CEBRA Project 180601: Alternative approaches to developing assurance about the regulatory compliance of consignments of plant products

Raphael Trouvé[1,2], Mark Ducey[3], Rose Souza-Richards[4], David Dall[5], and Andrew Robinson[1]

[1]CEBRA, The University of Melbourne
[2]SEFS, The University of Melbourne
[3]University of New Hampshire
[4]Ministry for Primary Industries, New Zealand
[5]Department of Agriculture and Water Resources, Australia

April 15, 2020

cebra
Centre of Excellence for
Biosecurity Risk Analysis

Contents

1 Executive summary **9**

2 Review of different inferential approaches to inspection **14**
2.1 Introduction . . . 14
2.2 Design-based and model-based inference: a simple example . . . 15
2.3 Design-based inference for consignments . . . 17
2.3.1 Simple Random Sampling . . . 17
2.3.1.1 Definition . . . 17
2.3.1.2 Estimating Prevalence . . . 18
2.3.1.3 Detecting Contamination . . . 19
2.3.1.4 Setting the confidence level, design prevalence, and sample size of an inspection . . . 22
2.3.1.5 Sampling Methodology . . . 22
2.3.2 Cluster and two-stage sampling . . . 24
2.4 Model-Based Inference . . . 25
2.4.1 Homogeneous populations and simple random sampling . . . 26
2.4.1.1 Estimating prevalence and detecting contamination . . . 27
2.4.1.2 The role of the sampling design . . . 29
2.4.2 Clustered populations or samples . . . 31
2.5 Bayesian inference . . . 33
2.5.1 Homogenous Populations and Simple Random Samples . . . 34
2.5.2 Clustered Populations or Samples . . . 37
2.6 Further Alternatives . . . 38
2.6.1 The Dempster-Shafer theory of evidence . . . 38
2.6.2 Imprecise Probabilities . . . 41

3 Case studies **44**
3.1 Simple random sampling . . . 44
3.1.1 Design-based inference for simple random sampling . . . 44
3.1.2 Model-based inference for simple random sampling . . . 45
3.1.3 Bayesian inference for simple random sampling . . . 45
3.1.3.1 Beta distribution as conjugate prior to the Binomial likelihood . . . 45
3.1.3.2 Bayesian inference using a noninformative prior. Case study with a 0.5% risk cutoff . . . 46
3.1.3.3 Bayesian inference using informative priors. Case study with a 0.5% risk cutoff . . . 46
3.1.3.4 Bayesian inference, case study with a 0.01% risk cutoff . . . 48
3.1.4 Dempster-Shafer theory of evidence for simple random sampling . . . 49
3.1.4.1 Dempster-Shafer theory, case study with a 0.5% risk cutoff (600 samples inspection) . . . 50

3.1.4.2 Dempster-Shafer theory, case study with a 0.01% risk cut-off (29956 samples inspection) . . . 53
3.1.5 Imprecise probabilities for simple random sampling . . . 53
3.2 Clustered sampling . . . 54
3.2.1 Design-based inference for clustered sampling . . . 54
3.2.2 Model-based inference for clustered sampling . . . 54
3.2.2.1 Using the Beta-binomial model for clustered data. . . . 54
3.2.2.2 Estimating the intra-cluster correlation coefficient ρ from past data . . . 56
3.2.3 Bayesian inference for clustered sampling . . . 60
3.2.4 Dempster-Shafer theory of evidence for clustered sampling . . . 63
3.2.5 Imprecise probabilities for clustered sampling . . . 63
3.3 Systems approach . . . 63
3.3.1 Design-based inference . . . 64
3.3.2 Model-based inference . . . 65
3.3.3 Bayesian inference . . . 65
3.3.4 Dempster-Shafer theory of evidence . . . 65

4 Adaptive Inspection Schemes **67**
4.1 Introduction . . . 67
4.2 Inspect Only Some Consignments . . . 67
4.2.1 Continuous Sampling Plans . . . 67
4.2.1.1 CSP-1 . . . 67
4.2.1.2 CSP-2 . . . 68
4.2.1.3 CSP-3 . . . 68
4.2.2 Skip-Lot Sampling Plans . . . 69
4.3 Inspect All Consignments; Vary Intensity . . . 70
4.3.1 MIL-STD-1916 . . . 70
4.4 Choosing operational parameters . . . 72

5 Discussion **74**

6 Summary, recommendations, and conclusions **76**

List of Figures

2.1 Posterior probability that the actual prevalence p is less than or equal to the design prevalence $p^* = 0.005$ as a function of sample size when no contamination has been detected, for two common noninformative prior distributions. 36

2.2 Upper 95% credible intervals for the prevalence p as a function of sample size when no contamination has been detected, for two common noninformative prior distributions. 37

3.1 a. Estimated distribution of p_j among consignments in the import plant pathway to Australia. b. Posterior distribution of infestation rate in an incoming consignment after finding zero BRM out of 183 samples during an inspection. 47

3.2 a. Infestation rate among consignments in a specific germplasm pathway. b. Posterior distribution of infestation rate in an accepted consignment of the pathway after finding zero BRM out of 2589 samples during an inspection. The vertical red lines represent p^*=0.01% (92% of the p_j values are left of the red line in a, while 95% of the values are left of the red line in b). Note that since the infestation rate of the pathway is already very low, the inspection doesn't reduce the infestation rate by much (posterior mean infestation rate of the inspected consignment is 2.07×10^-5, which is only 0.78 times lower than the prior mean of 2.63×10^-5). 49

3.3 a. Sample size for clustered data given by the beta-binomial model (approximate and exact solution). b. Ratio of sample size for the beta-binomial (approximate and exact) and the binomial distribution when varying ρ and n_k. 55

3.4 a. Sample size for clustered data given by the beta-binomial model (approximate and exact solution). b. Ratio of sample size for the beta-binomial (approximate and exact) and the binomial distribution when varying ρ and n_k 57

3.5 Conceptual diagram of the hierarchical Beta-Binomial model. 58

3.6 Mean and standard deviation of ρ from the hierarchical beta-binomial model estimated from potential plant pathways of different size. 59

3.7 Mean and standard deviation of ρ from the hierarchical beta-binomial model estimated from potential pathways of different size. 60

3.8 Posterior distribution of p_j after a clean inspection sample with a 0.5% risk cutoff. 62

3.9 Posterior distribution of p_j after a clean inspection sample with a 0.01% risk cutoff. 63

List of Tables

3.1 Possibility to use external information in different inference frameworks . . 44

4.1 Code letters for entry into the sampling tables for MIL-STD-1916 71
4.2 Attributes sampling plans for MIL-STD-1916. 71

Glossary

Term	Definition
Parameter	A parameter is a numerical attribute relating to the entire population of interest. In a border inspection setting, the numerical attribute that we wish to estimate is often the infestation rate of incoming consignments.
Infestation rate	The infestation rate (sometimes called prevalence) is the proportion of infested units (*i.e.*, units that contain Biosecurity Risk Material, BRM) in a consignment.
Sensitivity	The sensitivity of an inspection (sometimes called the confidence-level) is the minimum probability with which we wish to detect at least one BRM in the inspected sample given that the baseline contamination rate is at the design prevalence or higher. In the '600 samples rule', the sensitivity is 95%.
Design prevalence	The design prevalence (sometimes called risk-cutoff or detection level) is the lower limit of the infestation that we want to detect with a given sensitivity. In the '600 samples rule' the design prevalence is 0.5%. This should not be interpreted as a tolerance level: even if the estimated rate is below the level of detection, which it can be, so long as contamination is detected, the consignment will not be released.
Sample size	The sample size is the chosen number of units that will be inspected. This is usually denoted by the letter n. The sample size is typically chosen to have a given sensitivity to detect a given prevalence.
Prior	In Bayesian inference, the prior distribution represents our knowledge or belief of a parameter of interest before seeing the data.
Non-informative prior	A non-informative prior represents our ignorance of the value of the parameter. For example, if our parameter is a probability, it might be a uniform distribution on the 0—1 range. Note that most of the so-called non-informative priors still affect the posterior distribution of the parameter in some ways.
Informative prior	When we know something about the range of values that a parameter can take, it can be included in the prior distribution. For example, we might know from past data that the distribution of infestation rate among different consignments on the pathway follows a $Beta(.18, 8)$ distribution.

Term	Definition
Conjugate prior	In Bayesian inference, a prior is said to be conjugate for the likelihood function if when combining prior and likelihood, the posterior distribution is of the same probability distribution family than the prior distribution. An example relevant to biosecurity inspection is that combining a Beta prior distribution with a Binomial likelihood for the inspection data will lead to a posterior distribution that is also Beta distributed.
Expected value	The hypothetical mean value of a sampling distribution over many repetitions of the sampling.
Estimator	A rule or method of estimating a parameter of a population.
Bias	A bias is a systematic error. The bias of an estimator is the difference between the expected value of the estimator and the true value of the parameter being estimated.
Efficient estimator	An efficient estimator is the estimator that has the lowest variance among all unbiased estimator of the parameter.
BRM	Biosecurity Risk Material. A sampled unit that is infested by a pest, disease, or anything that is considered a biosecurity threat and that would render the consignment non-compliant.
Cluster sampling	With cluster sampling, we divide the population into separate groups, called clusters. Then, we randomly select clusters from the population, and we sample several units per cluster (*e.g.*, selecting individual fruits within selected crates of fruit within a container, rather than selecting individual fruits completely at random from the entire container).
Intra-cluster Correlation	The intra-cluster correlation coefficient (ICC, often written as ρ in equations) characterises the degree of similarity shared by units contained in the same cluster. For example, an ICC of one indicate that all units from the same cluster are exactly the same (*i.e.*, if one unit of a specific-cluster is infested, all the other units from the same cluster are infested and vice-versa). An ICC of zero indicates that two units sampled from the same cluster are no more similar than two units sampled from different clusters.

1 Executive summary

Consignments of plant products represent a significant potential pathway for invasive pests into Australia and New Zealand. Currently, the Department of Agriculture of Australia (DA) and the Ministry of Primary Industries of New Zealand (MPI) assess the risk associated with each consignment solely from its inspection data. However, testing seeds for pests and pathogens is commonly destructive, and concomitantly laborious and expensive. CEBRA projects 1806, 'Alternative approaches to developing assurance about the regulatory compliance of consignments of seeds' for DA and 180601 'Models for Border Inspection for pelleted seeds' for MPI, aimed to develop a statistical framework that will generalize the models that are applied in ISPM 31 (International Plant Protection Convention, 2008) to allow the provision of other sources of data in the decision making. This can include inspection history, or audit information and inspection results from parent lots that might arise from systems approaches.

The report is structured as follows. In chapter 2, we review design- and model-based inference frameworks that underlie ISPM 31 biosecurity inspection system, and allow decision making (*i.e.*, deciding if a consignment is compliant or not) by providing assurance about the correct sample size to be adopted with representative sampling to determine if a proportion of units that may be infested in a given consignment after inspection. Additionally, we review three alternative inference frameworks that might be used in biosecurity (*i.e.*, imprecise probability theory, Bayesian inference, and Dempster-Shafer theory of evidence), the latter two frameworks allow combining inspection data with external information when making inference. Typically, using external information will reduce the sample size required to make a decision on the compliance of a consignment. We review how these five frameworks interact with two typical type of data collected in biosecurity inspections (simple random samples vs. clustered samples). Typically, clustered sampling increases the sample size required to make a decision on the compliance of a consignment.

In chapter 3, we provide case studies for design, model-based, Bayesian inference, Dempster-Shafer, and imprecise probability theory for both simple random sampling and clustered sampling (when applicable) and their potential use in systems approach. In chapter 4, we review different adaptive inspection schemes which allow using external information by choosing to inspect or not inspect consignments based on the recent inspection history of the pathway. We conclude the report by summarizing the pro and cons of using these alternative frameworks.

Pro and cons of the different inference frameworks reviewed.

Several inference frameworks can be used to develop assurance about the regulatory compliance of consignments of germplasm. While some frameworks allow using external information when making inference (Bayesian, Dempster-Shafer, and to some extent, model-based inference) others do not (design-based inference, imprecise probability theory) (see table 3.1). Frameworks that do not allow using external information are of limited use for systems approach (analyzing systems approach data requires combining different sources

of evidence). Below, we summarize the pro and cons of the five framework that will be reviewed in this report.

Design-based inference

- This is the main type of inference used for border biosecurity inspection.
- In design-based inference, we can draw conclusions about the population from the sample because we know exactly how the sample was collected. No additional assumption is required which makes the method particularly objective.
- When the inspected units comes from a simple random sample, we can use the binomial sample size formula (Eq. 3.2) to compute sample size. This is the basis of the '600 samples' rule often used in biosecurity and also the basis for the 31,540 samples used for the plant product data supplied by New Zealand.
- When the data arrives in clusters but we still manage to do simple random sampling, we can also use Eq. 3.2 to compute sample size (simple random sampling protects against the detrimental effect of clustering on sensitivity and sample size).
- Does not allow the use of external information.

Model-based inference

- In model-based inference, we postulate a model that might have generated the data (*i.e.*, the inspection data might have been generated from a Binomial model), check the assumptions of the model, and make inference about the infestation rate.
- When the data comes from simple random sampling, model-based inference give the same sensitivity and sample size than design-based inference (Eq. 3.2).
- When there is clustering, we can use Eq. 3.9 to compute sample size. This requires estimating or fixing the intra-cluster correlation coefficient (ICC) of the pathway.
- Allows limited use of external information (for example, to estimate the ICC of the pathway from past data, section 3.2.2.2).

Bayesian inference

- In Bayesian inference, we postulate the potential values that the parameter of interest might take (prior information before seeing the data) as well as a model that might have generated the inspection data. We then combine the prior and the model with the inspection data to make our inference on the parameter of interest (typically the infestation rate of the consignment being inspected).
- When we use a non-informative uniform prior on the infestation rate of the consignment being inspected, Bayesian inference gives the same sample size as design-based and model-based inference for simple random sampling data (section 3.1.3) and as model-based inference for clustered sampling (section 3.2.3).

- The strength of Bayesian inference however is that it allows combining external information (informative prior) with inspection data (likelihood) to draw conclusion about the infestation rate of a consignment. Using informative prior (for example calibrated from past data on the pathway) allows to reduce sample size in both the simple random sampling and the clustered sampling cases (sections 3.1.3 and 3.2.3). In the case of a potential 0.5% pathway, the sample size can be reduced by a factor of around three compared to design-based inference. In the case of a potential 0.01% pathway, the sample size can be reduced by a factor of around 11.

- One issue that arises when using an informative prior is the assumption of stationarity (past data are representative of future data). We suggest monitoring and re-estimating the distribution of infestation rate among different consignments of the pathway regularly (perhaps every year). We can also use mixture priors to 'robustify' our prior.

- Another issue with Bayesian inference is that we do not always have analytical solutions for our estimates or our decision criteria. In the case of simple random sampling, we have an analytical distribution for the posterior p_j but we have to compute the sample size numerically. In the case of clustered sampling, both the posterior distribution of p_j and the sample size have to be computed numerically (by fitting the hierarchical model to a clean inspection data of different sizes and observing the effect on the posterior).

Dempster-Shafer theory of evidence

- Dempster-Shafer theory of evidence works directly on the decision scale (probability of compliance) rather than the infestation rate of the population. Dempster-Shafer theory is typically used to combine different lines of evidence when making inference. Each line of evidence can arise from a model and inspection data (*e.g.*, a Binomial model generated the observed inspection data) or can be completely subjective (*e.g.*, experts think that the proportion of compliant consignments in this specific pathway that used a systems approach is 90%).

- With only one source of evidence and in the simple random sampling case, the sample sizes are similar to those given by Bayesian inference with non-informative prior.

- The framework might be difficult to extend to support clustered sampling.

- The Dempster-Shafer framework allows combining external information when making inference. There are several ways to do so and perhaps not much to decide between them (see for example Rathman et al., 2018).

Imprecise probability theory

- Imprecise probability theory is a specific type of Bayesian analysis that was created to avoid having to fix a specific non-informative prior when we are ignorant about the value of the parameter of interest.

- The sample sizes are similar to Bayesian inference with a uniform prior in the case of simple random sampling.

- The framework might be difficult to extend to support clustered sampling.
- Does not allow the use of external information when making inference (Imprecise probability theory is all about non-informative priors).

Adaptive Inspection Schemes

- Adaptive inspection schemes provide a light-touch approach for implementing risk-based intervention.
- The sample sizes depend on recent inspection history.
- Reasonably easy to implement.
- Does not explicitly allow using external information when making inference but work-arounds are possible.

Conclusion and recommendations.

Of the five frameworks reviewed, Bayesian inference seems to be the most promising to allow incorporating sources of data other than the current inspection sample when making a decision. At this point, it is worth noting that there is a fundamental difference in the scope of the classical and Bayesian frameworks. Whereas the classical approach focuses on detection, the Bayesian approach focuses on estimation:

- In the classical approach to biosecurity inspection, the sample size is calculated to give a 95% confidence of detecting a consignment that has a 0.5% infestation rate.
- In the Bayesian approach, the sample size is calculated to give 95% confidence that the estimated infestation rate in an accepted consignment is less than 0.5%.

However and despite this perspective difference, Bayesian inference is compatible with current methods used in biosecurity: when using non-informative priors (*i.e.*, representing our ignorance of the infestation rate of the consignment before inspection), Bayesian results are similar to design and model-based inference (*e.g.*, after a clean '600 samples' inspection with a uniform prior, Bayesian methods infer that there are 95% chances that the infestation rate in an accepted consignment is below 0.5%). If available, Bayesian inference allows using information from external sources of data, which reduces the sample size required to make a decision on the compliance of a consignment. However, this comes at a cost: if future data are different from past data, we are no longer guaranteed to detect a given prevalence with a given sensitivity (as with the design-based inference procedure). There are different ways to penalize an informative prior. The most promising approach is to use a mixture prior that combines the informative prior with a uniform prior. This approach allows for the possibility that some of the future consignments might have an infestation rate higher than what we have seen in past data.

Data collected from a clustered population can result in noticeably reduced sensitivity for an inspection scheme. Keeping the sensitivity constant (with respect to simple random sample inspection) requires sampling more units. How many more units to sample will depends on the intra-cluster correlation coefficient ρ (*i.e.*, the degree of similarity among units sampled from the same cluster) and the number of units sampled per cluster n_k (the higher ρ and n_k, the higher we will need to increase the sample size to be to keep the

sensitivity constant) (see Fig. 3.3). When the infestation rate is relatively high (in the 0.5–2% range), it is possible to reliably estimate ρ for a pathway using model-based or Bayesian inference. However, when the infestation rate is very low (*e.g.*, in the case of NZ data, with a typical mean infestation rate of 0.003%), it is difficult to reliably estimate ρ from a pathway, even for large pathways (100 consignments).

Alternatively, adaptive inspection regimes might be a useful first step if early action is valuable and when we do not have enough data to apply Bayesian approaches.

2 Review of different inferential approaches to inspection

2.1 Introduction

Invasive species, including arthropods, plants, fungi, and microbial pathogens, pose a significant and growing risk to managed and native ecosystems worldwide. Hoffman and Broadhurst (2016) estimate the total annual cost of invasive species to the Australian economy in 2011–2012, including both losses and control expenditures, at nearly AU$14 billion, or approximately AU$560 per capita. Likewise, Colautti et al. (2006) estimate the costs associated with invasive species in Canada at CDN$34.5 billion, or at over CDN$1000 per capita, nearly 3% of GDP. Comparable costs for other major economies include USD$100.6 billion for the United States in 2003 (Pimentel et al., 2005), rising to over USD$200 billion by the end of that decade (Pimentel, 2011), USD$18.9 billion for China (Wan and Yang, 2016), and NZ$3.29 billion for New Zealand in 2009 (Giera and Bell, 2009). In both cases, losses plus expenditures total nearly 2% of gross domestic product (GDP) (Pimentel et al., 2005; Giera and Bell, 2009). Losses as a proportion of GDP are lower for many other developed economies, but inconsistent methodology and reporting makes it challenging to account for the full costs of invasions (Hoffman and Broadhurst, 2016).

Both the impacts of an incursion and the costs of control escalate dramatically once an invader has become established, so effective surveillance at the border is a critical component in national-scale efforts to reduce the social and ecological cost of invasive species (Whattam et al., 2014; Quinlan et al., 2015). Despite the values at stake, biosecurity surveillance programmes typically operate within tight budget constraints. For example, inspection is based on sampling, rather than an exhaustive inspection, for nearly all types of goods crossing international borders. When inspection is destructive, such as for the import of pelleted seeds, sampling is the only way to go if we don't want to destroy the whole consignment. The adoption of sampling necessitates a further decision, namely: how large of a sample should be taken to adequately manage the biosecurity risk of incursion? Existing international agreements, such as the ISPM-6 and ISPM-31 guidelines of the International Plant Protection Convention (International Plant Protection Convention, 2008; International Plant Protection Convention, 2016), recommend that biosecurity decisions be based on scientifically and statistically sound procedures, but provide little specific guidance on the procedures themselves.

Perhaps as a result, a wide range of applications rely on a small set of sampling and statistical approaches that have come to be accepted as common practice. One example is the 600-units sample, hereafter the 600 sample, which is designed to achieve a specified level of sensitivity (95%) when the prevalence of biosecurity risk material (BRM) is 0.5 percent within a consignment, assuming that inspection is carried out without error. The "600 sample" approach has been adopted as standard practice in a range of situations (Ransom, 2017; Ormsby, 2017), sometimes with adjustments for the number of units in a

consignment to be inspected (*e.g.*, New Zealand Ministry for Primary Industries, 2016).

The mathematical computations associated with a variety of common sampling situations have been well described in the literature (*e.g.*, Venette et al., 2002a). Unfortunately, in biosecurity as in other areas of application, the *conceptual* basis for drawing inferences from sample data is rarely specified (see, *e.g.*, Gregoire, 1998). This can lead to confusing or even contradictory interpretations of key terms such as bias and independence, with important consequences for what is considered an acceptable sampling design, or what is a valid inference once data have been collected. These distinctions become especially important when reality departs from an idealized case of simple random sampling within homogeneous consignments.

The goal of this chapter is to clarify the assumptions and implications of the two primary modes of inference from sample data, namely the design-based and the model-based approaches (Gregoire, 1998). Although key distinctions between these approaches have been well-characterized and debated in the literature on sampling theory (*e.g.*, Särndal et al., 1992; Little, 2004; Fuller, 2009; Chambers and Clark, 2012; Magnussen, 2015), the nature of that literature makes it less than fully accessible to many practitioners and even researchers in the biosecurity and risk assessment communities.

In order to make this material more accessible here, we attempt to be rigorous and correct while employing a minimum of notation, derivations, or proofs. First, we clearly define design- and model-based inference. Then, we explore the consequences of each paradigm, using the 600 sample as a starting point, beginning with the simple case (simple random sampling from a homogeneous consignment) and proceeding to consider the consequences of inhomogeneous consignments, and of drawing samples in clusters (*e.g.*, selecting individual fruits within selected crates of fruit within a container, rather than selecting individual fruits completely at random from the entire container). We briefly explore Bayesian and alternative inferential paradigms, with attention to their relationship to more traditional model-based approaches, such as Frequentism. Practical case studies on the consequences of using different inference framework in a biosecurity context are provided in chapter 3.

2.2 Design-based and model-based inference: a simple example

Simply put, the difference between design-based and model-based inference is in how inference from the sample is connected, conceptually, to the target process.

In design-based inference, we can draw conclusions about the process from the sample because we know exactly how the sample was collected — more specifically, the sample must be collected according to one of a number of designs, and the appropriate analysis carried out. In model-based inference we propose a probability distribution — a model — for the observed random variable, we test whether the model is correct, and then draw conclusions using the model.

An example that captures the flavour of the difference follows. Imagine we provide to you the following data, which represents a sample of biological/assigned sexes of schoolchildren: MFFFMFMFFFMFMFF. Your task is to estimate the underlying proportion of F's in the process from which these data were sampled. How to proceed?

One way would be to assume that the data follow a Binomial distribution, which requires (i) that the outcomes be independent of one another, (ii) that the probability of an F be a fixed (but unknown) quantity, and (iii) that the sample size be known in advance. If these three assumptions are true, then the data do follow a Binomial distribution, and we can proceed to use the standard approach. This is *model-based* inference.

But, any one of these assumptions may be false. We may have sampled human children sequentially in a schoolyard, in which case we would expect sex-based clustering, which would induce auto-correlation, contradicting the first assumption. We may have sampled multiple groups that have different proportions of F, also called heterogeneity, contradicting the second. And, the samples may have been collected according to the following stopping rule: sample until we reach the required number of F's (namely, 10), contradicting the third. In each case the Binomial distribution is wrong and may lead to misleading estimates, especially of the uncertainty. This is important for the current context because estimates of the uncertainty determine how much assurance can be gained from an inspected sample. We must check these assumptions in order to proceed confidently with model-based inference.

