimes used, and this makes it less likely that an animal will be infected (but is still not completely protective). These factors can be included in our list before the animal infection node:



**Step 4:  
Detection nodes**

We now have a description of the risk structure of the population. Next we need to describe how our surveillance system detects disease. For this example, we are doing a targeted survey. Specimens are collected from selected animals, and tested using first the Rose Bengal Test (RBT). This is then confirmed with the serum neutralisation test (SNT). Animals are considered positive if they test positive to both these tests. There are therefore two detection nodes:



**Step 5:  
Detection  
category nodes**

The next step is to identify any factors that influence the probability of detection. Our surveillance is being conducted in a country where transport is

difficult, and samples are tested at the central laboratory. Lengthy transport may result in poor sample quality, decreasing the sensitivity of the test. We could therefore divide the country into areas (*REMOTE* and *NOT REMOTE*) as an indicator of sample quality, and use different sensitivity values for our tests (lower sensitivity for remote areas). Our list now looks like this:

- HERD TYPE
- HERD SIZE
- HERD INFECTED
- RECENTLY ABORTED
- VACCINATED
- ANIMAL INFECTED
- AREA
- RBT RESULT
- SNT RESULT

**Step 6:**  
Group category nodes

The final step is to consider if we want to add a group category node, allowing us to analyse the results of the tree for different parts of the population. For instance we could add a *REGION* node so we have surveillance sensitivity estimates for each different region in the country. Group category nodes, if used, are normally added to the top of the tree.

- REGION
- HERD TYPE
- HERD SIZE
- HERD INFECTED
- RECENTLY ABORTED
- VACCINATED
- ANIMAL INFECTED
- AREA
- RBT RESULT
- SNT RESULT

**Step 7:**  
Node types and branches

It is a good idea at this stage to review the list. We may have forgotten some factors, or included some factors which, on consideration, are unlikely to play a significant role. For each node, you should then identify the type of node, and specify the branches, as shown below.

Node	Type	Branches
REGION	Group category	<i>REGION 1, 2, 3, ETC.</i>
HERD TYPE	Risk category	<i>BEEF, DAIRY</i>
HERD SIZE	Risk category	<i>SMALL, MEDIUM LARGE</i>
HERD INFECTED	Infection	<i>INFECTED, NOT INFECTED</i>

RECENTLY ABORTED	Risk category	<i>ABORTED, NOT ABORTED</i>
VACCINATED	Risk category	<i>VACCINATED, NOT VACCINATED</i>
ANIMAL INFECTED	Infection	<i>INFECTED, NOT INFECTED</i>
AREA	Detection category	<i>REMOTE, NOT REMOTE</i>
RBT RESULT	Detection	<i>POSITIVE, NEGATIVE</i>
SNT RESULT	Detection	<i>POSITIVE, NEGATIVE</i>

**Step 8:  
Branch  
probabilities**

Once the nodes and branches are defined, the next task is to calculate or estimate branch probabilities and other model parameters. This is discussed in Chapter 11.

**Step 9:  
Implement the  
tree**

The scenario tree then needs to be put into a format that allows it to be analysed. There are several options for doing this: on paper (for very simple trees); using a spreadsheet (suitable for simple and complex trees, but requires a great deal of time and care); or using specially designed software for scenario tree analysis (much simpler and suitable for simple to moderately complex trees – highly complex trees may need the flexibility of a spreadsheet).

**Step 10:  
Analysis**

Analysing the tree involves using the AND and OR probability rules. The probability of an animal being in each of the sub-populations is calculated by multiplying the probabilities for each of the branches in the limb. The probabilities for each limb that gives a positive outcome (infection detected) are added. This gives the *unit sensitivity (USE)* which is the average probability that a single animal that passes through the surveillance system will give a positive result.

In risk category nodes, there are special rules for handling the multiplication of probabilities. These are explained in Chapter 9. This book assumes that most analysis will be done using the specialised software, which simplifies these calculations. The software is described in Chapter 17.

## Tree-building rules

The example above described the general process of building a scenario tree. The following rules may be useful when building your own trees:

- Infection and detection nodes have two branches
- Category nodes have two or more branches
- A tree should be symmetrical – the nodes encountered are the same along every limb of the tree. This is not absolutely essential, but makes calculation easier.
- There must be at least one infection node. There are often two, rarely more than two.
- There must be at least one detection node, but there are often more.
- The last node in the tree must be a detection node.
- Risk category nodes must come before the infection node that they are influencing.

- Detection category nodes must come before the detection node that they are influencing.
- For each node, the probabilities of all branches must add up to one.
- Probabilities are conditional on all previous branches in the limb.

## Node order

---

The general order for nodes in a scenario tree is:

1. group category node (if required)
2. nodes relating to infection
  - a. zero or more risk category nodes describing risk factors operating at the herd level
  - b. zero or one herd infection node
  - c. zero or more risk category nodes describing risk factors operating at the animal level
  - d. zero or one animal infection node (as some trees may stop at the herd level)
3. nodes relating to detection
  - a. zero or more detection category nodes
  - b. one or more detection nodes describing the surveillance system

The order of risk category nodes when there are multiple nodes relating to the same infection node is not important. However, as the lower nodes are conditional on the higher ones, these conditional probabilities are often easier to estimate with one node order compared to another.

### Example

*SEX (MALE/FEMALE)* and *REGION (1/2/3/4)* are risk category nodes in a scenario tree. It is possible to include them in either order.

If *SEX* comes before *REGION* the probability for *MALE* branch of the *SEX* node is the proportion of males in the population. The probability for the *REGION 1* branch of the *REGION* node is the proportion of all males that are in region 1. While there is clearly a correct value for this proportion, it is not intuitively easy to grasp and may be difficult to calculate.

If *REGION* comes before *SEX*, the *REGION* proportions will be the proportion of all animals in each region. The proportion for the *MALE* branch of the *SEX* node under *REGION 1*, will now be the proportion of males in Region 1. This is conceptually easier to understand and the figures are likely to be more readily available in this form.

# Chapter 9 – Incorporating Risk into a Scenario Tree

There are sadistic scientists who hurry to hunt down errors instead of establishing the truth.

Marie Curie (1867 – 1934)

The purpose of a scenario tree is to describe how different parts of the population have different probabilities of being infected and being detected. Scenario trees allow us to analyse risk-based surveillance, targeted at groups that are more likely to be infected.

The previous chapter provided an overview of building a scenario tree and introduced the different types of nodes. This chapter focuses on the use of risk category nodes to incorporate risk into a scenario tree.

In common usage, ‘risk’ is defined as the likelihood of an adverse event occurring. However, in risk analysis, risk is a combination of both the likelihood and the consequences of an adverse event. In scenario-tree modelling, the term risk is used to describe only the likelihood of an event.

## Quantifying targeting in risk-based surveillance

---

Risk-based surveillance is an approach to disease surveillance that involves looking for disease where it is most likely to be present. Instead of representative surveillance (where we assume we know nothing about the risk of different sub-populations), we use our understanding of the disease to determine those animals that are most likely to be infected and concentrate our surveillance effort there. This is clearly more efficient – by examining the high risk groups you have a greater chance of finding the disease (if it is present) than by examining animals that are at lower risk.



To capture the benefit of this type of surveillance in the scenario tree, we need to understand exactly what we are doing. First, we are talking about identifying groups of animals that are at higher risk of disease. To do this, we need to answer the questions:

- Which animals? How do we define the group?
- How do these animals differ from the rest of the population? What is the difference in the risk?

Secondly, we need to understand how our surveillance is targeting this risk group. We need to be able to determine if animals in the risk group have a higher probability of being included in our surveillance than other animals in the population.

### *Describing differences in risk*

---

Consider a risk factor for bovine tuberculosis – the presence of infected wildlife in the area. To use this risk factor in a scenario tree, we need two things:

- 1) To define the high and low risk populations
- 2) To describe the differences in risk between them.

The first has almost been done – our high risk population consists of those farms that are in an area where wildlife are infected with tuberculosis. To use this definition, we would need maps of the distribution of infected wildlife, and maps of the cattle farms. Using these maps we could divide the population of cattle farms into those in the areas with infected wildlife and those in areas where there are no infected wildlife.

Quantifying the difference in risk is done using the relative risk (sometimes called the risk ratio, and abbreviated as *RR*). The relative risk describes the risk of being infected in one group, relative to the risk of being infected in the other group. Typically, this can be estimated by observational studies that measure the prevalence or incidence of disease.

**Relative risk describes how different parts of the population have different risk**

#### **Example**

A study of herd infection rates has been conducted in a country infected with bovine tuberculosis. The study found that the incidence of new herd breakdowns in the areas with wildlife was 3.6 breakdowns per 100 herds per year. In the areas without wildlife infection, the incidence was 1.2 breakdowns per 100 herds per year. The relative risk is the incidence in the high risk group relative to (or divided by) the incidence in the low risk group:

$$\begin{aligned} RR &= \frac{3.6}{1.2} \\ &= 3 \end{aligned}$$

This can be interpreted as meaning that herds in the high risk area are three times more likely to become infected than herds in the low risk area.

Relative risk is a ratio, and it can range between 0 and infinity. If the relative risk equals one, it means that the risk in the two populations is the same (i.e. the ‘risk factor’ is actually having no effect).

We use the relative risk to describe the difference in risk between different parts of the population, and to identify our high risk groups.

We can calculate a relative risk for risk factors that have more than two groups.

### Example

We want to divide the areas with infected wildlife into heavily and lightly infected areas which means we will have three risk groups: no wildlife infection, low levels of wildlife infection and high levels of wildlife infection. The incidence of herd breakdowns has been calculated for each of these areas:

No wildlife infection: 1.2 breakdowns per 100 herds per year

Low levels of wildlife infection: 2.7 breakdowns per 100 herds per year

High levels of wildlife infection: 4.3 breakdowns per 100 herds per year

To calculate the relative risk, we compare each of the risk groups to the group with the lowest risk (no wildlife infection). This gives relative risks of:

$$RR_{\text{no infection}} = \frac{1.2}{1.2} = 1$$

$$RR_{\text{low infection}} = \frac{2.7}{1.2} = 2.25$$

$$RR_{\text{high infection}} = \frac{4.3}{1.2} = 3.58$$

In many cases, suitable studies measuring the different incidence or prevalence of disease associated with the risk factor are not available. In these cases, it is necessary to estimate the relative risk, as discussed in Chapter 11.

### Putting relative risk into the tree

Now we know the different relative risks for the different groups in the population, we can use this in our risk category node. Consider a small part of a scenario tree:

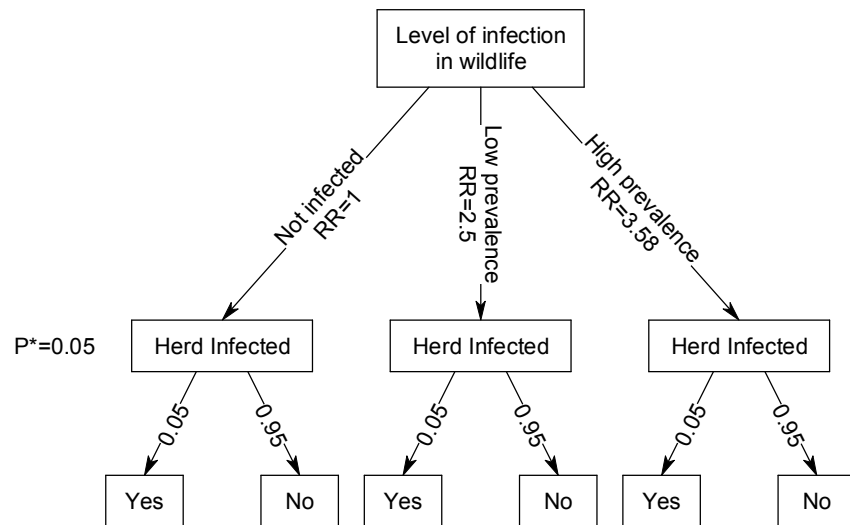


Figure 6: Example of a small section of a scenario tree showing a risk category node and three infection nodes

When analysing a scenario tree, we multiply the figures down each of the branches to determine the probability that an animal will be in a specific group. In this example the probability that a herd will be infected if it is in an area with no wildlife infection is 1 (the *RR*) times 0.05 (the design prevalence) which is 0.05.

However if the herd is in a high prevalence area, the probability is 3.58 times 0.05 which is 0.18.

In effect, by using the relative risks, we have changed the design prevalence (or probability that a herd or animal will be infected), and made it higher for our high risk groups. In this example, we are now saying that the chance of a herd being infected in the high risk area is 18% whereas it is only 5% in a low risk area. If the level of infection is that high, it will be much easier to find disease with our surveillance, so our surveillance will be much more sensitive in the high risk area.

This is the result that we want. The relative risk is used to adjust the standard design prevalence to show how much more likely it is that high risk herds or animals would be infected.

Unfortunately, there is a problem with this approach. The design prevalence sets the standard for our surveillance and is a measure of the average probability that an animal or herd will be infected. In this example, we said that the design prevalence is 0.05, or that on average, herds have a 5% probability of being infected. After adjusting for the risk of the different herds, the average design prevalence (assuming that each type of herd is equally represented in the population) is  $(0.05 + 0.11 + 0.18)/3 = 0.11$  or 11%. Using relative risk in this way has meant that the average probability of infection for the entire population has increased. We are no longer using our standard design prevalence.

While we have described the difference in risk, the values we use for risk in the model are changing our average design prevalence. These numbers need to be adjusted so that the average risk does not change. The approach to adjusting the numbers involves understanding population proportions and targeting. The relative risk describes the difference in the risk of the different populations, but we have not yet described how our surveillance is targeting these populations.

## Describing targeting

Risk-based surveillance targets high risk populations. Targeting means that some herds or animals are selected based on their risk. To understand targeting, it is necessary to compare it to representative sampling. In representative sampling (for instance, using random sampling), the proportions in the sample are the same as the proportions in the population. If 20% of the population has a high risk of being infected, a representative sample would have 20% of animals in the high risk group. In contrast, targeted sampling would have more than 20% of the sample from the high risk group.

To describe targeting, it is therefore necessary to compare two proportions: first to identify what proportion of the entire population is in the risk group, and then to identify what proportion of the sample is in the risk group. The difference between these two proportions describes the level of targeting.

Targeting is described by comparing the population proportion to the surveillance proportion

### Example

Ten percent of cattle herds are in areas with high levels of wildlife tuberculosis infection (our *population proportion* or *PrP* is 10%).

If 10% of animals in our surveillance (our *surveillance system component proportion* or *PrSSC*) are from the high risk group, there is no targeting.

If 50% of animals in our surveillance are from the high risk group, it is clear that we have concentrated our surveillance in this area and we employing targeted surveillance. This approach will give us a higher chance of detecting the disease than representative sampling.

If only 5% of animals in our surveillance come from the high risk group, we have under-represented this part of the population. This is called *negative targeting* or biased surveillance. Using this approach we are less likely to find disease than if we used representative sampling.

## Implementing risk in a scenario tree

Branches for most nodes have just one number associated with them – a probability or proportion for the surveillance system component. Out of all the farms or animals that are included in the surveillance system component, this is the proportion that fall into each branch.

However, it is now clear that for a full understanding of risk and targeting, the branches for our risk category nodes in the scenario tree require three pieces of information:

1. the relative risk for that branch (*RR*)
2. the proportion of the *population* in that branch (*PrP*)
3. the proportion of the *surveillance component* in that branch (*PrSSC*)

How are these three numbers used when calculating a scenario tree? This is explained in detail in the next section, but in brief:

1. The *PrSSC* is used in the same way as the proportions in other nodes.
2. The population proportion and relative risk are used to adjust the design prevalence that the risk category node refers to.
  - a. The population proportion and relative risk are combined into an *adjusted risk* that ensures that the average design prevalence is constant across the population
  - b. The adjusted risk is multiplied by the design prevalence to calculate the *effective probability of infection (EPI)*, which takes the place of the design prevalence.

## What you need to know

If you are implementing a scenario tree using the software described in Chapter 17, you can skip the next section as you already know everything you need to know: to capture the effective of risk-based surveillance in a scenario tree, you need to provide three different numbers for each branch of a risk category node:

- The relative risk of herds or animals in that group being infected (relative to the group with the lowest risk),
- The proportion of the population in that group, and
- The proportion of the surveillance system component in that group.

If you are implementing a scenario tree using a spreadsheet, or if you just want to understand how the tree works, then you should read the following section.

## Calculation of adjusted risk

This section is not required if you intend to use the freedom software

A relative risk is a ratio of the incidence or prevalence of infection in one group compared to another. If the prevalence in one group is 20% and in the other group is 10%, then the relative risk is 20/10 or 2:1. This ratio can be expressed in different ways and still have the same value:

$$20/10 = 2/1 = 4/2 = 1/0.5$$

The adjusted risk for a branch is a measure of the risk in that branch. The ratio of adjusted risks has the same value as the relative risk, but it adjusted risks expressed in different figures. This adjustment doesn't change the measure of risk, but is done so that the weighted average probability of infection across all groups remains the same as the design prevalence.

## The constraints

---

To calculate the adjusted risk, we must consider two constraints.

**Constraint 1:** The ratio of the adjusted risks must remain the same as the original relative risk.

This can be expressed by the formula:

$$\frac{AR_1}{AR_2} = \frac{R_1}{R_2}$$

Where:

$AR$  is the adjusted risk,

$R$  is the risk, and

$\frac{R_1}{R_2}$  is the relative risk

This means that if the relative risk, as shown above, is 2:1, then the ratio of the adjusted risks must be in the same ratio (e.g. 4/2, 6/3, 1.5/0.75)

**Constraint 2:** The average risk across the population is, by definition, equal to one. This average risk must not change.

As a formula this can be expressed as:

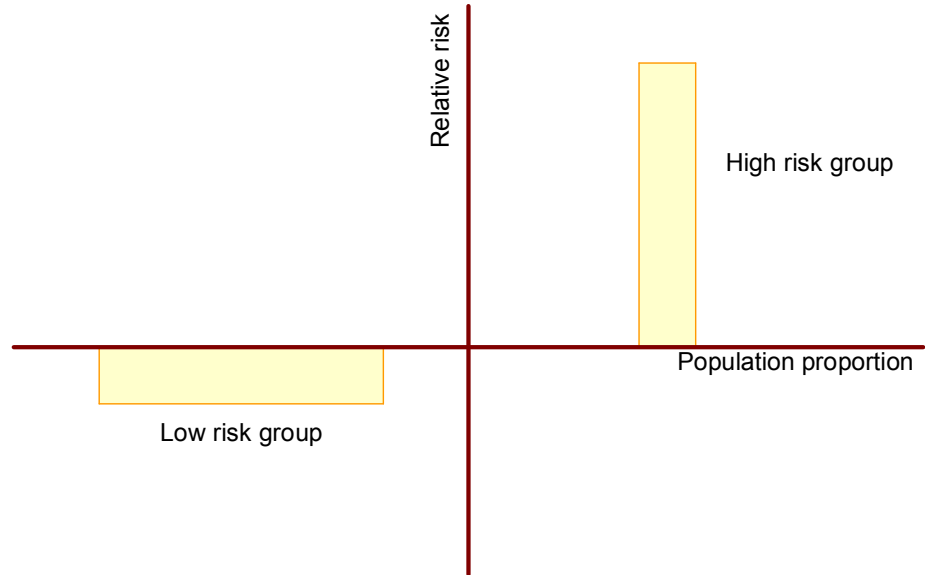
$$(AR_1 \times PrP_1) + (AR_2 \times PrP_2) = 1$$

Or more generally (when there are more than two risk groups):

$$\sum_{i=1}^I (AR_i \times PrP_i) = 1$$

where  $PrP$  is the population proportion.

The objective is to find the numbers for the adjusted risk that meet both these constraints. The solution is presented graphically below. The x-axis shows the population proportion and cuts the y-axis (relative risk) at the value of 1.



The correct values for adjusted risk occur when the area of the two rectangles are equal. In this example, there is a large proportion of animals in the low risk group. The risk in this group must be lower than one, to balance the risk in the other group, but as there are many animals in this group, the adjusted risk can be just a little below one. In the high risk group, there are very few animals. In order to balance the many animals in the low risk group, the adjusted risk has to be much greater than one.

### The solution

Mathematically, the solution is found using simultaneous equations, to solve for  $AR_2$  as shown. Constraint 1 above gives us:

$$\frac{AR_1}{AR_2} = \frac{R_1}{R_2}$$

Rearranging, we can equate this to  $AR_1$ .

$$AR_1 = \frac{AR_2 \times R_1}{R_2}$$

Our second constraint is:

$$(AR_1 \times PrP_1) + (AR_2 \times PrP_2) = 1$$

If we replace  $AR_1$  in this formula with the expression for  $AR_1$  from the previous formula, we get:

$$\left( \frac{AR_2 \times R_1}{R_2} \times PrP_1 \right) + (AR_2 \times PrP_2) = 1$$

To solve this for  $AR_2$ , we first extract  $AR_2$  from the terms on the left:

$$AR_2 \left[ \left( \frac{R_1}{R_2} \times PrP_1 \right) + PrP_2 \right] = 1$$

Reorganise the complex middle term to use the same denominator by multiplying  $PrP_2$  by  $R_2/R_2$ :

$$AR_2 \left[ \left( \frac{R_1 \times PrP_1}{R_2} \right) + \left( \frac{R_2 \times PrP_2}{R_2} \right) \right] = 1$$

Now divide both sides by the middle term and further simplify:

$$AR_2 = \frac{R_2}{(R_1 \times PrP_1) + (R_2 \times PrP_2)}$$

The value of  $R_2$  is normally 1 as this is the low risk (or reference) category. The formula can therefore be simplified to:

$$AR_2 = \frac{1}{(R_1 \times PrP_1) + PrP_2}$$

Finally, once  $AR_2$  is known,  $AR_1$  can be calculated:

$$AR_1 = AR_2 \times \frac{R_1}{R_2}$$

Or, if  $R_2$  is equal to 1,

$$AR_1 = AR_2 \times R_1$$

### Example

Ten percent of farms are from an area with a high prevalence of tuberculosis in wildlife (the population proportion for the high risk group,  $PrP_1$ ), and 90% ( $PrP_2$ ) are in areas that have no infection. The risk in the high prevalence area is three times greater than the area with no infection ( $R_1=3$ ,  $R_2=1$ ).

The adjusted risk for the low risk group is:

$$\begin{aligned} AR_2 &= \frac{1}{(R_1 \times PrP_1) + PrP_2} \\ &= \frac{1}{(3 \times 0.1) + 0.9} \\ &= \frac{1}{1.2} \\ &= 0.83 \end{aligned}$$

The adjusted risk in the high risk group is:

$$AR_1 = AR_2 \times R_1$$

$$AR_1 = \frac{1}{1.2} \times 3$$

$$= 2.5$$

Note how, in order to average to one, the adjusted risk in the low risk group is always less than one and the adjusted risk in the high risk group is always more than one.



# Chapter 10 Calculating sensitivity with a scenario tree

Things should be made as simple as possible, but not any simpler.

Albert Einstein (1879 – 1955)

At this point, we have seen an example of how to build a scenario tree, and how to capture risk within the tree. We are now able to use the tree to calculate the sensitivity of our surveillance system component. This is done in two stages:

- Calculate the average sensitivity of the surveillance system component for a *single animal*, or the component unit sensitivity (*CSeU*)
- Estimate the sensitivity for the entire surveillance system component (*CSe*) based on the total number of animals that are processed by the system.

## Calculation of unit sensitivity

---

Let us consider a simple example of a tree with one risk factor (*age*), one infection node and one detection node (an ELISA test). This may describe a surveillance system that involves collection of blood from animals, where age is a known risk factor (young animals at higher risk), and where most of the blood samples come from younger animals.

### *Building the scenario tree*

---

In this case, the structure of the scenario tree is very simple with just the three nodes.

Node	Type	Branches
AGE	Risk category	<i>YOUNG, OLD</i>
ANIMAL INFECTED	Infection	<i>INFECTED, NOT INFECTED</i>
ELISA RESULT	Detection	<i>POSITIVE, NEGATIVE</i>

### Organising the model parameters

It helps at this stage to carefully organise the parameters that will be used in the model. The first parameters will be those for the AGE risk category node. The risk in young animals is five times greater than in older animals. Young animals make up 20% of the population but 90% of the animals under surveillance. These figures are summarised below and the adjusted risk is calculated using the formulae shown in Chapter 9.

Risk factor	Branches	Relative risk	Population proportion	Surveillance proportion	Adjusted risk
AGE	<i>YOUNG</i>	5	0.2	0.9	2.78
	<i>OLD</i>	1	0.8	0.1	0.56

The infection and detection nodes require parameters as well. The design prevalence is 0.05 and the sensitivity of the ELISA is 95%.

Node	Branches	Probability Type	Value
INFECTION	<i>INFECTED</i>	Design prevalence	0.05
	<i>NOT INFECTED</i>		0.95
ELISA RESULT	<i>POSITIVE</i>	Sensitivity	0.95
	<i>NEGATIVE</i>		0.05

### Drawing the tree and adding parameters

In the examples we have used so far, we have drawn the tree starting from the top, with the branches leading down. This is useful to illustrate how the process works. However, calculation of the tree requires operations on many different numbers, and a spreadsheet is useful to perform these operations. The downward branching structure of the tree is difficult to represent in a spreadsheet, so it is more common to draw the tree starting from the left and moving across the page. An example of how our simple tree can be represented in a spreadsheet

is shown below. The columns represent nodes, and each row represents one possible limb (path through the tree).

Age		Infected		ELISA Result		
Branch	PrSSC	Branch	EPI	Branch	Se	Outcome
Old		Yes		Pos		Pos
				Neg		Neg
		No		Pos		Pos
				Neg		Neg
Young		Yes		Pos		Pos
				Neg		Neg
		No		Pos		Pos
				Neg		Neg

Once the tree structure is prepared, the parameter values can be entered. This is easier if the parameters have been well organised in the same spreadsheet as shown above.

For the AGE node, *PrSSC* represents the proportion in each group in the surveillance system.

For the INFECTED node, *EPI* represents the *effective probability of infection*. This is the adjusted risk (from the table above) multiplied by the design prevalence (0.05).

For the ELISA RESULT node, *Se* represents the sensitivity. Note that the sensitivity (probability of getting a positive test result) is zero where the animal is not infected (the *NO* branch of the infection node), as we are assuming perfect specificity.

The tree with the required parameters is shown below.

Age		Infected		ELISA Result		
Branch	PrSSC	Branch	EPI	Branch	Se	Outcome
Young	0.9	Yes	0.139	Pos	0.95	Pos
				Neg	0.05	Neg
		No	0.861	Pos	0	Pos
				Neg	1	Neg
Old	0.1	Yes	0.028	Pos	0.95	Pos
				Neg	0.05	Neg
		No	0.972	Pos	0	Pos
				Neg	1	Neg

### Calculating the tree

We are now ready to calculate the probability of each branch. This is done by multiplying each of the values across the table for each branch. For instance the probability for the first branch (young, infected, positive test result) is  $0.9 \times 0.139 \times 0.95 = 0.119$ , as shown below.

Age		Infected		ELISA result			
Branch	PrSSC	Branch	EPI	Branch	Se	Outcome	Probability
Young	0.9	Yes	0.139	Pos	0.95	Pos	0.119
				Neg	0.05	Neg	
		No	0.861	Pos	0	Pos	
				Neg	1	Neg	

Old	0.1	Yes	0.028	Pos	0.95	Pos	
				Neg	0.05	Neg	
		No	0.972	Pos	0	Pos	
				Neg	1	Neg	

The probability for the fourth branch would be  $0.9 \times 0.861 \times 1 = 0.775$ .

Age		Infected		ELISA result		Outcome	Probability
Branch	PrSSC	Branch	EPI	Branch	Se		
Young	0.9	Yes	0.139	Pos	0.95	Pos	0.119
				Neg	0.05	Neg	0.006
		No	0.861	Pos	0	Pos	0.000
				Neg	1	Neg	0.775
Old	0.1	Yes	0.028	Pos	0.95	Pos	
				Neg	0.05	Neg	
		No	0.972	Pos	0	Pos	
				Neg	1	Neg	

The completed calculations are shown below.

Age		Infected		ELISA result		Outcome	Probability
Branch	PrSSC	Branch	EPI	Branch	Se		
Young	0.9	Yes	0.139	Pos	0.95	Pos	0.119
				Neg	0.05	Neg	0.006
		No	0.861	Pos	0	Pos	0.000
				Neg	1	Neg	0.775
Old	0.1	Yes	0.028	Pos	0.95	Pos	0.003
				Neg	0.05	Neg	0.000
		No	0.972	Pos	0	Pos	0.000
				Neg	1	Neg	0.097

Before going any further it is always a good idea to check for errors. The tree represents the possible outcomes for a single animal in the surveillance system. Adding up the probabilities of all the possible outcomes (the last column) should always come to one.