Alternatively, we may know that the sequence of F and M arose from the following scenario. The names of 2000 students were recorded on a spreadsheet and a unique integer assigned to each. Fifteen random integers were selected from the unique integers, and the assigned sexes of the students thus selected were identified. Now, we know that no clustering is possible, because the sample was selected randomly from the list of names. We know that if there are sub-populations then the sampling occurred randomly across them, so whether or not they differ has no bearing on the statistical qualities of the estimate. (Indeed, unbeknown to us, someone could have constructed the original list of students by concatenating a list of male students with a list of female students; the genders could be completely segregated within the original list, and the validity of our inference would be unaffected.) And, we nominated a sample size of 15, so we knew the sample size in advance. We can use the Binomial distribution with confidence because the assumptions are satisfied by the design. This is *design-based* inference.

Although the difference between design- and model-based inference is conceptual, it carries important practical consequences. Within a model-based framework, we must be concerned with whether, or how well, the data and the mechanisms that generated the data conform with our assumed model. If that conformity is poor, then the inferences from our survey will be suspect. On the other hand, certain aspects of the design — such as the use of purposive (*i.e.*, subjective) or other non-probability sampling that would be anathema in the design-based context — may not create serious problems. (As Magnussen (2015) observes, however, claims that the design is completely irrelevant to model-based inference have done much to undermine confidence in its application.) From a design-based perspective, the correct implementation of a stated probability-driven sampling design is paramount. For example, in simple random sampling, sampling must be truly random; haphazard or "convenience" sampling seriously undermines confidence in any inferences. But, commonly few if any assumptions about the underlying structure of the population or the mechanisms that generated it are made. As we shall see below, the failure to specify which mode of inference is in operation has occasionally led to substantial confusion in recommendations about the design and implementation of sampling methods for biosecurity.

2.3 Design-based inference for consignments

Design-based inference is the older of the two dominant paradigms in survey sampling, with origins in seminal papers by Neyman (1934) and Neyman (1938). Neyman's work appeared at a time when the statistics community was grappling with serious questions about the meaning of "representative sampling" (Kruskal and Mosteller, 1980), and the design-based perspective quickly rose to dominance as dramatic advances in sampling theory were made during and after the Second World War. Classic texts on design-based inference include Hansen et al. (1953), Kish (1965), Sukhatme and Sukhatme (1970), and Cochran (1977); more recent texts include Gregoire and Valentine (2008) and Thompson (2012). The design-based perspective does not preclude the use of models, but it does restrict their role and influence. Design-based approaches that are informed by models are often called *model-assisted* (see, *e.g.*, Särndal et al., 1992).

In the design-based paradigm, the population parameter about which inference is to be made is fixed but unknown. The attributes of the sample units that comprise the population, including those about which we wish to draw inferences (such as their association with BRM) as well as those that might influence the outcome of sampling (such as their position and proximity to one another), are fixed. Randomness enters the sampling process through the selection of units into the sample, and the procedure for selecting units is governed by a design. The resulting probability distribution of inclusion or exclusion of sample units is often called the randomization distribution. The inclusion of the i^{th} unit can be described by a binary random variable δ_i that take the value $\delta_i = 1$ if the unit is included, and $\delta_i = 0$ if it is not; it is the δ_i, not the y_i (*i.e.*, the status non-infested = 0, infested = 1 of the unit i), that are considered as random. Inference is based on the use of estimating equations that are appropriate to the design.

A great deal of effort in design-based sampling theory is focused on proving the properties of the estimators (such as unbiasedness, and having unbiased estimates of variance) under a given design, and with few or no assumptions about the characteristics of the population. For example, under simple random sampling, the sample mean is known to provide an unbiased estimate of the population mean for a given attribute, without any assumption about the distribution of that attribute (such as normality) within the population as a whole (*e.g.*, Thompson, 2012, Chapter 2). Here, "unbiasedness" refers to a mathematical expectation over all possible outcomes of sampling under the given design, including those that were not observed, *i.e.*, if θ is a population parameter of interest, $\hat{\theta}$ is an estimate computed from sample data using an estimating equation, and

$$\mathrm{E}[\hat{\theta}] = \theta$$

where E[] indicates expectation over the possible samples under the design, then $\hat{\theta}$ is said to be *design-unbiased*.

2.3.1 Simple Random Sampling

2.3.1.1 Definition

A sampling design can be considered a simple random sample if it meets the following criteria:

1. The sample size n is fixed and known in advance.
2. Sample units are included in the sample by a chance mechanism.

3. Every sample unit has the same probability of being included in the sample.

4. Whether a given unit is included in the sample or not is independent of whether or not any other unit is included in the sample.

For example, suppose our population is a standard deck of 52 playing cards. The cards are well-shuffled, and the first 4 cards are drawn from the top of the deck. The shuffling is (for all practical purposes) a chance mechanism, and it creates a situation in which each card has an equal chance of being the first, second, third, or fourth card in the deck. The probability that the ace of spades will be in the sample of four cards, is the same as the probability that the ace of diamonds or the two of hearts or any other specific card is in the sample, namely 4/52. If we know that the first card actually is the ace of spades, the probability of any other card being among the remaining three is 3/51; the occurrence of the ace of spades does not, for example, make the ace of diamonds more probable to occur than the two of hearts, or vice versa. Thus, the inclusion of sample units (cards) is independent. In this example, we have simple random sampling without replacement as our design; had each card been put back into the deck and the deck reshuffled after each individual draw, we would have simple random sampling with replacement. In most biosecurity examples that use simple random sampling, sampling is without replacement — we would not inspect the same item twice.

Note that some apparently innocuous (and some less innocuous) sampling approaches do not satisfy the requirements of a simple random sample. For example, sampling every tenth unit that occurs in a sequence is a systematic sample, rather than a simple random sample. If the start of sampling is driven by a random choice (rather than, say, always choosing the first unit that occurs and then every tenth thereafter), the sample is still a probability sample. But, if we know that the second sample unit has been chosen, we know that the first, eleventh, twenty-first, and so on will not be, while the twelfth, twenty-second, and so on certainly will be. From a design-based perspective, systematic sampling violates the requirement of independence – possibly in a way that is advantageous, but nonetheless one that has consequences for estimation (and especially for the design-based estimation of variance or uncertainty; see Cochran (*e.g.*, 1977, Chapter 8), Thompson (2012, Chapter 12)), and Gregoire and Valentine (2008, Section 3.2.2).

Haphazard selection of sample units, selection of those that are most convenient, or selection of those that are most representative or most likely to be contaminated based on the judgment of the inspector, are not simple random sampling designs because the selection is not based on a chance probability. Purposive selection of the units subjectively judged most likely to be contaminated may be advantageous from the perspective of detecting BRM in a consignment, if the judgment of the inspector is reliable. However, if simple random sampling formulae are used, then estimates of the prevalence of BRM will be biased. It is possible (through profiling) to assign unequal probabilities of selection to different sample units, so that inspection targets higher-risk material; this is often advantageous, but from a design-based perspective it strictly requires different estimating equations than those used for simple random sampling (*e.g.*, Horvitz and Thompson, 1952).

2.3.1.2 Estimating Prevalence

Traditional design-based sampling textbooks often focus on the estimation of the population mean, and accounting for the uncertainty of that mean. In the context of biosecurity

inspection of a single consignment, the problem most easily cast in terms of estimating a population mean is that of estimating the prevalence of BRM within the consignment. Following a straightforward approach to estimating proportions (*e.g.*, Thompson, 2012, Chapter 5), suppose that there are N sample units (say, pieces of fruit) in a consignment (say, a shipping container). We will sample n fruits using simple random sampling. Let y_i, $i = 1 \ldots N$, be a binary variable: $y_i = 0$ if the i^{th} fruit is clean, and $y_i = 1$ if the i^{th} fruit is contaminated. It is easy to show that

$$\hat{p} = \frac{1}{n} \sum_{i \in n} y_i$$

where $i \in n$ indicates the individual units in the sample, is a design-unbiased estimator of the population prevalence p, *i.e.*,

$$p = \frac{1}{N} \sum_{i=1}^{N} y_i$$

and moreover that

$$\hat{\text{var}}(\hat{p}) = \left(\frac{N - n}{N} \right) \frac{\hat{p}(1 - \hat{p})}{n - 1}$$

is a design-unbiased estimator of the sample variance of $\hat{p}$. Confidence limits can be constructed by a normal approximation (by invoking the Central Limit Theorem and multiplying $\sqrt{\hat{\text{var}}(\hat{p})}$ by the appropriate value of t with $n-1$ degrees of freedom), or more exact confidence limits can be computed (albeit with some computational effort) based on the hypergeometric distribution (Thompson, 2012, Section 5.2).

2.3.1.3 Detecting Contamination

As simple and appealing as the estimation of prevalence within a consignment might be, it is usually not the most important question for biosecurity surveillance at the consignment level. Rather, the purpose of most consignment-level inspection is to determine whether or not the prevalence p exceeds, or is less than, some maximally-allowable design prevalence p^*. We might wish to say that a consignment has zero prevalence, *i.e.*, $p^* = 0$, but actually proving that would require 100% inspection of the consignment. Instead, a more realistic goal would be to state that $p \leq p^*$, where p^* is a specified, very small value, with a given (and hopefully high) level of confidence.

It is in this very context that the so-called "600 sample" was developed, along with its many variants. Generally speaking, suppose that we draw n sample units by simple random sampling, from a consignment containing N such units. The total number of contaminated units in the consignment is $Y = \sum_{i=1}^{N} y_i$. If we detect contamination in *any* unit in our sample, *i.e.*, if $\sum_{i \in n} y_i > 0$, then we will reject the consignment. But if $\sum_{i \in n} y_i = 0$, we accept it. Formally, this sampling plan is an acceptance sampling plan with zero acceptance number (Stephens, 2001); the original development of acceptance sampling methods by H.F. Dodge and colleagues during World War Two strictly followed Neyman's then-new design-based paradigm.

Under the design, if n units are drawn without replacement from a population of size N containing Y contaminated units, then the number of contaminated units $x = \sum_{i \in n} y_i$ in the sample follows a hypergeometric distribution

$$h\left(x; n, Y, N\right) = \frac{\binom{Y}{x}\binom{N-Y}{n-x}}{\binom{N}{n}} \tag{2.1}$$

where

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

and

$$\text{For } a \text{ integer: } \quad a! = a \times (a-1) \times (a-2) \times ... \times 3 \times 2 \times 1$$
$$\text{For } a \text{ non-integer } a! = \Gamma(a+1)$$

where ! is the factorial function and Γ is the gamma function.

The quantity $S = 1 - h(0; n, Y, N)$ is the probability of detecting at least one contaminated sample, and equals the sensitivity of the sampling design at a prevalence $p = Y/N$. If N is much larger than n (say by a factor of 20), or if the sample were drawn with replacement, then the distribution of x is similar to a binomial distribution, for which the computations are much simpler (though this is much less relevant now than in the early development of acceptance sampling). The binomial distribution is also "conservative" in that the required sample size n to yield a desired sensitivity is slightly larger than that computed from the hypergeometric. Under the binomial assumption, the sensitivity is simply

$$S = 1 - (1-p)^n \tag{2.2}$$

and rearranging gives a simple formula for the required sample size for a specified sensitivity, at a design prevalence p^*,

$$n = \frac{\ln(1-S)}{\ln(1-p^*)} \tag{2.3}$$

Values of n for selected values of N and sensitivity, at different values of the design prevalence, have been tabulated in a number of publications (*e.g.*, Stephens, 2001; International Plant Protection Convention, 2008). A common choice in practical biosecurity work is a design prevalence $p^* = 0.005$, or 1 in 200 units contaminated, and a sensitivity of 95%. In other words, at the design prevalence, only 5% of consignments are accepted when they should have been rejected. The actual sensitivity is higher whenever prevalence is higher also. Under the binomial (or large-consignment hypergeometric) assumption, the required sample size is $n = 598$, which is almost always rounded to $n = 600$ for practical work. Although we will emphasize the binomial in much of the following discussion, we note that Lane et al. (2018b) show that significant cost savings can often be achieved when the hypergeometric is appropriate and is used instead of the binomial to calculate required sample sizes, especially in the case on plant product lots which are considered by the relevant authority as small.

In an ideal world, the design prevalence and sensitivity might depend on the likely prevalence of BRM within a particular type of incoming material, and the risk associated with an incursion. However, the "600 sample" and its underlying $p^* = 0.005$ and $S = 0.95$ appear to have become an entrenched default for historical reasons, much in the same way that Fisher's tentative suggestion of $P \leq 0.05$ as a potential criterion for inference about experiments (Fisher, 1925) evolved into a *de facto* standard for generations of researchers.

Note that the assumptions involved in using Equation 2.1, from a design-based perspective, are quite limited: only that a simple random sample of size n has been drawn from a population of size N that contains Y contaminated units. (The assumptions involved for the binomial distribution are likewise limited to drawing a simple random sample, from a population with prevalence Y/N.) Notably, no assumptions have been made about the

proximity of the sample units with $y_i = 1$ to one another (*i.e.*, their spatial autocorrelation). It might be that the contaminated units are all packed into one corner of the consignment, or they might be dispersed uniformly through the consignment. Under the assumption of simple random sampling, each successive draw is an independent selection from the entire consignment and the spatial autocorrelation has no effect on our inference (simple random sampling 'protects' against autocorrelation or clustering). Likewise, no assumption has been made that the sample units are identical in their propensity to be contaminated; each sample unit either is contaminated ($y_i = 1$) or clean ($y_i = 0$), and those attributes are fixed, not random. Thus, from a design-based perspective, if simple random sampling has been employed and we will either accept or reject the entire consignment, the use of Equation 2.1 is valid whether the units in the consignment arise from a common line with a uniform propensity for contamination, from multiple lines with different propensities, or even are different types of fruit entirely.

Viewed from the design-based perspective, then, certain prescriptions in current national or international guidance seem misplaced. For example, International Plant Protection Convention (2008, page 7) states,

> A lot to be sampled should be a number of units of a single commodity identifiable by its homogeneity in factors such as: origin; grower; packing facility; species, variety, or degree of maturity; exporter; area of production; regulated pests and their characteristics; treatment at origin; type of processing ... Treating multiple commodities as a single lot for convenience may mean that statistical inferences can not be drawn from the results of the sampling.

While there may well be practical, political, or policy reasons for segregating different lines (and perhaps accepting or rejecting lines separately), within the design-based framework none of the factors listed is a barrier to making statistical inferences about the contamination of a consignment with BRM. This point has been identified and further clarified by Lane et al. (2018a), who point out that stratification with allocation of sample units proportional to stratum size always delivers at least the design specificity implied by Equations 2.1 and 2.2. Thus, ISPM-31 (International Plant Protection Convention, 2008) would appear to reject or preclude the use of simple random sampling in situations where, from a design-based perspective, its use is perfectly valid (though perhaps subject to improvement). Similarly, Venette et al. (2002a, page 150), write

> If individual items of the commodity (*e.g.*, heads of cabbage) were mixed sufficiently as the commodity was harvested and packed in an enormously large shipment and items were selected at random from the shipment, the likelihood of finding [a pest] may be approximated by simple binomial statistics.

The review by Venette et al. (2002a) is authoritative and justifiably influential, but here seems to imply that the binomial (and related results, such as Equation 2.2) apply only if a consignment is "sufficiently mixed". The implication has been carried forward by other authors. For example, Barron (2006) writes,

> However, predictions from the binomial distribution are based on the assumption that the prevalence of infestation is constant throughout the consignment (*i.e.*, there is no aggregation) and that simple random sampling is used so that sampling observations are independent (Venette et al., 2002a).

From a design-based perspective, pests or other BRM may be mixed throughout, or clustered, or even concentrated in one portion of the consignment, and the binomial

distribution remains an appropriate basis for inference under simple random sampling: aggregation is a non-issue. The required independence within the sample is guaranteed by random selection under the design; independence is consequence of *selection*, not proximity. As we will see below, from a model-based perspective, the implications are different, but the insistence on both homogeneity *and* a specified sampling regime will remain overly restrictive.

2.3.1.4 Setting the confidence level, design prevalence, and sample size of an inspection

Although in theory setting the level of confidence and the design-prevalence of an inspection allows determining the sample size of an inspection (Eq. 2.3), in practice the reverse process often prevails: the regulator chooses to inspect 600 units per consignment for historical reasons, or because it seems an appropriate amount of effort relative to other activities, and reports the corresponding 95% confidence level and 0.5% design prevalence.

A substantially more in-depth analysis would be needed to determine the optimal sample size. Below, we highlight several strategies to do so:

- Determining an acceptable leakage for a pathway, informed by a detailed pest risk analyses (*e.g.*, reducing the leakage sufficiently that it is impossible for pests to establish a minimal viable population or considering the economic costs of eradication), can provide a mechanism to calculate inspection sample size (Lane et al., 2018a).

- The sample size can be determined by optimising an objective function (*e.g.*, minimizing total leakage across different pathways, or minimizing the cost). For example, (Chen et al., 2017) allocated sampling effort among pathways to minimize the total leakage of infested units coming in the US. Pathway with high infestation rate ended up with a higher sample size than pathways with low infestation rate. In Camac et al. (2020), the optimal sample size is calculated based on a cost-benefit analysis. The cost-benefit analysis balanced the cost of increasing sampling effort with the cost caused by a pest incursion. Since the likelihood of establishment and the damages caused by an incursion are likely to be pest-dependent, this type of analysis require a deep understanding of the pathway and the biosecurity system involved. The approach might also be sensitive to the assumptions used to derive the costs.

- Inspection is not only a tool to stop pests at point of entry, but also a tool to monitor pathways risk and to help make informed decisions (*e.g.*, shutting down a pathway). From this perspective, we want to choose a sample size that is sufficient to detect sizeable changes in infestation rate and the presence of new quarantine pests with sufficiently low uncertainties.

2.3.1.5 Sampling Methodology

The validity of inference in design-based inference is conditioned on the design, meaning that if the design is followed then the inference is valid. Thus, adherence to the actual design in question is critical. Unfortunately, simple random sampling is not always so simple in practice, and the nature of the material involved in sampling consignments for

BRM can make operational implementation of simple random sampling quite challenging. At the same time, there may be opportunities associated with methods that do not conform to simple random sampling assumptions, that would strongly motivate their adoption regardless. For example, Barron (2006) points out that even though individual fruits may be identified as sampling units, sampling an entire carton or crate affords the opportunity to inspect the packing material for BRM as well. This suggests the use of cluster or two-stage sampling, which we will take up below.

Most formal presentations of simple random sampling follow the simplest possible setup, which we ourselves echo in Section 2.2. Specifically, a frame or list spanning the population of interest is available, and simple random sampling proceeds by selection of units from the list. There are instances in biosecurity sampling where such a list is available, or could be constructed on-the-spot. For example, airline passengers and their baggage are known to present a substantial hazard for BRM importation (*e.g.*, Liebhold et al., 2006; Hulme et al., 2008). The passenger manifest of an inbound aircraft comprises a frame from which a simple random sample (or other probability sample) of passengers can easily be drawn (*e.g.*, Lane et al., 2017). Where the number of incoming sample units is relatively small, for example with shipments of large, high-value items such as unprocessed tropical logs, it may be possible to construct a list upon arrival. However, when the number of incoming units is great, or incoming units arrive in bulk, a list or other convenient frame may be simply unavailable.

Fortunately, the absence of a list-based frame does not preclude the use of simple random sampling. For a few commodities that are subject to inspection, it is possible to perform physical mixing (akin to shuffling a deck of cards) so that the subsequent selection of individual sample units is, for all intents and purposes, uniform and random. For example, many nuts and seeds can be handled in this way (though large shipments may be further subdivided into sacks or other containers, suggesting a cluster sampling approach; see Subsection 2.3.2 below). Generally (especially in New Zealand), seeds are sampled in accordance to the internationally recognized ISTA sampling methodology, which intends to provide a representative sample for the seed consignment through randomness. Certainly physical mixing could not be used on delicate fruits, live plants, or other fragile commodities, however. For other commodities, the geometry of packing facilitates simple random sampling. For example, if young nursery seedlings are packed in flats in which the individuals are laid out with rectangular or hexagonal spacing, while the flats themselves are arranged in a shipping container in a regular fashion, then the location of a flat, combined with a row and column number for a position within a flat, constitutes a form of "address" that can be sampled randomly if the number and configuration of flats is known. (If a random draw in such a situation leads to an empty cell within a flat, or a position for a flat that is not actually occupied, that address would be rejected and a new one drawn; this can be viewed as a simple case of the acceptance-rejection method of von Neumann (1951).)

Finally, we would note that simple random sampling is possible even when the sample units themselves are packed hierarchically in cartons or other clusters that are of unequal size, but simple random sampling of the cartons is possible. For example, Barron (2006) reports that bananas are typically packed in cartons containing, on average, 16 "hands" or bunches; however, the actual number may vary. A typical consignment might consist of 20,000 or more rectangular cartons, packed in a shipping container. It might be possible to sample cartons by simple random sampling; but if one then selected a single hand of bananas from a chosen carton by simple random sampling, hands from cartons that

contain a larger number of hands would have a lower overall probability of selection than those containing fewer hands. This can be remedied by a slight modification of the acceptance-rejection method of von Neumann (1951). Specifically, having selected a carton, we then choose a random integer i from 1 to a number that must be equal to or larger than the number of hands that *could* occur in a carton. If i is less than or equal to the number of hands that actually occur in the selected carton, we choose the i^{th} hand for inspection. If i is greater than the number of hands in the selected carton, then we reject the choice entirely, and draw a new one (including both a new selection of carton, and a new value of i). The process is repeated until the desired total number of samples n is attained.

Notwithstanding the range of possible strategies for drawing a simple random sample, it may still be the case that the hierarchical structure of a consignment, its physical arrangement and packing, and/or the opportunity to combine inspection of packing materials with that of the nominal contents of the consignment will make other strategies, such as cluster sampling, far more attractive. From a design-based perspective, the move to a different design may be wise but it also requires the use of different estimating equations. Alternatively, the sheer mechanics of simple random sampling – or any type of probability sampling – may be so onerous that we must abandon the design-based framework entirely. The model-based alternative will be taken up in Section 2.4.

2.3.2 Cluster and two-stage sampling

Cluster sampling (or a related approach, two-stage sampling) arises naturally as an alternative in a variety of biosecurity contexts. Fruits packed within cartons, seeds within sacks, or even passengers traveling within family groups can all be considered as clusters of sampling units. It may be wiser to take advantage of such structure in the population, than to fight against it merely for the sake of preserving simple random sampling.

Formally, cluster or two-stage sampling recognizes hierarchical structure in the population by designating primary and secondary sampling units. For example, suppose a consignment consists of fruits packed within cartons, and our first step is to select a number of cartons at random, followed by selection of fruits within the chosen cartons. Then the primary sampling units are cartons, and the secondary sample units are fruits. In cluster sampling, once a carton is selected, then all of the fruits in that carton are included in the sample; the fruits so chosen thus form a cluster. In two-stage sampling, the fruits within a carton would be chosen by a further subsampling approach (for example, simple random sampling of fruits within that carton).

In the design-based framework, different estimating equations might be required for different combinations of sampling methods at each stage. Cluster sampling is treated in detail by Cochran (1977, Chapters 9 and 9A) and Thompson (2012, Chapter 12), while two-stage sampling is described by Cochran (1977, Chapters 10–11), Gregoire and Valentine (2008, Chapter 12) and Thompson (2012, Chapter 13).

As with simple random sampling, most design-based texts that address cluster or two-stage sampling concern themselves primarily with estimating the population mean, corresponding in our situation to the prevalence of BRM within a consignment. In that context, the partitioning of variance into within- and between-cluster components is critical, and tends to work to the disadvantage of cluster sampling when efficiency is assessed in terms of sampling variance as a function of the number of individual items inspected. As Thompson (2012, page 162) states,

> Cluster sampling is more often than not carried out for reasons of convenience or practicality rather than to obtain lowest variance for a given number of secondary units observed.

For some products, pairs of secondary sample units located within the same cluster are more likely to share the same contamination status ($y_i = 0$ or $y_i = 1$) than randomly chosen sample units from different clusters. In this case, the sampling variance associated with cluster (or two-stage) sampling is likely to be higher than that for simple random sampling, when the number of secondary sample units inspected is held constant. For cluster or two-stage sampling to be attractive, it must offer benefits in terms of cost, practicality, or the opportunity to inspect other aspects of the consignment (such as packing materials) simultaneously.