### Calculating the component unit sensitivity (CSeU)

The component unit sensitivity is the probability that a single unit (animal) passing through the surveillance system would be detected. This means that we are only interested in those animals that have a positive outcome (if the outcome is negative, the animal has not been detected).

The component unit sensitivity is therefore the sum of the probabilities of all the different limbs that can lead to a detection. These are highlighted below.

Age		Infected		ELISA result		Outcome	Probability
Branch	PrSSC	Branch	EPI	Branch	Se		
Young	0.9	Yes	0.139	Pos	0.95	Pos	0.119
				Neg	0.05	Neg	0.006
		No	0.861	Pos	0	Pos	0.000
				Neg	1	Neg	0.775
Old	0.1	Yes	0.028	Pos	0.95	Pos	0.003
				Neg	0.05	Neg	0.000

	No	0.972	Pos	0	Pos	0.000
			Neg	1	Neg	0.097

Only two limbs have a non-zero probability, so the component unit sensitivity is  $0.119 + 0.003 = 0.122$ . This means that the probability that the surveillance system will detect disease by examining one animal is (on average) 12.2%.

### Comparison with representative sampling

It is interesting to note at this point what the effect of our risk based surveillance has been on the sensitivity. If we had used representative sampling, so the average probability of infection was the same for all animals, the probability of getting a positive result from one animal (the unit sensitivity) would be:

$$\begin{aligned}
 CSeU &= P^* \times Se \\
 &= 0.05 \times 0.95 \\
 &= 0.0475
 \end{aligned}$$

By focusing our sampling on young animals (the group with the highest risk) we have almost tripled the unit sensitivity from 4.75% to 12.2%. This increase in sensitivity is the reason we use risk based surveillance.

### Calculation of component sensitivity (CSe)

The component unit sensitivity is interesting, but it is not the value we really want. We want to know how good our surveillance is at finding disease, not by examining a single animal, but for the whole surveillance system component, which is examining many animals.

This step is simple, as we can use a formula that we saw earlier. The CSe (surveillance system component sensitivity) is the probability that all animals in the surveillance system do not give a negative result:

$$CSe = 1 - (1 - CSeU)^n$$

If our surveillance system component (sampling blood from mostly young animals) involved the collection of 20 blood samples then:

$$\begin{aligned}
 CSe &= 1 - (1 - 0.122)^{20} \\
 &= 0.925
 \end{aligned}$$

We can compare the result of 92.5% to the sensitivity we would have had if we had used representative rather than risk-based sampling: 62.2%. Clearly targeting the high risk animals has provided a very big advantage in this case.

### What next?

This chapter has presented an example of a very simple scenario tree, and used it to calculate the sensitivity of a component of a surveillance system. This is the first complete example of the use of scenario trees presented in this book, and shows that they can be a useful and relatively simple tool. However, the example

shown here needs to be refined in a number of ways in order to make it more useful:

- Complexity
  - Most surveillance systems are more complex than the example shown in this chapter, with many more risk factors, two levels of infection nodes, and multiple detection nodes. The same basic principles apply to complex trees, but they are harder to implement.
- Model parameters
  - Scenario trees use many parameters to describe risk and proportions. These figures are often difficult to find, and a solution must be found when you don't know what the real values are.
- Uncertainty and variability
  - If we are not sure about some of the inputs, we need a way to express our uncertainty in our estimate of sensitivity.
- Clustering
  - This simple example assumed that all animals were at equal risk of being infected and each animal contributed the same amount of evidence that the population was free from infection. However, when disease clusters, testing many animals from the same herd provides less chance of finding the infection than the same number of animals spread over a large number of herds.

The next chapters will examine these issues to make the sensitivity estimates from the scenario tree more accurate and reliable.

# Chapter 11 – Probability

## Estimates in a Scenario Tree

An expert is a person who has made all the mistakes that can be made in a very narrow field.

Niels Bohr (1885 – 1962)

Scenario trees require many numbers. Each branch has a probability associated with it, and some nodes (risk category notes) have three numbers for each branch. As we work along each limb in the tree, each probability is conditional on all the previous probabilities in that limb. Where do all these numbers come from and how do we make sure that they are correct? This chapter provides some guidance in finding the right parameters for a scenario tree.

### Summary of required values

---

The purpose of a scenario tree is to calculate the sensitivity of a component of a surveillance system. Sensitivity is the probability that the surveillance system component will detect at least one infected animal, if the population is infected at the design prevalence. The result of the scenario tree analysis (sensitivity) is a probability, so all the branch parameters are probabilities as well.

The probability values are different depending on the type of node.

#### Infection node

Infection node:  
Design prevalence

An infection node has two branches: *INFECTED* and *NOT INFECTED*. The probability associated with the *INFECTED* branch is the design prevalence ( $P^*$ ). Prevalence is defined as the proportion of the population with a defined characteristic (in this case, the proportion that is infected). It can be thought of as

a probability as well: if an animal is drawn from the population at random, the probability that it will be infected is equal to the prevalence.

The approach to selecting an appropriate design prevalence was discussed in the section on page 39. To summarise, the way to choose, in order of preference, is:

- Use global standards (e.g. from the OIE code),
- Use regional standards (e.g. EU regulations),
- Check the requirements of your trading partners,
- Calculate based on your trading partner's stated acceptable level of protection (ALOP) – this approach is rarely feasible.
- Determine based on the biology of the disease (e.g. the minimum expected prevalence if infection is established).
- Determine based on practical considerations (what level of surveillance is affordable).
- Make an arbitrary choice based on common values (1%, 5% or 10%).

### Detection node

Detection node:  
Sensitivity

A detection node has two branches, representing *DETECTED* and *NOT DETECTED*. The probability associated with the *DETECTED* branch is a sensitivity. This is evident when the detection node refers to something like a laboratory test (e.g. a complement fixation test, CFT). However, detection nodes are also often used to describe other complex components of a surveillance system, such as the probability that a farmer will call the veterinarian if they notice that an animal is sick. This may also be thought of as a sensitivity.

### Risk category node

Risk category  
node: relative  
risk, population  
proportion and  
surveillance  
system  
component  
proportion

A risk category node is the most complex type of node in the scenario tree, as discussed in Chapter 9, as each branch has three figures associated with it.

#### Relative risk

The relative risk describes how some parts of the population are at higher risk than others. This is the only figure used in a scenario tree that is not a probability. It is a ratio that can take a value from zero to infinity. When adjusted (using the population proportion) to create the *adjusted risk (AR)* it is multiplied by the design prevalence to give the *effective probability of infection (EPI)* which, again, is a probability.

#### Population proportion (PrP)

The population proportion is used to change the relative risk to the adjusted risk. It represents the proportion of the entire population that is in the branch category. This is important as it allows the tree to take into targeting into account.

#### Surveillance system component proportion (PrSSC)

This is the proportion of animals in the surveillance system that fall into the branch category. Targeting is expressed by the difference between the population proportion and the SSC proportion.

With some surveillance systems, the *PrP* and the *PrSSC* are the same.



### Example

Consider surveillance based on a survey using random representative sampling. Twenty percent of the population are in a high risk group. The representative sampling will ensure that 20% of the sample (the *PrSSC*) is also in the high risk group. Representative sampling does not target high risk groups, so scenario tree analysis will give the same result as simpler methods of analysis.

### Example

A passive farmer reporting system may have coverage of the entire population, as every animal is owned by a farmer, and therefore every animal has a chance of being reported and detected if it becomes infected.

In this case the *PrSSC* is the same as the entire population, so the *PrSSC* is the same as the population proportion. Surveillance systems that have complete coverage of the population therefore do not take different risk groups into account. However, they do take differences in the probability of detection into account, so scenario tree analysis is very useful for these situations.

## Detection category node

Detection category node: SSC proportion (and population proportion)

A detection category node is used to divide the population into groups that have different probabilities of being detected. The branches of a detection category node are associated with the proportions of the surveillance system component that fall into that category.

Often, we will also want to note the population proportion for detection category nodes, to determine how good the sensitivity of our surveillance is compared to representative (non risk-based) sampling.

## Group category node

Group category node: SSC proportion (and population proportion)

The group category node is similar to a detection category node, in that it uses the surveillance system component proportion, but will often have the population proportion recorded for comparison purposes.

In summary, the values that may be required for node branches include:

- Relative risk
- Sensitivity
- Surveillance system component proportion
- Population proportion

## Sources of estimates

Sometimes figures for the probabilities are already available. Sometimes they are not, but data is available that allows the figures to be calculated or estimated. And sometimes, nothing is available. How do we deal with these different situations?

### Sensitivity

Sensitivity is the probability of getting a positive test result if the animal tested truly is infected. Sensitivity is used in detection nodes. Where a detection

node refers to the use of a laboratory test, sensitivity estimates may be available. For more complex detection nodes (for instance, the probability that a sick animal will be noticed by the farmer, or the probability that a veterinarian will take samples for laboratory analysis), there are unlikely to be existing estimates.

## **Laboratory tests**

### **Existing estimates for validated tests**

In some cases, there are published studies in which a laboratory test has been validated, and the sensitivity and specificity calculated. Even when this information has not been published, internal validation studies may have been carried out by the laboratory and the figures may be available directly from them. When these figures are available, it is reasonable to use them in the scenario tree model. However there are a couple of considerations that need to be taken into account.

- Sensitivity varies due to a variety of factors, including laboratory technique and factors associated with the population under study. If validation studies have been done in another part of the world and on different populations, the values for sensitivity may not be directly applicable to the local population.
- Where the key factors that influence sensitivity are known, these should appear in the model as detection category nodes. In this case, different values for sensitivity should be used for the different categories of animals. It is relatively rare for studies aimed at estimating sensitivity to calculate different values according to a range of influencing factors – instead they tend to estimate an average sensitivity across the population studied.
- Often, sensitivity estimates are published as point values (for instance, 93.5%). However, these estimates were calculated using sampling approaches, and there is therefore some element of random error associated with the estimates. Ideally, this should be reported along with the estimates, in the form of a 95% confidence interval (for instance, in the form 93.5% [87.2% - 98.5%]). Where confidence intervals are not available, they may be able to be calculated based on the sample size used to make the sensitivity estimate. The confidence interval describes the uncertainty around an estimate, which can be incorporated into the model as described in Chapter 12.

### **Generating new estimates**

Where no existing published or internal estimates exist for a laboratory test, the next option is to undertake the studies to generate the required estimates. The traditional study would involve identifying a number of truly positive animals (based on the use of a ‘gold standard’ test), and testing them with the test to be validated. The sensitivity is the proportion of these animals that have a positive test result.

A newer alternative study design (latent class analysis) makes it possible to estimate sensitivity and specificity without requiring the use of a gold standard. This requires the use of at least two different tests and two populations with different disease prevalence levels, although it does not require the true status of individual animals to be known. Occasionally, large stores of historical laboratory records are available that meet these requirements. This data may be able to be

analysed relatively quickly and cheaply, to produce good estimates of the test performance.

Both of these approaches pose a number of particular problems. The high cost and time involved in conducting a study often makes it impossible. However, more importantly, the reason for the sensitivity estimate is to support surveillance to demonstrate that the infection is not present in the country. If the country is free from the infection, there are no infected animals to test (and artificially infecting animals would be very dangerous). In fact, this creates a paradox: in order to estimate the sensitivity of the test correctly, it should be evaluated on the animals of interest (the local population), but when the population is free from infection, the test cannot be evaluated.

Normally, we are forced to use estimates from areas where the disease is present, either historical data in the country of interest (if the disease has been eradicated), or from other countries with roughly similar populations.

### **Expert opinion**

If suitable estimates of test sensitivity are not available from any source, there is still an approach to getting appropriate figures for use in the scenario tree. Even if the test has not been formally validated, it is likely that many scientists have used it in different situations for some time. These people are likely to have a reasonably good understanding of the performance of test. Formal approaches to gathering and analysing expert opinion offer a method for collecting sensitivity estimates when no other information is available. These approaches are discussed in detail in the next section.

### **Other detection probabilities**

Detection probabilities that are not associated with a laboratory test are most commonly found in trees using some aspect of passive reporting. A typical 'detection cascade' in a passive farmer reporting system may look like this:

- Infected animal shows clinical signs
- Farmer notices animal with clinical signs
- Farmer contacts veterinary services
- Veterinarian examines animal
- Samples taken for analysis
- Samples tested for disease in question

This is then normally followed by one or more laboratory tests to detect and then confirm the infection.

Other non-laboratory detection probabilities may be associated with activities like abattoir meat inspection.

### **Existing estimates**

The probabilities listed above have rarely been explicitly studied or quantified. The exception is perhaps the sensitivity of abattoir meat inspection for the detection of various diseases. It is unlikely that other useful information will be available in the published literature.

However, some figures could be available. For example the first in the list, the probability that an infected animal shows clinical signs, is likely to be included in general descriptions of the epidemiology of the disease, and could be included in both text books and published papers.

### Generating new estimates

Unlike the evaluation of laboratory tests, estimating the probabilities associated with these non-laboratory detection nodes may be feasible, even where the disease does not exist. One of the key advantages of a scenario tree is that it explicitly identifies the various probabilities involved in the detection system, and each of these can be studied separately.

Some of the above probabilities may be calculated from existing records. For instance, veterinary visit records may indicate how often a veterinarian collects samples for analysis from cases with a certain collection of presenting signs (consistent with the disease in question). Similarly the probability that a sample submitted from a possible case is tested may be able to be determined by examining laboratory testing records. While accessing and analysing these data sources may be difficult, they offer an approach to getting realistic probability estimates for some of the required parameters.

However for others (e.g. the probability that a farmer would notice clinical signs, or that they would contact the veterinary services), no records are likely to exist. It may be possible to conduct small studies to directly measure these probabilities. One approach is to convene a number of farmer meetings in different areas. At each meeting, farmers are shown a series of photographs or videos, and asked to answer a number of questions. The videos could include a mixture of scenes in which all the animals are healthy, and ones where one or more are showing signs of the disease. The questions may be:

- Do you notice anything unusual about this group of animals?
- If yes:
  - What is unusual?
  - Would you take any action as a result of this observation?
  - What action?

Such meetings should be conducted without giving any prior information to the participants (for instance, don't invite them to a meeting with an invitation indicating that the purpose is to study the detection of Classical Swine Fever).

Clearly, the responses to such questions in the setting of a meeting may not accurately reflect people's real behaviour, but it provides some indication and may be usefully applied in the scenario tree.

### Expert opinion

In the previous example, a study of farmer behaviour could be analysed in the same way as other surveys, with the precision of the estimate being related to sample size (the number of farmers included in the study). If a structured study such as that described above is not feasible, expert opinion provides an alternative source of data. While similar in some ways, gathering expert opinion is fundamentally different to a study such as that described above. In a traditional study, each participant provides a single observation. When using expert opinion, each expert, based on their experience, is considered to be able to represent many separate observations.

Often, the "experts" when collecting expert opinion are assumed to be some respected, well educated person – a laboratory scientist, university professor etc. The explicit probabilities required by a scenario tree model show that the appropriate experts may be very different for the different questions. For instance, the experts, when considering a farmer's ability to detect animals with disease, may be either a number of farmers themselves (with experience of the behaviour of their peers), or preferably, somebody who has extensive daily contact with

farmers, sees their animals and hears about their observations of disease. This may be somebody like a field veterinarian, a paraveterinary worker, or a trader. For each question, a different expert is likely to be required.

Details on the use of expert opinion are discussed in the next section.

## Proportions

In a scenario tree, branch probabilities for category nodes are based on proportions. They all require the proportion of animals in the surveillance system component that fall into the category represented by the branch, but it is generally useful to know the population proportion as well (and this is required for risk category nodes).

### Proportion of herds or proportion of animals?

It is important to note that the proportions used in category nodes refer to the units in the infection node immediately following. For instance, consider the following example list of nodes for a scenario tree to analyse brucellosis surveillance.

Node	Type	Branches	Proportion
REGION	Group category	<i>REGION 1, 2, 3</i>	Herd
HERD TYPE	Risk category	<i>BEEF, DAIRY</i>	Herd
HERD SIZE	Risk category	<i>SMALL, MEDIUM LARGE</i>	Herd
<b>HERD INFECTED</b>	<b>Infection</b>	<b><i>INFECTED, NOT INFECTED</i></b>	
RECENTLY ABORTED	Risk category	<i>ABORTED, NOT ABORTED</i>	Animal
VACCINATED	Risk category	<i>VACCINATED, NOT VACCINATED</i>	Animal
<b>ANIMAL INFECTED</b>	<b>Infection</b>	<b><i>INFECTED, NOT INFECTED</i></b>	
AREA	Detection category	<i>REMOTE, NOT REMOTE</i>	Animal
RBT RESULT	Detection	<i>POSITIVE, NEGATIVE</i>	
SNT RESULT	Detection	<i>POSITIVE, NEGATIVE</i>	

The two infection nodes (HERD and ANIMAL) have been highlighted. The first three nodes (REGION, HERD TYPE and HERD SIZE) are all category nodes, so their branches require proportions. As these three nodes are before the herd infection node, the proportions refer to the proportion of *herds* that fall into each group. For instance, if there are 10,000 herds in the country, and 4000 of these herds are in region 1, 3500 in region 2 and 2500 in region 3 then the probability for the *REGION 1* branch of the REGION node is 40%, in the *REGION 2* branch it will be 35% and 25% in *REGION 3*.

The herd size node also refers to the proportion of *herds* because it comes before the herd infection node.

The RECENTLY ABORTED node however, comes before the animal infection node, and therefore refers to the proportion of *animals* that have recently aborted. The VACCINATION node refers to the proportion of animals vaccinated.

After the animal level infection node, any detection category nodes refer to the animal level as well. Hence, the AREA node (*REMOTE* or *NOT REMOTE*) in the above node list refers to the proportion of animals that are in remote areas, or are in not remote areas. It would also have been possible (and maybe simpler) to include the AREA detection category node prior to the HERD node, as it is the herd which is located in a remote or not remote area. Even though the node relates to detection of individual animals, it can be placed at higher points in the tree if it is logical to do so.

### Conditional proportions

Remember too that all these proportions are conditional on the previous nodes, depending on which limb the node appears on. For instance, the branch probabilities for the AREA node (*REMOTE* and *NOT REMOTE*) are conditional on REGION, HERD TYPE, HERD SIZE, RECENTLY ABORTED and VACCINATED. In most cases, however, some of the nodes may be considered independent of previous nodes. For instance, the probability that an animal is in a remote area probably depends on the region (some regions have more remote areas than other regions), but may not be related to its vaccination or abortion status as these are independent of remoteness.

In practice, to determine the population proportions for the AREA node, one could:

1. Get a map of the country
2. Identify the three regions
3. Calculate the number of herds in each region (this provides the data for the branch probabilities for the REGION node).
4. Identify on the map those areas that are remote and those that are not (This may be simply done by marking a line at a given distance from the diagnostic laboratories (buffering). A more sophisticated approach would be to calculate 'travel time contours' based on road distance and speed – these are lines joining points that take the same time to travel to from the laboratory).
5. Based on region and area, divide the country into 6 parts (*REMOTE* and *NOT REMOTE* in each of the three regions).
6. In each of these six parts, calculate the total number of animals.
7. The category proportions can then be calculated.
  - a. For *REGION 1, REMOTE*, it is the number of animals in the remote area of region 1 divided by the total number of animals in region 1
  - b. For *REGION 1, NOT REMOTE*, it is 1 minus the proportion in the remote area.

This example shows how the AREA node was only conditional on one of the previous levels and could be considered independent of the others. For each node, it is necessary to determine on which of the previous nodes the node is conditional, and for which it is independent.

For example, RECENTLY ABORTED is probably conditional on HERD TYPE (*BEEF* or *DAIRY*) and may also be conditional on REGION. Laboratory abortion investigation records may provide some information about whether there are more abortions in one region than another – although difference may be due to bias because of different reporting rates between regions, so such data must be interpreted with caution.

The VACCINATED node is likely to be conditional on REGION and HERD TYPE, and the population proportion should be able to be estimated using veterinary service vaccination records or vaccine sales records. It may also be conditional on AREA (remoteness) in which case AREA is required higher in the tree.

### **Population and surveillance system component proportions**

The previous examples focused on the use of official statistics or other records to provide information about the population. SSC proportions are often simpler to calculate, as we normally have data about the animals that are included in our surveillance.

Ideally, surveillance data should contain a record for each animal in the surveillance system component, and each record should contain information on each of the nodes included in our tree. For instance, for each animal, the data set should contain the following information:

- A herd identifier linked to herd information including
  - The region the herd is in
  - The type of the herd (beef / dairy)
  - The number of animals in the herd
  - Whether the herd is in a remote area or not
- Whether the animal has recently aborted or not
- Whether the animal has been vaccinated or not
- The results for the RBT and SNT

If this information is available, the dataset can be quickly summarised to provide all the SSC proportions required. Another approach to using this type of complete data set will be discussed in Chapter 13.

Often, this type of complete data is not collected as part of the surveillance, so some values have to be estimated. One approach to estimating the SSC proportions is to assess whether there was any targeting or bias related to that factor (normally in discussion with the surveillance designers or field teams). For instance, did they attempt to preferentially collect samples from animals that had recently aborted, or did they try to avoid vaccinated animals? If not, then the population proportion can be used, on the assumption that without targeting for these factors, the animals would be roughly representative. If there was targeting, an estimate of the level of targeting will be required to estimate the SSC proportion.

### ***Relative risk***

---

Risk category nodes require estimates of the relative risk for each of the categories. These probably represent the most difficult values that are needed for a scenario tree.

For well-known risk factors, specific risk factor studies may have been conducted providing reliable estimates of the relative risk for different categories.

As with published sensitivity estimates, it is important to consider confidence intervals when using published estimates of relative risk.

In most cases, however, little information will be available. It is very difficult and expensive to conduct risk factor studies to measure the relative risk (and these can only be done when the disease is present). Expert opinion is usually necessary to estimate relative risks.

## Expert opinion

---

The use of expert opinion as a method of estimating parameters (such as the sensitivity of a test, or a relative risk) has often been viewed as undesirable and unreliable. This may be due to the common misconception that the approach is based on asking an expert (a wise person respected in their field) what they think the answer is, and the expert makes a guess at the answer.

The appropriate use of expert opinion is very different. Philosophically, it is based on the same type of approach as scenario tree modelling to demonstrate freedom from infection. To demonstrate freedom, we aim to use all available sources of evidence, even if they are complex and even if alone they do not contribute very much evidence. The principle is not to say that the evidence is imperfect and therefore to reject it, but to carefully understand the limitations of the different data sources, and to use that which is good (risk-based surveillance), and take into account that which is bad (the presence of bias).

Expert opinion is appropriate when no data based on direct structured observation is available. It may also be used to supplement information from small or potentially biased studies. The principle of the use of expert opinion is to acknowledge that, even if no study exists, there are usually a number of people who have a great deal of experience with the question of interest. Rather than say that this experience is not as well structured or as easily captured as an objective study, we try to capture this experience, while, at the same time, making sure that any limitations are clearly taken into account.

The three main rules for expert opinion are:

- ask the right experts
- ask them specific questions in a way that enables them to provide specific answers
- always capture uncertainty.

No matter how experienced an expert is, there is a significant chance that they will get the wrong answer. This doesn't matter too much if the answer is close to the real answer, but could be important if it is a long way from the real answer. Capturing uncertainty involves asking the experts not only to say what they think the answer is, but to indicate how confident they are about their answer. This is normally done by asking them to provide a confidence interval. If their estimate is wrong, the confidence interval indicates the range in which they are very sure the true estimate lies.

This approach is the same as that used with other parameters of the scenario tree that may be uncertain. For instance, a sensitivity estimate from a validation study is based on a sample of animals and therefore has a confidence interval that is related to the sample size. When we explicitly include confidence intervals in our scenario tree model, we accept that the result may be wrong, but we can offer a range in which we are very confident that the correct result lies. Chapter 12 provides a detailed description of how to incorporate uncertainty into a scenario



tree model, so for the moment we just need to ensure that experts provide confidence intervals.

A great deal of research into collecting and using expert opinion has been done, but this discussion will only look at a small number of common approaches that are generally practical and suitable to elicit the required parameters for scenario trees.

## *Gathering expert opinion*

---

### **Working with experts**

#### **Choosing experts**

As discussed previously, the right experts are usually not scientists and professors, and are almost certainly not the person building the scenario tree or the person in the office next door. The right experts are the people that have direct and significant experience with the specific question being asked, which means that the right experts are often different for each different question.

Experts' estimates are more likely to be applicable to the population of interest if there is broad experience of the population. This means that it is much better to have a group of experts than just one or two. This may be as few as five, but could be hundreds – although there is a point when the exercise stops being expert opinion and starts being a survey (for instance a survey of farmer behaviour).

#### **Interaction between experts**

Group dynamics can play an important role in the estimates provided by experts. Two main approaches are commonly used.

The first involves working with each expert independently. This may be with a one-on-one face to face interview, by telephone, by email, or by asking experts in a group setting to consider the question and provide their own written answers without consulting with the other experts. The advantage of this approach is that each person's opinion is not influenced by that of others, allowing the full range of experience to be captured. It avoids the danger of having one or two dominant personalities in a group who exert too much influence over the opinions of others. It also provides one result per expert, which can be used to assess the variation in responses and used as a measure of uncertainty.

The other option is to work in a group. The group is asked to discuss the question together. At the end of the discussion, you may ask the group to produce a single estimate based on consensus (with a confidence interval), or ask each expert to record their own estimate, in the light of the group discussions. Some of the advantages of the group approach include:

- Discussion within the group ensures that there is a common understanding of the question. When questions are answered individually, there is a risk that some will interpret the question slightly differently.
- Discussion also often prompts the memory of others within the group, and allows the question to be considered from a variety of points of view.

The best approach may vary for different situations and different expert groups.

## Questions

The way in which a question is asked has an important impact on the answers that may be received.

### Ensuring that the question is understood

It is very important that the experts fully understand the question that is being asked. This is not always as simple as it seems. For instance, when asking experts to estimate relative risk, it is possible or even likely that not all experts will be familiar with the concept of relative risk. This problem can be addressed in two ways.

First there may be a need for training and explanation. This is generally a good idea in any case, but is important when the questions use technical concepts. Relative risk is a common concept amongst epidemiologists, but may be poorly understood by many others involved in animal health. A brief explanation of what a relative risk is, how it is calculated and how it is used will improve the quality of responses. However, it may also be valuable to provide a list of examples of relative risks for known risk factors, which may be used as a comparison. Those that are inexperienced in the use of relative risks may think that values of 10 or 50 appear reasonable, without realising that most risk factors for many diseases have relative risks much lower than these (often in the range of 1.2 to 3).

The second approach is to ask questions in terms that are already understood by the experts. If the experts are not familiar with relative risks, then ask them for estimates that would enable a relative risk to be calculated. For instance:

“Imagine two groups, each of 100 animals. Group 1 has the risk factor, and group 2 does not have the risk factor. If the disease is present in the area, how many animals in group 1 would you expect to have the disease, and how many in group 2?”

The two prevalence estimates from the previous question can be used to calculate an estimate of the relative risk.

### Confidence intervals

For each question, it is important to capture the experts' uncertainty. This is normally done in two stages:

1. “What do you think the correct value is?”
2. “If you are wrong, what is the lowest possible value that could be correct, and what is the highest?”

This approach will provide a range in terms of the minimum and maximum possible values, with the most likely value somewhere in between. This is the most common approach used in expert opinion. In statistics, a 95% confidence interval is usually used, but this tends to be more difficult for many experts to imagine and to estimate.

## *Combining expert opinion*

---

When a group approach is used to collect expert opinion and the group is able to provide a single consensus estimate of the value (and of the confidence interval), these values can be used directly in the scenario tree. However, when experts provide values independently, there will be a number of different estimates and confidence intervals. We need an approach to combine or summarise these, to determine what should be used in the model.

## Uncertainty and variability

There are two reasons why an expert may not be able to give a precise single estimate of a value (for instance, a prevalence).

- They may not know the correct answer (they are uncertain),
- There may not be a single correct answer, as the prevalence may be different in different situations (there is variability in the answer).

The approach to combining estimates from different experts differs depending on whether the main reason for the confidence interval is uncertainty or whether it is variability. Of course, in many cases, both will be present.

### Uncertainty

When the variability in estimates is due to uncertainty, it implies that there is a single correct answer. Estimates from experts can be thought of as samples from a population of experts, each with some random error, but distributed around the true value.

The figure below shows an example of the results that 20 experts may provide, estimating the sensitivity of abattoir meat inspection for identifying paratuberculosis.