From the standpoint of detecting BRM and noncompliance, the situation is no better, and perhaps in some ways worse. As Barron (2006) illustrates through simulation, the sensitivity of a cluster sampling approach declines substantially from the nominal sensitivity given by Equation 2.2, when cluster sampling is employed and BRM is aggregated. The best way to describe the decline in sensitivity with changes in population pattern has not been extensively studied, and the work of Barron (2006) stands as a landmark in the biosecurity literature. From a design-based perspective, little can be said about this loss of sensitivity for a given consignment, especially if no BRM has been detected: the sample contains no information to distinguish between clustering of BRM, or its absence. In the worst case, BRM is completely clustered such that the proportion of primary sample units that are contaminated equals the prevalence p, and all secondary sample units in a contaminated primary unit are themselves contaminated, while a proportion $(1 - p)$ of primary units are entirely clean. In that case, the sensitivity is given by Equation 2.2, but with n replaced by the number of primary units sampled. In other words, to obtain a sensitivity of $S = 0.95$ at a design prevalence of $p^* = 0.005$, one would need to inspect 600 cartons of fruit, not 600 individual fruits. The only advantage is that if one knew this pattern of BRM was present in this type of consignment, it would only be necessary to inspect a single fruit in each selected carton. In practice, this is unknown; and the general insistence in design-based inference on avoiding unnecessary assumptions about the population in question also makes using background information, perhaps collected from similar consignments in the past, challenging. Using such information, so that experience can be accumulated and use to formulate more precise inferences over time, is a strength of model-based and especially of Bayesian inference.

2.4 Model-Based Inference

The seminal papers for model-based inference in survey sampling are Brewer (1963) and Royall (1970). Models had been used in statistics to test hypotheses in experimental and observational settings before that time, and there were antecedents within the survey sampling literature (*e.g.*, Mátern, 1960). But Brewer (1963) moved the model to the forefront in the context of sampling a finite population, and Royall (1970) directly challenged the assumption that random sampling was necessary for inference. This provocative paper kindled an especially violent episode in what Kish (1995) calls "The Hundred Years' Wars of Survey Sampling." (Of course this battle in the "war," and even the identities and aims of the warring factions, went nearly unnoticed outside the statistics literature, even in communities that rely on statistical inference.) The following decade saw an explosion of

methods for and applications of the model-based paradigm. Recent sampling texts with a model-based perspective include Valliant et al. (2000) and Chambers and Clark (2012).

In model-based inference, a probabilistic model provides the foundation for inference from the sample data. The attributes of the population are not treated as fixed, but rather as random outcomes generated by the assumed model. As Royall (1970) put it,

> If a fair coin is flipped, the probability that it will fall heads is one-half; if the coin was flipped yesterday, but the outcome has not yet been observed, the probability that it fell heads is still one-half. The state of uncertainty which applies to the outcome is unaltered by the single fact that the event which determines the outcome has already occurred.

Model-based inference is often considered to involve a hypothetical "superpopulation," since the actual population can be considered as just one random realization of the possible populations generated by the underlying model. Inference is based, not on the randomization distribution created through the inclusion or exclusion of units under a sampling design, but on the distributions assumed in the model itself. Although it was initially fiercely resisted by design-based traditionalists, the model-based perspective quickly found sympathy among a broader statistical audience. As Little (2004, page 546) put it,

> Survey sampling is perhaps unique in being the only area of current statistical activity where inferences are based primarily on the randomization distribution rather than on statistical models for the survey outcomes.

Gregoire (1998, page 1436) highlights the role of familiarity in the adoption of model-based inference, writing,

> . . . the presumption of a model . . . requires more assumptions than the design-based approach. But in this regard, it accords with nearly everything else one does in statistical estimation and prediction: a model is assumed based on prior experience and subject matter knowledge, the model is fitted to sample data according to some criterion . . . , the goodness of fit is checked, alterations are made if deemed warranted, and eventually the results of the fitted model are proclaimed.

Perhaps because the development of quantitative methods in biosecurity accelerated at the end of the 20th century and beginning of the 21st, and arguably drew on a broad range of expertise in statistics and epidemiology rather than an older foundation in survey sampling, model-based approaches have found a more comfortable home within the biosecurity community than in other agricultural and natural resource fields (Gregoire, 1998; Magnussen, 2015). However, the distinctions have not always been clear, and model-based inference brings different advantages as well as demands to the playing field.

2.4.1 Homogeneous populations and simple random sampling

In design-based inference, the randomization distribution played a fundamental role. In model-based inference, the relevant distribution arises from the random outcomes of individual observations assuming that the model is true (Valliant et al., 2000; Chambers and Clark, 2012). In modern model-based inference, the likelihood function plays an especially critical role in describing how well the data that have been observed fit the model for a given set of parameters. (The use of maximum likelihood in statistics dates to Fisher (1922); see Aldrich (1977) for an historical perspective on its development.) Particular

attention must be paid to developing a model that adequately captures the salient features of the system under investigation, not only in terms of capturing the expected values of the observations, but also how the observations can vary and how those variations may be correlated. Models often include one or more auxiliary variables; ideally these are known either for the entire population, or for a larger sample than that on which y_i will be measured. For example, if we are inspecting cartons of fruit within a container that includes cartons from multiple producers, then producer identity would be an auxiliary variable. Position within the container might be another useful auxiliary variable, particularly if we suspect that clean cartons have been preferentially packed toward the front of the container. Expert knowledge, prior experience, and data from past inspections can all play an important role in developing a model that includes the most important auxiliary variables for a particular commodity or pathway.

In the simplest case, either there are no auxiliary variables, or the auxiliary variables that are known are not related to the actual contamination of a unit y_i, or (just as critically from a model-based perspective) the *probability* that the i^{th} unit will be contaminated, p_i. In that case, the population can be considered as homogeneous. From a model-based perspective, if p_i is truly identical for every unit in the population, then it does not matter which units we select. For example, if we have 1000 identical trick coins that may not have $p_i = 0.5$ of showing heads, but we know p_i is the same for all of the coins, it will not matter whether we randomly sample 100 coins to toss, or toss one coin 100 times. This can create the impression that sample design is completely irrelevant to model-based inference, a misperception to which we will return below.

2.4.1.1 Estimating prevalence and detecting contamination

Suppose that we have a homogeneous population, having p_i identical for every member. Then if we constrain $Y = \sum_{i=1}^{N} y_i$ to equal $p_i N$ exactly, sampling any n units out of N will yield a distribution of $x = \sum_{i \in n} y_i$ values that follows the hypergeometric distribution $h(x; n, Y, N)$ (see Equation 2.1). In other words, if we wish to draw inferences about the actual prevalence within a specific consignment, we should employ the hypergeometric distribution. On the other hand, if we view p_i as the probability associated with a contamination process that generated the consignment under inspection, then the actual value Y is just the outcome of N random trials, and by chance it may be that $Y \neq p_i N$. Rather, we only believe $\mathrm{E}[Y] = p_i N$. In this case, we are taking a *superpopulation* perspective, and the distribution of X will follow the binomial distribution. In either case, when n is fixed, the maximum likelihood estimate of p_i is

$$\hat{p}_i = x/n \tag{2.4}$$

and the intuitive predictor of the total number of contaminated units in the consignment is the Lincoln–Petersen estimator, which follows

$$\hat{Y} = \frac{N}{n} x \tag{2.5}$$

The likelihood function for the hypergeometric (or binomial) distribution also allows us to construct standard errors and confidence limits for $\hat{p}_i$ (though the best procedures to tackle this simple problem are not as settled as one might suppose; see Brown et al. (*e.g.*, 2001) and Cai (2005). For a discussion within a biosecurity context, see Robinson et al. (2011)). Note that as N becomes large relative to n, the hypergeometric and binomial

distributions converge, and the difference between estimating a "population" p_i and a "superpopulation" one becomes numerically unimportant. This model is a straightforward application of the classic "urn model" where the goal is to estimate the proportion of balls of a given color that have been placed in an urn (*e.g.*, Chambers and Clark, 2012, Section 3.1). Equation 2.5 is referred to as the best linear unbiased predictor (BLUP) of Y, in that it is unbiased ($\mathrm{E}[\hat{Y}] = Y$) and has the lowest possible variance among all predictors that are linear in the observed values of y_i (Chambers and Clark, 2012, Section 3.3), assuming the model is true[1]. If, on the other hand, the y_i values are not independent (even though p_i is the same for every sample unit), for example, due to the contagious spread of a pest population through the consignment during the shipping process, then Equation 2.5 may be inefficient or even biased, and estimates of variance calculated under the assumption of an independent generating process will also be biased. In the presence of positive autocorrelation between y_i values, variances will typically be underestimated, and this will lead in turn to overestimating the confidence with which a consignment can be declared free of contamination.

For detecting contamination by BRM (or substantiating freedom from contamination) within a completely homogeneous population, the results for the binomial distribution in Equations 2.2 and 2.3 hold, provided N is much greater than n. Note that from a model-based perspective, again, it does not matter what the procedure may have been for selecting the samples, so long as the distribution of p_i and y_i is the same in the sample as in the full consignment, and as long as the realizations of the generating process are independent for all of the selected sample units. As a practical consequence, if the homogeneous model is credible, then the standard "600 sample" can be expected to deliver on the nominal $S = 0.95$ sensitivity at a design prevalence of $p^* = 0.005$, no matter how the sample units are selected. Simple random sampling, cluster sampling, or purposive sampling are all allowable – at least in principle.

Many consignments will consist of lines that vary in their p_i, either because the lines are different commodities, come from different growers or producers, or have experienced different environment or treatment before arriving at a port of entry. An important question is whether Equations 2.2 and 2.3 can be used for such consignments, which clearly violate the assumption of homogeneity. Once again, the results presented by Lane et al. (2018a) are instructive. Lane et al. (2018a) address the question within the implicitly design-based framework of ISPM 31 (International Plant Protection Convention, 2008), but the results can be generalized to a model-based framework. Suppose, following Lane et al. (2018a), that we are sampling from a set of k lines within a consignment, each with its own probability of contamination p_k and number of units N_k; basic considerations require

$$N = \sum_k N_k$$

and

$$p = \sum_k \frac{N_k}{N} p_k \tag{2.6}$$

where p, as before, is the overall prevalence of BRM within the consignment. Now, letting the population prevalence equal the design prevalence p^*, the sensitivity is (Lane et al.,

[1]Note that this is a different definition of bias than that employed in the design-based framework; there, the expectation was over possible samples, while here, it is over possible outcomes of the random variables (*i.e.*, observations). To distinguish the two properties, we call the latter model-unbiasedness

2018a)

$$S = 1 - \prod_k (1 - p_k)^{n_k} \tag{2.7}$$

Lane et al. (2018a) show that if $n_k \propto N_k$, a uniform probability with $p_k = p$ for all k minimizes the sensitivity S. In other words, heterogeneity in p_k causes the actual sensitivity to exceed the nominal sensitivity implied by Equation 2.2. The argument holds in the limit as n_k goes to one, *i.e.*, provided each unit in the sample is representative of an equal number of units in the population. In other words, so long as the empirical distribution of p_k in the sample is close to the actual distribution in the full population, S as given by Equation 2.2 is conservative, and assuming the binomial distribution to compute the sample size using Equation 2.3 will deliver at least the nominal (specified) sensitivity.

From a fully model-based perspective, consider the actual contamination of an individual sample unit y_i as a Bernoulli random variable with probability p_i. Now suppose that we draw a sample from the population by a noninformative design, and further that y_i and y_j are independent for $i \neq j$. It follows from the basic properties of expectation that $\mathrm{E}[y_i] = \mathrm{E}[p_i] = p$. Thus, in the absence of useful covariates, considering the outcome y_i as composed of a two-stage process (first selection of p_i, then determination of y_i conditional on p_i) is essentially a computationally-expensive but equivalent model to ignoring the within-population variability in p_i and basing the assessment directly on the binomial model.

The critical assumption made in the homogeneous population model is that the individual observations are independently and identically distributed (i.i.d.). Unfortunately, that is an assumption that is difficult if not impossible to verify, because only the y_i, and not the underlying p_i, are directly observable. Two questions are relevant. One is whether, epistemically, the i.i.d. assumption is credible. The second is whether it is possible for the data to falsify the assumption. Epistemically, we may have reason to question the assumption even in the absence of compelling, data-driven evidence. For example, if BRM within a consignment comprises a spatially-contagious pest that can migrate and spread within the consignment during shipping, then the i.i.d. assumption is suspect even if we lack consignment-specific data to disprove it. Likewise, if a consignment comprises multiple lines, each associated with some combination of factors (geographic origin, producer, subsequent handling and/or treatment) that might influence prevalence of BRM within the line, then we might suspect variation in p_i across lines and reject the simple binomial model out of hand (though as noted above, if sensitivity is the overriding issue, such a move may be misguided). It might also be that we have data — whether by design or by historical accident — including one or more candidate auxiliary variables as well as observed cleanliness or contamination. In that case, we can test (at least for past consignments) whether the auxiliary variables are correlated with the y_i (and hence p_i), or whether any clustering or spatial pattern is present. If so, then these may be used to enhance the sampling design and model used for future consignments. Confronting a hypothesis (the model) with data, falsifying elements of the model, using the results to improve the model, and repeating is one of the basic cycles within the scientific method.

2.4.1.2 The role of the sampling design

In the classic model-based approach, inference is conditional on the selected sample; the process by which the sample is selected plays a subordinate role. This has led to a

misperception that the sample design is completely unimportant to model-based inference. However, a simple example illustrates why this cannot be true. Suppose a consignment arrives in a shipping container, and out of convenience, the inspector ignores the sampling methodology and selects 600 units from the front of the container. But the shipper, recognizing this as a likely strategy, has intentionally placed only clean units near the front of the container, hiding units that may harbor contamination near the rear. Clearly, ignoring the sampling design and proceeding with the binomial model will usually lead to an underestimate of prevalence, and an overly confident assertion of freedom from contamination in this situation. Confusion over this issue may be one reason for a lack of trust in the model-based approach (Magnussen, 2015).

The necessary assumption in model-based inference is that the design is noninformative or ignorable (Chambers and Clark, 2012, Section 1.4). Specifically, for a design to be noninformative, the same model and parameters must be valid for both the sampled and non-sampled units in the population. Thus, inferences based on the sampled units (*e.g.*, parameter estimates) remain valid for the non-sampled units (Chambers and Clark, 2012, p. 10). Magnussen (2015) argues that for the design to be ignorable, the joint distribution of the y_i and the binary inclusion indicator δ_i (which may be no longer a random variable) should be independent, conditional on any covariate x_i. In the absence of covariates, a probability-based simple random sample is noninformative, but other designs might also be allowable (Chambers and Clark, 2012, p. 11).

As Chambers and Clark (2012, p. 11) write,

> The importance of non-informative sampling to the model-based approach to finite population inference cannot be overstressed.

However, sampling designs are often assumed to be noninformative when that assumption cannot withstand serious scrutiny. Haphazard or convenience samples often fail to be noninformative, as suggested by the container example. In vegetation ecology, the pernicious term "sampling without preconceived bias" (Ellenberg and Mueller-Dombois, 1974) is often used to disguise such methods; but ignorance can only explain, not excuse, the use of improper sampling methods even within a model-based framework. Of course, a key challenge — especially in the absence of covariates — is detecting and quantifying the influence of a non-ignorable sampling method on the results; without some information about the non-sampled units, it can be difficult to substantiate that the distribution of the y_i should be the same between the sampled and non-sampled fractions of the population. This argument extends even to the well-intentioned use of expert judgment in selecting a "representative" sample. Chambers and Clark (2012, p. 12) recommend the use of a probability sample "or some other non-subjective method," and go so far as to write,

> Designs where an expert chooses a set of units believed to be representative should be avoided.

They argue that since expert choice is likely to depend on covariates that are not explicitly included in the model, the sample distribution of y_i is likely to differ between the sampled and non-sampled fractions of the population. They suggest that where expert opinion is available, it should be used to inform the choice of covariates in the model, rather than the selection of sample units themselves. A model based on the kinds of prior experience and subject matter knowledge that biosecurity inspectors and their supporting biological team bring can be invaluable to the effectiveness of the inspection process. However, where that knowledge is imperfect, or contested, a probability-based sampling design may allow for

inferences that are robust to departures from an assumed model. A simple random sample — or a sufficiently close approximation that departures can probably be ignored — may represent a safe choice.

A somewhat different perspective is offered by Madden and Hughes (1995, p. 532) in the context of plant disease surveillance, who write,

> One cannot determine if the binomial distribution is appropriate if the data are collected as an unrestricted random sample of individual plants, because there is insufficient information on the observed distribution.

Madden and Hughes (1995) suggest cluster sampling, so that departures from the independence assumption of the binomial model can be detected and addressed. Against this point, one may argue that clustering of samples is likely to exacerbate the problem of non-independence. Moreover, tests for independence of contamination among sample units that are not hierarchically clustered have advanced in recent decades, so that cluster sampling is not strictly required (though it would be necessary to record information on the proximity of sample units to one another, to use such tests). On the other hand, as discussed in Subsection 2.3.2, cluster sampling presents a number of practical advantages, so we turn to clustered populations and samples next.

2.4.2 Clustered populations or samples

In a model-based context, the treatment of clustering typically emphasizes clustering within the population, rather than clustering as a feature of the sampling design (see, *e.g.*, Valliant et al. (2000, Chapter 8) and Chambers and Clark (2012, Chapter 6)). This is natural, since in the model-based approach it is the attributes of the population (*i.e.*, the y_i) that are considered random, rather than the inclusion of the units (*i.e.*, the δ_i). Therefore, it is natural to consider the clustering or correlation of attributes as part of the model. With that said, populations that have a hierarchical physical (and possibly biological) structure, such as consignments of fruits that are packed within crates inside shipping containers, may include population structure with scales and patterns that match those of the clusters that are advantageous for sampling. Thus, clustering of populations and the clustering of samples should be considered together.

Standard texts on model-based inference include the generalization of the BLUP estimator to populations with hierarchical clustering (see, *e.g.*, Chambers and Clark, 2012, Chapter 6). However, as with the BLUP for homogeneous populations, the treatment is most appropriate for estimating the mean of a normally-distributed variable. When the y_i are binary, the most popular model for a clustered population is the beta-binomial (Madden and Hughes, 1995; Hughes et al., 1996; Venette et al., 2002a). The beta-binomial model has long been used in a wide range of applications, and dates to an early paper by Skellam (1948). Early applications when the number of secondary sample units is constant across primary sample units was presented by Kemp and Kemp (1956) for vegetation quadrat data, by Chatfield and Goodhardt (1970) for consumer preference data, and by Griffiths (1973) for disease incidence within households. A modern maximum-likelihood approach, allowing unequal numbers of secondary samples within each primary sample, was first presented by Williams (1975). The correlated binomial model of Kupper and Haseman (1978), and the multiplicative binomial of Altham (1978), might be viable alternatives but do not seem to have been explored in a biosecurity context.

Under the beta-binomial model, the prevalences p_k associated with each cluster are assumed to follow a beta distribution, *i.e.*,

$$f(p_k) = \frac{p_k^{\alpha-1}(1-p_k)^{\beta-1}}{B(\alpha,\beta)} \tag{2.8}$$

where $B(\alpha,\beta)$ is the beta function. The beta distribution has a mean, corresponding to the population-level prevalence, of $p = \alpha/(\alpha+\beta)$. If we draw n_k sampling units by a noninformative procedure from the k^{th} cluster, then the unconditional probability of finding x_k contaminated units (*i.e.*, the probability without knowing the cluster-level prevalence p_k) is

$$bb(x; n_k, \alpha, \beta) = \binom{n_k}{x_k} \frac{B(\alpha + x_k, \beta + n_k - x_k)}{B(\alpha,\beta)} \tag{2.9}$$

The probability of a completely clean sample from an individual cluster is therefore

$$bb(0; n_k, \alpha, \beta) = \frac{B(\alpha, \beta + n_k)}{B(\alpha,\beta)} \tag{2.10}$$

It is straightforward to prove that $bb(0; n_k, \alpha, \beta) \geq (1-p)^{n_k}$, with equality only when $n_k = 1$ (*i.e.*, sample units are drawn from independent clusters) or in the limit as $(\alpha + \beta) \to \infty$ (*i.e.*, the variation in p_k between clusters vanishes, so that the cluster identity becomes noninformative). Recalling that the sensitivity under a binomial model is given by Equation 2.2,

$$S = 1 - (1-p)^n = 1 - [(1-p)^{n_k}]^{n/n_k}$$

it will be true, in general, that the sensitivity when allocating n_k sample units to each of n/n_k clusters, *i.e.*,

$$S = 1 - bb(0; n_k, \alpha, \beta)^{n/n_k} \tag{2.11}$$

is strictly lower than that under the binomial model. Of course, the effect is difficult to estimate without prior knowledge of α and β. In this context, it may be useful to consider reparameterization of the beta-binomial in terms of the overall prevalence p and the intracluster correlation coefficient ρ, which can be related to the original parameters by

$$\rho = \frac{1}{\alpha + \beta + 1}$$

Intuitively, ρ (which ranges from 0 to 1) measures the degree to which units from the same cluster are likely to share the same contamination status. Specifically, the probability that two units from the same cluster share the same contamination status is (Mak, 1988)

$$p_s = 1 - 2p(1-p)(1-\rho)$$

If prior data on consignments within a pathway were available, it would be possible to estimate ρ for the pathway; methods for estimating ρ and its uncertainty are discussed by Lui et al. (1996), Zou and Donner (2004), and Saha and Wang (2018). In practice, when prevalence is very low, it will be difficult or impossible to obtain reliable estimates of ρ. For example, when $p = 0.005$, we would expect that the fraction of clusters with BRM present would be at most $0.005n_k$, but perhaps as small as 0.005. The necessity of having multiple clusters with BRM present, in order to test whether contamination is

correlated within clusters, would imply having data on a large number of clusters indeed. In the absence of such data, two extreme alternatives suggest themselves, but neither is entirely attractive. One is to assume $\rho = 0$ in the absence of compelling evidence to the contrary, and proceed with the binomial model assuming independence. But then, the actual sensitivity may be less than the nominal sensitivity, and there may be substantial risk of BRM leakage. The other alternative is to assume the "worst case" $\rho = 1$. That would ensure that sensitivity would be at least at its nominal level, but the sample size requirement (for example, 600 clusters rather than 600 sample units for 95% sensitivity at $p^* = 0.005$) may entail considerable expense. Viewed from such a perspective, the price of complete ignorance about the characteristics of consignments on a pathway may be quite high, and the value of a well-designed study to discern those characteristics may likewise be high, despite the time and expense required (for an example on how to estimate ρ from a pathway and how much data is needed, the readers are referred to section 3.2.2.2. For a realistic pathway with p=0.02 and ρ=0.1, at least 30–40 consignments, each with 30 clusters and 20 units per cluster —*i.e.*, 600 units per consignment— are required to obtain a reliable estimate of ρ).

2.5 Bayesian inference

Bayesian inference is, in essence, a form of model-based inference with its origins in the work of Bayes (1763) and Laplace (1774). This broad class of approaches had previously been known simply as "inverse probability"; the term Bayesian only came into common use after World War II (Fienberg, 2006). The Bayesian approach stands in contrast to the conventional frequentist perspectives discussed so far, including both the design- and model-based perspectives, in its emphasis on treating all unknown quantities as random. As Lindley (1978), in discussing the works of de Finetti (1974) and de Finetti (1975), writes,

> In conventional, sampling-theory statistics the basic uncertainty that is admitted and from which all other uncertainties are derived is that concerning the data — the probability distribution of the data given the parameter. This is extraordinary because the data are just those things about which one is certain: they are there to be seen and analysed. The truly uncertain quantities are the parameters: You are uncertain about them and it is those that must be described probabilistically to make coherent sense.

The mid-20th century development of the Bayesian approach as a distinct branch of statistical thought was stimulated by the work of Alan Turing and his assistant I.J. Good at Bletchley Park, by wartime work at the Statistical Research Group at Columbia University, and by a growing awareness of the work on subjective probabilities by de Finetti (*e.g.*, de Finetti, 1937) and on the objective theory of inverse probability as propounded by Jeffreys (1939). The publication of *The Foundations of Statistical Inference* by Savage (1954) — a member of the Statistical Research Group team — can be seen as transformative (Fienberg, 2006).