Figure 7: Twenty experts' estimates of the sensitivity of abattoir inspection for detecting bovine paratuberculosis (simulated data)

If it is considered that there is one true value for the sensitivity, and it is assumed that the estimates of the experts are not biased, then we could summarise the results by taking the average of the estimates as an estimate of the true value. In this case the average is 0.37.

There are two methods we could use to describe the uncertainty. The first is to consider that any of the estimates could be correct and to use the lowest and highest estimates as the minimum and maximum possible values. This is the most conservative approach and would provide the widest range. The confidence interval is shown below. This is based on a beta-PERT probability distribution as discussed in the next chapter.

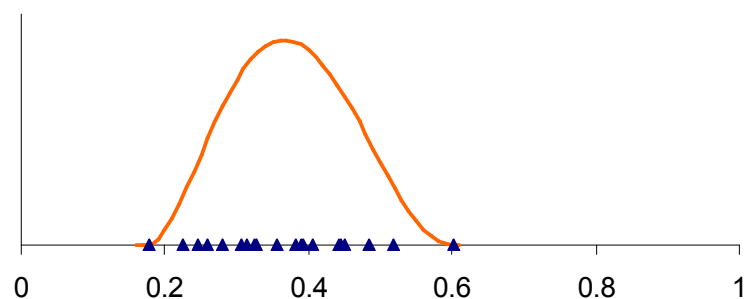


Figure 8: Expert's estimates with a PERT distribution describing uncertainty

The second would be to consider the confidence intervals provided by the experts, and take the mean of the lower and the mean of the upper ends of these intervals. Depending on the width of the experts separate confidence intervals, this may produce an overall confidence interval that is wider or narrower than the one shown above.

### Variability

If differences in expert opinion are considered to represent varying correct values for the parameter under different conditions, then the summary of the experts' views should retain these differences. Using the average is no longer appropriate. A common approach is to consider each expert's estimate and confidence interval as a distribution, and build up a composite distribution based on the views of all the experts. The details of how this is done will be discussed in the next chapter, but the figure below illustrates an example of the output. Each expert's opinion has contributed to the shape of the final curve.

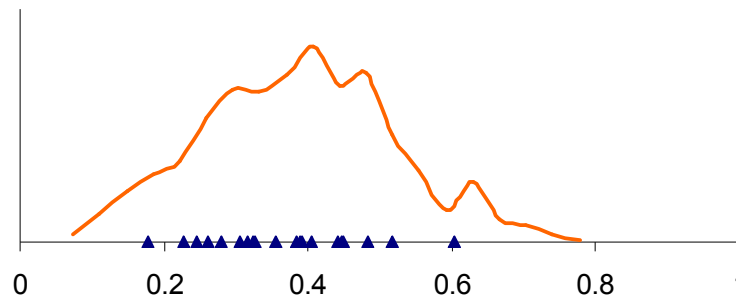


Figure 9: Expert's estimates with a composite distribution describing variability

### *Rinderpest example*

In a scenario tree for analysing livestock and wildlife surveillance for rinderpest, vaccination was considered an important factor. Vaccination had ceased at different times in the areas being considered, but for the purpose of this example, the last vaccination was given six years before the surveillance. Animals less than six years old were certain not to have been vaccinated, while animals older than six years may have been vaccinated.

Age was used as a risk category node with two branches, less than six years, and greater than or equal to six years. The surveillance targeted younger animals, so it was necessary to estimate the population proportion and the SSC proportion of animals less than six years of age.

After discussion with local experts (13 field veterinarians with extensive experience of all the species), it was agreed that it would be very difficult to directly estimate the proportion of each species less than six years of age. Instead, an indirect approach was used.

Experts were asked to describe the age structure of each species in terms of a survival curve. For each one-year age bracket, they were asked to estimate the proportion of animals born that survived to that age group. Below is an example of the information gathered from one expert.

Table 1: Example of one expert's estimates of the survival curve for different species

Years	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cattle	0.7	0.69	0.65	0.5	0.48	0.46	0.44	0.4	0.38	0.35	0.3	0.23	0.2	0.16	0.05
Sheep	0.65	0.4	0.2	0.18	0.1	0.04	0.02	0	0	0	0	0	0	0	0
Goat	0.65	0.4	0.2	0.18	0.1	0.04	0.02	0	0	0	0	0	0	0	0
Buffalos	0.85	0.8	0.83	0.8	0.75	0.6	0.57	0.45	0.4	0.34	0.33	0.29	0.25	0.2	0.15
Warthogs	0.6	0.4	0.2	0.18	0.1	0.04	0.02	0	0	0	0	0	0	0	0
Kudu	0.55	0.5	0.4	0.18	0.1	0.04	0.02	0	0	0	0	0	0	0	0
Giraffe	0.88	0.85	0.83	0.8	0.75	0.6	0.57	0.45	0.4	0.34	0.33	0.29	0.25	0.2	0.15
Eland	0.88	0.85	0.83	0.8	0.75	0.6	0.57	0.45	0.4	0.34	0.33	0.29	0.25	0.2	0.15
Gerenuks	0.65	0.59	0.4	0.35	0.3	0.21	0.12	0.07	0.04	0.02	0	0	0	0	0
Camels	0.88	0.85	0.83	0.8	0.75	0.6	0.57	0.45	0.4	0.34	0.33	0.29	0.25	0.2	0.15

As it was assumed that there was a single correct survival curve for each species, the estimates from each expert were averaged. This produced the following survival curves.

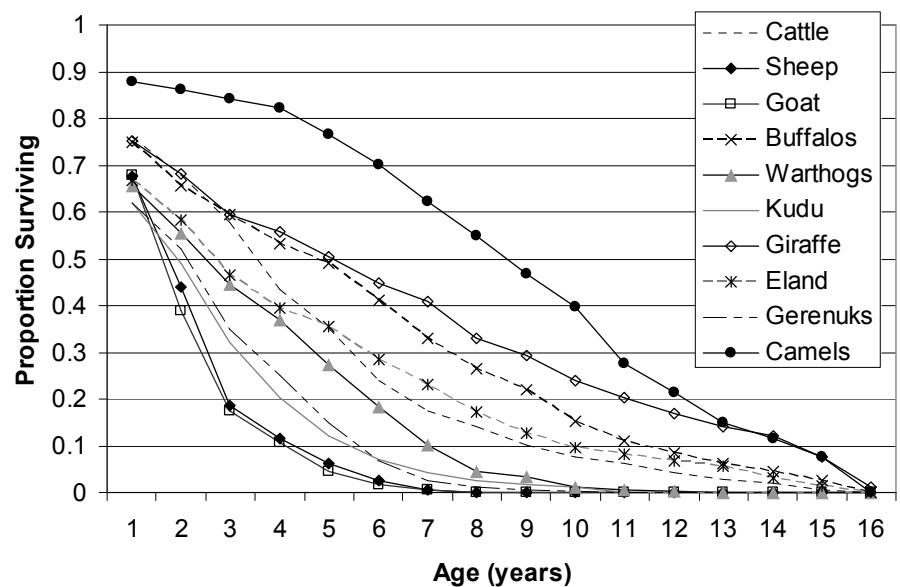


Figure 10: Estimated survival curves based on the means of 13 experts' estimates

The uncertainty around each of these curves was estimated based on the standard deviation of the estimates. The proportion of animals less than six years is the area under the curve left of the 6-year point on the x-axis, as a proportion of the total area under the curve for that species. The figure below shows the estimated proportions.

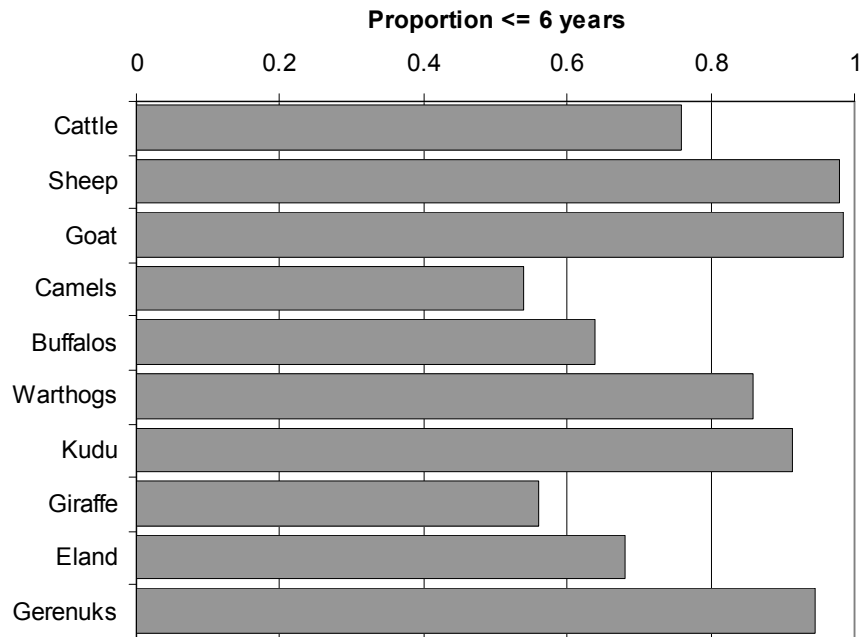


Figure 11: Estimated proportion of animals of different species less than six year of age, based on expert opinion

This example illustrates how it may be possible to gather the required estimates for a scenario tree by asking questions in a form that is easier for the experts to understand. Although the true survival curves for the different species in the study area are not known, the results in Figure 10 are certainly biologically believable and consistent with expectations.

# Chapter 12 – Incorporating Uncertainty

Do not expect to arrive at certainty in every subject which you pursue. There are a hundred things wherein we mortals. . . must be content with probability, where our best light and reasoning will reach no farther.

Isaac Watts (1674 – 1748)

## **Capturing uncertainty and variability in a model**

---

A model is a simplification of reality. Models are never completely perfect, and the objective is to create a model that provides answers that are good enough for effective decision making. If a model is too simple, it may miss important factors and not be able to provide useful information. If it is too complex, it may be too difficult to find all the parameters required.

Scenario trees are models of a surveillance system component. They attempt to capture the effect of all the major factors involved in the distribution of infection and the operation of the surveillance system, but cannot and should not include every smallest detail. As discussed in the previous chapter, it is often not possible to find the exact information for a parameter, as the data is not available. Instead, it is necessary to estimate the value required, recognising that it may be incorrect. If it is incorrect and we pretend that it is correct, then the model results will be wrong. However, if it is incorrect and we use confidence intervals to describe how close to the right answer we think we are, the model can take this uncertainty into account.

### Example

Consider a scenario tree model where all the parameters are known perfectly, except for one: the sensitivity of the laboratory test. A small study has been done to estimate the sensitivity, and produced an estimate of 96.5%, but the 95% confidence interval is from 81.2% to 98.6%.

If we analyse the scenario tree using the best estimate (96.5%) we will get a result for the sensitivity of our surveillance system component, say 88%. This may be close to the right answer, but it could also be wrong. We need to be able to communicate to those that are using the results of our analysis that the answer could be wrong, and describe how wrong. We have a confidence interval for our input, so it would be useful to have a confidence interval for our output as well.

One way to do this would be to run the model again, but this time, instead of using the best estimate, we could use the lower limit of the confidence interval (81.2%). This time, our scenario tree gives a different result, 83%. We could then run the model a third time, using the upper limit of the confidence interval (98.6%) and we would get a third result (91%).

Each time we run the model using a different input, we will get a different output. This example shows how the three different inputs (the bottom, middle and top of our confidence interval) can produce three different results that indicate the bottom, middle and top of the range of possible values for the sensitivity of our surveillance system.

This approach is good when there is just one parameter that has uncertainty, but normally there are many. If there are two parameters that have confidence intervals (the minimum for parameter 1 [ $\min_1$ ], the most likely for parameter 1 [ $ml_1$ ], and the maximum for parameter 1 [ $\max_1$ ] and the same for parameter 2), we could run the model multiple times to see what results we got for the following combinations:

- $ml_1$  and  $ml_2$
- $\min_1$  and  $ml_2$
- $\max_1$  and  $ml_2$
- $ml_1$  and  $\min_2$
- $ml_1$  and  $\max_2$
- $\max_1$  and  $\min_2$

As there are more uncertain parameters, there would be more and more different combinations, which makes this approach somewhat impractical.

The other problem with this approach is that all values between the minimum and the maximum are not equally likely to be correct. An expert's best estimate or the point estimate of a survey indicate the result that is most likely to be correct. The upper and lower limits could each, conceivably, be correct, but they are much less likely. Simply testing the upper and lower limits doesn't give an indication of what value is most likely to be the correct value.

Instead of just using three points to measure uncertainty, it is more effective to use a probability distribution. A probability distribution describes how likely each value is to be correct over a given range. In the example from the previous chapter, we summarised expert opinion by using the average value as the most likely correct result, and the lowest and highest of the experts' estimates as the minimum and maximum possible values. This is shown again in Figure 12. The line is a probability distribution which shows that the value of 0.37 is the most



likely. Values just above or just below 0.37 are also very likely, but as you get further away, the results are possible, but increasingly less likely.

If we use probability distributions as the input to our scenario tree model, then we can get a probability distribution as the output, describing how likely a range of different values are to be correct.

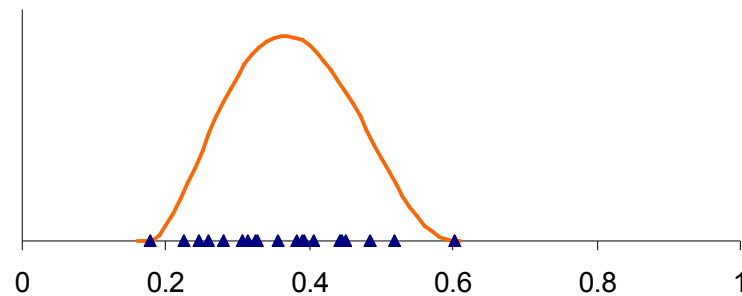


Figure 12: Expert estimates and a PERT probability distribution

The example in Chapter 10 showed how to calculate sensitivity with a scenario tree. At each step, we did calculations with numbers, multiplying them together to give another number as a result. If we want to use a distribution as the input to the scenario tree instead of a number, we need a tool that lets us do the same calculations on distributions. The tool is known as stochastic modelling.

## Stochastic modelling

The principle behind stochastic modelling is very simple. Just as we described above, if we run the scenario tree model using different inputs, we will get different results.

Stochastic modelling uses computer software to analyse the model again and again, each time using different inputs, and records the result of each analysis. Typically, a model may be run 1000 or 10,000 times, and each time there will be a different answer.

But what values are used for the inputs? In our example we chose the minimum, most likely and maximum value from the distribution. This gave us the limits but it wasn't able to show which values were more common and which were less common. In stochastic modelling, the input values are chosen *at random* from the input distributions. This approach gives the technique its other common name – Monte Carlo simulation, named after the famous casinos in Monte Carlo, where all activities are based on random chance.

The steps in running a stochastic model are:

1. Build your scenario tree.
2. Describe every parameter for which there is uncertainty or variability in terms of a distribution.
3. Tell the software to analyse the model for a set number of times (or iterations).
4. For each iteration, the model will randomly select a single number from each of the parameter distributions. For any other values in the model that use a fixed number, that number will be used for every iteration.
5. The result of each iteration is stored.

- When the iterations are finished, you can use the output values to draw a histogram that describes the output distribution.

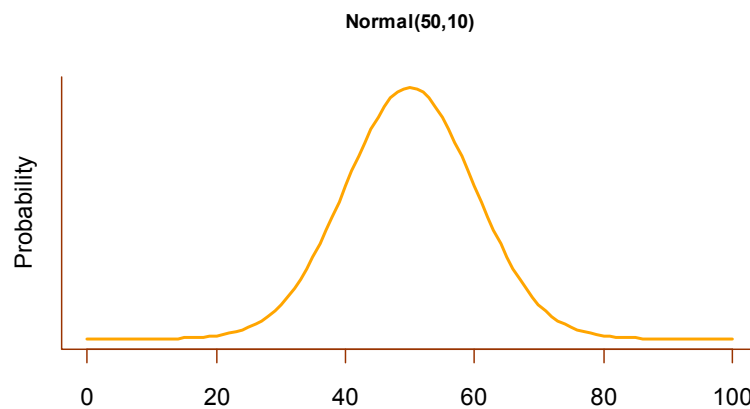
### ***Describing distributions***

---

Probability distributions are an important part of a stochastic model as they allow inputs to be described not as points but ranges of possible values each with a specified likelihood of being correct. Rather than specifying the probability of every value in the range, distributions are usually described in terms of a number of parameters.

#### **Normal**

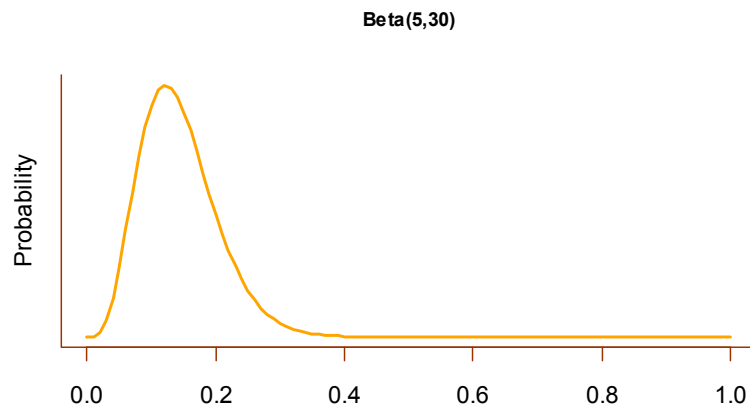
The most well known is the normal distribution, which is described in terms of the mean and the standard deviation, as shown below.



The normal distribution is used to describe many biological measurements such as weight or production. It is rarely used in scenario tree models.

#### **Beta**

A more common distribution for scenario tree models is the beta distribution. This is described by two parameters, alpha and beta, and is bounded in the range from zero to one. It is therefore very useful for modelling probabilities and proportions, such as prevalence, population proportions and sensitivity. An example of the beta distribution is shown below.



A beta distribution is also extremely useful for representing uncertainty about proportions generated from count data. The alpha and beta parameters can be calculated directly from the data:

$$\text{alpha} = x+1,$$

$$\text{beta} = n-x+1$$

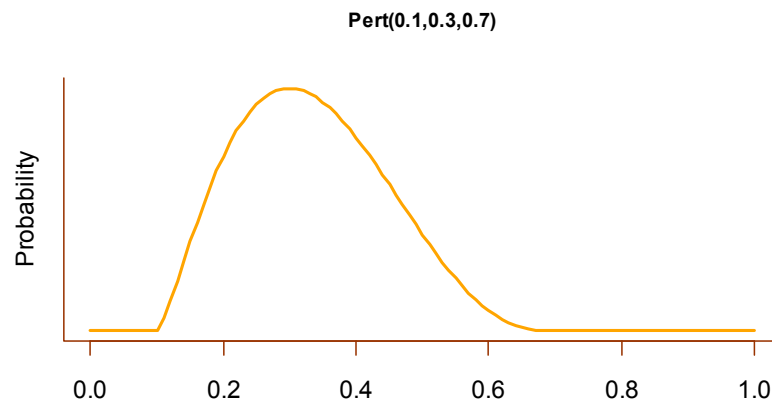
where

- $x$  is the number of successes and
- $n$  is the sample size.

They can also be calculated from the mode of the distribution, showing the highest point or most likely value, and a percentile (such as the 95% percentile) of the distribution, indicating the spread.

### PERT

A special form of the beta distribution has been developed to deal specifically with expert opinion. This is called the beta-PERT distribution and has the parameters minimum, most likely, and maximum. This is the most commonly used distribution in scenario tree modelling when expert opinion is used.



### Other distributions

A wide range of other distributions are available for particular purposes, including:

- Binomial – for a binary event (such as tossing a coin or becoming infected), this shows the likely number of successes in a given number of events.
- Discrete – this is an arbitrary distribution normally described by a data set or a histogram summarising a data set.
- Hypergeometric – this is used to model sampling from a population without replacement.
- Lognormal – this is a logarithmic transformation of the normal distribution and is used to describe skewed data such as herd or flock size, or the incubation period of disease.
- Triangular – this takes the same three parameters as a PERT distribution but joins the three with straight lines. The PERT is generally thought to provide a more realistic description of probabilities associated with expert opinion as it provides greater weight on values close to the mode and reduced weight in the tails.

### Rules of thumb

- If you are dealing with expert opinion, use the PERT distribution
- If you are dealing with other probabilities (sensitivity, proportions, prevalence), use the beta distribution. There are a number of tools available to calculate the alpha and beta parameters of a beta distribution if you know the mode and the 5<sup>th</sup> or 95<sup>th</sup> percentile values (see <http://epitools.ausvet.com.au>), or they can be calculated from the data as described above.
- For other distributions, only use them if you know they are appropriate for the data that you are analysing. If in doubt, consult a statistician.

## Software for stochastic modelling

---

Stochastic modelling requires specialised software that is able to randomly select values from defined input distributions and analyse the model repeated over many iterations. There are a number of software packages that are available for this purpose, but three will be mentioned here. The first two are add-ins for Microsoft Excel®, providing new formulae and menu items. The third is dedicated web-based software especially designed for scenario tree modelling.

### *Palisade @Risk*

---

@Risk is a well known powerful commercial software package available from <http://www.palisade.com/risk>. It is widely used and capable of performing all the functions required for scenario tree modelling. It is available for recent MS Windows operating systems and is accompanied by an extensive help system, so won't be described here any further.

### *PopTools*

---

PopTools provides an effective free alternative to @Risk. This software works in a very similar way, as an Excel plug-in with new formulae and menus. It was originally developed for ecological modelling but has a range of stochastic modelling tools that are more than capable of supporting the needs of scenario tree modelling. It also has an impressive array of other analytical tools and utilities. It is available for free download from <http://www.poptools.org>. The package includes a large collection of example spreadsheets to illustrate the different functions, as well as inbuilt help. More comprehensive help is available for separate purchase at a nominal fee.

The examples later in this chapter are based on the use of PopTools, but can be adapted to use with @Risk with minor modifications.

### *Freedom*

---

Both the previous software packages work as spreadsheet add-ins. This means that the scenario tree must be developed in a spreadsheet to be analysed. For small trees, this is relatively straightforward, but for large trees, this can become very complex and it is easy to make mistakes.

The "Freedom" web site at <http://freedom.ausvet.com.au> has a series of resources for scenario tree modelling, including dedicated web based software for the analysis of complex surveillance systems. This software guides the user through the development of a scenario tree, and analyses it using stochastic

modelling automatically. The use of this software is described in detail in Chapter 17.

## Example exercises

### Exercise 1: Combination of expert opinion

In Chapter 11 we discussed the use of expert opinion. This exercise will show how expert opinion can be combined, based on the assumption that variation is primarily caused by variability rather than uncertainty. This is a small stochastic model, but it is not a scenario tree model.

There are five different experts who have been asked to estimate the probability that a veterinarian would collect specimens from a case showing signs consistent with the disease of interest. Each expert has been asked to provide their most likely estimate as well as the minimum and maximum values. The data from the experts is shown below: In addition, the experts were asked to evaluate their own level of expertise related to the question, on a scale of 1 to 5. This is shown in the *weight* column.

	Min	Most likely	Max	Weight
<b>Expert 1</b>	0.2	0.3	0.4	5
<b>Expert 2</b>	0.25	0.3	0.55	5
<b>Expert 3</b>	0.65	0.75	0.9	1
<b>Expert 4</b>	0.4	0.5	0.7	3
<b>Expert 5</b>	0.3	0.55	0.75	2

This data is entered into Excel, as shown below. Colour coding is used to distinguish the values: **black** for labels, **blue** for input data, **orange** for random variables and **red** for output.

	A	B	C	D	E	F
1		<b>Min</b>	<b>Mode</b>	<b>Max</b>		<b>Weight</b>
2	<b>Expert 1</b>	0.2	0.3	0.4		5
3	<b>Expert 2</b>	0.25	0.3	0.55		5
4	<b>Expert 3</b>	0.65	0.75	0.9		1
5	<b>Expert 4</b>	0.4	0.5	0.7		3
6	<b>Expert 5</b>	0.3	0.55	0.75		2
7						
8						

The spreadsheet is now ready to be converted into a stochastic model. The first step is to enter formulae for the random variables. We will use a PERT distribution to describe the estimates for each of the experts. There are two ways to enter the formula:

Using the menu (this is easier when you are not used to the formulae)

1. Place the cursor in the cell where the formula should be entered (E2)

2. Click on the **PopTools** menu
3. Select **Random variable**
4. In the dialog select PERT as the distribution
5. Leave Length blank
6. The output cell should already be set to E2
7. Set the Min value to B2
8. Set the Likely value to C2
9. Set the Max value to D2
10. Leave the weight as 4
11. Click Go

By typing the formula yourself (this is faster when you are familiar with the formulae)

1. Place the cursor in the cell where the formula should be entered (E2)
2. Type: **=dPertDev(B2,C2,D2,4)**
3. Press enter

If PopTools is loaded and the formula has been entered correctly, you should now see a number in the cell E2. The number is a random value drawn from the PERT distribution, so will be different to the numbers shown here.

Copy the formula down to cells E3 to E6 so your spreadsheet looks like this:

	A	B	C	D	E	F
1		<b>Min</b>	<b>Mode</b>	<b>Max</b>	<b>Rand</b>	<b>Weight</b>
2	<b>Expert 1</b>	<b>0.2</b>	<b>0.3</b>	<b>0.4</b>	<b>0.264064</b>	<b>5</b>
3	<b>Expert 2</b>	<b>0.25</b>	<b>0.3</b>	<b>0.55</b>	<b>0.391712</b>	<b>5</b>
4	<b>Expert 3</b>	<b>0.65</b>	<b>0.75</b>	<b>0.9</b>	<b>0.730146</b>	<b>1</b>
5	<b>Expert 4</b>	<b>0.4</b>	<b>0.5</b>	<b>0.7</b>	<b>0.586328</b>	<b>3</b>
6	<b>Expert 5</b>	<b>0.3</b>	<b>0.55</b>	<b>0.75</b>	<b>0.618053</b>	<b>2</b>
7						
8						

The random numbers will be changed every time you recalculate the spreadsheet. This can be easily demonstrated by pressing the **F9** key. Each time you press, the numbers in column E change, representing new random numbers from the defined PERT distributions.

We have now described the input to our model in terms of distributions and entered the formulae to select random numbers from those distributions to be used for each iteration of the model. The next step is to do the model calculations to produce the output. The method we used to combine expert opinion is to randomly select a value from one expert at each iteration.

The experts gave themselves a weight between 1 and 5 to indicate their level of expertise. We will use these weights when we select which expert's opinion to select. Experts with a weight of 5 will be five times more likely to be selected than experts with a weight of 1. In this way, the opinion of our 'strong' experts will contribute more to our output than the opinion of our 'weak' experts, but all will have some contribution.

Enter the formula into cell E8 using either the **Random variable** menu option (Discrete distribution) or typing yourself:

**=DiscreteDev(E2:E6, F2:F6)**

Note that the parameters are both ranges.

The spreadsheet should now look something like the one below. Press the **F9** key again a few times to see how the numbers change when the sheet is recalculated.

	A	B	C	D	E	F
1		<b>Min</b>	<b>Mode</b>	<b>Max</b>	<b>Rand</b>	<b>Weight</b>
2	<b>Expert 1</b>	<b>0.2</b>	<b>0.3</b>	<b>0.4</b>	<b>0.264064</b>	<b>5</b>
3	<b>Expert 2</b>	<b>0.25</b>	<b>0.3</b>	<b>0.55</b>	<b>0.391712</b>	<b>5</b>
4	<b>Expert 3</b>	<b>0.65</b>	<b>0.75</b>	<b>0.9</b>	<b>0.730146</b>	<b>1</b>
5	<b>Expert 4</b>	<b>0.4</b>	<b>0.5</b>	<b>0.7</b>	<b>0.586328</b>	<b>3</b>
6	<b>Expert 5</b>	<b>0.3</b>	<b>0.55</b>	<b>0.75</b>	<b>0.618053</b>	<b>2</b>
7						
8				<b>Result</b>	<b>0.391712</b>	

Our model is now complete, with inputs expressed in the form of random variables from defined distributions, and an output value. The next step is the Monte Carlo simulation – analysing the model many times and collecting the result of each iteration.

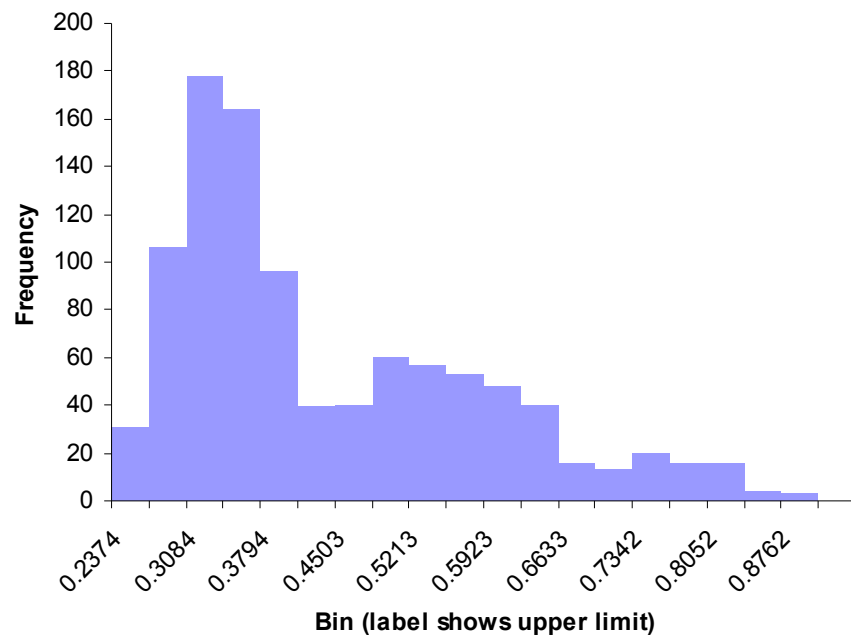
1. Click on the **PopTools** menu, select **Simulation tools** and then **Monte Carlo analysis**.
2. In the Monte Carlo analysis dialog you need to enter details of the simulation:
  - a. Dependent range: This is the output cell, E8. You are able to select multiple cells if you want to analyse a number of different outputs, but they must all be in a single column without blank spaces.
  - b. Test values: This is used to test whether your result is greater or less than some fixed value. Normally, this is left blank.
  - c. Lower percentile and upper percentile: these are used to provide statistics for the test values. Again, these can be left blank
  - d. Number of replicates. This determines how many times the model will be run. The default is 100, but for simple models, it is normally possible to run 1000 iterations quite quickly.
  - e. Output: This is where a brief summary of the results of the analysis will be produced. Specify a blank cell.
  - f. Test criterion: These are only used when test values are specified. Normally they can be ignored.
  - g. Random seed: This can be normally left as zero. If you set a fixed random seed (other than zero), every time you run the model with the same random seed, you will get exactly the same result (as the same sequence of random numbers are used). When zero is used, the computer's clock is used to generate a new random seed for each analysis.
  - h. Keep results: This allows you to store the result of each analysis on a new sheet in your spreadsheet. You should always make sure this box is selected so you can look at the output distribution.
  - i. Colour code for demo: This is for training purposes only – leave this blank.
  - j. Click Go and watch the counter indicate the iterations.

When the analysis is finished (it should be very quick for such a simple model) the summary results will be displayed, and a new page added to the spreadsheet, called **Monte Carlo results 1**. The summary results are not very revealing, but we can get more information by analysing the new spreadsheet as we can use it to see the output distribution.

To show the output distribution:

1. Open the new page, and highlight all the numbers in column B under Var 1.
2. Click on the PopTools menu, then Simulation Tools and then click Summary stats
3. The input range should already be set to the column of numbers
4. The test value can be left blank
5. For the output range, enter a blank cell.
6. Check the Sort range for percentiles box
7. Select 20 bins for histogram
8. Click Go

PopTools produces some more detailed summary statistics and then draws a histogram of the output results, which should be similar to the one shown below.



This histogram represents the combined estimates of the five experts. Note that it is not a smooth regular curve. The experts had quite different views, and each of these views is reflected in the output. This approach indicates that the probability that a vet would submit samples varies considerably – in some circumstances it is reasonably high, but most of the time it is quite low.

The great value of this output distribution is that it does not claim that there is a single correct value. It describes the variability, as well as the experts' uncertainty.



## Exercise 2: Analysis of a simple scenario tree

For this exercise, we will use the simple three-node scenario tree that we introduced in Chapter 10. This tree includes one risk category node (age) as young animals are at higher risk than older animals, an infection node, and a detection node. The calculations are the same as those used previously, but for this exercise, we will introduce uncertainty in some of the parameters:

- The relative risk for younger animals compared to older animals
- The population proportion of younger and older animals
- The SSC proportion of younger and older animals
- The sensitivity of the ELISA.

The layout of the spreadsheet is shown below, including model parameters, the scenario tree, and the results. For more complex models, these three sections are often divided between three pages. Remember the meaning of the colour coding which makes it easier to understand the model: **input values**, **random variables**, **normal formulae**, and **outputs**.

### Parameters

Age	Branch	Min	Most Likely	Max	Value
Relative risk	Young	1.5	5	7	4.04579626
	Old				1
Population proportion	Young	0.18	0.2	0.24	0.19426404
	Old				0.80573596
SSC proportion	Young	0.85	0.9	0.92	0.88348941
	Old				0.11651059
Adjusted Risk	Young	2.541826			
	Old	0.628264			
Design prevalence		0.05			
ELISA	Sensitivity	0.85	0.95	0.99	0.91843519
Animals in SSC		20			

### Scenario Tree

Age		Infected		ELISA result		Outcome	Probability
Branch	PrSSC	Branch	EPI	Branch	Se		
Old	0.1165	Yes	0.0314	Pos	0.9184	Pos	0.0034
				Neg	0.0816	Neg	0.0003
		No	0.9686	Pos	0	Pos	0.0000
				Neg	1.0000	Neg	0.1129
Young	0.8835	Yes	0.1271	Pos	0.9184	Pos	0.1031
				Neg	0.0816	Neg	0.0092
		No	0.8729	Pos	0	Pos	0.0000
				Neg	1.0000	Neg	0.7712

Check Sum: 1.0000

### Results

Unit Sensitivity (Actual)	0.106487
Unit Sensitivity (Representative)	0.045922
Component Se (Actual)	0.894799
Component Se (Representative)	0.609447
Sensitivity ratio	1.468215

The figures displayed represent a single iteration of the model. The next page shows the same spreadsheet layout, but cells containing formulae have been displayed with the formula rather than the result. Most of the formulae use well known functions but some deserve special note:

- The PopTools functions for a random variable from the PERT distribution in F3, F5, F7 and F12.
- The formula for calculating the adjusted risk in cells C10 and C9.
- The **SUMIF ()** function in C30. This calculates the sum of a column of numbers if the value in another column matches some criterion. The formula used **=SUMIF (G19 : G26 , "Pos" , H19 : H26)** adds each value in H19 to H26 if the corresponding value in G19:G26 is equal to "Pos".

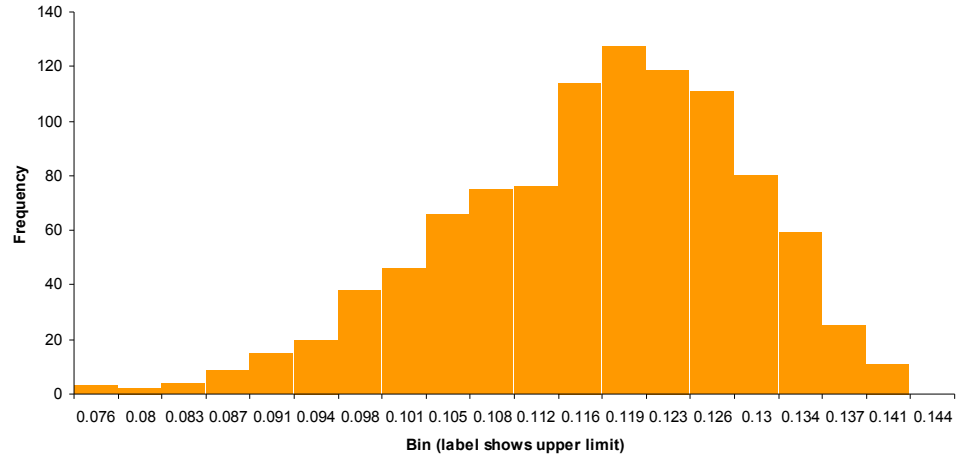
Try setting up this spreadsheet in Excel yourself and make sure that it is working properly.

	A	B	C	D	E	F	G	H
1	<b>Parameters</b>							
2	<b>Age</b>	Branch	Min	Most Likely	Max	Value		
3	Relative risk	Young	1.5	5	7	=dPertDev(C3,D3,E3,4)		
4		Old				1		
5	Population proportion	Young	0.18	0.2	0.24	=dPertDev(C5,D5,E5,4)		
6		Old				=1-F5		
7	SSC proportion	Young	0.85	0.9	0.92	=dPertDev(C7,D7,E7,4)		
8		Old				=1-F7		
9	Adjusted Risk	Young	=F3*C10					
10		Old	=1/((F3*F5)+(F4*F6))					
11	Design prevalence		0.05					
12	ELISA	Sensitivity	0.85	0.95	0.99	=dPertDev(C12,D12,E12,4)		
13	Animals in SSC		20					
14								
15								
16	<b>Scenario Tree</b>							
17	<b>Age</b>		<b>Infected</b>		<b>ELISA result</b>			
18	<b>Branch</b>	<b>PrSSC</b>	<b>Branch</b>	<b>EPI</b>	<b>Branch</b>	<b>Se</b>	<b>Outcome</b>	<b>Probability</b>
19	Old	=F8	Yes	=C11*C10	Pos	=\$F\$12	Pos	=F19*D19*B19
Neg					=1-F19	Neg	=F20*D19*B19	
No			=1-D19	Pos	0	Pos	=F21*D21*B19	
				Neg	=1-F21	Neg	=F22*D21*B19	
Young	=F7	Yes	=C11*C9	Pos	=\$F\$12	Pos	=F23*D23*B23	
				Neg	=1-F23	Neg	=F24*D23*B23	
		No	=1-D23	Pos	0	Pos	=F25*D25*B23	
				Neg	=1-F25	Neg	=F26*D25*B23	
							Check Sum: =SUM(H19:H26)	
29	<b>Results</b>							
30	Unit Sensitivity (Actual)		=SUMIF(G19:G26,"Pos",H19:H26)					
31	Unit Sensitivity (Representative)		=C11*F12					
32	Component Se (Actual)		=1-(1-C30)^C13					
33	Component Se (Representative)		=1-(1-C31)^C13					
34	Sensitivity ratio		=C32/C33					

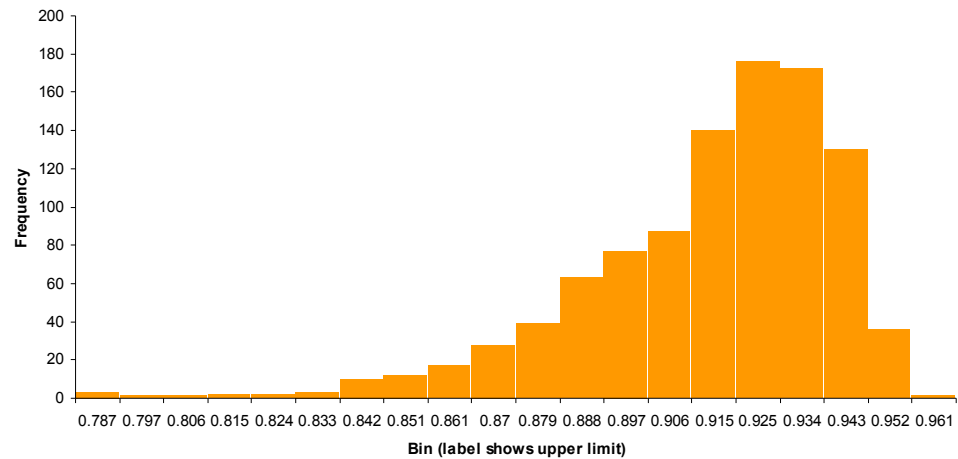
Once the model has been set up, it can be analysed stochastically. Use the PopTools, Simulation tools, Monte Carlo analysis menu to run the model 1000 times, using the values in C30 to C34 as the dependent range and making sure you check the 'Keep results' box. This will create a new page of Monte Carlo results, containing one column of results for each of the five output values, and 1000 rows, one for each iteration.

Use the PopTools, Simulation Tools, Summary stats tool to summarise each of these columns and graph the output. You should get something similar to the histograms shown below.

The first shows the distribution of the unit sensitivity, the probability of detecting disease if just a single animal were in our surveillance system component. It ranges between 7% and 14% with a mode of around 12%.



Let's look at the third variable, the surveillance system component sensitivity. This is normally the answer that we are most interested in. It shows that the sensitivity is about 92% but most of the values are in the range from 84% to 95%.

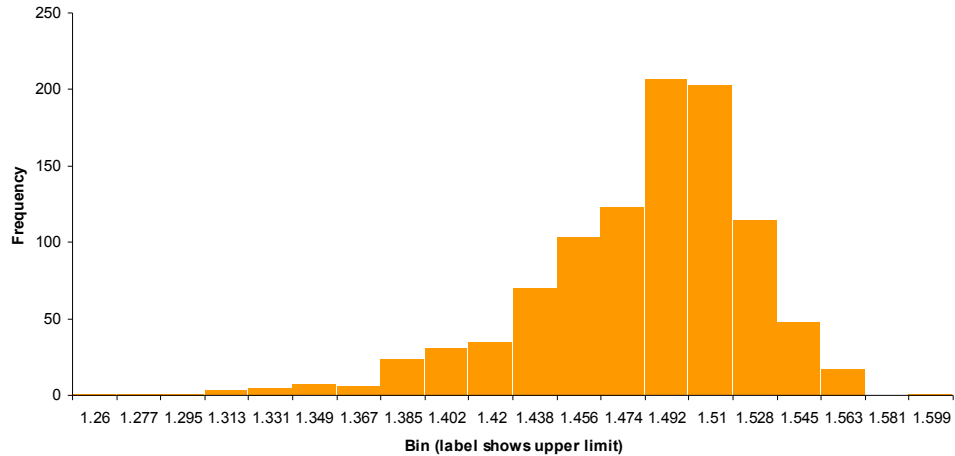


The fifth value has not yet been discussed in detail. It is often interesting to know how well surveillance is targeted. This can be done by comparing the sensitivity of the actual surveillance to a hypothetical surveillance system based on purely representative sampling. This value is known as the sensitivity ratio (*SR*)

$$\text{Sensitivity Ratio} = \frac{CSe_{\text{Actual}}}{CSe_{\text{Representative}}}$$

If the *SR* is equal to one, it means that the surveillance is as efficient as representative (e.g. random) sampling (it is well targeted). If the *SR* is greater than one, the surveillance is more efficient than representative sampling. If the *SR* is less than one, the surveillance is less efficient because it is poorly targeted or biased.

The distribution of the *SR* in our model is shown below. It is about 1.50, ranging from 1.3 to 1.56. This means that the sensitivity of the surveillance is about 1.5 times greater than representative surveillance using the same number of animals.



This example has been simplified to make it easier to understand. In reality, there are usually more risk factors to take into account, and certainly more follow-up tests before one concludes that the country or zone is infected. These normally result in a lower unit sensitivity. Also, most surveillance activities have many more than 20 animals, which means that the SSC sensitivity is often much higher.

## PopTools reference

---

These examples show that it can be quite simple to set up a stochastic model in a spreadsheet using PopTools, although larger models rapidly become more complex. This section provides some brief notes on using PopTools. See the help system and example spreadsheets for more information.

### Installation

---

To install PopTools, download the executable installation file from the PopTools web site at <http://www.poptools.org/download> and save it to your hard disk. Double click to start the installation process. In Windows Vista you may need to right-click on the file and select Run as Administrator for successful installation.

Once installed, Excel will open, with a spreadsheet containing a welcome message. To confirm that the system has been correctly installed, check to see if there is a new menu item "PopTools".

### Random variable functions

---

Some of the more commonly used distributions available in PopTools are listed below.

Function	Parameters	Distribution and notes
dBetaDev	Alpha, beta	Beta distribution (defined by alpha and beta)
dBetaMSDev	Mean, standard deviation	Beta distribution (defined by the mean and standard deviation)

Function	Parameters	Distribution and notes
dBinomialDev	Trials, probability	Binomial
dExpDev	Mean	Exponential
dHyperDev	Samples, affected, population	Hypergeometric
dLogNormalDev	Mean, standard deviation	Log normal
dNormalDev	Mean, standard deviation	Normal
dPertDev	Min, most likely, max, weight	PERT. Weight should always be set to 4 for consistency with other implementations of the PERT
dPoissonDev	Mean	Poisson
dRandInt	Lower, upper	A uniform random integer between the lower and upper bounds.
dRandReal	Lower, upper	A uniform random real number between the lower and upper bounds.
dTRand	Min, most likely, max	Triangular
DiscreteDev	Numbers, frequencies	Discrete distribution – selects a number at random from the list of numbers with a probability proportional to the frequencies.

Other less common distributions are also available including:

- Cauchy
- Correlated random variable
- Gamma
- Geometric
- Negative binomial
- Normal (integer)
- Pareto
- Weibull

### *Other useful functions*

In addition to the random variable functions, PopTools makes a large number of other specialised functions available. Many of these are designed to assist with matrix operations or statistical analysis. Some of the more useful general functions include.

- **FormulaText(ref)** – returns the formula in the referenced cell, displayed as text. This is useful for displaying how a spreadsheet works (it was used to produce the second diagram of the spreadsheet above) but can also be useful for documenting a spreadsheet.

- **F(value)** – displays the value if not an error, otherwise a blank cell. This is useful for hiding errors.
- **QSort(range)** – sorts a range of data. This is an array formula. Normal formulae only work on a single cell, but array formulae work on an array of cells. To enter an array formula:
  - Select a range of cells that the formula will occupy
  - Type the formula
  - Instead of pressing enter, press Shift-Control-Enter together
- **RandFix(true/false)** – When this is true, every random formula returns the expected value rather than a random value. This stops the model from behaving stochastically, and can be useful when checking for errors. This can also be achieved from the menus (PopTools, Fix random generator)

### Selected Menus and Dialogs

PopTools contains many features that are not directly relevant to scenario tree modelling. Some of these may be useful when calculating probability inputs for a model or for other related purposes. This section contains those features that are most likely to be of value.

#### Simulation tools menu

We have already used two of the items here – Monte Carlo simulation and Summary statistics.

#### Monte Carlo Simulation

Recalculates the current worksheet for the specified number of replicates. If the worksheet includes random variables, or a randomised range, a new result will be obtained for each replicate. The procedure counts the number of times that values in the dependent range exceed (or are less than - depending on the test criterion option) a range of test values, and also collects summary statistics.

Dependent range

Test values (optional)

Lower percentile

Upper percentile

Number of replicates

Output (choose 1 cell)

Test criterion

>  >=  <  <=  Range

Precision

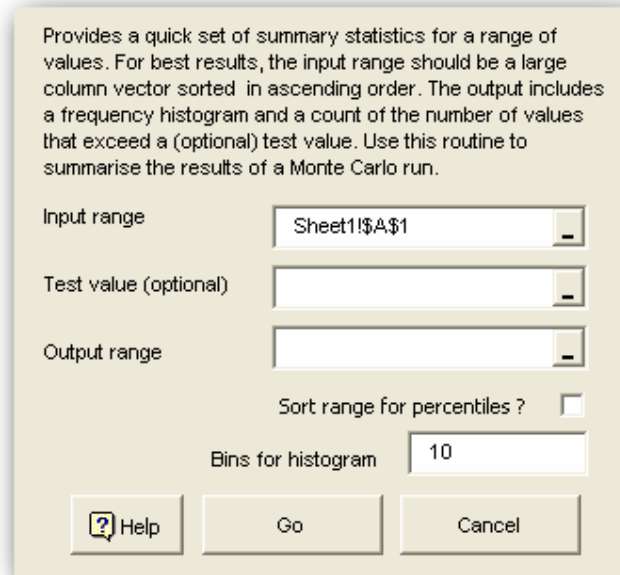
Random seed

Keep results

Colour code for demo

Inputs for the dialog have already been described on page 117.

## Summary Statistics



Inputs for the summary statistics dialog have been described on page 118

## Extra Stats menu

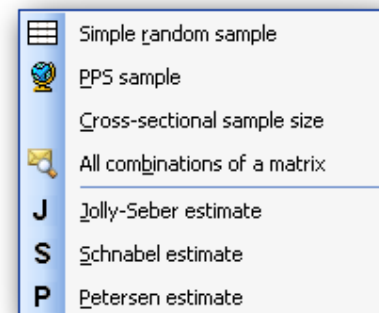
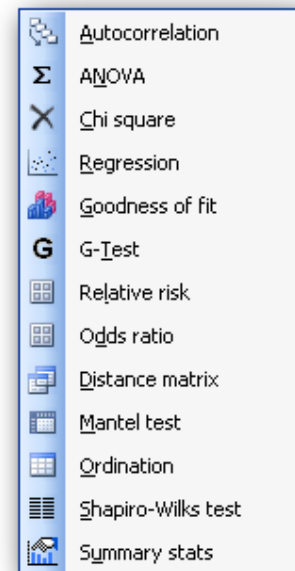
PopTools provides access to a range of statistical tools that are not readily available in Excel. The most commonly used ones for those working in disease surveillance are likely to be:

- ANOVA: Analysis of variance
- Chi square
- Regression
- Relative risk
- Odds ratio

## Sampling menu

This menu has tools for simulation of sampling and sample size calculation.

- Simple random sampling: this selects a simple random sample of values from a range of data in the spreadsheet.
- PPS sample: this selects a sample of data from the spreadsheet using probability proportional to size sampling.
- Cross-sectional sample size: Calculates the sample size for a prevalence survey.





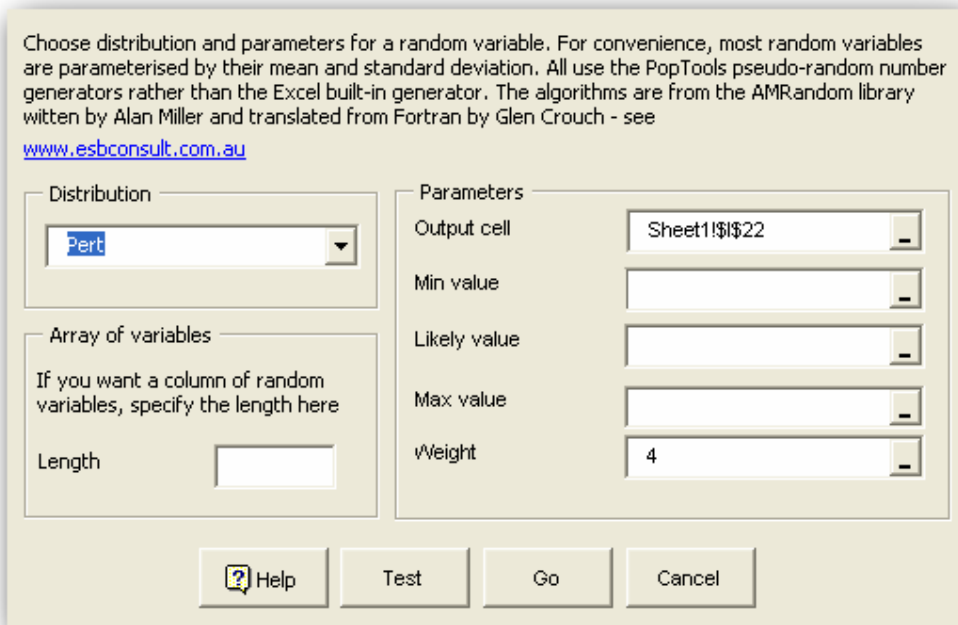
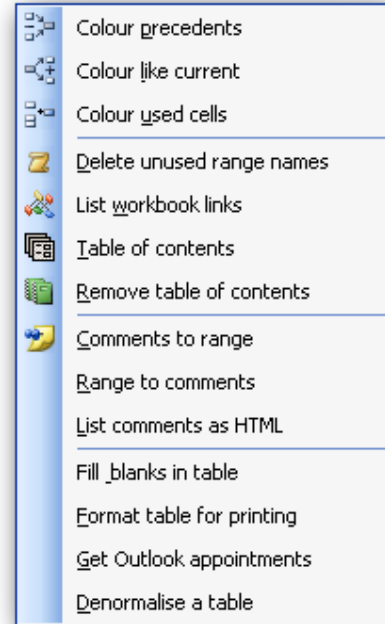
## Auditing menu

This contains a very useful list of functions for checking a spreadsheet and understanding how values are related and calculated.

## Random Variable

This dialog is a convenient way to select a random variable with a specified distribution for use in a stochastic model.

An example of entering data into this dialog was given on page 115.



# Chapter 13 – Clustering

Freedom and independence form my character.

Mustafa Kemal Ataturk (1881 - 1938)

## Clustering of disease and populations

---

Directly transmitted diseases usually form clusters. If an infected animal is brought into a population, other animals in the same herd are more likely to become infected than animals in other herds. Livestock movements or other factors may move the disease to other herds, and animals in those new herds will become infected, generating new clusters.

Most of the major diseases of livestock and directly transmitted, however some diseases do not cluster at the herd level. Vector borne diseases such as Bluetongue are not constrained by herd structures and fences, but are distributed wherever the vector is able to find a suitable habitat.

## *Lack of independence between animals*

---

The examples in previous chapters have shown how to calculate the unit sensitivity (*CSeU*, probability of detecting disease with a single animal in the surveillance system). This is then used to calculate the SSC sensitivity using the formula:

$$CSe = 1 - (1 - CSeU)^n$$

This approach assumes that animals are independent, which means that the probability of one animal being infected is not related to the probability that another animal in the same herd is infected. However, this is clearly not always true. When animals are grouped into herds, and disease clusters, if one animal is

infected, there are likely to be others that are also infected. The infection status of animals is therefore unlikely to be independent.

### Example

Consider one herd of 40 animals examined as part of a surveillance program that uses an animal-level design prevalence of 20% and a herd-level design prevalence of 5%. Before any surveillance is done in the herd we don't know if the herd is infected or not, but the herd-level design prevalence tells us that the probability that the herd is infected is 5%.

If we sample one animal and it tests negative, does this change the probability that the herd is infected?

Yes, after one negative test, we still don't know the herd status, but it is less likely to be infected. If we then test a second and a third animal, each time we get a negative result, we become more and more confident that the herd is uninfected.

After testing 20 animals from a herd of 40 animals, and all have tested negative, we can already be very confident that the disease is not present. Testing one more animal makes us more confident, but doesn't add as much new information as the first animal tested, because we are already very confident.

As the number of animals that test negative increases, the amount of new information that each new animal provides gets less and less, because we are already reasonably sure of the status of the herd.

In probability terms, the independent probability that an animal will test positive (when we have no prior test results from the herd) is different to the conditional probability that an animal will test positive, given that other animals in the same herd have already tested negative.

The formula we used to calculate the SSC sensitivity:

$$CSe = 1 - (1 - CSeU)^n$$

assumes that the new information provided by each animal is the same. However, we have just seen that if the animals come from the same herd, subsequent animals provide less new evidence. If they come from different herds, they provide more evidence.

For diseases that cluster, we need to find a different approach to calculating the SSC sensitivity, which is able to take the lack of independence between animals in the same herd into account.

### *Step-wise calculation of sensitivity*

---

We achieve this by calculating the sensitivity of the SSC in steps. First we calculate the sensitivity for each herd. Then, based on the herd-level sensitivity, we can calculate the SSC sensitivity. By treating each herd separately, we are able to take into account clustering and lack of independence between animals at the herd level.

To do this, we must know something about the herds. In the previous calculations based on the assumption of independence, the only figure that was used was  $n$ , the total number of animals in the SSC. When we take lack of independence into account, we should ideally know:

- Which herds in the population are part of the SSC
- Which animals in the SSC belong to which herd

- The size of each herd
- The risk characteristics of each of the animals
- The risk characteristics of each of the herds

This allows us to accurately calculate the herd-level sensitivity of every herd in the SSC. Where all these details are not known, a general description of the population (estimated number of herds, distribution of herd sizes, and estimated number of animals tested per herd) can be used to estimate the separate herd-level sensitivity for each herd in the SSC.

If you are using the web-based Freedom software, the details of these calculations will be handled automatically. If you are implementing the analysis in a spreadsheet, you should read the next section to better understand the approach to calculation of herd sensitivity.

## Herd level sensitivity calculation

The examples of scenario-trees shown previously are not able to take into account the lack of independence between animals within herds. A different approach is required to analyse information at the herd level.

### Spreadsheet layout example

Let us look at an example of how a spreadsheet could be organised to help with these herd-level calculations. Consider a surveillance system component for avian influenza in domestic chickens. A node list for a simplified scenario tree is provided below.