Central to Bayesian inference is the use of Bayes' Theorem to update prior knowledge or belief about unknown quantities, after observing the data. In the modern formulation, first put forward by Laplace (1812), the theorem states

$$\text{Prob}(\theta|X) = \frac{\text{Prob}(X|\theta)\text{Prob}(\theta)}{\text{Prob}(X)} \tag{2.12}$$

where θ represents the unknown (and therefore random, even if fixed) quantity or quantities, and X represents the observed data. Prob($\theta|X$) is the *posterior* distribution for θ, and forms the basis for inference. By contrast, Prob($X|\theta$) is the likelihood function — the probability of observing particular values of the data, given a hypothetical value of θ. Probabilities such as those in Equations 2.1 and 2.9 represent likelihoods. The probability Prob(θ) is the *prior* distribution for θ — *i.e.*, what is known or believed before the data have been observed. Finally, Prob(X), the marginal probability of X, is usually treated simply as a normalization function to ensure that the values of Prob(θ) sum (or integrate) to 1 over all possible values of θ. As Jeffreys (1939, Section 1.2) states succinctly, "The posterior probabilities of the hypotheses are proportional to the products of the prior probabilities and the likelihoods."

Bayes' Theorem, in itself, is not controversial: it is a straightforward statement about conditional probabilities. What was controversial in the emergence of Bayesian inference was the insistence that the likelihood function alone did not form a sufficient basis for coherent inference; the requirement for a prior distribution (even if that prior was putatively "noninformative"), and the assertion (especially by "subjectivist Bayesians", following de Finetti (1937)) that probabilities represented not only physical or aleatory probabilities (*e.g.*, the long-run probability of pulling white marbles from an urn) but also beliefs or epistemic probabilities (*e.g.*, ones' belief that a marble, already drawn from the urn, perhaps the only marble in the urn, and now hidden under a cup, is white). The requirement for a prior, and the incorporation of priors based on subjective beliefs into inference, seemed to many to contradict the goal of an objective, data-driven approach to statistical inference. The Bayesian approach departed dramatically from those developed and advocated by Fisher, Neyman, Pearson and others in the pre-war period, and the debate over Bayesian methods was often contentious. Not surprisingly — and compounded by a lack of computing power and software — Bayesian methods were slow to enter the relatively conservative subfield of survey sampling; the first fully Bayesian paper on the subject was probably that of Ericson (1969). Even 35 years later, Little (2004), espousing a Bayesian approach, wrote:

> Advocating Bayes for sample survey inference is "swimming upstream," because its subjectivist basis is anathema to many survey statisticians, who do not like modeling assumptions. But Bayesian methods run the gamut of subjectivity and can be as "objective" as any frequentist method when necessary; indeed, many frequentist answers can be replicated from a Bayesian perspective.

Nonetheless, applications of Bayesian approaches have begun to appear in the biosecurity arena, especially in complex problems requiring the integration of multiple types of data with expert opinion. A summary of examples and approaches can be found in Low-Choy (2015a) and Low-Choy (2015b).

2.5.1 Homogenous Populations and Simple Random Samples

The basic approach for calculating the posterior probability of a binomial proportion is one of the oldest results in Bayesian statistics, having been studied by Bayes (1763). Moreover, the calculations required are relatively simple. Since the prevalence p is unknown, it requires a prior distribution. Bayes (1763), and many others following, have assumed a uniform distribution for the prior. (There are other choices with different but compelling rationales; we shall return to the issue of selecting a prior below.) Note that the uniform distribution is a special case of the beta distribution, with $\alpha = \beta = 1$. When

a Beta(α, β) prior is updated with data reflecting x "successes" (in our case, detections of BRM) in n trials, the posterior distribution is Beta$(\alpha + x, \beta + n - x)$. The mean of the posterior distribution — which represents the most obvious point estimate of p, though not necessarily the only one — is $(\alpha + x)/(\alpha + \beta + n)$. And the probability that the unknown prevalence p is actually less than the design prevalence p^* is (*e.g.*, McBride and Johnstone, 2011)

$$\text{Prob}(p \leq p^* | x, n) = I(p^*; \alpha + x, \alpha + \beta + n) \tag{2.13}$$

where $I(z; a, b)$ denotes the regularized incomplete beta function

$$I(z; a, b) = \frac{1}{B(a, b)} \int_{t=0}^{z} t^{a-1}(1-t)^{b-1} \mathrm{d}t$$

Fortunately the incomplete beta function is available in nearly all comprehensive scientific and statistical software, so its computation is relatively straightforward.

The primary challenge here is the selection of a prior distribution, even if one confines the choice to those that are considered noninformative. A noninformative prior represents the idea that very little is known a priori and should lead to an inference that is ideally unaffected by information external to the current data. Berger (1985) lists four candidate distributions: the Bayes-Laplace prior, Beta$(1, 1)$; the Perks-Jeffreys prior, Beta$(0.5, 0.5)$; the Haldane prior, which is the limiting case of the beta distribution as both α and β approach zero; and the Zellner prior, which is not a beta distribution. The Haldane prior gives results that are "less than adequate" when $x = 0$ (Tuyl et al., 2008), which is the condition under which we wish to measure assurance, so we do not consider it further. When $x = 0$ (as when substantiating freedom from contamination by BRM) and n is large, the posterior arising from the Zellner prior is well-approximated by a $Beta(1, n + 2)$ distribution (Tuyl et al., 2008). Since this is the same posterior as would be produced by the Bayes-Laplace prior under the same circumstances, we focus on the remaining two priors that can be represented as beta distributions, namely the Bayes-Laplace and the Perks-Jeffreys. Between these two, there are evident differences in the posterior distributions and the resulting inferences for biosecurity. For example, Figure 2.1 shows the posterior probability that $p < 0.005$, for a range of sample sizes, when no contamination has been found. Even though these sample sizes would usually be considered "large," and we might expect the data to "speak for themselves" (McBride and Johnstone, 2011), differences in the posterior probability are evident even as the sample size approaches the typical 600-unit sample. This is because for our purposes the most important sampling event is when $x = 0$, and the prior distributions have greatest difference at that point. Figure 2.2 illustrates the same challenge, from the perspective of the credible interval (the Bayesian analog to the frequentist confidence interval). The credible intervals depicted are the upper one-sided credible intervals, and thus show the value that p is believed to lie with a probability of 0.95. Again, the differences are evident even at operational sample sizes. Notably, while the Bayes-Laplace prior requires a sample of $n = 597$ to deliver 95% certainty that $p \leq p^*$, in agreement with the usual frequentist sample size calculation, the Perks-Jeffreys prior requires only $n = 383$. Unfortunately, as Berger (1985) pointed out, there is no compelling reason to choose one noninformative prior over the other. Or, perhaps it is more correct to say that there are compelling reasons, but those reasons disagree, with some authors strongly favoring the uniform Bayes-Laplace (*e.g.*, Geisser, 1984; Tuyl et al., 2008; Tuyl et al., 2009) while others favor the Perks-Jeffreys (*e.g.*, Box and Tiao, 1973; Bernardo and Smith, 1994; McBride and Johnstone, 2011; Berger et al., 2015). We note that the paper

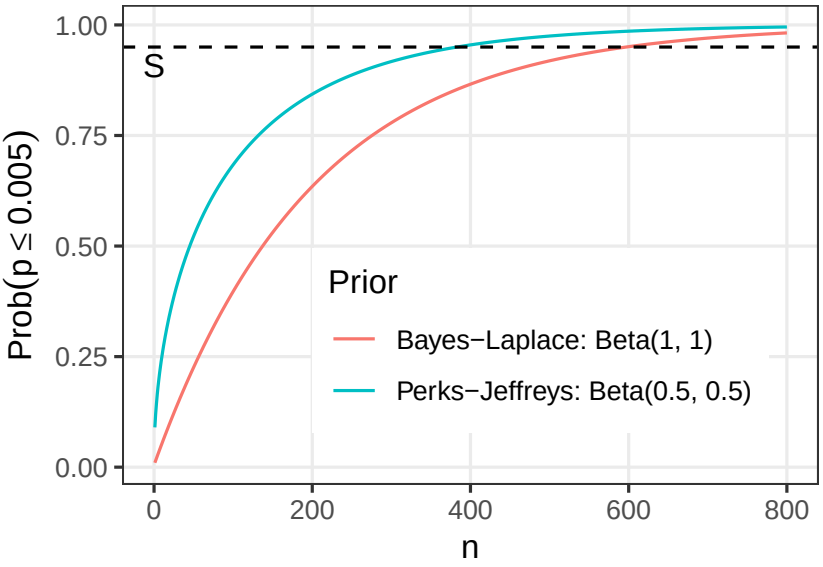

Figure 2.1: Posterior probability that the actual prevalence p is less than or equal to the design prevalence $p^* = 0.005$ as a function of sample size when no contamination has been detected, for two common noninformative prior distributions. Bayes-Laplace prior corresponds to a Beta$(1, 1)$ prior, while Perks-Jeffreys corresponds to a Beta$(0.5, 0.5)$ prior.

by Tuyl et al. (2008) is particularly relevant to the biosecurity situation, because it focuses on the suitability of different priors when $x = 0$ in a binomial trial. Conversely, that of McBride and Johnstone (2011) is also posed in an invasive species context, and while less theoretical than both Tuyl et al. (2008) and Tuyl et al. (2009), suggests the opposite choice. Given the disagreement among Bayesian statisticians about which noninformative prior is appropriate to the situation, could we blame a cost-conscious inspector for choosing the Perks-Jeffreys prior and thereby reducing the number of inspections needed by over 35% at one stroke? In fact, Tuyl et al. (2008) primarily recommend the Bayes-Laplace prior as a reference for sensitivity analysis, suggesting that an *informative* prior taken from the Beta$(1, \beta)$ family, with $\beta > 1$, should be used in situations where $x \approx 0$. The use of an informative prior would require specification in terms of previous data (*i.e.*, an Empirical Bayes approach; Martz and Lian, 1974) or by elicitation (*e.g.*, Low-Choy, 2012; Martin et al., 2012). To our knowledge the practical consequences of such a choice, either in terms of design or of inference after inspecting one or many consignments, has not been fully explored in the literature.

In section 3.1.3, we provide a case study illustrating the effect of using noninformative and informative priors in the biosecurity context with simple random sampling. In section 3.1.3 and in practice, we will use 1) a uniform prior when we have no data on the pathway as it is compatible with and gives similar sample sizes to current practice (*e.g.*, a 600 units sample allows to be 95% sure that the estimated infestation rate in accepted consignments is below 0.5%); 2) an informative prior estimated from recent past data; or better 3) a weighted mixture of uniform and informative priors that make use of external data when available, but also recognizes that past data are not always representative of future, allowing for the possibility of encountering consignments with a higher infestation rates than what we have seen in the past.

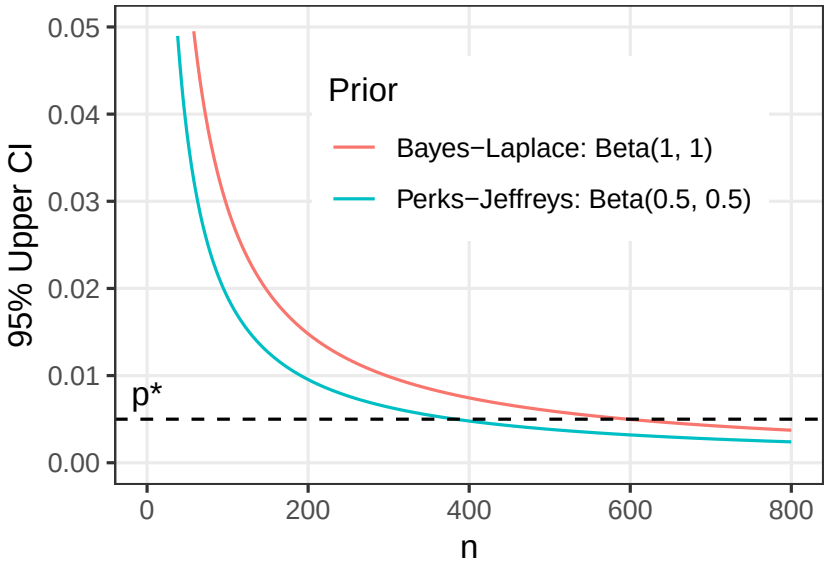

Figure 2.2: Upper 95% credible intervals for the prevalence p as a function of sample size when no contamination has been detected, for two common noninformative prior distributions.

2.5.2 Clustered Populations or Samples

Many of the concerns over clustering in the population, and issues surrounding cluster or two-stage sampling, are the same for Bayesian inference as for frequentist model-based inference. So is the most obvious choice of a model: the beta-binomial. The difference is that from a Bayesian perspective, there are now two parameters (α and β, or equivalently p and ρ) that require prior distributions.

Defining a noninformative prior for the beta-binomial is not as straightforward as for the binomial. Following Yang and Berger (1998), the Fisher information matrix for the beta-binomial is

$$I(\alpha,\beta) = \begin{bmatrix} \psi^{(1)}(\alpha) - \psi^{(1)}(\alpha+\beta) & -\psi^{(1)}(\alpha+\beta) \\ -\psi^{(1)}(\alpha+\beta) & \psi^{(1)}(\beta) - \psi(1)(\alpha+\beta) \end{bmatrix}$$

where $\psi^{(1)}(z)$ is the polygamma function of order 1, *i.e.*,

$$\psi^{(1)}(z) = \sum_{k=0}^{\infty} (z+k)^{-2}$$

The Jeffreys prior for the beta-binomial is then the square root of the determinant of $I(\alpha,\beta)$. By contrast, Gelman et al. (2014, p. 110–111) recommend a noninformative prior

$$\text{Prob}(\alpha,\beta) \propto (\alpha+\beta)^{-5/2}$$

Unlike the ordinary binomial, updating the beta-binomial cannot proceed by elementary calculation (since the prior is not conjugate). Bayesian estimation and inference for the beta-binomial is described by Lee and Sabavala (1987), Lee and Lio (1999), and Everson and Bradlow (2002). To our knowledge, the influence of the prior on the posterior for the beta-binomial model, when all or nearly all of the samples indicate freedom from BRM, has not been studied.

In section 3.2.3, we provide a case study illustrating the effect of using noninformative and informative priors in the biosecurity context with clustered sampling.

2.6 Further Alternatives

The frequentist design- and model-based paradigms, along with the Bayesian, represent the primary modes of statistical inference in survey sampling. To our knowledge, they are the only ones that have been adopted in biosecurity applications, and as the review above suggests, there is still territory to be explored even within those "well-understood" lands. But they are not the only approaches that could be considered. In this section, we take up some alternatives — ideas from beyond the borderlands, as it were — that may yet provide some benefit in applications.

The class of methods considered here is often described as dealing with *non-additive measures*, in that they violate the principle (promoted to an axiom by Savage, 1954) that measures of uncertainty, taken over a mutually-exclusive and exhaustive set of events, should sum to 1. That axiom is satisfied by ordinary probabilities. However, a number of researchers have questioned the necessity, or even wisdom, of that requirement in all situations. For example, Good (1976, p. 129), in summarizing subjectivist Bayesian principles he had developed over the previous decades, writes

> In practice one's judgments are not sharp, so that to use the most familiar axioms it is necessary to work with judgments of inequalities. For example, these might be judgments of inequalities between probabilities, between utilities, expected utilities, weights of evidence ... or any other convenient function of probabilities and utilities.

Work identifying the shortcomings of additive probability models, especially in decision problems involving incomplete information or even ignorance, and attempts to find alternatives, date at least to the work of Keynes (1921). It would be impossible to summarize all of that work in a rigorous fashion here. Rather, we focus on two main alternatives: the Dempster-Shafer theory of evidence (Shafer, 1976), which first rose to prominence in the artificial intelligence community, and the theory of inference from imprecise probabilities and previsions, as first outlined by Walley (1991). The latter theory can be approached either from Bayesian or frequentist perspectives.

2.6.1 The Dempster-Shafer theory of evidence

The Dempster-Shafer theory of evidence had its origins in the statistical work of Dempster (1966), Dempster (1967a), Dempster (1967b), Dempster (1968a), and Dempster (1968b), whose intent was to provide additional flexibility in the specification of uncertainty in probabilistic models and hypothesis testing. Shafer (1976) further developed and clarified the theory, and provided both a philosophical foundation and mathematical extensions. Most applications of the Dempster-Shafer theory have focused on uncertain reasoning in artificial intelligence, expert systems, and pattern recognition problems (Shafer and Pearl, 1990; Denoeux, 2016). It has been less widely used in the area of general statistical inference (Denoeux, 2014). However, some attempts have been made to employ it as a general approach to simplified decision problems; in the natural resources arena, it has been suggested to be useful in situations where data are sparse, absent, or inconsistent (*e.g.*, Caselton and Luo, 1992; Ducey, 2001). The Dempster-Shafer theory is widely known and used in certain application areas, having generated thousands of further papers in the past four decades (Denoeux, 2016). However, it has not, to our knowledge, been exploited within the biosecurity arena.

At the heart of the Dempster-Shafer approach is the *basic probability assignment*, which assigns belief over the possible outcomes of an uncertain event. Let Ω be the set of all

possible outcomes of the event. For example, the event might be whether or not the inspection status of the next consignment on a pathway will be clean, noncompliant due to faulty paperwork, or noncompliant due to the presence of BRM. (For simplicity, let us suppose these are mutually exclusive, though of course in reality a consignment could have bad paperwork *and* BRM present.) Let $\mathcal{A}$ be a set in Ω: perhaps one of the singletons *clean*, *bad paper*, or *BRM present*, or perhaps a set formed by the union of more than one of those choices. The basic probability assignment satisfies

$$\begin{aligned} m(\emptyset) &= 0 \\ m(\mathcal{A}) \geq 0,\ \forall \mathcal{A} &\subseteq \Omega \\ \sum_{\mathcal{A}\subseteq\Omega} m(\mathcal{A}) &= 1 \end{aligned}$$

Note that belief can be assigned to any non-empty subset of Ω, not just to singletons, and that $m(\mathcal{A})$ represents the assignment of belief precisely to the set $\mathcal{A}$; thus, the quantities $m(bad\,paper)$, $m(BRM\,present)$, and $m(bad\,paper \cup BRM\,present)$ are distinct quantities and bear no necessary relation to one another, apart from not violating the constraints given above.

The belief function measures the total assignment of belief to each subset $\mathcal{A}$, *i.e.*,

$$\mathrm{Bel}(\mathcal{A}) = \sum_{\mathcal{B}\subseteq\mathcal{A}} m(\mathcal{B})$$

The plausibility function represents the total assignment of belief to outcomes that do not directly exclude $\mathcal{A}$, *i.e.*,

$$\mathrm{Pl}(\mathcal{A}) = \sum_{\mathcal{B}:\mathcal{A}\cap\mathcal{B}\neq\emptyset} m(\mathcal{B})$$

The belief and probability functions share direct relationships:

$$\begin{aligned} \mathrm{Bel}(\mathcal{A}) &= 1 - \mathrm{Pl}(\mathcal{A}^C) \\ \mathrm{Pl}(\mathcal{A}) &= 1 - \mathrm{Bel}(\mathcal{A}^C) \end{aligned}$$

where $\mathcal{A}^C$ is the complement of $\mathcal{A}$. Although Shafer (1976) emphasizes an epistemic interpretation, others have suggested that the belief and plausibility functions can be considered as lower and upper bounds on the probability of the outcomes within a decision-theoretic context (Dempster and Kong, 1987; Caselton and Luo, 1992).

A key difference from the Bayesian approach is the handling of complete ignorance. Because the Bayesian approach can only assign belief to singletons, the challenge of formulating a noninformative prior is one of assigning (very precise) probability masses to the (very precise) singleton subsets of Ω. Such assignments carry (very precise) behavioral implications even though they are based on a lack of information. By contrast, in the Dempster-Shafer approach, one assigns $m(\Omega) = 1$, and no mass to any smaller subset of Ω. This results in the *vacuous belief function*

$$\begin{aligned} \mathrm{Bel}(\Omega) &= 1 \\ \mathrm{Bel}(\mathcal{A} : \mathcal{A} \subset \Omega) &= 0 \end{aligned}$$

and the corresponding plausibility function

$$\begin{aligned} \mathrm{Pl}(\mathcal{A} : \mathcal{A} \subseteq \Omega) &= 1 \\ \mathrm{Pl}(\emptyset) &= 0 \end{aligned}$$

In other words, under a state of complete ignorance, any outcome in Ω is plausible, but we do not specifically believe any outcome is more likely than any other. This seems a more intuitive description of ignorance than, for example, insisting that all outcomes are in fact equally likely. If we know only that an urn contains red and blue marbles, do we really believe the probability of drawing a red marble is 1/2? And should that change if we know that the blue marbles come in two distinguishable colors, navy and aqua? Coarsening or refinement of Ω poses challenges for the representation of ignorance in the Bayesian approach, but not in the Dempster-Shafer theory.

When two independent bodies of belief or evidence are to be combined, the approach follows Dempster's rule of combination, as first laid out by Dempster (1967a). (A related combination rule when Ω is not discrete, but continuous, is presented by Shafer (2016).) In the discrete case, if m_1 and m_2 are independent probability assignments, we may compute the combined probability assignment $m(\mathcal{C})$ for any non-empty set $\mathcal{C}$ as

$$m(\mathcal{C}) = \frac{\sum_{\mathcal{A},\mathcal{B}:\,\mathcal{A}\cap\mathcal{B}=\mathcal{C}} m_1(\mathcal{A})m_2(\mathcal{B})}{\sum_{\mathcal{A},\mathcal{B}:\,\mathcal{A}\cap\mathcal{B}\neq\emptyset} m_1(\mathcal{A})m_2(\mathcal{B})} \tag{2.14}$$

where the top line of the equation simply combines the probability of any combinations of A and B that are not inconsistent with C. The bottom line effectively rescales the total probability of all non-empty sets back to 1 by excluding any combinations of A and B which are inconsistent. Readers are referred to Shafer (1976, Chapter 4) for a thorough and reasonably lucid treatment. Luckily, for certain special cases, the rule gives rise to much simpler results, and binomial sampling is one of these, as originally outlined by Dempster (1966) and fully developed by Dempster (1968b). If one begins from a vacuous belief function about p, and observes x successes in a binomial trial with sample size n, then a lower bound to the posterior distribution for p is given by a beta distribution with $\alpha = x$ and $\beta = n+1-x$. The corresponding upper bound is also a beta distribution with $\beta = n+1-x$, but with $\alpha = x+1$.

These lower and upper probability distributions can be interpreted behaviorally in decision problems (Dempster and Kong, 1987). In situations with linear utilities, only the upper and lower expected values of p are needed; these follow directly from the parameters of the beta distribution, *i.e.*,

$$\begin{aligned} \overline{p} &= (x+1)/(n+1) \\ \underline{p} &= x/(n+1) \end{aligned}$$

for the upper and lower values, respectively. In the case of substantiating freedom from BRM, we are more concerned with the belief in, and plausibility of, the actual prevalence exceeding the design prevalence (*i.e.*, $p > p^*$). Letting $F(x;\alpha,\beta)$ denote the cumulative distribution function for the beta distribution, we may calculate these two quantities as

$$\begin{aligned} \mathrm{Bel}(p > p^*) &= 1 - F(p^*, x, n+1) \\ \mathrm{Pl}(p > p^*) &= 1 - F(p^*, x+1, n+1) \end{aligned}$$

So, for example, having observed $x = 0$ fruits infested with BRM in a sample of $n = 600$ fruits from a consignment, our *belief* that the prevalence exceeds 0.005 in that consignment is zero (the beta distribution becomes degenerate when $\alpha = 0$, but we take the limiting value). This is sensible — there is nothing in the evidence to indicate there is actually any contamination. On the other hand, the *plausibility* that the prevalence exceeds 0.005, equals 0.0497. In other words, the evidence doesn't entirely rule out a higher prevalence,

but it is implausible. The situation changes if, in the next consignment, we observe $x = 2$ contaminated fruits in a sample of $n = 600$. Now, our belief that the prevalence exceeds 0.005 equals 0.1954: the evidence doesn't compel us to believe the proposition is true. But the plausibility is 0.4170 (or, conversely, the belief that the prevalence is less than 0.005 has fallen to 0.5830). Our confidence that this consignment is clean, is not high. We reject it.

2.6.2 Imprecise Probabilities

Although the Dempster-Shafer theory offers a flexible model for decision making under incomplete information and ambiguity (Walley, 1996b), from some perspectives its reliance on Dempster's rule of combination (and a related rule of conditioning) is its Achilles' heel. As a number of authors have shown, in certain circumstances the combination of belief functions can lead to behavioral implications that are illogical (Voorbraak, 1981; Walley, 1996b; Halpern and Fagin, 1992; Kyburg Jr., 1987; Pearl, 1990; Walley, 1987). Halpern and Fagin (1992) cite a number of earlier examples that appear in less-accessible proceedings, and emphasize the distinction between an interpretation of the belief function as generalized evidence (in which case the rule may be acceptable) versus generalized probability (in which case it is suspect). Such implications are especially problematic if the decision-theoretic perspective of Dempster and Kong (1987) is adopted.