Node	Type	Branches
Adjacent to wetlands	Risk category	Yes, No
Flock type	Risk category	Commercial, Backyard
Flock infected	Infection	Infected, uninfected
Animal infected	Infection	Infected, uninfected
Initial test	Detection	Test positive, negative
Confirmatory test	Detection	Test positive, negative

As with the earlier spreadsheet example, it helps to lay out the parameters clearly so you can access them for later calculations. For simplicity, this example will not be a stochastic model. The parameters are shown below:

	A	B	C	D	E	F
1	<b>Parameters</b>	<b>Branch</b>	<b>RR</b>	<b>PrP</b>	<b>PrSSC</b>	<b>AR</b>
2	Adjacent to wetlands	Yes	1.5	0.05	0.8	1.463415
3		No	1	0.95	0.2	0.97561
4	Flock type	Backyard	4	0.9	0.95	1.081081
5		Commercial	1	0.1	0.05	0.27027
6						
7	Flock infected		0.01			
8	Animal infected		0.05			
9	Initial test		0.98			
10	Confirmatory test		0.9			
11	Combined Test Se		0.882			
12	P(Animal Pos)		0.0441			

In addition to the calculated figures for adjusted risk (*AR*, discussed in Chapter 9), two other figures have been calculated for convenience:

- the combined sensitivity of the two tests used, ( $Se_1 \times Se_2$  or, in the spreadsheet =B9 \* B10) and
- the probability that an animal will provide a positive test result ( $P_A^* \times Se$ , or =B8 \* B9 \* B10).

Instead of the model being displayed as a tree, we can use a table with one row per flock. The information on the line should include everything we need to calculate the flock sensitivity, and the probability of getting a positive test result from the flock. Our example surveillance system component only has 20 flocks, which are shown below.

	A	B	C	D	E	F	G	H
20	Flock ID	n	N	Wetland?	Backyard?	SeH	EPI	P(neg result)
21	1	23	33000	1	1	0.645606	0.015821	0.989786
22	2	25	45000	1	1	0.676174	0.015821	0.989302
23	3	16	28000	1	1	0.514041	0.015821	0.991868
24	4	12	27000	1	0	0.417964	0.003955	0.998347
25	5	10	43000	1	1	0.363022	0.015821	0.994257
26	6	28	22000	0	1	0.717155	0.010547	0.992436
27	7	6	33000	0	1	0.237087	0.010547	0.997499
28	8	17	39000	0	0	0.535472	0.002637	0.998588
29	9	10	18000	1	1	0.363022	0.015821	0.994257
30	10	28	11000	1	1	0.717155	0.015821	0.988654
31	11	15	47000	1	1	0.491622	0.015821	0.992222
32	12	14	42000	1	1	0.468168	0.015821	0.992593
33	13	25	16000	1	1	0.676174	0.015821	0.989302
34	14	24	41000	1	1	0.661235	0.015821	0.989539
35	15	10	48000	1	1	0.363022	0.015821	0.994257
36	16	14	36000	0	1	0.468168	0.010547	0.995062
37	17	22	39000	1	1	0.629256	0.015821	0.990045
38	18	18	28000	1	1	0.555958	0.015821	0.991204
39	19	11	14000	1	1	0.391112	0.015821	0.993812
40	20	30	31000	0	1	0.741552	0.010547	0.992179

The data columns in this table are:

- The flock ID is a unique number to identify each flock.
- n is the number of animals tested from the flock
- N is the total number of birds in the flock
- Wetland: 1 means that the flock is near a wetland, and 0 means that it is not. Most flocks are near a wetland as the surveillance targeted these flocks
- Backyard: 1 means that the flock is a backyard flock, and 0 means that it is a commercial flock

In general terms, these columns should include a herd or flock identifier, one column for each herd-level risk factors indicating the branch that the herd falls into, the total number of animals in the herd and the number of animals tested. This structure can be further extended to include different risk groups for animals within a herd, with an extra row for each risk group.

In addition to the data columns, the table also includes three calculated columns:

F) SeH is the flock-level sensitivity that has been achieved by testing the  $n$  animals in the flock. The different ways of calculating the herd- or flock-level sensitivity are discussed in detail in the next section. This example uses the binomial formula that we are already familiar with.

$$SeH = 1 - (1 - P_A^* \times SeA)^n$$

This is included in the spreadsheet as:

**=1 - (1-B12) ^B21**

G) EPI or the effective probability of infection. This captures the adjusted risk values for the two risk category nodes and uses them to adjust the flock-level design prevalence.

$$EPI = P_H^* \times AR_1 \times AR_2$$

This can be extended for as many risk category nodes as are required. The spreadsheet formula is:

**=IF(D21, \$F\$2, \$F\$3) \* IF(E21, \$F\$4, \$F\$5) \* \$B\$7**

In this formula, **IF()** statements are used to select the correct adjusted risk values for the flock's risk group. The first checks the value of D21 to see if the flock is near a wetland or not, and chooses the appropriate adjusted risk.

H) P(Neg result) is the probability that the flock will have all negative results from the testing. This is one minus the probability of getting at least one positive result. This, in turn, is the probability that the flock is infected (the EPIH) times the probability of it being detected if it is infected (the flock-level sensitivity, SeH).

$$Pr(\text{negative herd result}) = 1 - (EPIH \times SeH)$$

Or, in the spreadsheet:

**=1 - (F21 \* G21)**

Once the probability that each flock will produced a negative result in the surveillance has been calculated, we can calculate the SSC sensitivity. This is the probability that at least one flock will give a positive result, or one minus the probability that all will give negative results. To calculate the probability that all flocks will give negative results, we simply multiply together the probabilities of negative results for each flock. The general formula is:

$$SSCSe = 1 - \prod_{i=1}^I (1 - EPIH_i \times SeH_i)$$

and in the spreadsheet

**=1 - (PRODUCT(H21:H40))**

In our example, this gives us an estimated component sensitivity of 13.5%. For comparison purposes, we can analyse the same component using the scenario tree and assuming independence between animals. The spreadsheet implementation of the scenario tree is shown below.

Wetland		Type		Flock infected		Animal Infected		Test results			Probability		
Branch	PrSSC	Branch	PrSSC	Branch	EPI	Branch	P*A	Branch	Se	Outcome			
Yes	0.8000	Backyard	0.9500	Yes	0.0158	Yes	0.0500	Pos	0.8820	Pos	0.0005		
								Neg	0.1180	Neg	0.0001		
						No	0.9500	Pos	0	Pos	0.0000		
						Neg	1.0000	Neg	0.0114				
				No	0.9842	Yes	0.0500	Pos	0.0000	Pos	0.0000		
										Neg	1.0000	Neg	0.0374
		No	0.9500					Pos	0	Pos	0.0000		
								Neg	1.0000	Neg	0.7106		
				Commercial	0.0500	Yes	0.0040	Yes	0.0500	Pos	0.8820	Pos	0.0000
										Neg	0.1180	Neg	0.0000
								No	0.9500	Pos	0	Pos	0.0000
								Neg	1.0000	Neg	0.0002		
No	0.9960					Yes	0.0500	Pos	0.0000	Pos	0.0000		
										Neg	1.0000	Neg	0.0020
				No	0.9500			Pos	0	Pos	0.0000		
								Neg	1.0000	Neg	0.0378		
No	0.2000			Backyard	0.9500	Yes	0.0105	Yes	0.0500	Pos	0.8820	Pos	0.0001
										Neg	0.1180	Neg	0.0000
								No	0.9500	Pos	0	Pos	0.0000
										Neg	1.0000	Neg	0.0019
		No	0.9895			Yes	0.0500	Pos	0.0000	Pos	0.0000		
										Neg	1.0000	Neg	0.0094
				No	0.9500			Pos	0	Pos	0.0000		
						Neg	1.0000	Neg	0.1786				
		Commercial	0.0500	Yes	0.0026	Yes	0.0500	Pos	0.8820	Pos	0.0000		
								Neg	0.1180	Neg	0.0000		
						No	0.9500	Pos	0	Pos	0.0000		
						Neg	1.0000	Neg	0.0000				
No	0.9974			Yes	0.0500	Pos	0.0000	Pos	0.0000				
								Neg	1.0000	Neg	0.0005		
		No	0.9500			Pos	0	Pos	0.0000				
						Neg	1.0000	Neg	0.0095				

Check 1.0000  
 CSeU 0.000627  
 CSe (independent) 0.201045

As can be seen from this analysis, the component sensitivity, when assuming independence is 20.1%, compared to 13.5% when we take clustering into account. Failing to account for lack of independence among animals will mean that the sensitivity is overestimated. This is because we assume that we are getting the same amount of new information from every animal that is tested. However, if we test many animals from the same herd, we are getting less and less information from each new animal, so our overall sensitivity is lower.

The size of the difference between sensitivity when we take lack of independence into account and when we don't depends on whether the collection of the surveillance data is also clustered. If there have been many animals taken from a small number of herds, it will make a big difference. If there have only been a few animals taken from many different herds, then accounting for lack of independence may make almost no difference at all.

### Herd-level sensitivity formulae

The approach to calculating the sensitivity at the herd level varies slightly depending on the size of the herd and the number of animals tested from each herd. This section discusses the various options. The principles described here can be extended to higher grouping levels if they exist (for example, there may be three levels of infection nodes for intensive animal production – animal, house and farm), or to the entire surveillance system component.

#### **Small proportion of herd tested**

When the proportion of the herd that is tested is small, it is reasonable to assume that sampling without replacement does not significantly change the probability that the next animal selected will be infected animal. In this case, we can use the simpler binomial formula to estimate the herd-level sensitivity. If there are no animal-level risk nodes, the formula is:

$$SeH_h = 1 - (1 - P_A^* \times SeA)^{n_h}$$

Where:

- $SeH_h$  is the herd level sensitivity for the  $h^{th}$  herd
- $P_A^*$  is the design prevalence at the animal level
- $SeA$  is the animal-level sensitivity
- $n_h$  is the number of animals tested from the  $h^{th}$  herd

If there is one or more risk category nodes related to the animal-level infection node, there will be a number of different groups of animals within the herd with different risks of being infected. In this case, the group-level sensitivities are separately calculated and multiplied together to give the herd-level sensitivity, using the following formula which assumes  $J$  different risk-groups of animals within the herd.

$$SeH_h = 1 - \prod_{j=1}^J (1 - EPIA_j \times SeA_j)^{n_j}$$

#### **Large proportion of herd tested**

When the proportion of the herd that is tested is large, the assumptions of the binomial formula used above are no longer valid. Instead, it is more appropriate to use the binomial approximation to the hypergeometric distribution to calculate sensitivity. The formula for the calculation is:

$$SeH_h = 1 - (1 - SeA_{Av} \times \frac{n_h}{N_h})^{EPIA \times N_h}$$

Where:

- $SeA_{Av}$  is the average animal-level sensitivity for the herd



- EPIA is the effective probability of infection of animals in the herd
- $N_h$  is the total number of animals in the herd, and
- $n_h$  is the number of animals tested from the herd

### **All animals tested**

When the entire herd is tested, if there are infected animals in the herd, we can guarantee that those animals will be tested. The herd level sensitivity is therefore based on the animal-level sensitivity and the number of infected animals in the herd (the animal-level design prevalence). The formula is:

$$SeH_h = 1 - (1 - SeA_{Av})^{d_h}$$

Where:

- $d_h$  is the number of infected animals in the herd,  $EPIA \times N_h$  rounded up to an integer

# Chapter 14 – Combining Multiple Surveillance Components

There is only one way in which a person acquires a new idea; by combination or association of two or more ideas he already has into a new juxtaposition in such a manner as to discover a relationship among them of which he was not previously aware.

Francis A. Carter

Scenario trees give us the capacity to analyse complex risk-based surveillance to estimate its sensitivity. Normally, a surveillance system is made of a number of different components that provide different types of evidence that the disease is not present, or different approaches for the early detection of a disease.

## Example

Consider the possible sources of evidence for bovine tuberculosis status.

- Routine herd tuberculin tests
- Movement or export testing
- Abattoir meat inspection for granulomas
- Passive clinical surveillance
- Human health surveillance detecting *M. bovis*

Each of these systems is a component of an overall surveillance system, and each has a different capacity to detect the presence of the disease. Using scenario trees we are able to estimate the sensitivity of each of the components, which is useful. However, we are also interested in the system as a whole – what is the sensitivity of all our surveillance, considering each of the components together?

This chapter presents techniques to combine different components of a surveillance system to answer this question.

## Simple example

Sensitivities are probabilities, and by this stage, we are already experts at combining probabilities. If we have two surveillance components, each with their own sensitivity, we can use probability theory to combine them into a single sensitivity.

### Example

Consider a surveillance system with two components – component 1 is a structured survey and component 2 is abattoir meat inspection. We have analysed each component and calculated the sensitivity:

$$CSe_1 = 82\%$$

$$CSe_2 = 45\%$$

Component sensitivity is the probability that we would detect the disease using that component. The combined surveillance system sensitivity ( $SSe$ ) is the probability that we would detect disease in at least one of the components. Using logic that is now familiar, this can be calculated as the one minus the probability of not detecting disease in any of the components, giving us the simple formula for the sensitivity of the surveillance system:

$$SSe = 1 - ((1 - CSe_1) \times (1 - CSe_2))$$

If we are combining  $I$  different components, this can be generalised to:

$$SSe = 1 - \prod_{i=1}^I (1 - CSe_i)$$

For this example the combined sensitivity would be:

$$\begin{aligned} SSe &= 1 - (1 - 0.82) \times (1 - 0.45) \\ &= 0.901 \end{aligned}$$

This approach is simple, and is appropriate for combining some components in a surveillance system. Unfortunately, there are many situations where this approach over-estimates the combined sensitivity due to overlapping of surveillance components.

## Overlapping surveillance components

---

In the previous example we had two components – structured surveillance and abattoir meat inspection. Some herds are beef herds and the primary purpose of production is to sell animals for slaughter, so many animals go through the abattoir. Other farms are dairy farms, or may be breeding farms, and therefore send very few animals to the abattoir. Surveillance sensitivity for the abattoir component is therefore better in some farms than others.

For the structured survey, some farms are selected and some farms are not – the sensitivity in those farms that are not selected is zero, but we collect useful information from those farms that are selected.

The problem occurs when farms are present in both surveillance components. If a farm is a beef farm, and sends lots of animals to the abattoir, then it will make a significant contribution to the overall sensitivity of the abattoir surveillance system. However, if it is also selected in the structured survey, it is also contributing information to the sensitivity of that component. The difficulty is that the information the farm provides is not new information. If the farm has been submitting animals to the abattoir all year and they all test negative, this provides quite a bit of evidence that the farm is not infected. Testing the farm again as part of the structured surveillance doesn't provide as much new information as if we had tested a farm that doesn't send any animals to the abattoir.

Where herds are included in more than one component of a surveillance system, the components are not independent. It is necessary to take this lack of independence into account when analysing the data, or we will overestimate the combined sensitivity of the system.

## Accounting for the overlap

---

In Chapter 13 we looked at analysing a scenario tree herd-by-herd to take clustering into account. This approach gives us an opportunity to account for the overlap between surveillance system components at the herd level as well.

Normally, when we analyse a single herd in a SSC, we have a number of pieces of information:

- The probability that the herd is infected, (from the design prevalence) which is the same for all herds
- The risk factors applying to that herd, expressed in terms of the adjusted risk
- The sensitivity for that herd, based on the number of animals tested and the animal-level sensitivity.

The result of the analysis is an estimation of the probability that the herd will give a negative result. We can also estimate the probability, after testing, that the herd is infected. This can be done using Bayes' theorem, as discussed in Chapter 2. The prior probability of being infected is given by the design prevalence, the new information is the surveillance that has been carried out, and the posterior probability says how likely the herd is to be infected based on the surveillance. As all our surveillance results are negative, the posterior probability will be lower than the prior (the design prevalence).

This posterior is a description of our state of knowledge about the herd after the surveillance has been done. We may have assumed that all herds had the same prior probability of being infected, but after testing some of the herds, we are

**Bayes' theorem is used to calculate the posterior probability of infection**

**The prior for one component is the posterior of the previous**

more confident that they are free. In contrast, we have no information about the untested herds, so we must still assume that the probability that they are infected is equal to the design prevalence.

When it comes to the next component of the surveillance system, normally it is analysed in the same way. However, instead of starting with the assumption that we know nothing about the state of the herds (and therefore using the design prevalence as an estimate of the probability that each herd is infected), if we have already done some surveillance, we now have better information about the state of some herds. Those herds that have already been tested with negative results have a lower probability of being infected than the untested herds. We can start our analysis of the second component by using the updated information about herd status, based on the results of the first component. In Bayesian terms, this means that instead of using the design prevalence as our prior for the probability that each herd is infected, we use the posterior estimate from the first surveillance component.

If three or more components are analysed, this chain can continue. For each component, the prior probability that each herd is infected is the posterior probability from the analysis of the previous component.

When we use this approach, herds that have been tested in another component will have a lower prior probability of infection, which means that we will be less likely to find any infection in that herd, and the contribution that that herd makes to the component sensitivity will be less. Herds that have not been previously tested will continue to have the design prevalence as their prior probability of infection, and so will contribute the same amount to the component sensitivity as if we had analysed the component independently.

To calculate the system sensitivity, we use the same approach presented in the simple example at the start of this chapter to combine the sensitivity of each component. This now gives a more accurate measure of system sensitivity because the contribution of the overlapping herds has already been removed, so the components may now be considered independent.

### Spreadsheet example

If you use the web-based Freedom software, and provide herd level data that allows the herds to be matched between surveillance system components, the software is able to take overlap between components into account and calculate the combined sensitivity.

If you are using a spreadsheet, the easiest approach is to include all the surveillance system components on the same sheet, one next to the other, as shown in the example below.

The process (referenced by columns) involves:

- A) Identify every herd uniquely. Every herd that appears in any component of the surveillance system should be listed.

For each component:

- B) The prior probability that the herd is infected. For the first component analysed, this is the herd-level design prevalence (B2). For the subsequent components, this is the posterior probability that the herd is infected from the previous component (see columns H and N)
- C) The number of animals tested. If the herd was not included in the component, then the number of animals tested is zero

- D) The herd-level sensitivity, as calculated in previous examples. This example has been simplified so no animal or herd level risk factors have been included. The herd-level sensitivity in D10 is therefore:  

$$=1 - (1 - B5)^{C10}$$
- E) The effective probability of infection, as calculated in previous examples. This simple spreadsheet has no risk factors, so this is equal to the prior probability that the herd is infected (B10).
- F) The probability that the herd will have a negative results in the surveillance, as calculated in previous examples. This is one minus the probability that it is infected (the EPI) times the probability that the infection will be detected (the herd-level sensitivity):  

$$=1 - (D10 * E10)$$

Spreadsheet 1: Accounting for overlap between surveillance system components using Bayesian revision

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
1	<b>Parameters</b>			<b>Result</b>															
2	P*H	0.05		SSe	0.2505														
3	P*A	0.2																	
4	SeA	0.56																	
5	CSe																		
6	U	0.112																	
7	<b>Calculations</b>																		
8																			
9	<b>ID</b>																		
10	1	0.05	25	0.949	0.05	0.003	0.003	10	0.695	0.003	0.998	0.001	0.001	0	0.000	0.001	1.000	0.001	
11	2	0.05	0	0	0.05	0.050	0.050	14	0.810	0.050	0.959	0.010	0.010	19	0.895	0.010	0.991	0.001	
12	3	0.05	0	0	0.05	0.050	0.050	0	0.000	0.050	1.000	0.050	0.050	20	0.907	0.050	0.955	0.005	
13	4	0.05	10	0.695	0.05	0.016	0.016	3	0.300	0.016	0.995	0.011	0.011	23	0.935	0.011	0.990	0.001	
14	5	0.05	28	0.964	0.05	0.002	0.002	8	0.613	0.002	0.999	0.001	0.001	0	0.000	0.001	1.000	0.001	
15	6	0.05	0	0	0.05	0.050	0.050	12	0.760	0.050	0.962	0.012	0.012	1	0.112	0.012	0.999	0.011	
16																			
17				CSe1	0.1249					CSe2	0.084114					CSe3	0.0649		

Spreadsheet 2: The same calculations without accounting for overlap, using a constant prior probability of infection

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	
7	<b>Calculations</b>																		
8																			
9	<b>ID</b>																		
10	1	0.05	25	0.949	0.05	0.003	0.003	10	0.695	0.050	0.965	0.016	0.016	0	0.000	0.050	1.000	0.050	
11	2	0.05	0	0	0.05	0.050	0.050	14	0.810	0.050	0.959	0.010	0.010	19	0.895	0.050	0.955	0.005	
12	3	0.05	0	0	0.05	0.050	0.050	0	0.000	0.050	1.000	0.050	0.050	20	0.907	0.050	0.955	0.005	
13	4	0.05	10	0.695	0.05	0.016	0.016	3	0.300	0.050	0.985	0.036	0.036	23	0.935	0.050	0.953	0.003	
14	5	0.05	28	0.964	0.05	0.002	0.002	8	0.613	0.050	0.969	0.020	0.020	0	0.000	0.050	1.000	0.050	
15	6	0.05	0	0	0.05	0.050	0.050	12	0.760	0.050	0.962	0.012	0.012	1	0.112	0.050	0.994	0.045	
16																			
17				CSe1	0.1249					CSe2	0.149311					CSe3	0.1356		

G) The posterior probability that the herd is infected. This is an application of Bayes' theorem, analogous to the negative predictive value in clinical testing. Bayes' theorem was discussed on page 20. This is calculated as:

$$\Pr(D+ | T-) = 1 - \frac{(1-P)Sp}{(1-P)Sp + P(1-Se)}$$

When Sp is equal to 1 this simplifies to:

$$\Pr(D+ | T-) = 1 - \frac{1-P}{1-P \times Se}$$

Or in the spreadsheet (G10):

$$=1 - (1-E10) / (1-E10 * D10)$$

The component sensitivities are calculated in cells F17, L17 and R17 as:

$$=1 - \text{PRODUCT}(F10:F15)$$

The surveillance system (combined) sensitivity is calculated in cell F2 as:

$$=1 - (1-F17) * (1-L17) * (1-R17)$$

The system sensitivity is 25.05%. For comparison purposes, the second sheet shows the same calculations performed without taking into account the overlap between the components. This is done by setting the prior probability that each herd is infected to be the design prevalence for all three components (rather than the posterior from the previous component). The system sensitivity in this case is 35%.

In this case, there was significant overlap between the components, so the analysis resulted in a significant decrease in the estimate of system sensitivity. If there were less overlap, the difference would be less.

Compare the component sensitivities between the two approaches. For component 1, there is no difference, as in both cases, the analysis used the herd-level design prevalence as the prior. However, the sensitivities for component 2 and 3 are progressively lower. Using this approach we sometimes find that those components analysed last contribute almost nothing, as there is almost complete overlap with previously analysed components.

Changing the order of components in the analysis will not change the final estimate, but will change the sensitivity of the individual components.



# Chapter 15 – Probability of Freedom

If there be two subsequent events, the probability of the 2<sup>d</sup> b/N and the probability of both together P/N, and it being 1<sup>st</sup> discovered that the 2<sup>d</sup> event has also happened, the probability I am right is P/b.

Reverend Thomas Bayes (1702 – 1761)  
[first formulation of Bayes' theorem in

*Essay towards solving a Problem in the Doctrine of Chances*]

## Sensitivity versus freedom

---

The main measure of the quality of surveillance to demonstrate freedom or for early detection of disease is the sensitivity of the surveillance. Scenario tree analysis allows us to estimate the sensitivity of complex surveillance systems, and Chapter 14 introduced approaches that enable us to combine the sensitivity of multiple surveillance components.

Surveillance sensitivity and design prevalence are commonly used as standards for surveillance. Sensitivity is therefore a useful measure, but the concepts of sensitivity are sometimes difficult to communicate, especially to people with a non-technical background.

When dealing with freedom from infection, the first question that may be asked is “Is the country free from infection?” or “How confident are we that we are free?” The analyses that have been used thus far in this book have attempted to answer that question by estimating sensitivity. To put the answer into words “The probability that the surveillance would be able to detect infection, assuming that the population is infected at a level specified by the design prevalence, is

Sensitivity is hard to understand for non-technical people

Probability of freedom is easier to understand

x%.” For a non-technical person, this answer may be very hard to interpret. There seems to be a paradox: the question is about the country being free, and the answer assumes that the country is infected.

Most people would find it much easier to understand an answer expressed in terms of the probability of freedom. For instance: “The probability that the country is free from infection is x%.” Communicating the results of analysis in these terms make it much easier for people to understand what is being said, and has a number of other benefits as well.

## Calculation of the probability of freedom from infection

Surveillance sensitivity is a conditional probability – the probability that the surveillance system would find the disease, given that the country is infected at a specified design prevalence. In probability terms this can be written as:

$$P(T+ | D+)$$

Where:

- T+ stands for test positive, or the surveillance produces a positive outcomes, and
- D+ stands for disease positive, or the country is truly infected at a specified level.

The probability of freedom can also be expressed in these terms:

$$P(D- | T-)$$

or the probability that the country is free from infection (D-) given that the surveillance has not produced a positive result (T-).

When you compare these two probability statements, there are a couple of important things to notice:

- They are not the same. Sensitivity cannot be interpreted as probability of freedom.
- The conditionality is reversed. Sensitivity is conditional on the population being infected, while probability of freedom is conditional on negative surveillance results.
- The probability of freedom at the country level looks rather like a negative predictive value at the animal level.

Probability of freedom is analogous to the negative predictive value of a test

The negative predictive value of a diagnostic test is the probability that an animal is truly negative, given that we got a negative test result. At the country level, we are interested in the probability that the country is truly negative, given that we got negative results from our surveillance.

Just as with predictive values, we can use Bayes’ theorem (page 29) to calculate the probability of freedom:

$$P(\text{free}) = \frac{\text{True Negative}}{\text{True Negative} + \text{False Negative}}$$

$$= \frac{(1 - P) \times Sp}{(1 - P) \times Sp + P \times (1 - Se)}$$

Where:

- Sp and Se are the sensitivity and specificity of the surveillance system, and
- P is the *prior* probability that the country was infected.

If the specificity of the surveillance system is 100%, this simplifies to:

$$P(\text{free}) = \frac{1 - P}{1 - (P \times Se)}$$

## Selecting a prior

In the formula for negative predictive value used for individual animal diagnostic testing, the prior probability of infection is estimated by the prevalence of the disease. However, when we are working at the country level, how do we know the value for the prior?

One possibility would be the prevalence at a country level. Consider the population to be all countries in the world, and the prevalence is the proportion of countries that are infected. Most would agree that this is not a reasonable estimate of the probability that any particular country is infected, especially one that is claiming to be free, as geographic region and biosecurity play an important role in a country's disease status.

Instead, the prior should reflect the county's particular situation. If one believes that the country is free, and then undertakes surveillance to help support this claim, then the prior probability that the country is infected should be quite low.

### Example

We have undertaken surveillance that has a sensitivity of 75%. If we think that the prior probability of infection was 10% (which means that the prior probability of freedom was 90%), our posterior probability of freedom will be 97.3%.

If we had chosen a prior of only 20% probability of being free, our posterior (with the same surveillance sensitivity) would have been 50% instead of 97.3%.

When reporting the probability that a country is free from infection based on surveillance, the choice of the prior can make a very big difference. If a country wishes to indicate that it is very likely to be free, it will choose a high prior probability of being free. However, trading partners may not agree with such an optimistic view of things, and may challenge the prior.

In this way, the prior is similar to the design prevalence. It has a big impact on the result, but is difficult to choose objectively. In order to avoid disagreements about the interpretation of the analysis of surveillance, there needs to be an objective and mutually agreeable method of selecting a prior.

One useful approach is to base the prior on previous information. The prior describes the probability of being free *before* the current surveillance was done, but it can be based on earlier surveillance. This means that the value for the prior for the current year can be justified by analysis of last year's surveillance data. This solution looks good until one asks what the prior should be for the analysis of last year's surveillance. The obvious answer is to base it on an analysis of the year before. And so on.

Regardless of how many years of surveillance data are available, there will always be a starting point where there is either no previous surveillance, or it was known that the country was infected. In both cases, we are still left with the question of what value we should use for the prior at the beginning of our surveillance to demonstrate freedom from infection.

If the surveillance has started after an eradication program, it means that in the year before, cases were detected, so the probability of infection was 100% and the probability of freedom was 0%. In the following year, no cases were detected so the country may be free. The simplest approach to address the concerns of the stakeholders is to use a standard 'compromise' value for the prior probability of freedom before the first round of surveillance. This value is 50%.