Walley (1991) constructed an alternative approach to upper and lower probability, along with a related concept, that of upper and lower previsions constructed as prices for uncertain gambles. Walley (1991) followed Savage (1954) in seeking an axiomatic basis that would ensure consistency and behavioral rationality, while maintaining the philosophical and mathematical openness to indeterminacy that had characterized Keynes (1921), Dempster (1966), and Shafer (1976). The theory is constructed with behavioral implications at its cornerstone. Key principles link the subjectivist and objectivist interpretation of probabilities. Among these are coherence, avoidance of sure loss, and a principle of direct inference: if one knew the (aleatory) chances of an event, one would adopt those as one's (epistemic) belief about the probability of the event. Taken together, these principles give rise to an approach for combining interval-valued probability measures that can be interpreted as a generalization of Bayes' rule. Although Walley (1991) remains authoritative for the foundation of the theory, a great deal of subsequent work has built upon his framework; for a more recent review, albeit one with a great deal of mathematical formalism, see Augustin et al. (2014).

The theory of imprecise probabilities bears some resemblance to Dempster-Shafer theory in the way that complete prior ignorance is represented. Let $\mathcal{A}$ be a non-empty event (or set of events) in a sample space Ω. Then, under complete ignorance, belief about $\mathcal{A}$ is represented by the lower and upper probabilities

$$\underline{P}(\mathcal{A}) = 0$$
$$\overline{P}(\mathcal{A}) = 1$$

These probabilities have a behavioral implication: $\underline{P}$ is the maximum rate at which there is compelling reason to bet for $\mathcal{A}$, while $\overline{P}$ is the minimum rate at which there is compelling reason to bet against $\mathcal{A}$. Initially, there is no compelling reason to bet for or against $\mathcal{A}$; as information relevant to an event is acquired, we would expect $\underline{P}$ and $\overline{P}$ to converge. The vacuous probability assignment uniquely obeys several useful principles. It obeys the embeddedness principle that the assignment of probabilities should not depend

on the sample space in which $\mathcal{A}$ is embedded. For example, if $\mathcal{A}$ is "the next BRM encountered will be an arthropod," $\underline{P}$ and $\overline{P}$ do not depend on whether the possible outcomes are *arthropod, non-arthropod* or *arthropod, fungus, plant* or indeed how many types of arthropod or fungus or plant we might recognize and include as possibilities. This kind of invariance is very challenging to represent using the classic Bayesian non-informative priors. It also satisfies the symmetry principle: under a state of ignorance, all outcomes receive the same probability assignment. These conditions seem necessary to a proper description of ignorance, but as Walley (1996a, p. 5) states, "It is clear that Bayesian model can satisfy both principles."

In the case of binomial sampling, imprecise probability theory gives rise to results that remain superficially similar to those of Dempster-Shafer theory. Walley (1991, Section 5.3) presents an imprecise beta model for the posterior distribution of the parameter of the binomial (in our case, the prevalence p), having observed x successes in n trials. Walley (1991) begins by assuming a convex set of prior distributions, of sufficient breadth to give rise to the vacuous probability assignment, and also containing all the usual Bayesian noninformative priors as special cases. Under this assumption, the posterior lower and upper expectations for p are

$$\underline{P}(\mathcal{A}) = \frac{x}{n+s}$$
$$\overline{P}(\mathcal{A}) = \frac{x+s}{n+s}$$

where s is a hyperparameter describing the strength of the original (vacuous) prior. When $s = 1$, the results agree with those of the Dempster-Shafer theory (Dempster, 1966; Shafer, 1976). However, Walley (1991) and Walley (1996a) argue for $s = 2$ as encompassing all the usual "noninformative" Bayesian priors (including the improper Haldane prior), as well as ensuring that the resulting credible intervals are (at least) corresponding confidence intervals in the frequentist sense. The full posterior lower distribution for p is given by the beta distribution with parameters $\alpha = x$ and $\beta = n + s - x$; the corresponding upper distribution is given by $\alpha = x + s$ and $\beta = n - x$.

The use of a set of prior distributions would suggest that the imprecise beta model is essentially a robust Bayes approach (*e.g.*, Berger, 1990). However, to the robust Bayes analyst, the set of priors is only a tool for checking whether the initial prior is unduly influential on the result. Ultimately, the robust Bayesian will choose a single probability distribution as the prior, and achieve a correspondingly precise result. Under imprecise probability theory, no selection of prior within the set is ultimately made; the result of the analysis includes the final imprecision as an essential element. Moreover, as Walley (1991) notes, the use of a set of priors is one way to construct a coherent imprecise posterior probability, but it is not the only one.

Walley (1996a) extends the imprecise beta model to the case of multiple outcomes (the imprecise Dirichlet model), which in turn opens the door to a potentially broad set of settings for statistical inference. Further exploration of the basic model, in efforts to rely solely upon the likelihood function and avoid the need to define a prior distribution (either implicitly or explicitly), may be found in Walley and Moral (1999) and Walley (2002). Walley (2000) provides some unifying theory and terminology.

To date, there have been few direct applications of imprecise probability methods to biosecurity applications, though Coolen and Elsaeiti (2009) present work on acceptance sampling that may be applicable. Given the sensitivity of standard Bayesian methods to

the prior when little or no BRM is encountered, and the foundation of imprecise probability methods within a probabilistic framework, this imprecise probability perspective may be deserving of further exploration within biosecurity problems.

3 Case studies

In this section, we illustrate the use of design-based, model-based, Bayesian inference, Dempster-Shafer, and imprecise probability theory for non-clustered (simple random sampling) and clustered data (data collected in clusters or batches)[1] using two cases studies: a typical plant import pathway to Australia with a prevalence risk cutoff of 0.5% (*i.e.*, which is typically dealt with using a 600 sample inspection) and a typical plant import pathway to New Zealand with a prevalence risk cutoff of 0.01% (*i.e.*, which is typically dealt with using a 31,540 sample inspection). We will refer to the two pathways by their risk cutoffs to be clear that the differences arise from the nature of the risk of the pathways rather than from their national source.

The five inference frameworks differ in their ability to use information other than the inspection sample when making inference on the infestation rate of the inspected consignment (Table 3.1). Of the five frameworks, only Bayesian inference allows using external information to increase the precision of the estimates for both simple random sampling and clustered sampling case (to our knowledge, Dempster-Shafer theory has not been extended to clustered sampling).

3.1 Simple random sampling

We do simple random sampling when each unit in a consignment has the same probability of being sampled. Simple random sampling forms the basis for most of the statistical framework that has been developed in biosecurity (IPPC, 2008).

3.1.1 Design-based inference for simple random sampling

For both design-based (simple random sampling) and model-based inference (absence of clustering in the data collection or intra-cluster correlation coefficient of zero), the sensitivity S of a binomial test (probability of failing an inspection) is only a function of the infestation rate p of the consignment and the number of samples n of the consignment:

Table 3.1: Possibility to use external information in different inference frameworks

Inference framework	Simple random sampling	Clustered sampling
Design-based	No	No
Model-based	No	Yes
Bayesian	Yes	Yes
Dempster-Shafer	Yes	?
Imprecise probabilities	No	No

[1] In order to make the case studies chapter (chapter 3) self-contained, we re-derive or repeat equations that were given in the review chapter (chapter 2) when necessary.

$$S = 1 - (1 - p)^n \tag{3.1}$$

We can rearrange Eq. 3.1 to solve for n as a function of S and p^*, the design-based infestation rate that we are willing to accept:

$$n = \frac{\ln(1 - S)}{\ln(1 - p^*)} \tag{3.2}$$

For example, for a given p^* of 0.005 and S of 0.95, n equals 598. This n equals 598 is often rounded to 600 and forms the basis of the '600 samples rule' often used in biosecurity (Venette et al., 2002b; IPPC, 2008). In plain English, this means that in the long run, inspecting 600 samples per consignment will filter ∼95% of consignments that have an infestation rate of 0.5%. When the infestation rate is below 0.5%, the acceptance sampling procedure will filter fewer consignments and vice versa (see Eq. 3.1, which gives the sensitivity of the test when varying p and fixing n = 600). Alternatively, if we want to filter 95% of the consignments that have a 0.01% infestation rate, the sample size will need to be 29,956 (which has been rounded to 31,540 samples in the NZ case, as they assume a detectability of 95%, so that p^* in Eq. 3.2 is replaced by $0.95 \times p^*$).

In design-based sampling, no external information is used for decision making.

3.1.2 Model-based inference for simple random sampling

When the data arise from simple random sampling and is not clustered, design-based and model-based inference give the same answer. We can use Eq. 3.1 to estimate the sensitivity of the inspection and Eq. 3.2 to estimate its sample size (we need to sample n ∼600 units to filter 95% of consignments that have an infestation rate of 0.5% and 29,956 units to filter 95% of consignments that have an infestation rate of 0.01%).

In model-based sampling for non-clustered data, no external information is used for decision making.

3.1.3 Bayesian inference for simple random sampling

While in design-based inference, the acceptance sampling procedure is built up from the design-sensitivity S of the test, in Bayesian inference, acceptance sampling can be seen as an estimation problem: we accept an incoming consignment j if we are 95% sure that its infestation rate p_j is below a fixed threshold p^* of 0.5%.

In a Bayesian model-based setting, we combine prior information on potential p_j values for the consignment and data on the number of infected samples of the consignment to estimate the posterior probability of the consignment infestation rate p_j. If 95% of the posterior distribution of the consignment is below 0.5%, then consignment j is deemed compliant.

3.1.3.1 Beta distribution as conjugate prior to the Binomial likelihood

If we assume that the number of infested sampled units in consignment j follows a Binomial distribution, a natural choice of prior for the potential distribution of p_j before seeing consignment j's data is the Beta distribution. The Beta distribution is bounded in the [0–1] range and is flexible enough to represent different distributional shapes. The Beta distribution is also the conjugate prior to the Binomial likelihood. This means that

combining a Beta prior on p_j with a Binomial likelihood for the inspection data will produce a posterior distribution of p_j that is also Beta distributed. Specifically, the prior $p_j \sim Beta(\alpha, \beta)$ gets updated to the posterior $p_j \sim Beta(\alpha + x, \beta + n - x)$ after finding x infested units out of n samples when inspecting consignment j.

3.1.3.2 Bayesian inference using a noninformative prior. Case study with a 0.5% risk cutoff

If we use a uniform prior $Beta(1, 1)$ on p_j (*i.e.*, any values of p_j in [0–1] is equally likely a priori), we need to sample 597 units free of BRM to have 95% confidence that p_j is below 0.5%.[2]. That is, when we use a uniform prior on p_j, the Bayesian inference procedure gives a similar answer as the design-based and model-based procedure (*i.e.*, the '600 samples' rule).

3.1.3.3 Bayesian inference using informative priors. Case study with a 0.5% risk cutoff

However, a uniform prior on the potential distribution of p_j is often unrealistic. Most of the consignments have an infestation rate well below 100%. We can use a more sensible prior for future consignments in the pathway by improving our knowledge of the pathway, *e.g.*, by estimating the potential distribution of p_j among consignments. If each consignment has its own infestation rate[3] and the distribution of infestation rate among consignments in the pathway follows a Beta distribution, then we can estimate this Beta distribution by fitting a beta-binomial model to past data on the pathway[4]:

$$\begin{aligned} x_j &\sim Binomial(p_j, n_j) \\ p_j &\sim Beta(\alpha, \beta) \end{aligned} \tag{3.3}$$

where x_j is the number of infected units out of n_j sampled units in consignment j, p_j is the infestation rate of consignment j, and α and β are parameters to be estimated.

Assuming that the infestation rate for future consignments in the pathway also follow the estimated $p_j \sim Beta(\alpha, \beta)$, our best guess at p_j before seeing consignment's j data is $Beta(\alpha, \beta)$. This prior information will increase the precision of p_j even after seeing the data. It will often reduce the number of sampled units needed to be 95% sure that the infestation rate of the inspected consignment p_j is below 0.5%.

For example, in the import plant germplasm pathway for Australia, the distribution of infestation rate among consignments follows a Beta distribution with mean $p = 0.022$ and $\rho = 0.11$ ($\alpha = 0.17$, $\beta = 8$). This corresponds to $\sim 60\%$ of the consignments having

[2]The probability mass of a distribution that is below a given threshold p can be computed in most statistical software. In R, we would use the 'pbeta' function. We can increase the number of samples n until the 95% of the posterior distribution is below 0.005. For example, after finding zero BRM out of 596 samples, 94.9% of the posterior distribution pbeta(0.005, 1, 1+596) = 0.949 is below 0.005, which doesn't reach our criteria. After finding zero BRM out of 597 samples, the 95% of the posterior distribution is pbeta(0.005, 1, 1+597) = 0.950 is below 0.005. When we use a uniform prior on p_j, the sample size is n=597.

[3]This is called heterogeneity or overdispersion and it seems relatively common in quarantine biosecurity data

[4]A number of software can be used for this application. Here we used the 'vglm' function from the 'VGAM' package in R, which is parametrized in terms of its mean p and its ICC ρ. The relationship linking parameters p and ρ and the usual parameters α and β of the Beta distribution is: $p = \frac{\alpha}{\alpha+\beta}$ and $\rho = \frac{1}{\alpha+\beta+1}$.

an infestation rate below 0.5% (Fig. 3.1). Using $Beta(0.17, 8)$ distribution as prior, we can compute the number of sampled units free of BRM that needs to be sampled in order to be 95% sure that the infestation rate of new consignments in the pathway are below 0.5%. With a $Beta(0.17, 8)$ prior (95% percentile equals 0.12), it is reached for $p_j \sim Beta(0.17, 8 + 183)$, that is, when we find zero BRM out of 183 sampled units (95% percentile equals 0.005, Fig. 3.1).

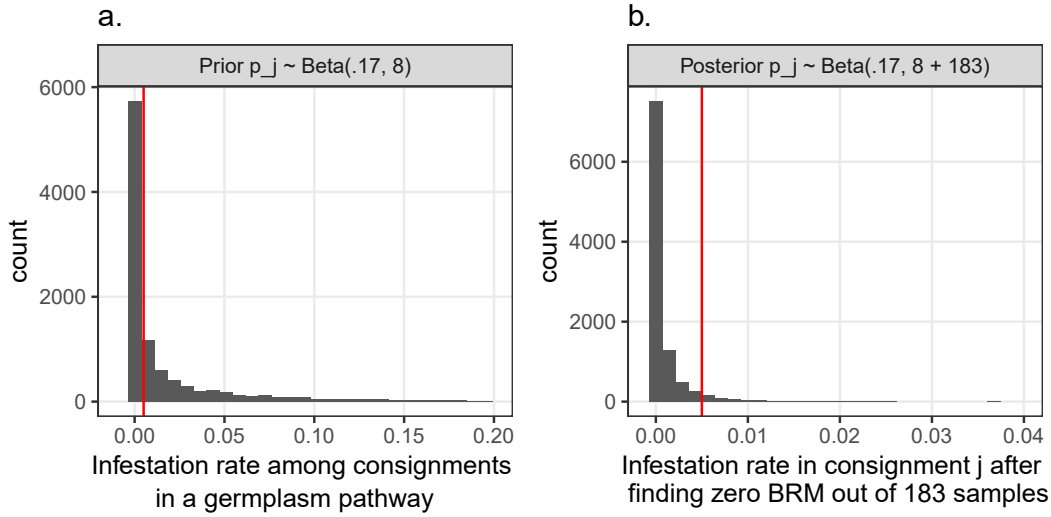

Figure 3.1: a. Estimated distribution of p_j among consignments in the import plant pathway to Australia. b. Posterior distribution of infestation rate in an incoming consignment after finding zero BRM out of 183 samples during an inspection. The vertical red lines represent p^*=0.005 (60% of the p_j values are left of the red line in a, while 95% of the values are left of the red line in b). Note that the scale of the x-axis is zoomed-in for the posterior distribution.

A note of caution: using past data to estimate the distribution of infestation rate among consignments in the pathway assumes that the distribution does not change with time. This is unlikely to be true. We advise regular re-estimation of this distribution, as well as using 'robustified' priors (see section below) to allow for higher infestation rates values than what has been detected in past data.

In Bayesian inference, external sources of information other than past data can be used. However, if we want to use them, these external source of information have to be expressed as a prior. If, for convenience, we want to keep using the Beta-Binomial model (as it is analytically tractable), the prior has to be a Beta distribution. For example, we might use a different prior distribution for pathways that have a systems approach vs. pathways that do not (see section 3.3).

Using more robust informative priors. We can also penalize our informative prior so that it is more broad than implied by the distribution of p_j in the pathway. A natural way 'robustify' our informative prior is to use a prior made of a mixture of a Beta and a uniform distributions (Schmidli et al., 2014). When using a mixture of conjugate priors, the posterior distribution is simply the weighted posterior of each individual component of the mixture (Dalal and Hall, 1983; Diaconis and Ylvisaker, 1985), which is analytically tractable: In our example, after sample n units free of BRM, a mixture prior $\phi Beta(.17, 8) + (1 - \phi)Beta(1, 1)$, where ϕ represents the weight associated with the past

data prior (*i.e.*, how much we trust past data), will lead to a posterior that is distributed $\phi Beta(.17, 8+n) + (1-\phi)Beta(1, 1+n)$. In this specific example and with a weight ϕ of 0.5, we will need to sample 472 units free of BRM before 95% of the mass of the posterior distribution of p_j for the inspected consignment is below the threshold infestation rate of 0.5%[5].

A second way to avoid giving too much weight to past data is to follow Tuyl et al. (2008) advice and limit informative priors to $Beta(1, \beta)$ distributions. In our example, the $Beta(1, \beta)$ distribution that is most similar to the $Beta(.17, 8)$ distribution in terms of its mean is the $Beta(1, 47)$ distribution. With a $Beta(1, 47)$ prior, the sample size will be 551 units (a $Beta(1, 47)$ prior is effectively equivalent to having a beta-binomial model with uniform prior and considering that we have already inspected the equivalent of 46 units free of BRM before even starting the inspection). An alternative way to use Tuyl et al. (2008) robust prior is to find the $Beta(1, \beta)$ distribution that the same 95% quantile as the $Beta(0.17, 8)$ distribution, *i.e.*, the $Beta(1, 24)$ distribution. With a $Beta(1, 25)$ prior, we will need to sample 573 units before 95% of the posterior distribution of the infestation of the inspected consignment is below 0.5%. In these later case studies, the gain from using an informative prior vs. an noninformative prior is relatively small.

3.1.3.4 Bayesian inference, case study with a 0.01% risk cutoff

This is the current design prevalence and sensitivity aimed for to import certain products in New Zealand. In a Bayesian inference framework with simple random sampling and using a uniform prior, we will need to sample 29955 units free of BRM (rounded to 31540 in the protocol) before 95% of posterior mass of the infestation rate of the consignment is below 0.01%. In a typical import plant pathway to New Zealand, the distribution of infestation rate among consignments follows a $Beta(0.253, 9623)$ distribution. This corresponds to $\sim$ 92% of the consignments having an infestation rate below 0.01% (Fig. 3.2a.). Using $Beta(0.253, 9623)$ distribution as prior, we can compute the number of sampled units free of BRM that needs to be sampled in order to be 95% sure that the infestation rate in consignments that are accepted is below 0.01%. With a $Beta(0.253, 9623)$ prior, it is reached for $p_j \sim Beta(0.253, 9623 + 2580)$, that is, when we find zero BRM out of 2580 sampled units (95% percentile equals 0.01%) (Fig. 3.2b.).

Using a robust informative prior. If we penalize the informative prior using a mixture prior that combines Beta(.253, 9623) and Beta(1, 1) with a weight of 0.5, we will need to sample 23386 units free of BRM before 95% of the mass of the posterior distribution of p_j for the inspected consignment is below the threshold infestation rate of 0.01%[6].

If we follow the Tuyl et al. (2008) robust prior approach, the $Beta(1, \beta)$ distribution that has the same mean than the $Beta(0.253, 9623)$ distribution is $Beta(1, 38035)$. This poses an issue as this prior distribution already has 97.8% of its mass below 0.01% (we would accept consignments from this pathway without inspecting any unit). An alternative way to use Tuyl et al. (2008) robust prior is to find the $Beta(1, \beta)$ distribution that the same 95% quantile as the $Beta(0.253, 9623)$ distribution, *i.e.*, the $Beta(1, 23600)$ distribution. With a $Beta(1, 23600)$ prior, we will need to sample 6356 units before 95% of the posterior distribution of the infestation of the inspected consignment is below 0.01%, which is more reasonable but might also be too optimistic.

[5] In R, 0.5 * pbeta(0.005, .17, 8 + 472) + 0.5 * pbeta(0.005, 1, 1 + 472)
[6] In R, 0.5 * pbeta(0.0001, .253, 9623 + 23386) + 0.5 * pbeta(0.0001, 1, 1 + 23386)

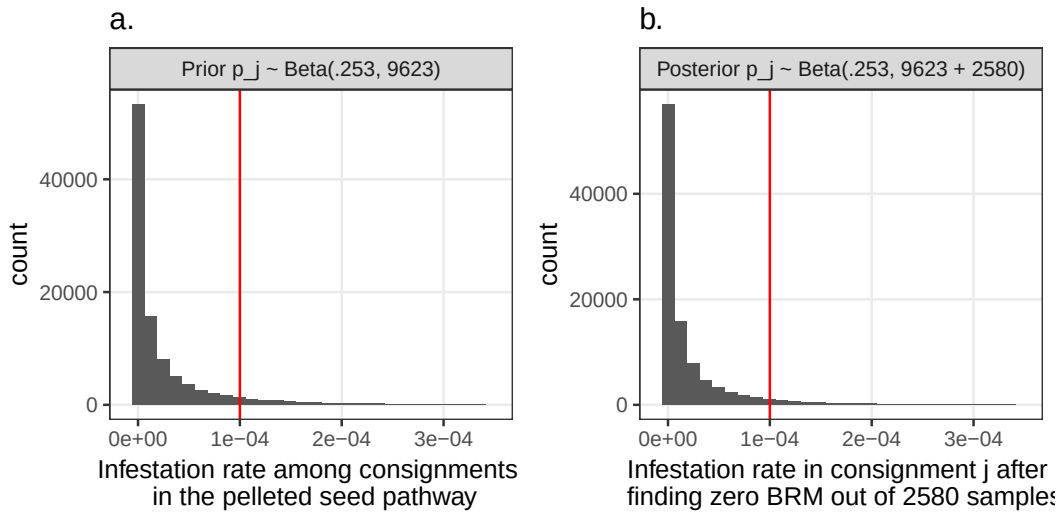

Figure 3.2: a. Infestation rate among consignments in a specific germplasm pathway. b. Posterior distribution of infestation rate in an accepted consignment of the pathway after finding zero BRM out of 2589 samples during an inspection. The vertical red lines represent p^*=0.01% (92% of the p_j values are left of the red line in a, while 95% of the values are left of the red line in b). Note that since the infestation rate of the pathway is already very low, the inspection doesn't reduce the infestation rate by much (posterior mean infestation rate of the inspected consignment is 2.07×10^-5, which is only 0.78 times lower than the prior mean of 2.63×10^-5).

3.1.4 Dempster-Shafer theory of evidence for simple random sampling

An alternative framework that allows combining several sources of information when making inference is the Dempster-Shafer theory of evidence. When using the Dempster-Shafer theory of evidence in a biosecurity setting, it might be easier to work directly at the scale of the decision (*i.e.*, whether the infestation rate of a consignment is below 0.5%) rather than on the infestation rate within each consignment. Below, we follow the case study and steps given in Rathman et al. (2018). In a biosecurity inspection setting, our decision set might have three focal elements: C (compliant), NC (non-compliant), and C, NC (we don't know).

Inference from one source of information. Following Rathman et al. (2018), the first step of the analysis consists of assigning a probability $Pr(\{C\})$ (*i.e.*, the probability that the consignment is compliant) to focal element $\{C\}$ and $Pr(\{NC\})$ (*i.e.*, the probability that the consignment is non-compliant) to focal element $\{NC\}$. We also need to assess the relative reliability (rel) of these probability assignments (how reliable is the source of information that allowed us to fix these probabilities?). The basic probability mass associated with each focal element is then computed by multiplying its initial probability assignment by the reliability of the information. The mass for the remaining focal element $m(\{N, NC\})$ (*i.e.*, the probability of not knowing the correct status of the consignment) is then calculated so that the mass for the three focal groups sums to one.

$$
\begin{aligned}
m(\{C\}) &= Pr(\{C\}) \times rel \\
m(\{NC\}) &= Pr(\{NC\}) \times rel \\
m(\{C, NC\}) &= 1 - m(\{C\}) - m(\{NC\})
\end{aligned}
$$

The probability masses are in turn used to compute the Belief (lower probability bound) and Plausibility (upper probability bound) associated with each focal element:

$$
\begin{aligned}
Bel(\{C\}) &= m(\{C\}) \\
Pls(\{C\}) &= m(\{C\}) + m(\{C, NC\}) \\
Bel(\{NC\}) &= m(\{NC\}) \\
Pls(\{NC\}) &= m(\{NC\}) + m(\{C, NC\})
\end{aligned}
$$

3.1.4.1 Dempster-Shafer theory, case study with a 0.5% risk cutoff (600 samples inspection)

After a clean 600 samples inspection, we will be 95% confidence that the consignment is compliant. However, we are not 5% confident that the consignment is non-compliant (since we found no infested samples), rather we are 5% confident of not knowing the correct status of the consignment. We consider the reliability to be 100%.

$$
\begin{aligned}
rel &= 1 \\
Pr(\{C\}) &= 0.95 \\
Pr(\{NC\}) &= 0
\end{aligned}
\tag{3.4}
$$

The probability mass associated with a clean 600 samples inspection are:

$$
\begin{aligned}
m_1(\{C\}) &= Pr(\{C\}) \times rel = 1 \times 0.95 = 0.95 \\
m_1(\{NC\}) &= Pr(\{NC\}) \times rel = 0 \times 0.95 = 0 \\
m_1(\{C, NC\}) &= 1 - m_1(\{C\}) - m_1(\{NC\}) = 1 - 0.95 - 0 = 0.05
\end{aligned}
$$

As expected from the literature ((Dempster, 1968b) and section 2.6.1), the belief (lower probability bound) and plausibility (upper probability bound) in compliance will be 95% and 100%, respectively. The belief and plausibility in non-compliance will be zero and 5%, respectively:

$$
\begin{aligned}
Bel(\{C\}) &= m_1(\{C\}) = 0.95 \\
Pls(\{C\}) &= m_1(\{C\}) + m_1(\{C, NC\}) = 1 \\
Bel(\{NC\}) &= m_1(\{NC\}) = 0 \\
Pls(\{NC\}) &= m_1(\{NC\}) + m_1(\{C, NC\}) = 0.05
\end{aligned}
$$

Adding an external source of data Additionally, if we know from past data that 62% of the consignments in the pathway were compliant and 38% were non-compliant (*e.g.*, if the distribution of infestation rate in the pathway is $p_j \sim Beta(.17, 8)$), we can assign basic probabilities $Pr(\{C\}) = 0.62$ and $Pr(\{NC\}) = 0.38$ to the focal elements. Because this represents past data on the pathway, we assume the reliability to be say 70%. The probability mass associated with this second source of information are:

$$\begin{aligned} m_2(\{C\}) &= 0.62 \times 0.7 = 0.434 \\ m_2(\{NC\}) &= 0.38 \times 0.7 = 0.266 \\ m_2(\{C, NC\}) &= 1 - 0.434 - 0.266 = 0.3 \end{aligned}$$

Combining the two sources of information. Following Rathman et al. (2018), we compute the ground probability masses by combining basic probability masses from different sources:

$$\begin{aligned} q(\{C\}) &= m_1(\{C\}) \times m_2(\{C\}) + m_1(\{C\}) \times m_2(\{C, NC\}) \\ &\quad + m_1(\{C, NC\}) \times m_2(\{C\}) \\ &= 0.95 \times 0.434 + 0.95 \times 0.3 + 0.05 \times 0.434 = 0.719 \\ q(\{NC\}) &= 0 \times 0.266 + 0 \times 0.3 + 0.05 \times 0.266 = 0.0133 \\ q(\{C, NC\}) &= m_1(\{C, NC\}) \times m_2(\{C, NC\}) \\ &= 0.05 \times 0.3 = 0.015 \\ q(\{\emptyset\}) &= m_1(\{C\}) \times m_2(\{NC\}) \\ &\quad + m_1(\{NC\}) \times m_2(\{C\}) \\ &= 0.95 \times 0.266 + 0 \times 0.434 = 0.2527 \end{aligned}$$

In the original Dempster-Shafer combination rule, the joint basic probability mass m_D associated with each joint focal element is obtained by:

$$\begin{aligned} m_D(\{C\}) &= \frac{q(\{C\})}{1 - q(\{\emptyset\})} = 0.719/(1 - 0.2527) = 0.9621303 \\ m_D(\{NC\}) &= \frac{q(\{NC\})}{1 - q(\{\emptyset\})} = 0.0133/(1 - 0.2527) = 0.0177974 \\ m_D(\{C, NC\}) &= \frac{q(\{C, NC\})}{1 - q(\{\emptyset\})} = 0.015/(1 - 0.2527) = 0.02007226 \\ m_D(\{\emptyset\}) &= 0 \end{aligned}$$

Dempster's combination rule results expressed as belief and plausibility functions:

$$\begin{aligned} Bel(\{C\}) &= m_D(\{C\}) = 0.9621303 \\ Pls(\{C\}) &= m_D(\{C\}) + m_D(\{C, NC\}) = 0.9822026 \\ Bel(\{NC\}) &= m_D(\{NC\}) = 0.0180711 \\ Pls(\{NC\}) &= m_D(\{NC\}) + m_D(\{C, NC\}) = 0.03822136 \end{aligned}$$

After combining two sources of data, the Belief for compliance is 0.9618449 and the Plausibility for compliance is 0.9819602: we accept the consignment.

An important feature of the ground probability mass relationships is that they are associative; *i.e.*, the order in which the required pairwise operations are performed does not matter. Thus, we can apply the procedure sequentially to handle several sources of evidence or update an existing combined structure with a new source of evidence (combine m_1 and m_2 to form m_D, then combine m_D and m_3 to form m_{D2}...).

Note that these calculations can be simplified by using the "EvCombR" package in R:

```
# Load packages
  library(EvCombR)
# Construct a state space
  state_space <- c("C", "NC") # Compliant and non-compliant set
# First data source (inspection data)
  # Set parameters
    # A clean 600 samples inspection
    C1 <- 0.95 # Probability of compliance
    NC1 <- 0 # Probability of non-compliance
    rel1 <- 1 # Reliability of the inspection
    m1C <- C1 * rel1 # Probability mass of compliance
    m1NC <- NC1 * rel1 # Probability mass of non-compliance
    m1C_NC <- 1 - m1C - m1NC # # Probability mass of "We don't know" category
  # Construct mass functions
    # Belief is m_1C; Plausibility is m_1C +  m_1C_NC
    m1 <- mass(list("C"=m1C, "NC"=m1NC, "C/NC"=m1C_NC), state_space)
# Second data source (past data, with p_j~Beta(.17, 8).
# 62% of the consignment are compliant and 38% are non-compliant
  # Set parameters
    C2 <- pbeta(0.005, .17, 8) # ~62% of the consignments are compliant
    NC2 <- 1 - pbeta(0.005, .17, 8) # ~38% are non-compliants
    rel2 <- 0.7 # Reliability of say 70%, as this is past data
    m2C <- C2 * rel2
    m2NC <- NC2 * rel2
    m2C_NC <- 1 - m2C - m2NC
  # Construct mass functions
    m2 <- mass(list("C"=m2C, "NC"=m2NC, "C/NC"=m2C_NC), state_space)
# Combine the mass functions by using Dempster's combination rule
  dComb(m1, m2)
```

Now, we can vary the number of samples in the inspection to see how many sample are required to have a Belief of compliance of 0.95. When combined with the information on past data on the pathway, a probability of compliance of 93.4% for the inspection data alone is enough to provide a compliance of 0.9500192 for the combined data. This corresponds to a clean inspection of 542 samples (93.4% of the posterior probability mass is below 0.5%): using external information within the Dempster-Shafer inference framework allow reducing sample size from 597 to 542 (reducing sample size by 55 units or, equivalently, by a factor of 0.91). This result (542 samples) assumes the reliability of the historical data is 70%. If the reliability is 100% then around 493 samples are needed and if the reliability is zero, 597 samples.

3.1.4.2 Dempster-Shafer theory, case study with a 0.01% risk cutoff (29956 samples inspection)

In the binomial simple random sampling, zero finds from a sample size of 29956 provides assurance to be 95% confident that the infestation rate is below 0.01%. As we work on the probability of compliance and not on the infestation rate itself, the computations for this number would be identical to those of the 600 samples above (a Belief of compliance of 0.95 and a Plausibility of one). The difference is in the underlying sample size that this represents (600 samples for a design prevalence of 0.5% vs. 29956 samples for a design prevalence of 0.01%).

From past data on the pathway, we know that say 85% of the consignments were compliants. If we applied a reliability of 70% for this past data and combined the mass functions using the Dempster's combination rule, we obtain a Belief of 0.978 and a Plausibility of 0.994 for compliance.

When combined with the information on past data on the pathway, a probability of compliance of 93.4% for the inspection data alone is enough to provide a Belief of 0.9508576 for the combined data. This corresponds to a clean inspection of 21896 samples (93.4% of the posterior probability mass is below 0.01%): using external information within the Dempster-Shafer inference framework allows reducing sample size from 29956 to 21896 (reducing sample size by 8060 units or, equivalently, by a factor of 0.73).

3.1.5 Imprecise probabilities for simple random sampling

In the case of binomial sampling encountered in biosecurity inspection, the full posterior lower distribution for the infestation rate p of the inspected consignment with x BRMs out of n samples is $Beta(x, n + s - x)$; the corresponding upper distribution is given by $Beta(x + s, n - x)$, where s is a hyperparameter describing the strength of the original (vacuous) prior (see section 2.6.2). Since we are interested in the probability of exceeding an infestation rate, we focus specifically on the upper distribution which after a clean inspection sample (x=0) is $Beta(s, n)$.

When $s = 1$, the upper posterior distribution is $Beta(1, n)$ and the results are similar to the results given by Bayesian inference with a uniform prior and Dempster-Shafer theory: the sample size which has 95% of its upper posterior distribution mass below 0.5% is obtained for $n = 598$ ($n = 597$ for Bayesian inference with uniform prior); the sample size which has 95% of its upper posterior distribution mass below 0.01% is obtained for $n = 29956$ ($n = 29955$ for Bayesian inference with uniform prior).

However, Walley (1991) and Walley (1996a) argue for using $s = 2$. In such a case, the sample size which has 95% of its upper posterior distribution mass below 0.5% is obtained for $n = 946$ and the sample size which has 95% of its upper posterior distribution mass below 0.01% is obtained for $n = 47436$. When choosing $s = 2$, the sample size given by imprecise probability theory is higher than in the other inference frameworks.

Note that imprecise probability theory does not allow the use of external information through the use of a prior (indeed its main goal is to avoid specifying a prior that might be too informative).

3.2 Clustered sampling

It is common for biosecurity data within a consignment to be clustered (*e.g.*, , we sample multiple seeds per crates from several crates in the consignment).

3.2.1 Design-based inference for clustered sampling

If the data comes in cluster but we still manage to do simple random sampling (*i.e.*, each unit is equally likely to be sampled) we can ignore the clustered nature of the original data and to uses the formula for non-clustered data (Eqs. 3.1 and 3.2): inspecting 600 units leads to a 95% sensitivity for a 0.5% infestation rate.

However, if the data is sampled in clusters but we have a clean inspection (*i.e.*, zero BRM), the design-based estimators will have learned nothing about the potential value of the intra-cluster correlation coefficient ρ, and thus we will not be able to draw conclusions about the mean infestation rate p of the consignment. We will instead need to resort to using external information and calibrate ρ from past data on the pathway. This is possible in either the model-based or Bayesian framework.

3.2.2 Model-based inference for clustered sampling

When the items and hence sampling is clustered, we have to account for the clustering in the statistical procedure. Clustering reduces inspection sensitivity (it reduces the chances of finding infested units in an inspection). This means that if we want to keep the sensitivity constant, we will need to increase the sample size relative to the sample size of the binomial sampling. How much more to sample depends on the degree of aggregation of the pest (the ICC) and on the number of units sampled per cluster.

3.2.2.1 Using the Beta-binomial model for clustered data.

The default model for Binomial clustered sampling is the Beta-Binomial model[7]. Under the Beta-Binomial model, each cluster has its own prevalence p_k and the distribution of prevalence among clusters is assumed to follow a Beta distribution, *i.e.*,

$$f(p_k) = \frac{p_k^{\alpha-1}(1-p_k)^{\beta-1}}{B(\alpha,\beta)} \tag{3.5}$$

where $B(\alpha,\beta)$ is the beta function[8]. The beta distribution can be described by its mean $p = \alpha/(\alpha+\beta)$, corresponding to the population-level infestation rate, and an intracluster correlation coefficient (ICC) $\rho = 1/(\alpha+\beta+1)$ which describes the similarity among units sampled from the same cluster.

[7]Alternative and more complex models do exist. For example, we might model clustered sampling using a zero-inflated approach, which model two processes: the proportion of cluster that is infested (which can be modelled by *e.g.*, a Binomial distribution) and the infestation rate in cluster that are infested (which can be modelled by *e.g.*, a Beta-Binomial distribution). Since we did not have data on the number of infested samples per cluster and number of samples per cluster, we could not test which of the zero-inflated model or Beta-Binomial model was more adequate, and kept the simpler Beta-Binomial as default model. It would be useful to record this type of data to test which model is more appropriate for a range of pathways.

[8]The beta function can be found in different software such as R

If we draw n_k sampling units by a noninformative procedure from the k^{th} cluster, the unconditional probability of finding x_k contaminated units (*i.e.*, without knowing the cluster-level prevalence p_k) is

$$bb(x; n_k, \alpha, \beta) = \binom{n_k}{x_k} \frac{B(\alpha + x_k, \beta + n_k - x_k)}{B(\alpha, \beta)} \tag{3.6}$$

The probability of a completely clean sample from an individual cluster is therefore

$$bb(x = 0; n_k, \alpha, \beta) = \frac{B(\alpha, \beta + n_k)}{B(\alpha, \beta)} \tag{3.7}$$

giving, as the overall sensitivity S for a sample of n units, from n/n_k clusters with n_k in each unit:

$$S = 1 - bb(x = 0; n_k, \alpha, \beta)^{n/n_k} \tag{3.8}$$

We can re-arrange Eq. 3.8 to derive an exact sample size formula for n given S, α, β, and n_k, the number of units sampled per crate[9]. The formula follows:

$$n = \frac{n_k \ln(1 - S)}{\ln B(\alpha, \beta + n_k) - \ln B(\alpha, \beta)} \tag{3.9}$$

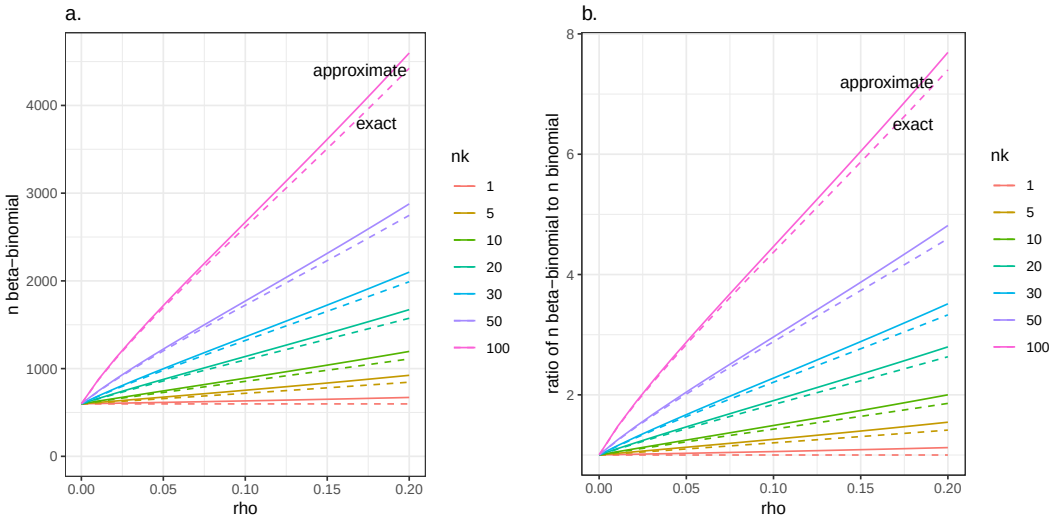

Figure 3.3: a. Sample size for clustered data given by the beta-binomial model (approximate and exact solution). b. Ratio of sample size for the beta-binomial (approximate and exact) and the binomial distribution when varying ρ and n_k. To compare with the 600 samples rule, sensitivity was fixed to 0.95 and p^* to 0.005. As can be seen, the approximate equation is not consistent when n_k=1 (*i.e.*, when cluster sampling converge to simple random sampling).

[9]To our knowledge, the exact formula derived in this report (Eq. 3.9) is original. Previously, an approximate formula based on the negative-binomial approximation to the beta-binomial was used (Venette et al., 2002b). For the sake of comparison, we show sample size predictions for both the exact and approximate formula in Fig. 3.3.

Case study with a risk cutoff of 0.5%. Typically, keeping the sensitivity of the test constant with increasing ρ values requires sampling more units (Fig. 3.3). For example, in the absence of clustering, a 95% chance of detecting an infestation rate of 0.5% requires sampling $\sim$ 600 units (Eq. 3.2) whereas it can require sampling $\sim$ 860 units (10 units per crate out of 86 crates) when ρ equals 0.1 and n_k=10 units per crate (Eq. 3.9). The sensitivity of the test also depends on how many units per crates are sampled. When we sample only one unit per crate ($n_k = 1$), the Beta-Binomial result collapses to the Binomial case (Fig. 3.3), illustrating how simple random sampling of crates protects against clustering. Increasing the number of units sampled per crate decreases the sensitivity of the test. Keeping the sensitivity constant requires to increase the overall number of units to sample in the consignment. For example, when ρ equals 0.1, reaching a sensitivity of 95% for $p^* = 0.5\%$ requires sampling 600 units when $n_k = 1$ (600 crates), 860 units when $n_k = 10$ (86 crates), 1100 units when $n_k = 20$ (55 crates), and 2600 units when $n_k = 100$ (26 crates). Choosing between a number of crates and number of units per crate will depends on the relative costs and convenience of sampling a new crates vs. sampling more units within the same crate. This will requires having an optimization procedure minimizing the cost.

Case study with a risk cutoff of 0.01%. This is the case for the data from New Zealand. The sample size given by the binomial equation is 29955, which in practice is rounded up to 31,540. We consider a range of sample per cluster n_k of 1–1000. As seen in Fig. 3.4, if we sample from 30 clusters only (*i.e.*, $n_k \sim 1000$), the effective sample size can be much lower than expected for simple random sampling. For example, if the ICC is 0.1 and n_k=1000, we will need to sample 23 times more units than under the binomial sample size for simple random sampling, *i.e.*, close to 700,000 seeds to reach a 95% sensitivity of detecting a 0.01% infestation rate (this is 1,000 seeds per cluster from 700 clusters). If we keep a sample size n of 31,540 units, the sensitivity at a design prevalence of 0.01% is equivalent to sampling 31,540 / 23 units, *i.e.*, a sensitivity of 13%. Equivalently, the risk-cutoff for which we would have a 95% sensitivity would be for a design prevalence of $\sim$ 0.22%.

A better result can be achieved if we sample 100 seeds per cluster. In this case, we would need to sample 130729 units ($\sim$ 4.4 times more than for the binomial sample size) to have a 95% sensitivity of detecting a 0.01% infestation rate (100 seeds per cluster from 1364 clusters). If we keep a sample size n of 31,540 units, the sensitivity at a design prevalence of 0.01% is equivalent to sampling 31,540 / 4.4 units, *i.e.*, a sensitivity of 51%. Equivalently, the he risk-cutoff for which we would have a 95% sensitivity would be for a design prevalence of $\sim$ 0.042%.

3.2.2.2 Estimating the intra-cluster correlation coefficient ρ from past data

One of the pre-requisites to applying Eq. 3.9 to clustered data is knowing the intra-cluster correlation coefficient ρ of a pathway. The intra-cluster correlation coefficient informs us on how similar the infestation rates among crates are. When ρ equals zero, the infestation rate is the same for all crates and we can ignore the fact that we do clustered sampling. When ρ increases, the variability in infestation rate among crates increases. In a biosecurity setting, we would want to estimate ρ from past data on the pathway and use Eq. 3.9 to fix the number of sample units per consignment of the pathway given S, n_k, p^*, and ρ. One question then is how much data do we need to reliably estimate ρ for a

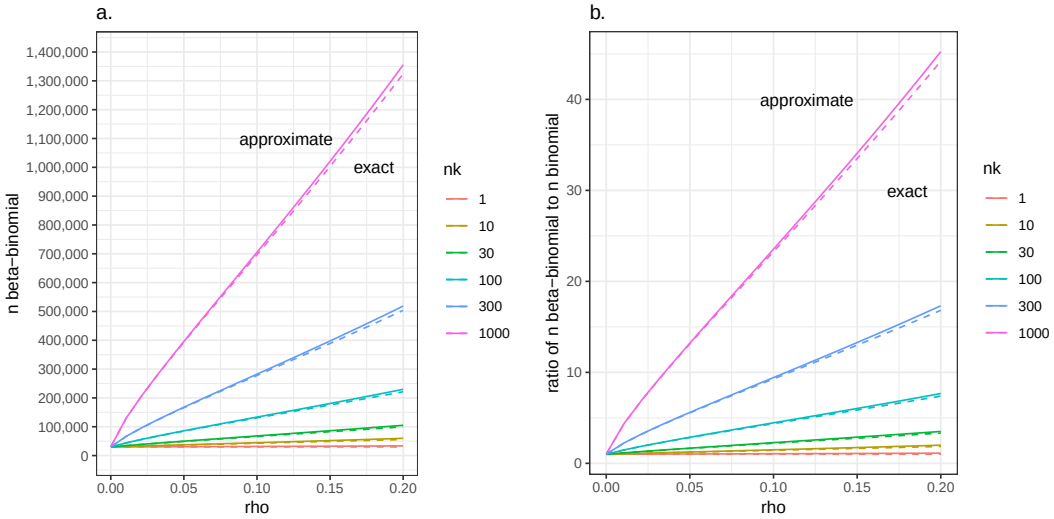

Figure 3.4: a. Sample size for clustered data given by the beta-binomial model (approximate and exact solution). b. Ratio of sample size for the beta-binomial (approximate and exact) and the binomial distribution when varying ρ and n_k. To compare with the 31,540 samples rule, sensitivity was fixed to 0.95 and p^* to 0.01%.

given pathway? As we currently do not have a database reporting the number of infected samples for each cluster of each consignment, we will rely on simulating datasets to get an idea of the number of samples required to estimate ρ with sufficient precision.