The advantage of this approach is that, while the prior has a big impact on the posterior in any one year, for a series of analyses in which the posterior of one is used as the prior of the next, the starting prior quickly loses its influence on the result. This means that, as long as the surveillance is reasonably sensitive, the probability of freedom after, say 5 years of analysis, is reasonably independent of the starting prior probability of infection being present. For example, with the surveillance described above, the posterior probability of freedom would be greater than 99.5% after 5 years, regardless of whether a prior probability of infection of 10%, 50% or 80% had been used.

# Chapter 16 – Incorporating Historical Surveillance Data

History is the witness that testifies to the passing of time; it illumines reality, vitalizes memory, provides guidance in daily life and brings us tidings of antiquity.

**Cicero (106 BC - 43 BC)**

The previous chapter suggested that the best way to choose a suitable prior probability of freedom when calculating the current probability of freedom is to use the posterior probability from the previous year or years. This approach opens up the possibility of incorporating historical data into our analysis of surveillance.

## Value of historical data

---

For many diseases there is often a certain amount of historical surveillance data that can provide evidence for freedom from infection. Passive clinical surveillance with an absence of reports consistent with the infection is often available, but other types of surveillance may also exist.

**Surveillance data loses value as it ages**

Clearly current surveillance is useful. Surveillance from last year is probably useful as well. But is surveillance that was done 20 years ago relevant to the current disease situation? Most would consider such old information to be irrelevant. This demonstrates the principle that surveillance data loses its value as it gets older. It would not be valid to analyse surveillance done 20 years ago and claim that it has the same value in demonstrating current freedom.

On the other hand, we would probably be more confident about the status of a country that has undertaken surveillance every year for the last 20 years with consistently negative results, compared to one that has only done surveillance for

the first time during the current year. Historical data has some value, which can accumulate over time, but this value decreases with age.

### Example

What is it that makes old surveillance data less valuable? Consider the example of a single farm and a hypothetical disease that can only be introduced into the farm through the introduction of live animals. The farm undertook detailed surveillance of all its animals 20 years ago and demonstrated, to a very high level of confidence, that the disease was not present.

If the farm is a closed herd, breeds its own replacements and never introduces animals from outside the herd, there has been no opportunity to introduce the infection from outside. The confidence in freedom is the same today as it was 20 years ago as there is no way the farm could have lost its free status.

On the other hand, if the farm sells 20% of its animals every year, and buys in a further 20% to replace them, there has been a constant risk of introducing new disease. The surveillance from 20 years ago tells us nothing about the current status.

The reason for the decrease in value of historical data is the risk of introduction of new disease that would change the free status of the population. Where the risk of introduction of disease is small, historical data retains more of its value. Where it is great, the value quickly vanishes.

Remember the analogy of the scales that was used on page 63. This showed that the probability that a population is free from infection is a balance between the surveillance evidence that accumulates over time and the risk of introduction of new disease into the herd. When the surveillance evidence outweighs the risk of introduction, the evidence for freedom accumulates and the probability rises towards 100%. When the risk of introduction outweighs the surveillance, the probability of freedom decreases towards 0%.

## Risk of introduction

Risk analysis is used to estimate the probability of introduction of infection

The risk of introduction of disease is measured as the probability that disease will enter the herd during a certain time period (the time period of analysis).

The good news is that there is a well defined methodology that enables us to estimate the risk of introduction, and that many of the techniques used are very similar to the techniques that have been discussed in this book. The methodology is quantitative risk analysis.

The bad news is that, as with the creation of a scenario tree model, performing a thorough quantitative risk analysis can be a very challenging and time consuming task. In the ideal situation, a risk analysis has already been undertaken, and the results can be used directly. Often this is not the case and either a detailed risk analysis is required, or a quick simple risk analysis can be undertaken.

Risk analysis methodology has been extensively described and is beyond the scope of this book. We will limit ourselves to the following comments:

- A full risk analysis involves a number of steps including hazard identification, risk assessment, and risk mitigation. For the purposes of assessing the probability of introduction of infection, only a small

part of this overall process is required – the release and exposure assessment.

- This is based on the use of risk pathway diagrams (similar to a scenario tree) which describe the steps that must occur for infection to move from the population of origin and become established in the target population.
- Quantitative risk assessment is based on the multiplication of probabilities for each pathway, as has been discussed for scenario trees.
- Uncertainty can be incorporated into risk pathway models using the same approach as that taken for scenario tree models. Input probabilities are described as distributions and stochastic modelling generates an output distribution. The output distribution can then be used in the scenario tree model. For quick, simple risk pathway analyses, it is important to be realistic about uncertainty and to use stochastic modelling.

## Calculation of posterior probability of freedom

---

If you are analysing historical surveillance data using the web-based software, the system is able to perform the calculations automatically. You need to provide the following information:

- The date of each surveillance observation (so the data can be divided into multiple periods)
- The length and number of periods to be analysed
- The probability of introduction of infection during each period

Calculations using a spreadsheet are illustrated below.

### *Time period of analysis*

---

Any analysis of surveillance data requires a definition of the time period over which data is analysed. If the surveillance is based on a time-limited activity (such as a structured survey), then the time period is the duration of the activity. However, much surveillance is ongoing or sporadic. For instance, abattoir meat inspection surveillance is happening every day so there is a constant stream of data available.

One of the most important factors influencing the sensitivity of surveillance is the number of animals that pass through the surveillance system. When there is an ongoing stream of surveillance data, it must be divided into discrete time periods to analyse it. The longer the time period for analysis, the greater the number of animals that will be included in the analysis. Choosing a long time period therefore increases the apparent sensitivity, while short time periods have a lower sensitivity.

While it may be tempting to use a long time period to give a higher sensitivity, this is rarely the best approach. Consider analysis of foot and mouth disease (FMD) surveillance data for a period of five years. It could be analysed as a single dataset for the five years, five time periods or one year each, or many one month time periods. The sensitivity per month would be very much lower than the 5-year sensitivity. However, as this chapter has shown, it is possible to combine the monthly data together to generate an overall probability of freedom. Normally, the estimate based on combined short time periods will be a little lower

than an estimate based on analysis of all the data as one time period, but they will be similar.

The reason for the difference is that analysing the data as short time periods allows the value of the older data to be discounted according to the risk of introduction of infection. When analysing the data as a single time period, the data five years ago and the data yesterday are both treated as if they have the same value.

It is therefore better to analyse the data as multiple relatively short time periods rather than a single long time period. But how long should these short time periods be? The answer depends largely on the nature of the disease. For rapidly spreading diseases with short incubation periods such as FMD, newly introduced disease can spread and reach the design prevalence very quickly. A short period of analysis is appropriate – normally a month but it could be as short as a week. For slow diseases such as tuberculosis or bovine spongiform encephalopathy a period of analysis of a year is usually used.

### Spreadsheet implementation

Incorporation of historical surveillance data involves the repeated calculation of the probability of freedom starting with the earliest time period, using Bayes' theorem. For the each period, the posterior probability of infection for the previous period is used as the prior probability of infection for the current period. To account for the risk of introduction of disease, the posterior probability of infection is adjusted to account for the possibility that disease may have been introduced during that period.

The calculation is based on the surveillance system sensitivity and the risk of introduction for each time period. The spreadsheet below illustrates how the calculations can be set up.

	A	B	C	D	E	F	G	H	I
	Time period	SSSe	P(intro)	Prior P(inf)	Prior P(free)	Post P(free)	Post P(inf)	Post P(inf) Adjusted	Post P(free) adjusted
1	0	0.459	0.043	0.5	0.5	0.648929	0.351071	0.378975	0.621025
2	1	0.597	0.03	0.378975	0.621025	0.802615	0.197385	0.221463	0.778537
3	2	0.634	0.044	0.221463	0.778537	0.905705	0.094295	0.134146	0.865854
4	3	0.699	0.018	0.134146	0.865854	0.955444	0.044556	0.061754	0.938246
5	4	0.761	0.023	0.061754	0.938246	0.984513	0.015487	0.038131	0.961869
6	5	0.411	0.023	0.038131	0.961869	0.977183	0.022817	0.045292	0.954708
7	6	0.638	0.014	0.045292	0.954708	0.983116	0.016884	0.030647	0.969353
8	7	0.485	0.01	0.030647	0.969353	0.983979	0.016021	0.025861	0.974139
9	8	0.616	0.02	0.025861	0.974139	0.989909	0.010091	0.02989	0.97011
10	9	0.405	0.012	0.02989	0.97011	0.981998	0.018002	0.029786	0.970214
11	10	0.509	0.043	0.029786	0.970214	0.98515	0.01485	0.057212	0.942788
12	11	0.451	0.017	0.057212	0.942788	0.967759	0.032241	0.048693	0.951307
13	12	0.553	0.032	0.048693	0.951307	0.977632	0.022368	0.053652	0.946348
14	13	0.536	0.049	0.053652	0.946348	0.974368	0.025632	0.073376	0.926624
15	14	0.417	0.032	0.073376	0.926624	0.955872	0.044128	0.074716	0.925284
16	15	0.513	0.011	0.074716	0.925284	0.962163	0.037837	0.048421	0.951579

The columns and formulae are as follows:

- A. The period number. Periods can be any length, but the values in B and C must be calculated based on the selected time period
- B. The surveillance system sensitivity, based on the analysis of one or more components



- C. The probability of introduction of disease, based on published data or risk analysis.
- D. The prior probability that the country is infected. For the first time period, this is set at a standard starting value of 0.5. For subsequent time periods, it is the adjusted posterior probability of infection for the previous time period. For example, cell D3 contains **=I2**
- E. The prior probability of freedom. This is just one minus the prior probability of infection. Cell E2 contains **=1-D2**
- F. The posterior probability of freedom calculated using Bayes theorem. This is the key result that we are after for each time period. The formula (assuming perfect specificity) and its simplification are:

$$\begin{aligned} \text{Post P(free)} &= \frac{1 - P}{(1 - P) + P \times (1 - SSe)} \\ &= \frac{1 - P}{1 - P \times SSe} \end{aligned}$$

The implementation in cell F2 is **= (1-D2) / (1- (D2\*B2) )**

- G. The posterior probability of being infected. This is simply one minus the posterior probability of being free. Cell G2 contains **=1-F2**
- H. The posterior probability of infection at the beginning of the *next* time period adjusted for the probability of introduction of infection. This value is used as the prior for the next time period. This calculation is based on the generalisation of the OR probability rule for non-exclusive outcomes. The country could be infected because:
  - it was not free to start with, or
  - it became infected during the time period or
  - both occurred.

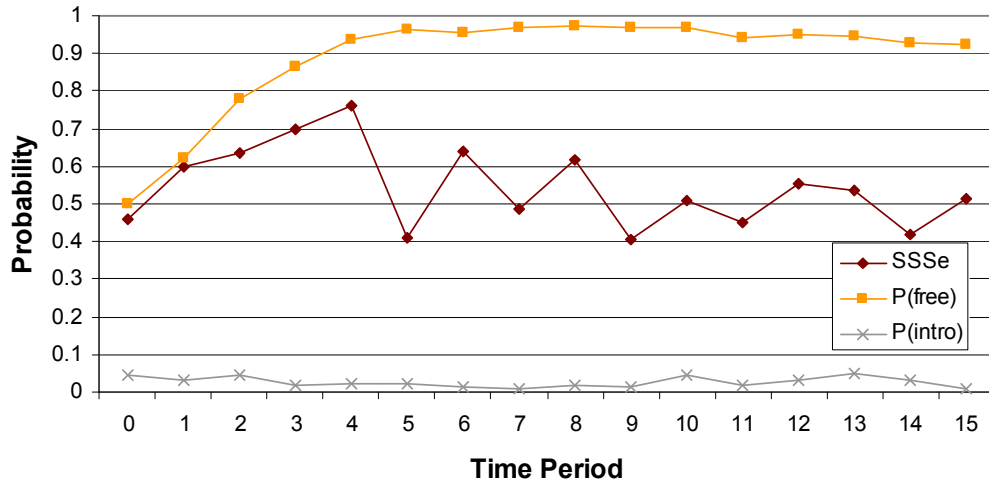
This is calculated as the sum of the two possibilities minus the overlap, or the probability that both outcomes have happened. This can be expressed as:

$$P(A \text{ or } B) = P(A) + P(B) - P(A) \times P(B)$$

In the spreadsheet cell H2 contains **=G2+C2- (G2\*C2)**

- I. The adjusted posterior probability of being free at the start of the next time period is 1 minus the adjusted probability of being infected. I2 contains **=1-H2**.

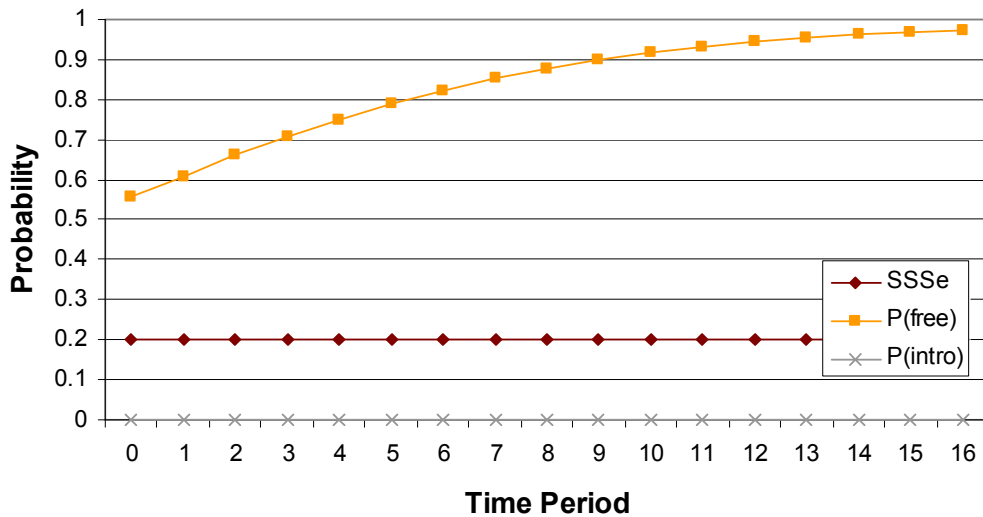
The data from the spreadsheet have been plotted below.



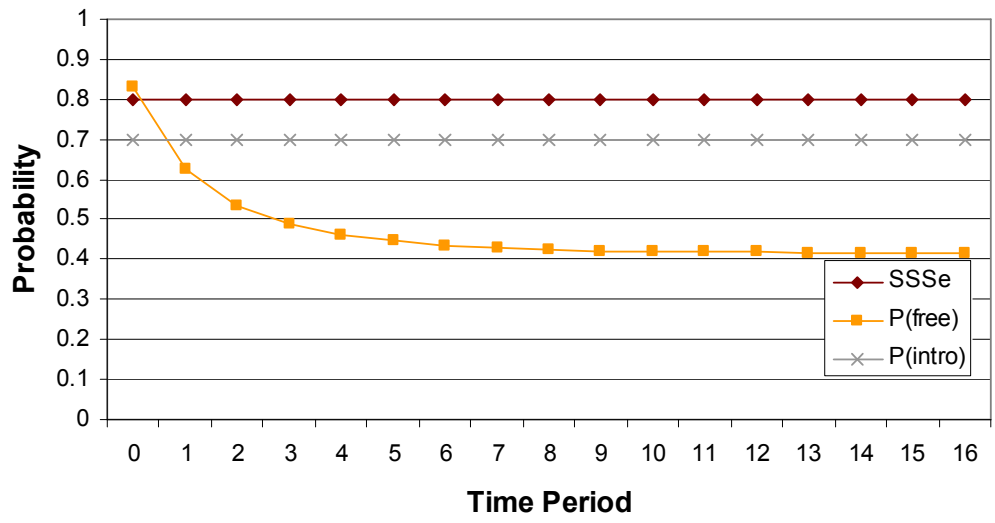
## Examples

The ability to combine historical surveillance data provides a great deal of flexibility when seeking to demonstrate freedom from infection. If the risk of introduction is low then inexpensive surveillance with relatively low sensitivity can generate a high probability of freedom if it is continued over a long enough time.

The figure below shows surveillance with 20% sensitivity and a 0.1% probability of introduction of disease.



If the risk of introduction of disease is high, even very sensitive surveillance may be inadequate to achieve a high probability of freedom. The figure below shows surveillance with a sensitivity of 80% but a risk of introduction of 70%.



# Chapter 17 – Freedom Software

We've heard that a million monkeys banging on a million typewriters will eventually reproduce the entire works of Shakespeare. Now, thanks to the Internet, we know this is not true.

**Robert Wilensky**

Analysing surveillance to demonstrate freedom from infection using scenario trees is possible using a spreadsheet with a Monte Carlo simulation add-in. However the process is complicated and it is easy to make mistakes, especially with a large model.

In order to make scenario tree modelling easier, web-based software has been created to handle all the calculations. To undertake a valid analysis of surveillance, it is still necessary to have a good understanding of the principles of scenario tree modelling, develop an appropriate model structure and provide the correct input parameters, but the software handles all the complex calculations.

## Overview

---

The system is centred on the analysis of components of a surveillance system by building a scenario tree. You can create as many scenario trees as you want.

To build a scenario tree, first you define the node list and assign default probabilities to each of the branches. The second step is to edit the branch probabilities in detail to take into account the different conditional probabilities in each branch.

You are then given the option of uploading data, which will allow you to use the actual herd structure to take clustering into account, as well as the dates of the surveillance, to analyse multiple time periods. If this data is not available, a simple analysis is possible

Finally, the scenario tree is analysed using stochastic modelling. This can take some time, so an email notification with a link to the results is sent once the analysis is complete.

A number of components of the one surveillance system can then be combined, taking into account overlap, to calculate the overall system sensitivity.

## Getting started

The software is available on the “Analysis of Complex Surveillance Systems” website, which can be accessed at:

<http://freedom.ausvet.com.au>

When you log in you will see the welcome page as shown below. This page provides some background to the system and gives you access to a variety of resources.

Home Help **Analysis of Complex Surveillance Systems**

### Scenario Trees to Quantify Confidence in Freedom from Disease

**Introduction**

Demonstrating that a population is free from disease (or from a pathogen) is a difficult task, based on probabilistic assessments of accumulated evidence. The statistical theory has been established for the design and analysis of structured surveys using random sampling to gather evidence for freedom from disease. Such surveys are often expensive and wasteful of resources as they ignore existing evidence.

Scenario tree modelling provides an alternative method of analysis of surveillance data to provide quantitative measures of the sensitivity of a surveillance system and the probability of disease freedom. This method allows complex surveillance systems (based on non-random selection) to be modeled and analysed, enabling the use of evidence from existing surveillance systems to be used to support claims of freedom from disease.

Scenario tree modelling uses stochastic simulation, and can be implemented in spreadsheets. However this task is time consuming, error prone, and requires stochastic add-in software. This site allows users to develop their own scenario tree model through a simple on-line interface, and notifies the user by email when simulations have been completed.

**Getting Started**

What's it about? [Guide to the Methodology](#)  
How to do it? [Web site User's Manual](#)  
General Information [Courses, links, downloads, examples, etc.](#)  
Documentation Home [Links to all documentation](#)  
Talk about it [Freedom Discussion Group](#)

*Log in to use on-line software*

User name   
Password    
New user? [Create a new user account](#)

**Quantifying Confidence in Disease Freedom**  
This site was created by [AusVet Animal Health Services](#)  
Project funded by the [Australian Biosecurity Cooperative Research Centre](#)  
Project leader: [Tony Martin](#), Website Copyright © 2004-2009  
Portions of code used in this site are based on [TreeMenu v1.3](#) JavaScript Tree Menu  
This site uses [PHP](#), [MySQL](#), the [R](#) statistical language and [PmWiki](#). Images courtesy of FreeFoto.com

## Log in and privacy

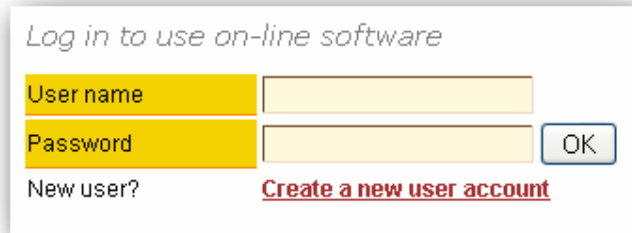
The software is free to use but you first have to log in. The reason for this is so you can store private scenario-tree models on the web site. Surveillance data is often analysed for trade or other politically sensitive purposes. When learning to use the system or analysing data, it is common to produce a number of practice models that may not truly reflect the real disease situation. If all the analyses were publicly accessible, trading partners would be able to examine and risk misinterpretation of each others' analyses. To avoid these problems, users are

required to log in to the system, and all analyses are accessible only by the person that created them (unless they explicitly request that it be made public).

The only information that is stored when you create an account to log in is your username, password and email address. The email address is used to notify you when lengthy analyses have been completed. The address is never released to anybody else, and you will not receive unsolicited email from the system.

## First log in

The first time you use the system, you need to create a user account. On the bottom of the home page, click on [Create a new user account](#).



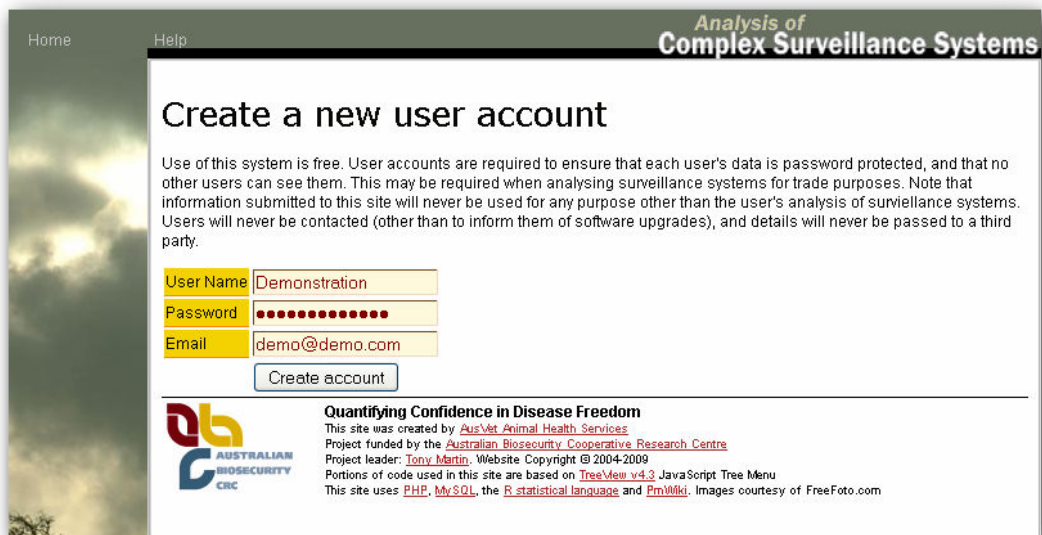
Log in to use on-line software

User name

Password

New user? [Create a new user account](#)

You will then be prompted to enter your details. Make a note of your user name and password for your next log in. Click  to finish the process.



Home Help **Analysis of Complex Surveillance Systems**

## Create a new user account

Use of this system is free. User accounts are required to ensure that each user's data is password protected, and that no other users can see them. This may be required when analysing surveillance systems for trade purposes. Note that information submitted to this site will never be used for any purpose other than the user's analysis of surveillance systems. Users will never be contacted (other than to inform them of software upgrades), and details will never be passed to a third party.

User Name

Password

Email

**Quantifying Confidence in Disease Freedom**  
This site was created by [AusVet Animal Health Services](#)  
Project funded by the [Australian Biosecurity Cooperative Research Centre](#)  
Project leader: [Tony Martin](#). Website Copyright © 2004-2009  
Portions of code used in this site are based on [TreeView v4.3](#) JavaScript Tree Menu  
This site uses [PHP](#), [MySQL](#), the [R statistical language](#) and [PmWiki](#). Images courtesy of FreeFoto.com

You will receive a message to indicate that account has been created. You are now ready to start using the system.

## Building a scenario tree

### Information required

Before you start using the software to build your scenario tree, it is a good idea to have everything prepared. You will need:

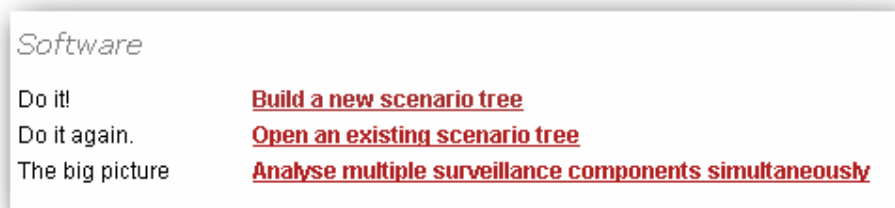
- A node list, identifying all the nodes, their types and the branches for each

- All the parameters for each branch. This includes:
  - For category nodes
    - SSC proportion
    - Population proportion
    - Relative risk (for risk category nodes)
  - For infection nodes, the design prevalence
  - For detection nodes, the sensitivity
- Each of these parameters (except the design prevalence) may, if required, be entered as a distribution, in which case you will need the parameters to define the distribution.
- If you have a dataset available, you will need this available to upload. The format is discussed below.
- Information on the number of animals and herds processed in the surveillance.

### Setting up the analysis

---

When you have logged in, the home page will display a list of options for the software.



The first allows you to create a new scenario tree. The second allows you to reopen an existing model for further editing. The third manages the combination of multiple surveillance components to calculate system sensitivity.

To start building the scenario tree click on [\*\*Build a new scenario tree\*\*](#). The Create Scenario Tree page will be displayed. This has three parts:

1. Tree description: provide some basic information about the tree
2. Node list: this is where you build the structure of the tree
3. Build tree button: This is the step where the full tree is created.

The screenshot shows a form titled "1. Tree Description" with the following fields:

- Tree Name:** A text input field with a question mark icon to its left.
- Description:** A large text area with a question mark icon to its left.
- Context:** A dropdown menu with a question mark icon to its left, currently showing "Animal health".
- Public:** A checkbox with a question mark icon to its left, which is currently unchecked.

Enter a brief **name** for the tree. This should allow you to distinguish your tree from other trees. Normally, you would include the disease and region of interest, and possibly information about the version.

Enter a **description** of the tree. This is optional.

The **context** indicates the field you are working in. Options include:

- Animal health
- Plant health
- Pests
- Human health
- Zoonoses.

The terminology on some of the pages will change according to the context, to make the system easier to understand.

If you check the **public** box, your scenario tree will be available for others to see. If you don't check it, it will be available only to you.

### Example

For this exercise, we will use the simple scenario tree described in Chapter 10.

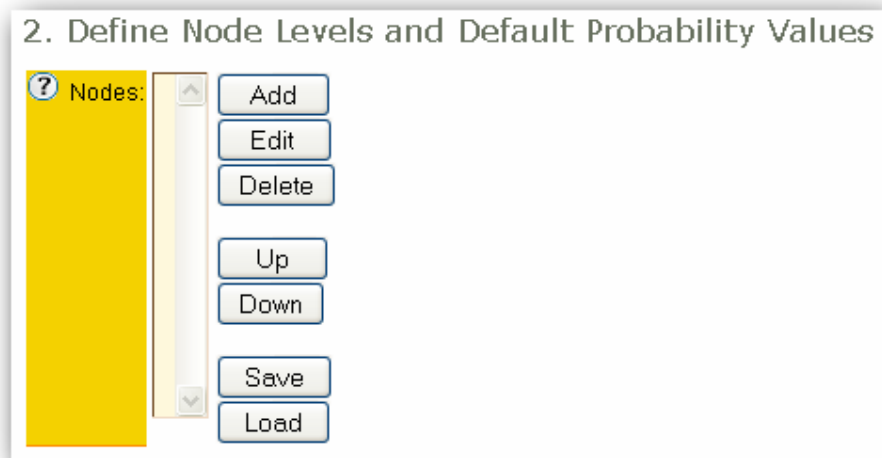
Enter the name as: **Demonstration tree**

The context is: **Animal health**

The node list will consist of age, animal infected and ELISA result.

## Building the node list

The next step is to define the list of nodes. The nodes will be displayed in the (initially empty) list box.



To add a node, click **Add**. A new pop-up window will open allowing you to enter the node details. Make sure your browser is configured to allow pop-up windows.



## Node definition

The screenshot shows a dialog box titled "Define node". It contains four fields, each with a question mark icon to its left:

- Node Name:** A text input field containing the word "Age".
- Node Type:** A dropdown menu with "Risk Category" selected.
- Number of Branches:** A text input field containing the number "2".
- Notes:** A text area containing the text "Animals less than six months of age are more susceptible to the disease." in red font.

At the bottom of the dialog are two buttons: "OK" and "Cancel".

The **node name** describes the node. It should be relatively brief. Remember that it should imply a question that is answered by the different branches.

The **node type** has been discussed on page 70. The options are:

- Detection category
- Risk category
- Detection
- Infection

For group category nodes, you should use a detection category node that contains the same information.

Specify the **number of branches** for category nodes (this option is not included for infection or detection nodes as there are always two branches).

Enter any **notes** that may help remember details about the node. This could include references to source information.

Click **OK** when the information for the node is complete to be taken to the branch definition window.

## Branch definition

The branch definition window contains one row for each branch.

Enter a **name** for the branch. Remember that this takes the form of an answer to the question posed by the name of the node.

Enter the branch **parameters**. These are the default parameters that will be used for all limbs of the tree. In many cases, the branch probabilities will be conditional on previous nodes, and should therefore be different on different limbs. The conditional probabilities will be edited in a later step.

These are different depending on the type of branch, as discussed below. Note that the parameters for the last branch are automatically calculated.

Enter a brief **note** about the branch if required.

### Infection nodes

#### Define Branches

Node: Animal infected

No.	Branch Name	Design Prevalence	Notes
1	Infected		
2	Uninfected	N/A	

OK Cancel

There is only one parameter for an infection node - the **design prevalence**. This should always be a fixed value so there is no option to enter a distribution.

### Detection nodes

#### Define Branches

Node: ELISA

No.	Branch Name	Sensitivity	Notes
1	Positive		
2	Negative	N/A	

OK Cancel

There is only one parameter for detection nodes – the sensitivity. This can be entered as a distribution as discussed below.

### Detection category nodes

#### Define Branches

Node: Sample quality

No.	Branch Name	Category proportion		Notes
		SSC	Population	
1	Good quality			
2	Medium			
3	Poor quality	N/A	N/A	

OK Cancel

For detection category nodes, two parameters are required: the proportion for each branch in the SSC, and the proportion in the population. Both of these can be entered as distributions. It is possible to have two or more detection category branches.

Note that if the population proportion is blank, it is automatically set to the value of the SSC proportion. This makes specification of the parameters simpler for components that have comprehensive coverage (such as passive disease reporting systems) and for which the SSC and population proportions are the

same. To specify a different proportion, you should edit the automatically entered value.

### Risk category nodes

No.	Branch Name	Category proportion		Relative Risk	Notes
		SSC	Population		
1	Young			Reference (RR=1)	
2	Old	N/A	N/A	Reference (RR=1)	

Branches for risk category nodes require three parameters: the SSC and population proportions (as with the detection category nodes) and the relative risk.

When specifying the relative risk, the value for any branch is always relative to the lowest risk branch which must have a risk equal to 1. This is known as the reference category. The rules for setting the relative risk are:

- You can use any of the branches as the reference category (set it to **Reference (RR=1)**)
- There must be only one reference category. All the other branches should have the relative risk explicitly set
- The risk in other branches should be greater than or equal to one.
- Risks in other branches may be entered as distributions

### Specifying probabilities and distributions

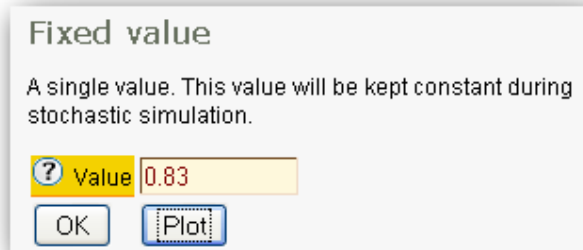
For all probabilities that may be entered as a distribution a drop-down data entry box is displayed. This lists the different distributions available within the software.

To enter a parameter, click on the drop-down box, and select the appropriate distribution from the list. A new window will pop up allowing you to enter the parameters for the distribution.

### Fixed value

This option allows you to enter a single fixed value that is not a distribution. This should be used for parameter for which there is no uncertainty or variability (for example, if the exact proportion of beef and dairy farms in the population is known from official farm registrations, then this exact value can be used).

Note that all proportions should be entered as values between zero and one and not as percentages.



**Fixed value**

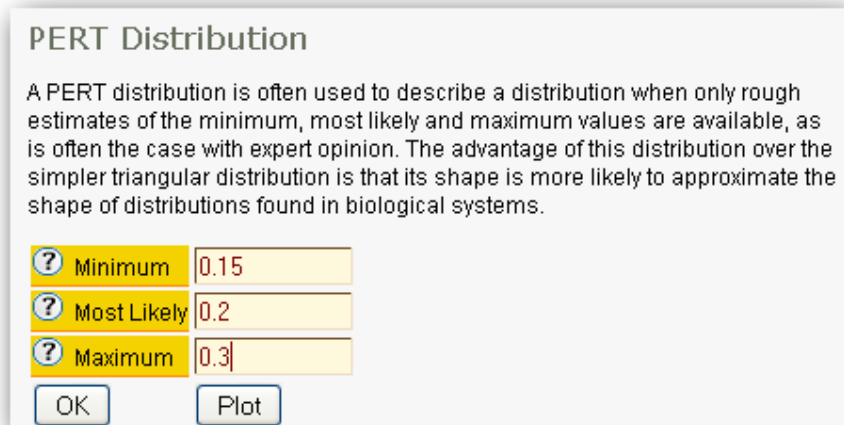
A single value. This value will be kept constant during stochastic simulation.

? Value 0.83

OK Plot

### Distributions

All other options allow you to enter parameters defining a distribution. The parameters required will be different depending on the distribution chosen. For example, the PERT distribution, shown below, requires the minimum, most likely and maximum values.



**PERT Distribution**

A PERT distribution is often used to describe a distribution when only rough estimates of the minimum, most likely and maximum values are available, as is often the case with expert opinion. The advantage of this distribution over the simpler triangular distribution is that its shape is more likely to approximate the shape of distributions found in biological systems.

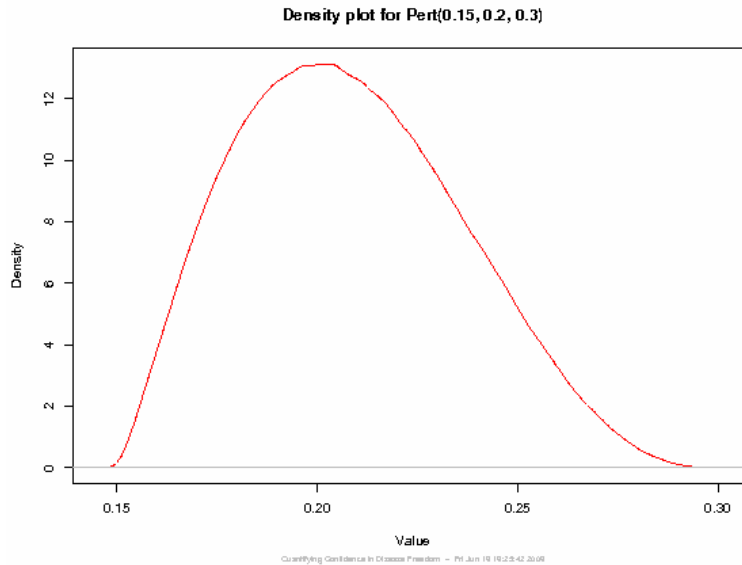
? Minimum 0.15

? Most Likely 0.2

? Maximum 0.3

OK Plot

It is possible to view the shape of the defined distribution by clicking **Plot**. A graph will be displayed in a new pop-up window.

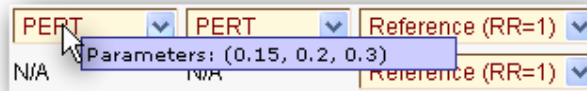


## Editing probability parameters

---

### Viewing the current parameters

When a distribution has been entered into a drop-down box, only the name of the distribution is displayed. To see what the parameters are, place your mouse over the drop-down box (without clicking). A small window will show the current parameters.



### Changing the current parameters

To edit the parameters of the distribution without changing the type of distribution, just click once on the drop-down window. This will open the window to define the distribution parameters.

If you need to change the type of the distribution, the method is a little different depending on the browser that you are using. For some browsers (such as **Firefox**), click and hold the mouse button down, then drag the cursor to the new distribution and release.

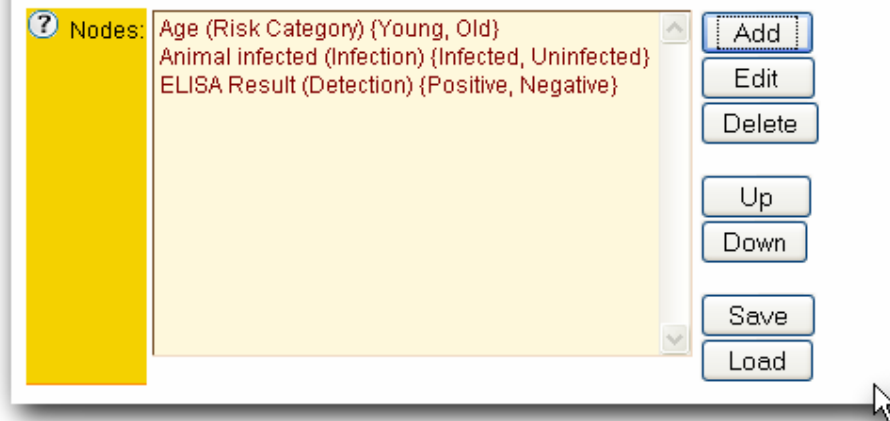
For other browsers (such as some versions of **Internet Explorer**), dragging with the mouse may not work. In this case you need to use the keyboard to select the right distribution. Use the Tab key to move to the drop-down list that you want to change, then use the up- or down-arrow keys to change the distribution type. When the right type is showing, you can click on the list to open the window to define the distribution parameters.

## Editing the node list

---

When the node and branch parameters have been defined, clicking **OK** will save the parameters to the node list on the main page. Continue to add new nodes in the same way to build up your complete node list.

## 2. Define Node Levels and Default Probability Values

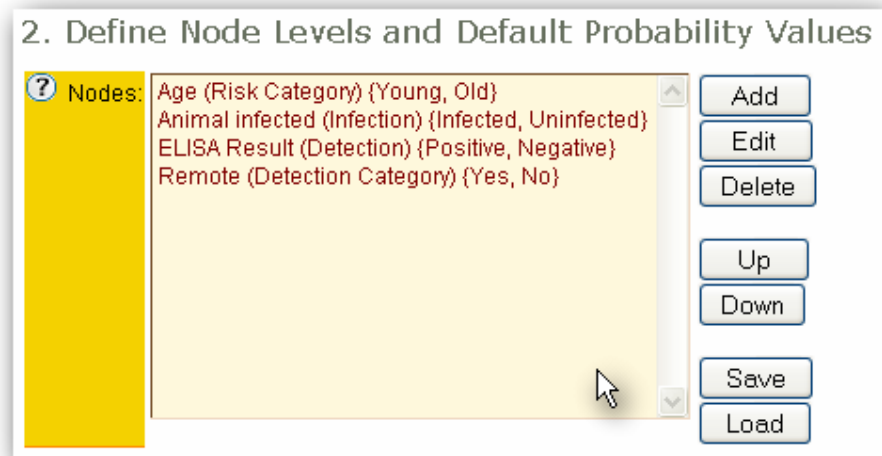


The buttons on the right of the node box allow you to edit the node list.

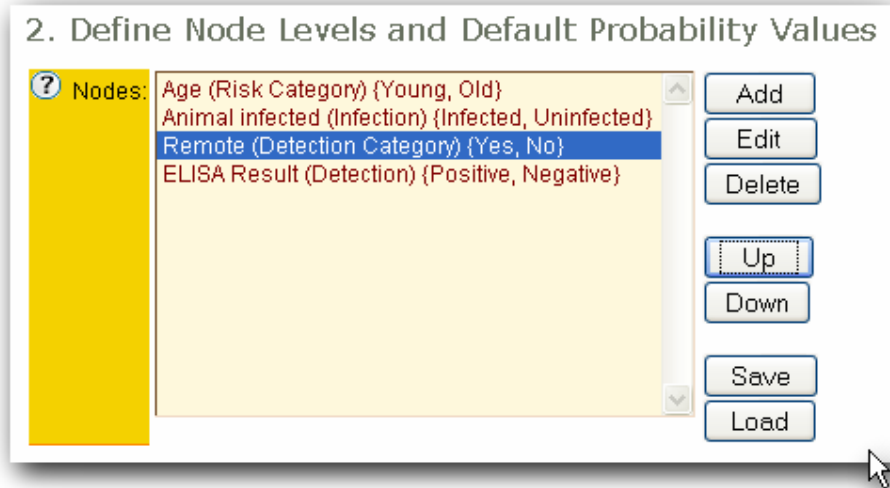
### Example

After consideration we realise that the quality of the sample has an impact on the sensitivity of the ELISA. We therefore want to add a node based on location (remote or not remote) that indicates how long it takes for samples to be transported.

We define the new node by clicking **Add** and entering the parameters. When we are done, the node list will be as shown below.



The problem is that *Remote* should come before *ELISA result*. To change the order of the nodes you can highlight the node we want to move by clicking on it and then use the **Up** or **Down** buttons.



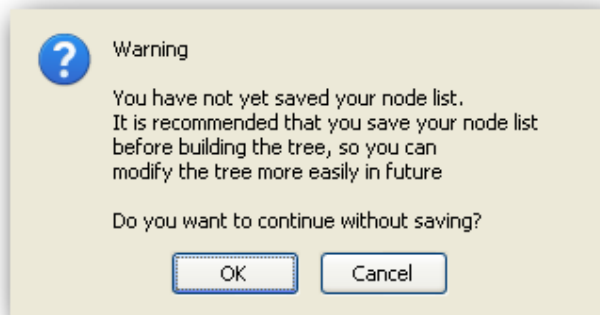
The **Load** and **Save** buttons can be used to save a text file of the tree node definition to your local hard disk, and the load it later. This is normally not necessary, as the structure will be saved to the server hard disk and can be edited or modified on the system. These buttons can therefore be ignored.

### Building the scenario tree

Up until this point, the node list and parameters have been stored in your computer's memory. If you shut down your browser, all the data would be lost.

To save the data and to create the scenario tree, click **Build Tree**. This converts the node list into a tree structure and saves the structure to the database on the server, ready for analysis. Once you have done this, the data is safe. You can log out and return to your analysis later if required. For large trees, this step can take some time, so be patient.

When you click **Build Tree** you may get the following warning.



As it is possible to modify the tree later, it is not necessary to save the node list. You can safely click **OK**.

When the tree has been built, the summary page is displayed.

## Tree Summary

Node Levels	4
Total Nodes	15
Total branches	31

Generating tree: 0.0412 s  
Writing structure to database: 0.0005 s

### *Displaying the tree structure*

---

Clicking  allows you to look at the full tree structure with all the current parameters. This step is optional. You will be given another opportunity to view and download the full tree structure after you have edited the conditional probabilities.

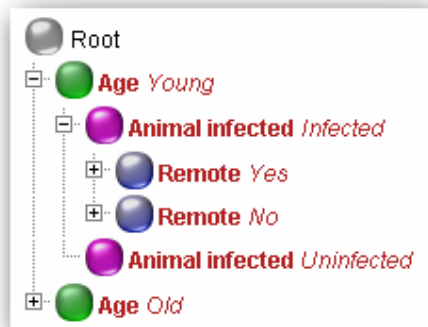
### **Refining conditional probabilities**

---

When you click on  the tree is displayed in an expandable structure. This allows you to navigate your way down individual limbs to find any nodes that have conditional probabilities that need to be corrected.



To expand or collapse a branch, click on the '+' or '-' to the left of the branch. Clicking the  button will display all the nodes at once.

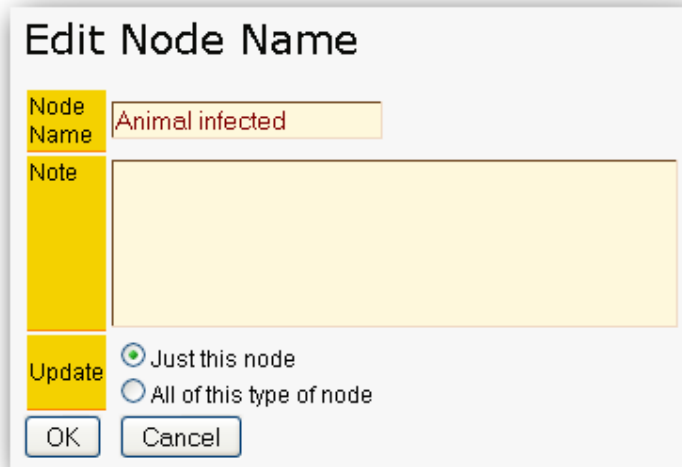


To edit the **Node Name** click on the node name (in bold). This is not normally necessary, but if you want to change the name for some branches for



clarity, it is possible. This also allows you to add a different note to some or all nodes. A pop-up box is displayed to make the changes.

It is possible to specify which nodes the changes will be applied to. To change only to the currently selected node, check the **Just this node** radio button. To change all occurrences of this node for every branch of the tree, check the **All of this type** of node radio button.



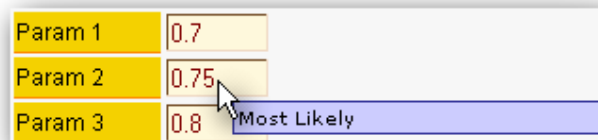
To edit the branch, click on the branch name (in italics). This is what is most commonly required to change the conditional branch probabilities.

### Example

In our tree we currently have a single distribution describing the sensitivity of the ELISA. However, we have added a detection category node (Remote) because sample quality from remote areas is lower and therefore the sensitivity is lower. We need to edit the sensitivity in those branches that are below the *Yes* branch of the Remote node, to enter a lower sensitivity.

When you click on the name of the branch, a window pops up to edit the branch. At the top of the window the path to that particular branch is explained with the node name and branch name for all previous nodes, ending in the current branch.

Below that, the branch details are displayed for editing. These edits will apply only to the specific branch being edited. The distribution parameters are labelled Param 1, Param 2 and Param 3. If you can't remember what the parameters for your distribution are, place the mouse over the field and a small message will pop up with the name of the parameter.



**Example**

We have used a PERT (0.9, 0.95, 0.98) distribution to describe the sensitivity of the ELISA. For this branch (samples from remote areas), we will change it to PERT (0.7, 0.75, 0.8).

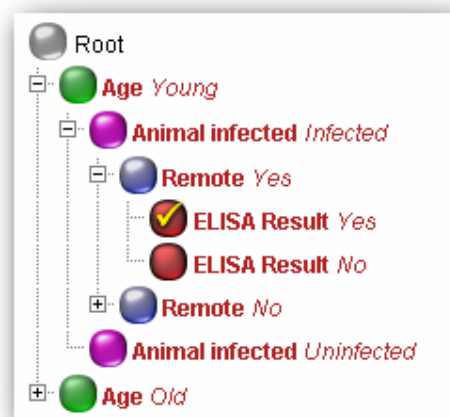
**Edit Branch Probabilities**

**Age, Young**  
**Animal infected, Infected**  
**Remote, Yes**  
**ELISA Result, Yes**

Branch Name	Yes
Value	Probability
Distribution	PERT
Param 1	0.7
Param 2	0.75
Param 3	0.8
Notes	

OK Cancel

When you click OK, the database is updated, and the tree display places a check mark on the node that has been edited. This helps you keep track of which nodes have been updated and which may still need to be done.



**Example**

We have updated one node on the Young→infected→remote branch. There is another node on the Old→infected→remote branch that needs to be updated.

Note that the sensitivity on the uninfected branches is not relevant, so there is no need to update the values for the Young→uninfected→remote and Old→uninfected→remote branches.

When all the required edits have been made to the tree structure, there are three options at the bottom of page:

- Upload Data. This option will be discussed later.
- Export Full Tree. View the full tree structure.
- Run Model. Perform the stochastic modelling.

## Exporting the tree structure

When all the parameters for the tree have been entered the definition of the model is complete. In order to reproduce the model, for documentation or publication purposes, it is important that all the parameters used are recorded.

The **Export Full Tree** button gives you several options for viewing and saving all the details of the tree structure (the model definition). First, the information is presented in plain text on the web page.

```
Root
[Age] Old,16,2: [Relative Risk: Fixed value 1 ]
[Animal infected] Uninfected,24,3: Automatically calculated
[Remote] No,28,1: Automatically Calculated
[ELISA Result] No,30,4: Automatically Calculated
[ELISA Result] Yes,29,4: [Probability: PERT 0.9 0.95 0.98]
[Remote] Yes,25,1: [Population proportion: Fixed value 0.2 ] [Proportion processed in Surveillance System Component: Fixed value 0.8 ]
[ELISA Result] No,27,4: Automatically Calculated
[ELISA Result] Yes,26,4: [Probability: PERT 0.9 0.95 0.98]
[Animal infected] Infected,17,3: [Design Prevalence: Fixed value 0.05 ]
[Age] Young,1,2: [Relative Risk: PERT 2 3 5] [Population proportion: PERT 0.15 0.2 0.3] [Proportion processed in Surveillance System Component: PERT 0.5 0.7 0.8]
[Animal infected] Uninfected,9,3: Automatically calculated
[Remote] No,13,1: Automatically calculated
[ELISA Result] No,15,4: Automatically calculated
[ELISA Result] Yes,14,4: [Probability: PERT 0.9 0.95 0.98]
[Remote] Yes,10,1: [Population proportion: Fixed value 0.2 ] [Proportion processed in Surveillance System Component: Fixed value 0.8 ]
[ELISA Result] No,12,4: Automatically calculated
[ELISA Result] Yes,11,4: [Probability: PERT 0.9 0.95 0.98]
[Animal infected] Infected,2,3: [Design Prevalence: Fixed value 0.05 ]
```

The buttons at the top of the page give two options for download of the model definition. Export Text File downloads the data in tab-delimited format, suitable for opening in a spreadsheet.

The **Export PDF** creates a formatted PDF version of the model definition. This may be convenient for filing or distributing to colleagues.

The exported file for our example tree looks like this:

**Age Old (Risk Category)**

Relative Risk: Fixed value (1)

**Animal infected Uninfected (Infection)**

Probability: Automatically Calculated

**Remote No (Detection Category)**

Probability: Automatically Calculated

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.9, 0.95, 0.98)

**Remote Yes (Detection Category)**

Population proportion: Fixed value (0.2)

Proportion processed in Surveillance System Component: Fixed value (0.8)

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.9, 0.95, 0.98)

**Animal infected Infected (Infection)**

Design Prevalence: Fixed value (0.05)

**Remote No (Detection Category)**

Probability: Automatically Calculated

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.9, 0.95, 0.98)

**Remote Yes (Detection Category)**

Population proportion: Fixed value (0.2)

Proportion processed in Surveillance System Component: Fixed value (0.8)

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.7, 0.75, 0.8)

**Age Young (Risk Category)**

Relative Risk: PERT (2, 3, 5)

Population proportion: PERT (0.15, 0.2, 0.3)

Proportion processed in Surveillance System Component: PERT (0.5, 0.7, 0.8)

**Animal infected Uninfected (Infection)**

Probability: Automatically Calculated

**Remote No (Detection Category)**

Probability: Automatically Calculated

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.9, 0.95, 0.98)

**Remote Yes (Detection Category)**

Population proportion: Fixed value (0.2)

Proportion processed in Surveillance System Component: Fixed value (0.8)

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.9, 0.95, 0.98)

**Animal infected Infected (Infection)**

Design Prevalence: Fixed value (0.05)

**Remote No (Detection Category)**

Probability: Automatically Calculated

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.9, 0.95, 0.98)

**Remote Yes (Detection Category)**

Population proportion: Fixed value (0.2)

Proportion processed in Surveillance System Component: Fixed value (0.8)

**ELISA Result Negative (Detection)**

Probability: Automatically Calculated

**ELISA Result Positive (Detection)**

Probability: PERT (0.7, 0.75, 0.8)

Click **Back** to return to the Edit Tree page, ready to run the model.

## Running the model

Now that all the scenario tree parameters have been entered, we are ready to run the model in its simplest form. More complex situations will be discussed later.

To run the model, click the Run Model button at the bottom of the Edit Tree page. This gives us the screen to enter the simulation parameters.

**Set Simulation Parameters**

Total Units Processed	<input type="text" value="20"/>
Total groups in the population	<input type="radio"/> Entire population examined <input type="radio"/> Specify total number: <input type="text"/> <input checked="" type="radio"/> Unknown
Prior probability of freedom	Fixed value ▾
Number of iterations	5 ▾

The **total units processed** is the total number of animals in the surveillance system. We will use 20 for our example.

The **total groups in the population** is, in our terminology, the total number of herds in the entire population. In our example, this is unknown.

The **prior probability of freedom** was discussed on page 145. We will use a fixed value of 0.5 for this example.

The **number of iterations** determines how many times the model will be run to build up the output distribution. When the model is first defined, it is a good idea to run it with a small number of iterations to make sure that everything is working as expected. Later, when any refinements have been made, the final version can be simulated with up to 1000 iterations. Remember that this is a web application and the server is being shared by other users. Stochastic simulation takes a lot of processing power so please don't perform a large number of big simulations when smaller ones would suffice.

The buttons at the bottom of the page give you the option to:

- Save parameters and start simulation: this runs the models. The simulation parameters you have just set will also be saved for the next time you run the model.
- Just save parameters: the model will not be run.
- Edit tree: return to the previous page to make further changes to the tree.

When you are ready, click **Save parameters and start simulation**. The model will start running and a message is displayed. When the model has finished, you will be sent an email with a link to the results. The reason for this is that some models can take a very long time to run, and the web system times-out waiting for it to finish. With a very small model, like that in our example, it should take less than a minute, so check your email to see if there is a message. With a large complex model with many iterations, it may take several hours.

The email will look something like this:

Subject: Scenario Tree Model Results  
 From: Freedom from Disease

This is an automatically generated email from the Analysis of Complex Surveillance Systems (Scenario Tree Modeling) web site.

Scenario Tree Name: Demonstration tree  
 Run by: Angus  
 Description: Demonstration scenario tree

Your requested model simulation has been completed and the results are now available for viewing at:

[http://freedom.ausvet.com.au/content.php?page=show\\_results&treeid=212&uid=c3afe256687d8b8c8062f869dc34d212](http://freedom.ausvet.com.au/content.php?page=show_results&treeid=212&uid=c3afe256687d8b8c8062f869dc34d212)

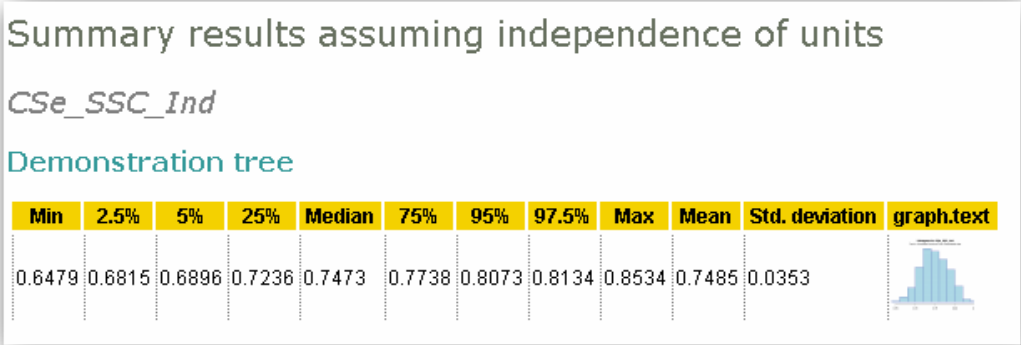
Note that these results will only be retained for the next 24 hours, after which time they will be deleted from the system. Please save a copy of any results you wish to keep on your hard disc.

When you click on the link, a new web page will open with the results of your analysis.

## Interpreting the output

The output of the analysis consists of summary tables, graphs and data sheets to download. The output page is long and has many different pieces of information, some of which are important, and some of which are only relevant in specific situations.

The most important information is given in the section called Summary results. The results for our example analysis will be used to illustrate the concepts.



The heading indicates that these results are based on the assumption of the independence of units. This means that clustering was not taken into account, as no herd-level data was provided. We will look at uploading herd-level data in the next section.