Case study with a risk cutoff of 0.5% While in theory it might be possible to estimate ρ from only one consignment, in practice it will require a consignment and an inspection that are much larger than usual (the typical sampling size of 600 units per consignment with only one consignment doesn't allow to reliably estimate ρ, cf. Fig. 3.6. Exploratory analyses showed that getting reliable ρ estimates from only one consignment might require sampling at least n_k=50 units per crate from n_{crates}=200 crates, *i.e.*, 10,000 samples). Additionally, this particular consignment might not be representative of the pathway. Hence, in a situation where we have a pathway containing several consignments and each consignment contains several crates, we suggest combining data from different consignments to estimate ρ. However, in such a case, it would be necessary to account for the heterogeneity among consignments before estimating the heterogeneity among crates within consignments (the ICC=ρ). Below, we detail a possible implementation of such a hierarchical Beta-Binomial model with two levels of heterogeneity (Fig. 3.5): the first level represents the variability of infestation rate p_j among consignments (Eq. 3.10 which we modelled using a normal distribution with standard deviation σ on the logit scale[10]. The second level represents the variability of infestation rate p_{jk} among crates within a consignment j which we modelled using a beta-distribution. As a simplification, we assumed the intra-cluster correlation coefficient ρ describing the heterogeneity among crates within

[10]In theory, it should be possible to use a hierarchical Beta Beta-Binomial model, with a first Beta distribution modelling the heterogeneity of infestation rate among consignments (cf. section 3.1.3) and a second Beta distribution modelling the heterogeneity among crates within one consignment (cf. this section). However, the hierarchical Beta Beta-Binomial model proved to be unpractical and we had convergence issues fitting the model. Hence we instead resorted to using the hierarchical logistic Beta-Binomial model

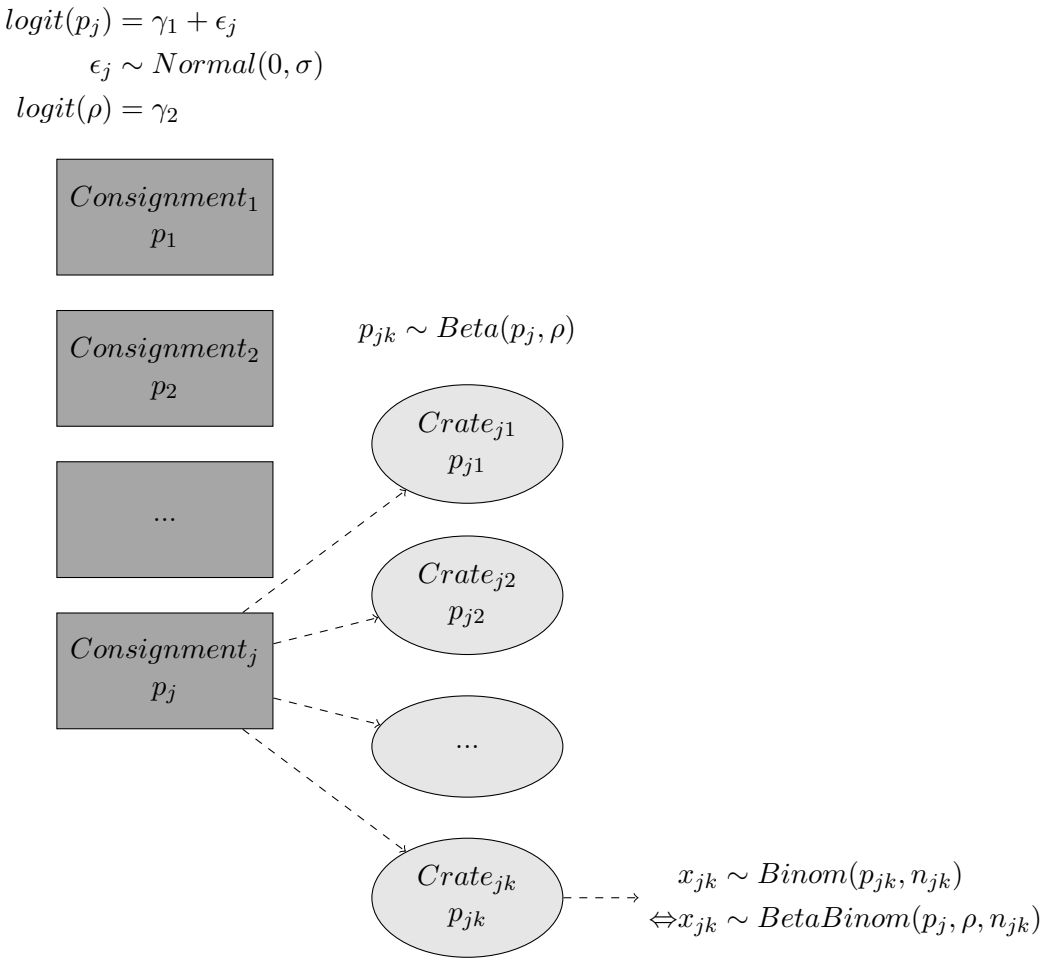

Figure 3.5: Conceptual diagram of the hierarchical Beta-Binomial model. We have a pathway made up of several consignments and each consignment has several crates. Each consignment has its own infestation rate p_j, sampled from the pathway distribution. Within each consignment, each crate also has its own infestation rate p_{jk} sampled from a Beta distribution with mean p_j and intra-cluster correlation coefficient ρ. Note that ρ is assumed to be the same for all consignments of the pathway.

a consignment to be the same among all consignments. The distribution of infected units x_{jk} in crate k in consignment j thus follows a Beta-Binomial distribution with mean p_j, ICC ρ, and number of sample per crate n_{jk}. The model follows:

$$\begin{aligned} x_{jk} &= BetaBinom(p_j, \rho, n_{jk}) \\ logit(p_j) &= \gamma_1 + \epsilon_j \\ \epsilon_j &\sim Normal(0, \sigma) \\ logit(\rho) &= \gamma_2 \end{aligned} \tag{3.10}$$

We simulated data from a potential pathway using this model, and then fitted the model to the simulated data to see if we were able to recover the parameters. As there is no off-the-shelf solution for estimating heterogeneity among crates within consignments (ρ) when there is heterogeneity among consignments, we implemented the model in the Stan language. We did the simulations for 1, 10, 20, 30, ..., and 100 consignments, each

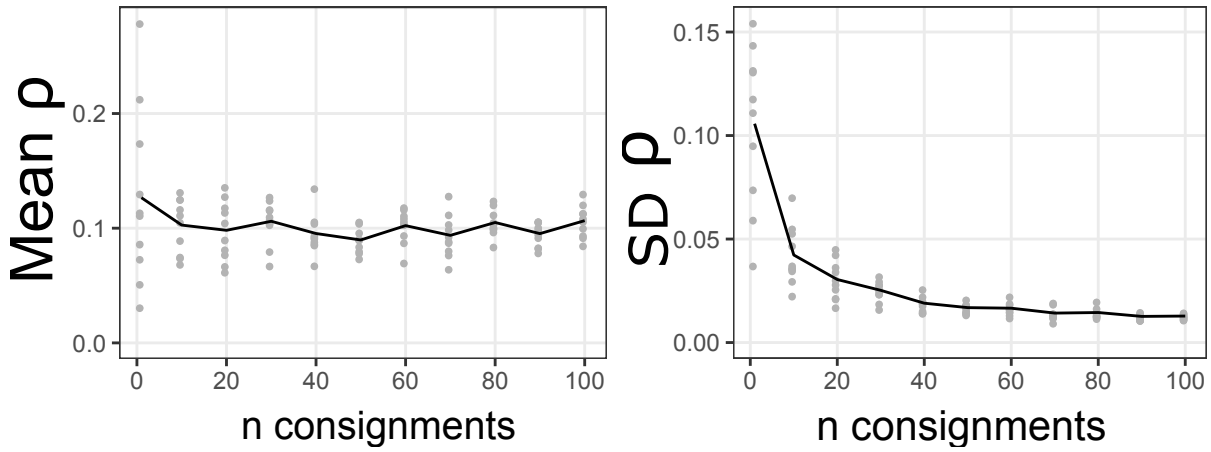

Figure 3.6: Mean and standard deviation of ρ from the hierarchical beta-binomial model estimated from potential plant pathways of different size. We used Eq. 3.10 to simulate a typical pathway for the plant data and used parameters logit(p)=-5.15, ρ=0.1, and σ=1.7. We then estimate ρ by fitting Eq. 3.10 to the simulated pathway using Stan.

with 20 crates and 30 units sampled per crate (*i.e.*, 600 samples per consignment). The pathway we simulated was similar to a typical import plant pathway to Australia. The distribution of infestation rates among different consignments in a typical plant pathway follows a $Beta(.17, 8)$ that we approximated using a normal distribution with mean of -5.15 and standard deviation σ=1.7 on the logit scale[11]. We replicated the simulation 10 times for each pathway size.

With only one consignment of a typical size (20 crates, each with 30 units sampled per crate), it seems difficult to reliably estimate ρ values: there is a high variability in the mean estimate of ρ across the 10 simulations and the standard deviation associated with each replicate is particularly high (the coefficient of variation mean/sd is close to 100%) (Fig. 3.6). However, aggregating data from several consignments seems to help. Pathways with 30–40 consignments or more are able to give relatively precise estimates of ρ (Fig. 3.6). Given these results, we suggest inspecting at least 30 consignments for estimating ρ from a pathway.

For smaller pathways, it might even be possible to go one step further and build a hierarchical model across pathway, with each pathway having its own p, ρ and σ (thus we would get a distribution of ρ for different pathways).

Case study with a risk cutoff of 0.01% We simulated data from a potential pathway using Eq. 3.10, and then fitted the model to the simulated data to see if we were able to recover the parameters. We did the simulations for 1, 10, 20, 30, ..., and 100 consignments, each with 30 crates and 1000 units sampled per crate (*i.e.*, 30,000 samples per consignment). The distribution of infestation rate among different consignments in a typical pathway imported in New Zealand follows a $Beta(.253, 9623)$ that we approximated using a normal distribution with mean of -11.4 and standard deviation σ=1.3 on the logit

[11] Both $Beta(.17, 8)$ and $\frac{1}{1+exp\left(-N(-5.15,\ 1.7)\right)}$ distributions have a similar mean and standard deviation on the probability scale, however, their skewness is different so they do represent slightly different models

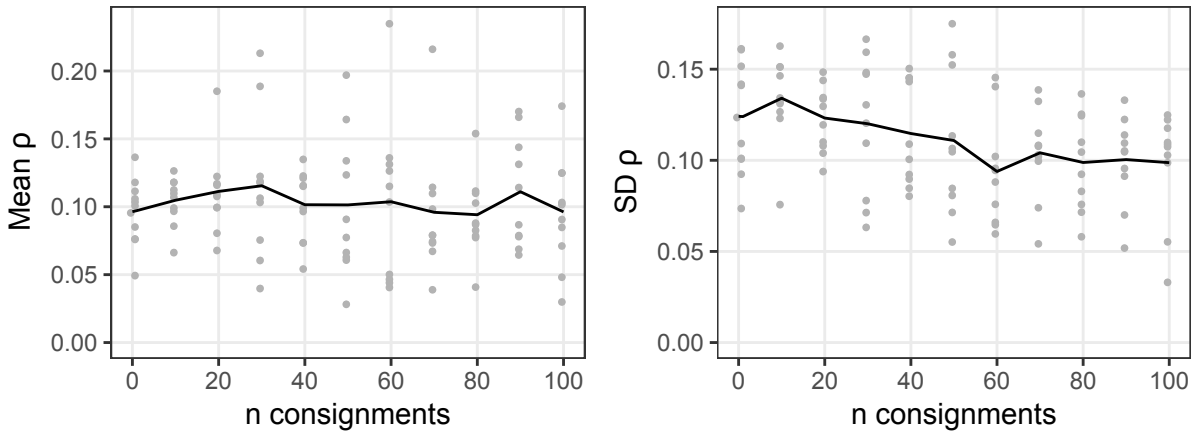

Figure 3.7: Mean and standard deviation of ρ from the hierarchical beta-binomial model estimated from potential pathways of different size. We used Eq. 3.10 to simulate a typical pathway for the plant data and used parameters logit(p)=-11.4, ρ=0.1, and σ=1.3. We then estimate ρ by fitting Eq. 3.10 to the simulated pathway using Stan.

scale[12]. We replicated the simulation 10 times for each pathway size.

While the mean estimate of ρ seems well identified (Fig. 3.7), the posterior does not actually do much better than the prior that we used for fitting the model (a normal distribution of -3 with standard deviation of 1.5 in the logit scale, which has a mean of 0.097 and standard deviation of 0.12). This means that we did not learn much from the data as can be seen by the particularly large standard deviation in the posterior distribution of ρ, even for large simulated pathways (Fig. 3.7b.). To see how much more data might be needed to reliably estimate ρ from a pathway, we also simulated a pathway with 1000 consignments. However, the uncertainties in the posterior distribution of ρ were still fairly large (SD=0.036, 95% credible interval of ρ = 0.038–0.173). We suspect that the infestation rate of the pathway is too low and that too few crates within consignments have BRMs to be able to learn about ρ.

3.2.3 Bayesian inference for clustered sampling

Bayesian inference can be used in two different ways when doing clustered sampling: The first approach is only partly Bayesian and consists of using Bayesian inference to estimate the parameter *rho* of the pathway. We then proceed similarly to model-based inference and compute the sample size by fixing the sensitivity of the test and a given threshold cutoff using Eq. 3.9. Indeed, as we could not fit the hierarchical model using classical models in the model-based inference for clustered data section (section 3.2.2.2), we relied on Bayesian estimation to estimate ρ of the pathway (Bayesian inference is especially well suited to fitting hierarchical models).

The second approach is the equivalent to what we did in section 3.1.3 but for clustered sampling: using some prior distribution for the pathway (heterogeneity at the consignment level and heterogeneity at the crate within consignment level), we can compute the number

[12]Both $Beta(0.253, 9623)$ and $\frac{1}{1+exp\left(-N(-11.4,\ 1.3)\right)}$ distributions have a similar mean and standard deviation on the probability scale, however, their skewness is different so they do represent slightly different models

of units free of BRM that needs to be sampled from a consignment in order to be 95% sure that the infestation rate of the consignment is below a certain risk-cutoff. However, unlike in the simple random sampling case, there is no closed form solution for Bayesian clustered sampling. Still, we can compute the posterior distribution of p_j after a clean inspection by implementing the model in a Bayesian probabilistic programming language (*e.g.*, in our case, Stan).

Case study with a risk cutoff of 0.5%

Using an noninformative uniform prior $p_j \sim Beta(1, 1)$ to describe the distribution of infestation rate among different consignments of the pathway and fixing ρ=0.1 as prior for the distribution of infestation rate among different crates within a consignment (*e.g.*, ρ might have been estimated from past data on the pathway as in section 3.2.2.2), we need to inspect $\sim$ 45 crates, each with 30 sampled units (*i.e.*, 1350 units) to have 95% of the posterior distribution of infestation rate p_j of the inspected consignment below 0.5% (Fig. 3.8a.). While the inference framework is different, the answer is essentially the same as for model-based inference for clustered sampling with $\rho = 0.1$ and $n_k = 30$ (Eq. 3.9, n=1321 units). As for model-based inference for clustered sampling (Fig. 3.3), sample size will decrease for lower ρ and increase for higher ρ.

If we replace the uniform prior by an informative prior $p_j \sim Beta(.17, 8)$, we now only need to inspect $\sim$ 15 crates, each with 30 sampled units (*i.e.*, 450 units free of BRM) to have 95% of the posterior distribution of infestation rate p_j of the inspected consignment below 0.5% (Fig. 3.8b.). Using Bayesian inference with an informative prior on clustered data allows reducing the sample size by a factor of three (450 units) compared to using model-based inference or using Bayesian inference with noninformative prior (1350 units). This gain is similar to the gain between informative and noninformative prior in the simple random sampling setting (183 vs. 600 units, see section 3.1.3). Similar to Bayesian inference for simple random sampling, the reduction in sample size is due to combining inspection data for a specific consignment with external information on the pathway. However, the same safeguards also apply and we should be cautious of the stationarity assumption implied by using prior information estimated from past data on the pathway. For Bayesian with informative priors, clustered sampling (with $\rho = 0.1$) required sampling $\sim$ 450 units vs. 183 units for SRS (ratio of $\sim$2.46).

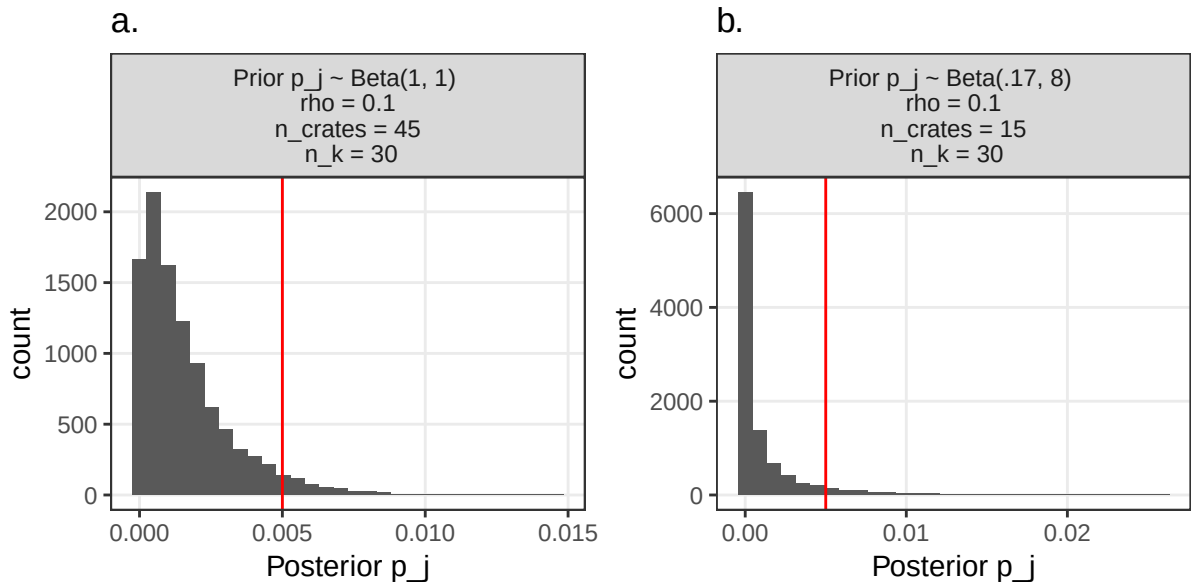

Figure 3.8: Posterior distribution of p_j after a clean inspection sample with a 0.5% risk cutoff. a. We used an noninformative prior on the distribution of infestation among consignments in the pathway and a pathway ρ of 0.1 (ρ can be estimated from past data on the pathway, see section 3.2.2.2). After inspecting 45 crates, each with 30 samples (1350 samples free of BRM), 95% of the posterior mass distribution of the infestation rate of the inspected consignment is below 0.5%. b. We used an informative prior on the distribution of infestation among consignments in the pathway ($p_j \sim Beta(.17, 8)$) and a pathway ρ of 0.1. After inspecting 15 crates, each with 30 samples (450 samples free of BRM), 95% of the posterior mass distribution of the infestation rate of the inspected consignment is below 0.5%. The models were fitted using Stan.

Case study with a risk cutoff of 0.01%

Using a noninformative uniform prior $p_j \sim Beta(1, 1)$ to describe the distribution of infestation rate among different consignments and fixing ρ=0.1 as prior for the distribution of infestation rate among different crates within a consignment, we need to inspect $\sim$ 700 crates, each with 1000 sampled units (*i.e.*, 700,000 units free of BRM) to have 95% of the posterior distribution of infestation rate p_j of the inspected consignment below 0.01% (Fig. 3.9a.). If we replace the uniform prior by an informative prior $p_j \sim Beta(.253, 9623)$, we now only need to inspect $\sim$ 60 crates, each with 1000 sampled units (*i.e.*, 60,000 units free of BRM) to have 95% of the posterior distribution of infestation rate p_j of the inspected consignment below 0.01% (Fig. 3.9b.). Using Bayesian inference with an informative prior on clustered data allows reducing the sample size by a factor of 11 (60,000 units) compared to using model-based inference or using Bayesian inference with noninformative prior (700,000 units). This gain is similar to the gain between informative and noninformative prior in the simple random sampling setting (2,580 vs. 29,955 units, see section 3.1.3). Similar to Bayesian inference for simple random sampling, the reduction in sample size is due to combining inspection data for a specific consignment with external information on the pathway. However, the same safeguards also apply and we should be cautious of the stationarity assumption implied by using prior information estimated from past data on the pathway.

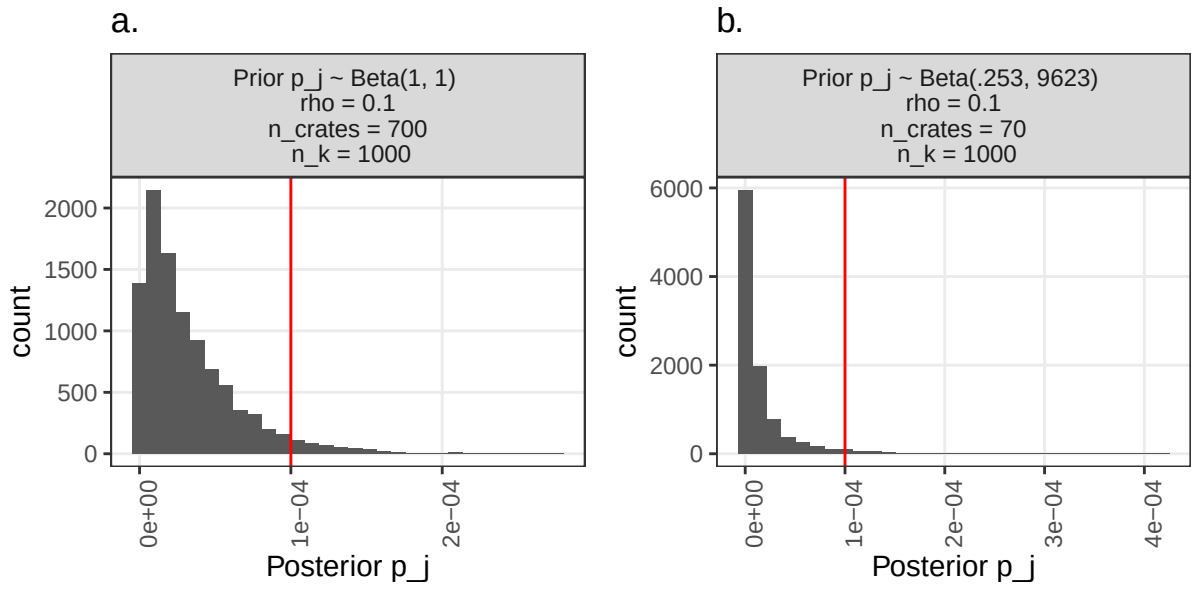

Figure 3.9: Posterior distribution of p_j after a clean inspection sample with a 0.01% risk cutoff. a. We used an noninformative prior on the distribution of infestation among consignments in the pathway and a pathway ρ of 0.1 (ρ can be estimated from past data on the pathway, see section 3.2.2.2). After inspecting 700 crates, each with 1000 samples (700,000 samples free of BRM), 95% of the posterior mass distribution of the infestation rate of the inspected consignment is below 0.01%. b. We used an informative prior on the distribution of infestation among consignments in the pathway ($p_j \sim Beta(.253, 9623)$) and a pathway ρ of 0.1. After inspecting 70 crates, each with 1000 samples (70,000 samples free of BRM), 96% of the posterior mass distribution of the infestation rate of the inspected consignment is below 0.01%. The models were fitted using Stan.

3.2.4 Dempster-Shafer theory of evidence for clustered sampling

We found no articles in the literature on the application of Dempster-Shafer theory of evidence when the data is clustered. While it does not mean that the Dempster-Shafer theory cannot deal with clustered data, more research might be needed before we are able to apply the framework to this situation.

3.2.5 Imprecise probabilities for clustered sampling

To our knowledge, the theory of imprecise probability has not been developed when clustered data. Furthermore, imprecise probability theory does not allow using external information when doing inference.

3.3 Systems approach

Systems approach is an integrated method for addressing biosecurity threats that combines two or more independent risk-reducing measures on a pathway (IPPC, 2002). The independence of the risk reduction measures comes from using different actions (*e.g.*, pest monitoring, insecticide spraying, cold storage, heat treatment, intermediate inspections...)

that are applied at different production stages (pre-harvest, harvest, post-harvest...) (see vanKlinken et al., n.d., for a review). Since we typically expect pathways following a system approach to have lower infestation rates than similar pathways not following a system approach, this might allow reducing the sample size of a border inspection.

Note that a four-year national Australian project started in 2018 and led by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) currently reviews the use of systems approach in biosecurity and develops new methods for quantifying the risk reduction associated with using systems approach (vanKlinken et al., n.d.). Thus, in our report, we will not try to be comprehensive with regards to how systems approach data can be analyzed (especially since the effect of many of the actions *e.g.*, insecticide spraying or cold storage, might be pathway or pest specific). Rather, we will focus on how different inference frameworks (design, model-based, Bayesian inference, and Dempster-Shafer theory of evidence) might deal with this type of data. Specifically, we will focus on one risk-reduction action: the presence of intermediate inspections in a pathway (one advantage of focusing on intermediate inspections is that the effect is not pathway specific) and for which we have analytical formula.

Characterizing the effect of other types of systems approach interventions (cold storage, bait trapping, insecticide spraying...), might require additional data, *e.g.*, control-treatment experiments or expert elicitation (Jarrad et al., 2011; Hemming et al., 2018) to quantify how they would reduce the infestation rate of a consignment. These information might in turn be used differently in different inference framework (*e.g.*, external information cannot be used in Design-based inference, it can be used as a prior in Bayesian inference, and it be used in combination similarly to any other source of information in Dempster-Shafer theory of evidence).

3.3.1 Design-based inference

The simplest way to combine results from an intermediate inspection and a final border inspection is to assume (or ensure) that the units sampled in the intermediate and in the final border inspection are independent (*i.e.*, assume that different units are sampled in both inspections, which is not unreasonable given that each inspection typically sample only a small fraction of the population and which is assured when sampling is destructive). In such a case, the number of BRMs resulting from two independent random samples of size n_1 and n_2 sampled from the same population give a similar result to a single inspection of size $n_1 + n_2$ (the sensitivity of an inspection with a sample size of $n_1 + n_2$ is given in Eq. 3.1).

Hence, if we inspect $n_1 = 200$ units in the intermediate inspection, the final border inspection only needs to sample $n_2 = 397$ units to have a 95% chance of detecting a prevalence of 0.5%.