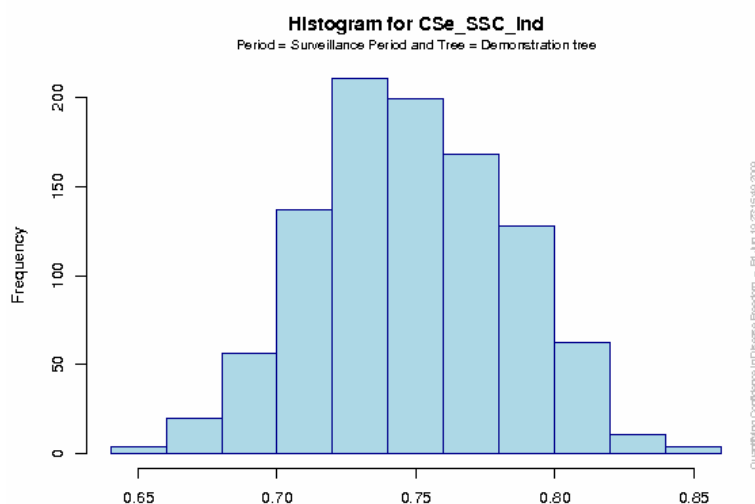
In this example, the results summarise a number of parameters, given abbreviated titles:

CSe_SSC_Ind	The component sensitivity of the surveillance system component, assuming independence between animals
CSe_Rep_Ind	The component sensitivity of a hypothetical representative surveillance system component, assuming independence between animals
Se Ratio_Ind	The sensitivity ratio, or ratio between the sensitivity of the actual component and of the hypothetical representative component. See page 122. Assumes independence between animals.
PFree_Ind	The probability that the population is free. Assumes independence between animals.

When herd-level data is available, additional outputs will be produced

Each parameter is summarised with a series of summary measures and a graph. The minimum, percentiles, maximum, mean and standard deviation describe points on, or summaries of, the output distribution. However, the graph is normally the easiest way to understand the output.

Click on the graph to display a larger image.



For our example, the key results (median values) were:

- Component sensitivity: 75%
- Sensitivity ratio: 1.24
- Probability of freedom: 80%

Many of the other outputs on the results page relate to the analysis of multiple trees or multiple time periods. These will be discussed later, after we have looked at how to perform these types of analyses.

## Modifying an existing scenario tree

It is possible to use an existing scenario tree (either one that you have created or a public tree) as the basis for a new tree. This process involves cloning an

existing tree, modifying the node structure, and editing the conditional probabilities.

### Opening or cloning existing trees

After logging in, you can see a list of available projects by clicking on **Open an existing scenario tree**. You can also access this page by clicking on **Open** at the top of the page.

The list includes your private trees at the top, as well as a list of trees that have been marked as public below. For each tree, you can **Open** or **Clone**. You can also **Delete** trees that you have created.

### Open a Scenario Tree

Your Private Trees	Created	Description	
<a href="#">BC ISA - Auditing</a>	<a href="#">21/11/2008</a>		<input type="button" value="Open"/> <input type="button" value="Clone"/> <input type="button" value="Delete"/>
<a href="#">Demonstration tree</a>	<a href="#">19/06/2009</a>	<a href="#">Demonstration scenario tree</a>	<input type="button" value="Open"/> <input type="button" value="Clone"/> <input type="button" value="Delete"/>
<a href="#">Test Tree 2 (cloned)</a>	<a href="#">11/11/2008</a>	<a href="#">CSF test tree with re-ordered nodes</a>	<input type="button" value="Open"/> <input type="button" value="Clone"/> <input type="button" value="Delete"/>
<a href="#">WSD tree (cloned)</a>	<a href="#">11/11/2008</a>	<a href="#">Tree for white spot disease surveillance in Qld, based on original from Qld training workshop</a>	<input type="button" value="Open"/> <input type="button" value="Clone"/> <input type="button" value="Delete"/>
Public Trees	Modified	Description	
<a href="#">BJD Clinical</a>	<a href="#">21/11/2008</a>	<a href="#">WA clinical diagnostic system</a>	<input type="button" value="Open"/> <input type="button" value="Clone"/>
<a href="#">BJD Survey</a>	<a href="#">21/11/2008</a>	<a href="#">1995 serological survey for BJD in WA</a>	<input type="button" value="Open"/> <input type="button" value="Clone"/>

Cloning a tree extracts the node list from an existing tree, using the default probabilities from one of the branches of the tree.

#### **Example**

The demonstration tree that we have created now contains four nodes: age, animal infected, remote, and ELISA result. There is no grouping node at the herd level and no herd-level risk factor nodes. We will clone this tree, and add a risk category node (region) and a herd infection node.

Click on the **Clone** button next to the demonstration tree. The node list will be displayed.

Edit the name of the tree to help distinguish it from the earlier version (by default “(cloned)” is added to the name).

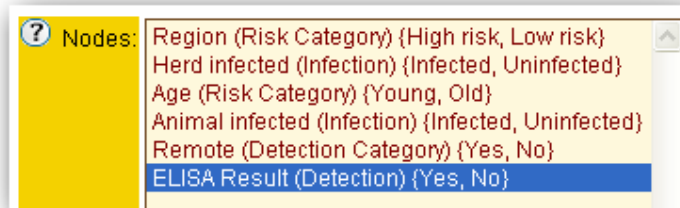
Add the region risk category node with two regions (high risk, low risk). The values for the high risk branch should be based on PERT distributions with the following parameters.

Value	Minimum	Most likely	Maximum
SSC proportion	0.48	0.5	0.52
Population proportion	0.08	0.1	0.15
Relative risk	2	2.3	2.5

Add a herd infection node with a design prevalence of 1%. Move the two new nodes to the top of the list.

The node list should then look like this.





Build the tree and remember to edit the ELISA sensitivity for the remote branches (these are the only conditional probabilities in our tree). The values should be PERT(0.9, 0.95, 0.98) for the *No* branch and PERT(0.7, 0.75, 0.8) for the *Yes* branch. There will be four different branches to edit this time.

The tree is now ready for analysis. This time, we will upload some data.

## Uploading data

The more advanced analyses involve accounting for clustering (lack of independence between animals within the same herd) and analysis of surveillance data over multiple time periods. For both of these, we need to upload a dataset indicating information at the herd level and the date of each observation.

The process of uploading data involves first setting up the data to upload in a spreadsheet, then copying it to the web page for upload, and finally identifying columns and variables that correspond to the tree structure.

### Preparing the data for upload

The data to be uploaded should be prepared in a spreadsheet. There should be no missing data and no empty rows. The first row should be a header row with the names of the columns. The rest of the rows each contain data about one herd. The columns should be

- A group identifier (e.g. herd ID). This should be a unique number for every herd. If multiple components of a surveillance system are going to be analysed, then herd ID should match between the different datasets. IDs may include text or numbers.
- One column for any category nodes that appear above the level of the first infection node (normally the herd infection node). For instance, if herd type is a risk category node in our tree, there should be a column labelled herd type. Each row would contain the type of the herd, dairy or beef. It is not necessary to put in a column for every category node, only those nodes that you have data available for.
- The number of animals in the herd that passed through the surveillance system component.
- The total number of animals in the herd. This is optional so should only be included if this information is available.
- The date of the observation in European format (dd/mm/yyyy). This is only required if multiple time periods are to be analysed.

#### Example

Prepare a dataset for use with our new scenario tree. This should contain the fields shown below. In this example, we have used a simulated dataset containing

600 records, spread over 6 months. You may use a genuine dataset or simulate your own data.

Farm ID	Date	Region	n	Herd Size
1	1/01/2004	Other	16	98
2	1/01/2004	SJ	7	27
3	1/01/2004	Other	2	6
4	1/01/2004	Other	3	13
5	1/01/2004	Other	44	300
6	1/01/2004	Other	54	440
7	1/01/2004	Other	198	1183
8	1/01/2004	Other	13	130
9	1/01/2004	Other	66	368
...	...	...	...	...

### Submitting the data

At the bottom of the Edit Tree page, click **Upload Data**. The upload page consists of some instructions on the data format, and a blank text box.

Copy the data from your spreadsheet and paste it in the text box. Make sure you include the header row (with the column names). There must not be any blank rows, blank columns or missing data. In Excel on Windows systems, you can highlight the data block and press Ctrl-C. On the web page, click in the text box and press Ctrl-V to past the copied data.

Once the data is pasted, do not try to edit it. The data may look messy due to lines wrapping, but this should not be changed. Just click **Upload Data** to submit the data.

### Identifying the columns

Data columns in the spreadsheet may be in any order and have any column heading. The system needs to be told which columns in the uploaded data correspond to which values and nodes in your tree. The next page allows you to instruct the software how the uploaded data is organised.

? Group Identifier	Farm ID
? Number of units processed	n
? Time (date field)	Date
? Group size	HerdSize
? Region	Region
? Age	-- Not Used --
? Remote	-- Not Used --

OK Back

Choose the column from your uploaded data that corresponds to the items listed. There will often be nodes for which no data is available (e.g. animal-level nodes) so these can be left as “Not Used”. For our example data, the form should look like that shown above.

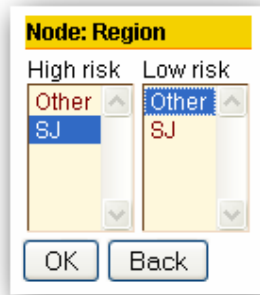
Click **OK** to continue to the next step.

## Classifying categories

---

The data we uploaded had a column for the region node. This divided herds into those that are in a high-risk region and those that are in a low risk region. The dataset used codes for the region: *SJ* (an abbreviation of the high risk region) and *Other* (for all the rest of the regions). The system needs to know which code corresponds to the high risk and which to the low risk areas.

For each category node, the system will display a series of list boxes containing the codes in the uploaded data for that node. You are required to identify which code or codes correspond to which branch of the category node.



In our example, click on *SJ* in the high risk list and *Other* in the low risk list to indicate the meaning of the different codes. If we had used abbreviations for all the different regions, there would be several different codes that correspond to the low risk branch. You can select as many classifications in the list as required by holding down the control key and clicking on items in the list.

Click **OK** to continue. You will receive a message confirming that the data has been uploaded and classified. The data is stored on the web server and will be used for future analyses. If you wish to change the data, you can upload a different dataset and the existing data will be replaced. Click **Run Model** to continue to the simulation parameters page.

## Simulation parameters

---

When data with herd details and dates has been uploaded, a number of new options appear in the simulation parameters page.

**Set Simulation Parameters**

**Total Units Processed**  
 Use actual number from uploaded data  
 Use this number (ignore uploaded data):

**Total groups in the population**  
 Entire population examined  
 Specify total number:   
 Unknown

**Multiple time periods**  
 Analyse as a single time period  
 Analyse multiple time periods

**Time periods for analysis**  
**Start date:**    
**End date:**    
**Period:**    
**Number of periods:**

**Risk of introduction**  
 Enter either:  
 1. a single value as the constant probability of introduction of disease over all time periods (between 0 and 1) or  
 2. 6 values (between 0 and 1, each on a new line) being the probability of introduction of disease for each of the time periods (either copy from a spreadsheet or enter with each value on a new line). There should be no header row, just one value per line.

**Prior probability of freedom**

**Number of iterations**

The uploaded data contains information on the **total units processed** – the number of animals processed per herd, and the herds included in the surveillance system component. By default, the analysis will be based on this number of animals, but you can choose to specify a different number of animals.

If dates are included in the uploaded data, you can specify whether you want to analyse the data for **multiple time periods** or as a single period. For this example we will analyse each month separately.

It is then necessary to provide details about the **time periods for analysis**. The **start** and **end dates** are automatically determined from the data, but you can choose to analyse the data for a shorter period.

The length of the analysis **period** can be specified. Normally it is either Month or Year.

Based on the dates and the specified analysis period, the **number of periods** for analysis is calculated.

The **risk of introduction** is used to take into account the decreasing value of historical data. If the risk is approximately constant, you can enter a single probability here, which will be applied to all time periods. If the risk varies by time period, you can enter a probability for each time period on a new line.

Click **Save parameters and start simulation** to run the model in the normal way. An email will be sent when the simulation is complete.

## Outputs for multiple time periods and clustered data







When the analysis is performed with uploaded data, the results are rather more detailed. Uploaded data should always contain information on clustering

(the number of animals processed per herd), so the results now take the lack of independence between animals in the same herd into account.

It is also possible to include dates in uploaded data and to analyse the data over multiple time periods. When this is done, separate results are provided for each time period, as well as an estimate of the combined probability of freedom over the multiple time periods.

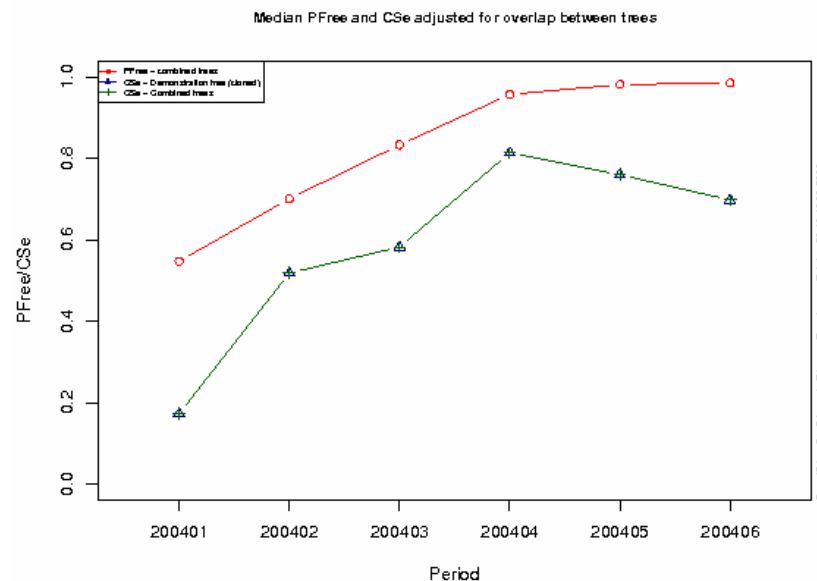
*CSe\_adjusted*

**Demonstration tree (cloned)**

	Min	2.5%	5%	25%	Median	75%	95%	97.5%	Max	Mean	Std. deviation	graph.text
200401	0.1661	0.1678	0.1685	0.1705	0.1723	0.174	0.1758	0.1763	0.1777	0.1723	0.0023	
200402	0.4992	0.5055	0.5076	0.5134	0.518	0.5226	0.529	0.5303	0.5346	0.5179	0.0065	
200403	0.5639	0.5704	0.5718	0.5777	0.5821	0.5868	0.5926	0.5939	0.5979	0.5821	0.0063	
200404	0.7988	0.8036	0.805	0.8096	0.8135	0.8171	0.8219	0.8228	0.8258	0.8134	0.0052	
200405	0.7431	0.7488	0.7502	0.7553	0.7596	0.7637	0.7691	0.7701	0.7736	0.7595	0.0057	
200406	0.6802	0.6863	0.6877	0.6932	0.6977	0.7022	0.7077	0.7089	0.7127	0.6976	0.0061	

This figure shows the component sensitivity adjusted for clustering at the herd level. As before summary values and graphs are available.

One of the more important outputs is a series of graphs showing the probability of freedom based on the combination of data over time, taking into account the risk of introduction.



Some of the output titles indicate that the result is either *\_adjusted* or *\_independent*. These refer to adjustment for overlap between multiple components of the surveillance system. At the moment, we have just examined a single component, so the results are the same. We will look at combining multiple components in the next section.

## Combining multiple components

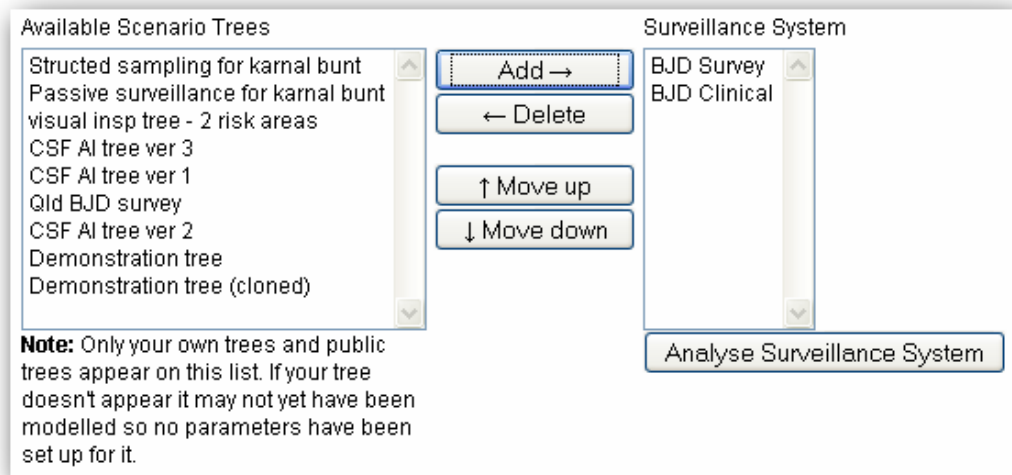
The approach to combining multiple components of a surveillance system was described in Chapter 14. The software is also able to perform these calculations.

This analysis requires that:

- A scenario tree has been created and analysed for each of the components to be combined.
- Data has been uploaded, with the same herd identifiers used between the different data sets. This is necessary so the system can check to see if the same herd appears in two or more components.
- The component scenario tree structures are compatible. For instance, if one tree has herd and animal infection nodes, all the trees must have these nodes.
- The design prevalence in each infection node must be the same between trees.

Analysis can be done for a single time period or over multiple time periods.

To start the analysis from the home page, click on **Analyse multiple surveillance components simultaneously**. You will then be given the chance to select which prepared scenario trees you wish to include as components of the surveillance system.



Select a tree from the list on the left, then click **Add →** to add that tree to the list on the right. You can change the order for the analysis of the trees by selecting a tree on the right and clicking **Move up** or click **Move down**.

When the list is complete, click **Analyse Surveillance System** to start the analysis. You will be sent an email to inform you when the analysis is complete.

The output is similar to that described previously, however they will include separate analysis of each component, and results based on adjusting for the

overlap between systems, as well as results based on the assumption of independence between systems (for comparison).

# Appendices

## Glossary

---

Term	Meaning
Acceptable level of protection	In risk analysis, the level of protection that a country sets, against which the results of risk analyses are judged.
Active surveillance	Surveillance in which the primary users of the surveillance data (usually the veterinary authorities) initiate and design the data collection
Adjusted risk	This is a measure of risk in specified branch of a risk category node. The ratio of the adjusted risks between two branches has the same value as the relative risk, but the adjusted risks are adjusted to ensure that the average risk in the population remains equal to one.
Bayes' theorem	<p>A probability formula that enables prior knowledge to be combined with new information to give an updated (posterior) probability estimate. In probability notation, it is expressed as:</p> $\Pr(A   B) = \frac{\Pr(B   A) \times \Pr(A)}{\Pr(B)}$



Term	Meaning
Bias	If a survey procedure were repeated using the same methodology many times, bias is the difference between the true value and the mean of the estimated values.
Branch	In a scenario tree, one of several outcomes from a node. For instance, a node <i>age</i> may have two branches: <i>old</i> and <i>young</i> .
Category node	A node in a scenario tree that divides the population into two or more categories according to some criterion. The main types of category nodes are <i>risk category nodes</i> (which divide the population into groups that have different probabilities of being infected) and detection category nodes (which divide the population into groups with different probabilities of being detected).
Census	A survey or other surveillance activity that examines every member of the population (in contrast to a sample)
Clustering	The phenomenon whereby disease is not evenly distributed in a population but forms pockets of high prevalence amongst areas of no disease
Component sensitivity	The sensitivity of a component of a surveillance system – the probability that the component will detect disease if the population is infected at or above the design prevalence. Also called surveillance system component sensitivity (SSCSe)
Conditional probability	The probability of an event occurring given that another event is known to have occurred.
Design prevalence	A hypothetical prevalence of disease which against which the surveillance system is evaluated.
Detection category node	A node in a scenario tree that divides the population into groups with different probabilities of being detected.
Detection node	A node in a scenario tree relating to a step in a surveillance procedure that is necessary for a case of disease to be detected. The probability of success at this step is the sensitivity.
Effective probability of infection	The design prevalence multiplied by the adjusted risk, giving the probability that a particular group in a scenario tree will be infected
General surveillance	Surveillance that is able to detect many or any disease (in contrast to surveillance that is targeted at detecting only one disease)

Term	Meaning
Herd sensitivity, Group sensitivity	The probability of detecting at least one infected animal when a herd is examined.
Independence	In probability, two events are independent if the outcome of one is not influenced by the outcome of the other
Infection node	In a scenario tree, a node that describes the probability that an animal or group of animals will be infected. The probability in the <i>infected</i> branch of an infected node is the design prevalence.
Monte Carlo simulation	An analytical technique involving repeating an analysis many times, using different input parameters drawn randomly from defined distributions.
Node	In a scenario tree, a node represents a factor that may take two or more values, each with assigned probabilities.
Passive surveillance	An activity in which the primary purpose for the collection of the data is <i>not</i> surveillance.
Population proportion	The proportion of the entire population that has some defined characteristic of interest (in contrast to the surveillance system component proportion)
Posterior probability	An estimate of the probability of event, calculated using Bayes' theorem, based on prior knowledge and new information.
Prior probability	When using Bayes' theorem, an estimate of the probability of an event occurring, before new information about the event has been collected.
Random error	In sampling, error due to the random effect of selecting one animal or another. Random error leads to lack of precision that can be minimised by using a large sample size
Relative risk	The probability of an event occurring in one part of the population, divided by the probability of it occurring in another. Also known as risk ratio.
Risk	The probability of an adverse event occurring. In this book, risk is defined simply as a probability, in contrast to its use in risk analysis, where it is likelihood combined with consequences.
Risk category node	In a scenario tree, a node that classifies the population into two or more groups each with a different risk of being infected.
Scenario-tree	A branching quantitative model used for the analysis of surveillance systems components.

Term	Meaning
Sensitivity	The probability of getting the right answer from a test on an <i>infected</i> population. The true positive rate.
Specificity	The probability of getting the right answer from a test on an <i>uninfected</i> population. The true negative rate.
Stochastic	Describes a process involving chance.
Surveillance system	When applied to surveillance for a particular disease, the collection of activities that produce data that contribute to our understanding about the status of that disease.
Surveillance system component	A component of a surveillance system. A single activity that produced data about disease status. Abbreviated as SSC.
Surveillance system component proportion	The proportion of animals or herds in a surveillance system component that have a characteristic of interest, in contrast to the population proportion.
Surveillance system sensitivity	The probability that the surveillance system would detect disease if the population is infected at or above the design prevalence
Syndrome	A defined collection of clinical signs possible with other epidemiological information.
Systematic error	An error in surveys or surveillance that results in the expected value (mean value of many repetitions of the activity) is different from the true population value. Systematic error causes bias or lack of accuracy, and may be caused by sampling bias, measurement bias, analysis bias or confounding.
Targeted surveillance	<ol style="list-style-type: none"> <li>1. Surveillance aimed at detecting a specific disease, as opposed to general surveillance.</li> <li>2. Surveillance targeted at a portion of the population. Risk-based surveillance.</li> </ol> <p>The two different usages is unfortunate, but context usually makes it possible to determine which meaning is intended.</p>
Unit sensitivity	A general term for the probability that a single unit passing through the surveillance system would be detected as being infected, assuming that the population is infected at or above the design prevalence.

## Abbreviations and symbols

Symbol	Meaning
AR	Adjusted risk. This is a measure of risk in specified branch of a risk category node. The ratio of the adjusted risks between two branches has the same value as the relative risk, but ARs are adjusted to ensure that the average risk in the population remains equal to one.
CSe	Surveillance system component sensitivity (SSCSe)
CSeU	Component unit sensitivity. This is a general term for the probability that a single unit passing through the surveillance system would be detected as being infected, assuming that the population is infected at the design prevalence.
EPI	The effective probability of infection. This is the design prevalence multiplied by the adjusted risk.
GSe	Group sensitivity. The probability that disease would be detected in a group of animals.
$P(x)$	The probability of event $x$
$P^*$	The design prevalence. This is a hypothetical prevalence of disease against which the surveillance system is evaluated.
$P^*_A$	Animal-level design prevalence. This is the proportion of animals infected within an infected herd.
$P^*_H$	Herd-level design prevalence. This is the proportion of herds infected within the population.
PrP	The population proportion. This is the proportion of animals in the study population that fall into a specified group as defined by a category node.
PrSSC	The surveillance system component proportion. This is the proportion of animals in the SSC that fall into a specified group as defined by a category node.
RR	Relative risk (also known as risk ratio).
Se	Sensitivity (true positive rate)
SeA	The animal-level sensitivity. This is the same as the unit sensitivity when the animal is the unit of analysis (which is the most common case in livestock applications).
SeH	Herd sensitivity. This is the same as group sensitivity when the herd is the grouping level (which is the most common case in livestock applications).
Sp	Specificity (true negative rate)
SS	Surveillance system
SSC	Surveillance system component
SSe	Surveillance system sensitivity. The probability that the surveillance system would detect disease if the population is infected at or above the design prevalence

# Index

- @Risk, 114
- Abattoir meat inspection, 3
- Abattoir surveillance, 14
- Acceptable level of protection, 42
- Adjusted risk, 82
- Bayes' theorem, 29, 138
- Beta distribution, 112
- Beta-PERT distribution, 113
- Bias, 6, 11
- Binomial distribution, 113
- Blood, 15
- Branch, 69
- Census, 5
- Classifications, 7
- Clustering, 54, 70, 128
- Component. *See* Surveillance system component
- Component sensitivity, 91
- Components
  - combining, 136
  - overlapping, 138
- Comprehensive coverage, 6
- Conditional probabilities, 26, 29
- Conditional proportions, 100
- Confidence, 39
- Cost, 12
- Country, 2
- Coverage, 5
- Design prevalence, 40, 45
  - integer, 44
  - selecting, 41
- Detecting disease, 4
- Detection probabilities, 97
- Diagnosis, 7
- Discrete distribution, 113
- Disease, 2
- Disease control, 4
- Disease eradication, 4
- Distribution of disease, 4
- Early warning, 44
- Effective probability of infection, 82
- Expert opinion, 97, 98, 102
  - choosing experts, 103
  - combination of, 115
  - combining, 104
- Farmer disease reporting system, 3
- Freedom from disease, 4, 34
  - absolute and relative, 45
  - probability of, 144
- Freedom software, 114, 154
- Historical surveillance, 147
- Hypergeometric, 51, 113
- Independence
  - between animals, 128
- Independent, 23
- Indirect indicators, 9
- Indirect surveillance, 16
- Infection, 2
- Limb, 69
- Lognormal distribution, 113

- Meat inspection, 14
- Model parameters, 88
- Monte Carlo simulation, 111
- Negative reporting, 9, 18
- Node, 69
  - category, 71
  - detection, 70
  - detection category, 72
  - group category, 72
  - infection, 70
  - order of, 77
  - risk category, 72
  - types of, 70
- Normal distribution, 112
- Outcome, 69
- Participatory disease surveillance, 18
- Passive disease reporting system, 12
- PDS. *See* participatory disease surveillance
- Period of analysis, 149
- Popper, Karl, vii
- PopTools, 114
  - installation, 123
  - random variable functions, 123
  - reference, 123
- Population proportion, 94
- Posterior, 30, 149
- Practicality, 12
- Precision, 12
- Prior, 30, 145
- Probability, 20
- Probability distributions, 28, 112
- Probability rules
  - AND, 22, 28
  - NOT, 25
  - OR, 24, 28
- Proportions, 99
- Random error, 11
- Random variables, 21
- Relative risk, 79, 94, 101
- Representative sampling, 91
- Representative surveys, 47
- Representativeness, 5
- Risk
  - describing, 78
  - incorporating, 78
- Risk factors, 10
- Risk of introduction, 148
- Sample size, 52
- Sampling, 5, 36
- Scenario-tree
  - branch probabilities, 69
  - building, 73, 87
  - calculating, 89
  - introduction, 66
  - purpose of, 68
  - stochastic analysis, 119
- Sensitivity, 11, 27, 31, 38, 50
  - herd-level, 130
  - formulae, 134
  - step-wise calculation, 129
  - survey, 48
- Sentinel herds, 5, 15
- Signs, 8
- Small populations, 51
- Specificity, 31, 38, 52
- Specimens, 7, 15
- Stochastic modelling, 111
- Surveillance
  - active, 3
  - Disease focus, 3
  - general, 4
  - historical, 147
  - origin of information, 3
  - passive, 3
  - purpose, 4
  - quality of, 10
  - risk-based, 57
  - sensitivity, 59, 87
  - specificity of, 39
  - targeted, 3
- Surveillance system, 2
  - complex, 61
- Surveillance system component, 2
- Surveillance system component proportion, 94
- Survey design, 48
  - optimising, 56
- Surveys
  - structured, 61
- Syndromes, 8
- Syndromic surveillance, 8, 16
- Systematic error, 11
- Targeting, 78, 81
- Tests, 31, 96
  - combination of, 33
- Tissues, 15
- Triangular distribution, 113
- Two stage surveys, 54
- Uncertainty, 105, 109
- Unit sensitivity, 87, 90
- Variability, 105, 106, 109
- Variation, 59
- Vector surveillance, 10

Zero reporting. *See* negative reporting