While in theory it is possible to relax the independence assumption by using partial resampling estimators (see Ware and Cunia, 1962), it leads to additional issues in practice. For example, we would need to know what is the proportion of units that are sampled in both inspections. Additionally, part of the improved efficiency that comes from using partial resampling inspections instead of independent inspections, comes from being able to estimate the covariance between two successive inspections on the same sampling units. However, since we will only get clean inspection samples with zero BRMs, it is not possible to estimate this covariance (the issue is similar to the one encountered when using design-based inference for clustered sampling, where the design-based estimator was not

able to estimate the intra-cluster correlation coefficient from clean inspection data, see section 3.2.2.2).

One obvious issue with this method (which is also common to other inference frameworks) is that if the intermediate inspection sampled 600 units, then we would not need to sample any units in the final border inspection to declare the consignment compliant. This is obviously too optimistic as it ignores potential issues with the reliability of the first inspection (was the intermediate inspection done properly?) and also the potential for re-infestation during transport. To circumvent this issue, we might chose to lower the effective sample size of the intermediate inspection (*e.g.*, for an intermediate inspection of 600 units, we might consider that it corresponds to 300 units inspected), we might chose to always carry a small sample size inspection for the final border inspection (say we always sample 100 units), or it might be necessary to run audits or specific analyses to see if the systems approach was implemented properly.

3.3.2 Model-based inference

The reasoning that was developed for design-based inference can also be developed for model-based inference: If an intermediate inspection of size n_1 with a population prevalence of p_j is inspected, the number k_1 of BRMs will follow the Binomial distribution $k_1 \sim Binom(n_1, p_j)$. Similarly, if a border inspection of size n with a population prevalence p_j, the number k_2 of BRMs will follow the Binomial distribution $k_2 \sim Binom(n_2, p_j)$. The number k of BRMs when combining both inspection will follow the binomial distribution $k \sim Binom(n_1+n_2, p_j)$, which will have a 95% chances of detecting a 0.5% infestation rate when $n_1 + n_2 = 598$.

3.3.3 Bayesian inference

In a Bayesian inference framework, including the effect of an intermediate inspection in the pathway is straightforward: If the infestation rate in a consignment before inspection is described by $p_j \sim Beta(\alpha, \beta)$, the posterior distribution after finding zero BRM out of n inspected samples is $p_j \sim Beta(\alpha, \beta + n)$. For example, if we start with a $Beta(1, 1)$ prior (any infestation rate in the 0–1 range is equally likely a priori) and conduct a clean 200 sampled units intermediate inspection, then the posterior distribution is $p_j \sim Beta(1, 201)$. The $Beta(1, 201)$ distribution can in turn be used as prior for the final border inspection. In such a case, the final inspection will only need to sample 397 units as the posterior $p_j \sim Beta(1, 201 + 397)$ has 95% of its posterior mass below 0.5%. Thus the answer using a non-informative prior is essentially the same as for design and model-based inference.

Similar to what we wrote in section 3.1.3, we might need to penalize the prior information to account for the fact that the intermediate inspection might not be 100% reliable or for potential re-infestations of the pathway. Alternatively, we might try creating a more mechanistic model of how this re-infestation might happen at different production stages of the pathway (which seems to be the approach taken in CSIRO's project, vanKlinken et al., n.d.).

3.3.4 Dempster-Shafer theory of evidence

As we are combining evidence from different inspections (*i.e.*, different sources of information), a similar methodology to what was presented in section 3.1.4 can be applied

here: For example, after a first clean 200 samples inspection, we are 63.49% sure that the infestation rate is below 0.5% (in R, $pbeta(0.005, 1, 1 + 200) \sim 0.64$, focal element $\{C\}$ associated with a compliant consignment) and 36.51% confident of not knowing the correct status of the consignment (focal element $\{C, NC\}$). The focal element $\{NC\}$ for non-compliance gets a probability of zero as we did not find any infested samples. After a second clean 397 samples inspection, we are 86.4% sure that the infestation rate is below 0.5% (in R, $pbeta(0.005, 1, 1+397) \sim 0.86$) and 13.6% confident of not knowing the correct status of the consignment. Combining both sources of information (*i.e.*, a combined sample size of 597) with the Dempster's combination rule and using 100% reliability for both inspections (using the 'dComb' function from the 'EvCombR' package in R), we would be 95% sure that the infestation rate of the consignment is below 0.5%, which provides us with enough assurance to mark the consignment as compliant.

Additionally, if we think that the results from the first inspection are not 100% reliable (whether because the inspection was not done properly or because there is a chance of re-infestation following the first inspection), it is possible to reduce the reliability of the first inspection before combining both sources of information. For example, if we fix the reliability of the first inspection to be 70% (and keep the reliability of the second inspection to be 100%), we will need to sample 480 samples (instead of the original 397) in the second inspection be 95% sure that the consignment is compliant after combining both sources of data using Dempster's combination rule.

4 Adaptive Inspection Schemes

4.1 Introduction

Although they do not formally fall within the coverage of this project, adaptive inspection schemes provide plans that can be used to choose to inspect or not inspect consignments based on the recent inspection history of the pathway. In a sense such plans use recent history to provide assurance about the pathway, and therefore also the next consignment. We therefore cover them here briefly for completeness.

Such schemes have been in use for some time in both Australia (the Department of Agriculture's Compliance-Based Inspection Scheme, CBIS[1]) and New Zealand (risk-based inspection of low-risk fresh produce such as green beans imported from Australia).

We briefly review several schemes gathered under two sampling principles, namely: (i) do not inspect all consignments, however when we inspect we do so with constant within-consignment intensity; and (ii) inspect all consignments but alter the intensity.

4.2 Inspect Only Some Consignments

Adaptive sampling plans involve electing to not inspect some consignments. Consequently, they should only be used on pathways in which some leakage can be tolerated (but see Section 5). The Department of Agriculture has applied such plans in pathways that show low contamination rates, such as dried apricots and hulled sesame seeds.

The principle value of these plans is that they will naturally tend to allocate more inspection effort to domains of the pathway that have the higher interception rate, without requiring predictive statistical modeling or machine learning. The price to pay is that the plans will always experience a delay in response to changes in the pathway contamination rate, whereas predictive modeling allows for the possibility of anticipation.

The original variants of the adaptive sampling theme were the *continuous sampling plans* developed in the 1940's and thereafter by mathematicians to make quality control more efficient.

Very readable accounts can be found in Stephens (2001) and Montgomery (2009).

4.2.1 Continuous Sampling Plans

Numerous continuous sampling plans have been developed in the past 70 years, for example Stephens (2001) lists nine variants.

4.2.1.1 CSP-1

The original continuous sampling plan, CSP-1, was proposed by Dodge (1943) as a way of managing a continuous flow of consignments offered to the inspector. Briefly, the

[1] http://agriculture.gov.au/import/goods/plant-products/risk-return

algorithm works as follows.

Rule 1 Inspect 100% of the consignments, and when c consecutive consignments are found clear of defects, switch to Rule 2.

Rule 2 Inspect a random sample of f% of the consignments, and when any such inspection identifies a defect, switch to Rule 1.

The operation of the plan is then set by choosing the variables c and f. Dodge (1943) developed a number of performance indicators that could be used to guide the choice of these parameters, namely the average fraction of the total units inspected in the long run (AFI), the expected fraction of defective units in the long run (AOQ), and its mode, called the AOQL, which can be interpreted as the highest AOQ for all possible values of the underlying contamination rate.

One argument against the use of CSP-1 is that it does not distinguish between temporary and permanent increases in the failure rate. That is, a single interception is sufficient to flip the pathway from fractional inspection to 100% inspection. This suggests that the plan does not tolerate isolated instances of contamination, which contradicts the principle that they should only be applied to pathways in which some leakage can be tolerated.

4.2.1.2 CSP-2

Dodge and Torrey (1951) introduced two additional CSP plans, namely CSP-2 and CSP-3. The difference between CSP-2 and CSP-1 is that the plan does not return to 100% inspection when an isolated incident of contamination is found.

Rule 1 Inspect 100% of the consignments, and when c consecutive consignments are found clear of defects, switch to Rule 2.

Rule 2 Inspect a random sample of f% of the consignments, and when any such inspection identifies a defect, switch to Rule 3.

Rule 3 Inspect a random sample of f% of the consignments, until k units have been inspected. If any are contaminated, apply Rule 1. If not, then apply Rule 2.

The advantage of CSP-2 relative to CSP-1 is that it allows occasional and isolated contamination to pass through without triggering a return to 100% inspection. Dodge and Torrey, 1951 showed that CSP-2 is generally more economical than CSP-1 when the arriving contamination rate p is smaller than AOQL.

4.2.1.3 CSP-3

As noted above, Dodge and Torrey (1951) also introduced CSP-3. This plan follows CSP-2 in that it allows for isolated incidents of contamination, but in a sense it takes out some insurance by briefly increasing the inspection rate after detecting contamination. The rule is as follows.

Rule 1 Inspect 100% of the consignments, and when c consecutive consignments are found clear of defects, switch to Rule 2.

Rule 2 Inspect a random sample of f% of the consignments, and when any such inspection identifies a defect, switch to Rule 3.

Rule 3 Inspect the next b (usually, 4) consignments. If any are contaminated, then apply Rule 1. If not, then apply Rule 4.

Rule 4 Inspect a random sample of $f\%$ of the consignments, until $k - b$ consignments have been inspected. If any are contaminated, then apply Rule 1. If not, then apply Rule 2.

CSP-3 is used by the Department of Agriculture for the Compliance-Based Inspection Scheme[2]. However, CSP-1 was preferred for a recent CEBRA project because it is simpler and easier to communicate to stakeholders, who are thus likely to develop a clearer understanding of the incentive properties of the inspection rule (Susie Hester *pers. comm.*, also see Rossiter et al., 2018). Furthermore, Rossiter and Hester (2017) suggested that the CSP-1 algorithm would be preferable from the biosecurity regulator's perspective, particularly where the consequences of biosecurity risk material leakage are perceived to be relatively large. Under the scenarios modelled, CSP-1 and CSP-3 had higher payoffs to the regulator than mandatory inspection, and of those two, CSP-1 had the highest payoff to the regulator.

4.2.2 Skip-Lot Sampling Plans

Skip-Lot Sampling Inspection is an extension of the Continuous Sampling Plan that supports sampling and inspection of sub-consignment items, such as the inspection of articles of fruit within a consignment. Montgomery (2009) noted:

> "[...] one should be careful to use skip-lot sampling plans only for situations in which there is a sufficient history of supplier quality to ensure that the quality of submitted lots is very good. Furthermore, if the supplier's process is highly erratic and there is a great deal of variability from lot to lot, skip-lot sampling plans are inappropriate. They seem to work best when the supplier's processes are in a state of statistical control and when the process capability is adequate to ensure virtually defect-free production."

Dodge (1955) introduced Skip-Lot Sampling by applying the principles of CSP-1 to a series of consignments defined as batches of material, in which the user examines only a single unit from each consignment of units. This sampling plan is useful when the consignments are small or inspection is slow and costly.

Perry (1973) extended skip-lot sampling to SkSP-2, which operates as follows. First, instead of examining only a single unit from each selected consignment, we nominate a *reference plan*, that is, we select a sample of units from the consignment according to a selected sample design — for example, simple random sampling: selecting and inspecting a simple random sample of n units from each consignment and counting the number of defective units d; if $d = 0$ then accept the consignment, otherwise reject the consignment.

Rule 1 Inspect every consignment using the reference plan until c consecutive consignments have been accepted. Then use Rule 2.

Rule 2 Apply the reference plan to a fraction f of the consignments submitted until any consignment is rejected, at which point return to Rule 1.

[2] http://agriculture.gov.au/import/goods/plant-products/risk-return

Confusingly enough, both SkSP and SkSP-2 are extensions of CSP-1. There are numerous variations on skip-lot sampling, for example Vijayaraghavan (2000)'s SkSP-3 is a variant of CSP-2.

4.3 Inspect All Consignments; Vary Intensity

The second type of plan can be thought of as a generalization of the first. In the first type, consignments were inspected or not inspected based on the recent performance of the pathway. In the second type, all consignments are inspected, but the intensity of inspection varies according to recent history. We cover one such plan, namely the Lot-by-Lot Attributes Single Sampling Procedure prescribed by MIL-STD-1916 (Defense, 1996; Stephens, 2001).

4.3.1 MIL-STD-1916

The Lot-by-Lot Attributes Single Sampling Procedure prescribed by MIL-STD-1916 involves inspecting every consignment but with intensity that depends on recent history. There are seven levels of verification, namely I – VII. The user starts by selecting a verification level.

According to the standard, the verification level should depend on the nature of the characteristic that is the target of the inspection: critical, major, or minor. For example a *critical* characteristic is one that "...judgment and experience indicate must be met to avoid hazardous or unsafe conditions for individuals using, maintaining, or depending upon the product; or that judgment and experience indicate must be met to assure performance of the tactical function of a major item such as a ship, aircraft, tank, missile, or space vehicle." A *major* characteristic is "A characteristic, other than critical, that must be met to avoid failure or material reduction of usability of the unit of product for intended purpose.", and a *minor* characteristic is "A characteristic, other than critical or major, whose departure from its specification requirement is not likely to reduce materially the usability of the unit of product for its intended purpose or whose departure from established standards has little bearing on the effective use or operation of the unit." (Defense, 1996).

The supporting documentation goes on to say "For Critical Characteristics - VL VII should always be used. This inspection is a verification of the automated screening or fail-safe manufacturing operation implemented in accordance with paragraph 4.4 of MIL-STD-1916. Majors should typically use VL levels between III and VI. Minors should typically use VL levels between I and III. The more important the characteristic is, the higher the VL. Lower VL's may also be considered where relatively small sample sizes are necessary and large sampling risks can or must be tolerated as, for example, when inspection costs are high. If no VL is specified, then VL IV for majors and VL II for minors should be used." (Defense, 1999). From our point of view, biosecurity risks can most likely be characterised as *critical*; consequently for biosecurity inspection the invocation of MIL-STD-1916 would be at verification level VII.

Each level of verification corresponds to a sample size for any given consignment size. The sample size classes corresponding to the consignment size and the verification level are presented in Table 4.1.

The rule starts at a level elected by the regulator. Each level corresponds with three sampling plans, namely *reduced*, *normal*, and *tightened*. Each sampling plan is charac-

Table 4.1: Code letters for entry into the sampling tables for MIL-STD-1916 (from Defense, 1996).

Lot Size	VII	VI	V	IV	III	II	I
2–170	A	A	A	A	A	A	A
171–288	A	A	A	A	A	A	B
289–544	A	A	A	A	A	B	C
454–960	A	A	A	A	B	C	D
961–1632	A	A	A	B	C	D	E
1633–3072	A	A	B	C	D	E	E
3073–5440	A	B	C	D	E	E	E
5441–9216	B	C	D	E	E	E	E
9217–17408	C	D	E	E	E	E	E
17409–30720	D	E	E	E	E	E	E
30721 and above	E	E	E	E	E	E	E

terised by a sample size. In Table 4.2, the tightened/reduced plan can be determined as the verification level to the left/right of the specified normal verification level respectively. Tightened inspection of VL-VII is T, and reduced inspection of VL-I is R. For a pathway comprising large consignments such as lots of 32,000 seeds for sowing, sampling against a critical characteristic (requiring VL VII) corresponds to code E. In this pathway, Normal sampling is 3072 seeds, Reduced is 1280, and Tightened is 8192.

Table 4.2: Attributes sampling plans for MIL-STD-1916 (from Defense, 1996). When the lot size is less than or equal to the sample size, 100 percent attributes inspection is required.

Code	Tight	VII	VI	V	IV	III	II	I	Reduced
A	3072	1280	512	192	80	32	12	5	3
B	4096	1536	640	256	96	40	16	6	3
C	5120	2048	768	320	128	48	20	8	3
D	6144	2560	1024	384	160	64	24	10	4
E	8192	3072	1280	512	192	80	32	12	5

The switching rules between the three plans are then applied as follows (Defense, 1996):

Normal to Tightened When normal inspection is in effect, tightened inspection shall be instituted when 2 consignments have been withheld from acceptance within the last 5 or fewer consignments.

Tightened to Normal When tightened inspection is in effect, normal inspection may be instituted when the following conditions are both satisfied:

1. The cause for producing the nonconformances is corrected.
2. 5 consecutive consignments are accepted.

Normal to Reduced When normal inspection is in effect, reduced inspection may be instituted when the following conditions are all satisfied:

1. 10 consecutive consignments are accepted while on normal inspection.
2. Production is at a steady rate.
3. The contractor's quality system is considered satisfactory by the Government.
4. Reduced inspection is considered desirable by the Government.

Reduced to Normal When reduced inspection is in effect, normal inspection shall be instituted when one of the following conditions occur.

1. A consignment is withheld from acceptance.
2. Production becomes irregular or delayed.
3. The contractor's quality system is unsatisfactory.
4. Other conditions warrant that normal inspection be re-instituted.

Failing under Tightened Finally, If sampling inspection of lots or batches remains in tightened inspection due to discovery of nonconformances, the Government reserves the right to discontinue acceptance of the product until the causes of nonconformances are eliminated or other means acceptable to the procuring agency have been instituted. When sampling inspection is restarted after discontinuation of acceptance, it shall be at the tightened inspection stage.

The acceptance number (that is, acceptable number of contaminated items per consignment) for this regime is 0, which is somewhat controversial. Stephens (2001) argues that this prescription gives the decision rule unattractive properties.

It is worth noting that the documentation of MIL-STD-1916 provides a number of other useful prescriptions, for example detailed requirements and specifications to which providers must adhere (Defense, 1996).

4.4 Choosing operational parameters

This chapter provides an overview of several approaches to using recent inspection history to assess the pathway level of assurance. Two types of approaches were covered: (i) not inspecting all consignments, but those that are inspected are inspected using the same sample design (the reference plan); and (ii) inspecting all consignments but varying the intensity of the inspection. The latter approach is arguably a generalization of the first if it includes the possibility that the within-consignment sampling intensity could be 0.

No matter which approach is used, it is necessary to select operational parameters that dictate how the system should be applied. All approaches are provided with documented performance indicators that provide insight into the performance of the inspection regime based on assumptions about the inspection efficacy and the contamination rate of the pathway.

Having chosen an approach, it is then necessary to select one or more of these performance indicators that then guide selection of combinations of operational parameters. For example, usage of the CSP family of plans can be guided by the AOQ or the AOQL (see Section 4.2.1). This means that combinations of the clearance number c and monitoring

fraction f can be selected by deciding on a value of AOQL and using a look-up table to identify the corresponding values of f and c.

A further complication is that the effects of f and c on the performance indicators interact with one another, so several different combinations of these parameters will achieve the same AOQL. In order to select between these combinations we have to interpret the effect of the parameter choices on the management of pathway biosecurity risk. Each operational parameter can be classified according to its effect on the sampling process and therefore its interpretation in terms of managing biosecurity risk.

Regardless of the system, most operational parameters can be classified into one of three types, namely (i) the number of inspected and compliant consignments that it takes to change the proscribed inspection levels down, (ii) the number of inspected and non-compliant consignments that it takes to change the proscribed inspection levels up, and (iii) the amount of effort (number of consignments to be inspected or number of units per consignment) at any given level.

For example, in CSP-3, there are four parameters, namely: the monitoring fraction f, the clearance number c, the number of inspected consignments within which two detected contaminations results in switching to 100% inspection k, and the number of consignments that must be inspected after the detection of any contamination, b. Each has at least one direct operational interpretation, as follows.

c the clearance number sets the number of consecutive compliant consignments that it takes to convince us that the pathway can be monitored instead of fully inspected. This can be interpreted as a required level of assurance (We need to see 20 compliant consignments in a row before we're convinced) or an incentive (the stringent requirement becomes an incentive for stakeholders to ensure that the pathway achieves and maintains a high level of compliance), or both.

f the monitoring fraction sets the amount of effort that is imposed on the pathway when it is in sampling mode. It can be interpreted as a required level of assurance (We need to see 25% of the consignments as a warranty that the pathway contamination rate and risk haven't changed) or an incentive (the reduction in sampling effort translates to a reward for high biosecurity compliance by the pathway stakeholders), or both.

k the minimum spacing of two detected contaminated consignments provides a sensitivity to what constitutes a change in pathway status as opposed to an isolated incident. Having seen one contaminated consignment, we are effectively on alert. How many clear consignments do we need to see before we can relax again?

b the duration of temporary 100% inspection allows us to set a preference for how long we scrutinize the pathway after detecting non-compliance. It provides a greater sensitivity to discerning between isolated incidents and important changes in pathway status.

5 Discussion

The previous chapters include a considerable amount of discussion, but a few topics need further exploration.

It is important to recognize that although ideally the effect of border intervention would be to totally prohibit entry by contaminated items, the reality is that all border intervention can do is to intercept a portion of the contamination, hence reducing the arrival rate. This would be true even if every consignment were inspected and every unit examined within every consignment; some infections are sub-clinical, some pathways simply cannot be regulated, such as wind and tide, and so on.

Indeed, the most stringent border intervention operation known to the authors is that imposed by Chevron Corporation for entry into Barrow Island, a Class A environmental heritage area off the coast of Western Australia. Chevron has close to total control over the pathway, wrapping equipment before shipping, and undertaking careful inspection of all passengers. Nonetheless, the careful post-border surveillance exercise ongoing at that location shows that leakage still occurs. It is impossible to imagine that a federal regulator would have the resources, or indeed the imprimatur, to impose so stringent a system. Hence, leakage should be accepted as being inevitable, echoing Beale et al.'s "Zero risk is unattainable and undesirable," (Beale et al., 2008).

Consequently, there is no chance that a border intervention operation will entirely eliminate risk. Instead, the goals of a border intervention operation should be: (i) verify the overall compliance to biosecurity regulation and policy of the pathway, and (ii) where necessary as part of an end-to-end biosecurity regulatory framework, reduce the approach rate sufficiently that it is impossible for pests to establish a minimal viable population (MVP). Note that this prescription is not quite the same as keeping the pathway contamination rate below a given level because contamination is variable in nature. Half an MVP of pest A and half of an MVP of pest B will not result in an establishment.

This view of border intervention has important implications. A key element to the application of risk-based reduced intervention regimes is: because intervention will be reduced, the probability of contamination leakage will necessarily increase — at least on the part of the pathway that experiences the effort reduction. Unfortunately this element of risk-based intervention is enough to deter some regulators, who are squeamish about the political or biological implications of taking any action that increases risk, even knowing that doing so might provide an opportunity to decrease risks elsewhere. As noted above, even the most stringent border intervention program can not and will not guarantee zero leakage. That is: leakage will happen regardless of the border activity, and a risk-based approach may increase the leakage marginally, but at a substantial cost savings.

Another aspect to tolerating leakage that is not often appreciated is the necessity of gathering information about portions of the operation that one does not consider risky. Given a risk profile, it's tempting to assume that (a) the risk profile is known exactly, and (b) the risk profile will never change, and so it is safe and reasonable to devote all the limited resources available to reducing the known risks. However, risks are never known exactly in operational settings, and risks are not necessarily constant, so it is

critically important that the regulator develop some mechanism for assessing the quality and timeliness of the information that underpins the risk profiles. In short, this means that a small amount of inspection needs to be done on all pathways to guard against changes in infestation rates, as a hedge against possible fraud, and just in case the original assessment was wrong.

6 Summary, recommendations, and conclusions

Several inference frameworks can be used to develop assurance about the regulatory compliance of consignments of germplasm. While some frameworks allow using external information when making inference (Bayesian, Dempster-Shafer, and to some extent, model-based inference) others do not (design-based inference, imprecise probability theory) (see table 3.1). Frameworks that do not allow using external information are of limited use for systems approach (analyzing systems approach data requires combining different sources of evidence). Below, we summarize the pro and cons of the five framework reviewed in this report.

Design-based inference

- This is the main type of inference used for border biosecurity inspection.
- In design-based inference, we can draw conclusions about the population from the sample because we know exactly how the sample was collected. No additional assumption is required which makes the method particularly objective.
- When the inspected units comes from a simple random sample, we can use the binomial sample size formula (Eq. 3.2) to compute sample size. This is the basis of the '600 samples' rule often used in biosecurity and also the basis for the 31,540 samples used for the plant product data supplied by New Zealand.
- When the data arrives in clusters but we still manage to do simple random sampling, we can also use Eq. 3.2 to compute sample size (simple random sampling protects against the detrimental effect of clustering on sensitivity and sample size).
- Does not allow using external information.

Model-based inference

- In model-based inference, we postulate a model that might have generated the data (*i.e.*, the inspection data might have been generated from a Binomial model), check the assumptions of the model, and make inference about the infestation rate.
- When the data comes from simple random sampling, model-based inference give the same sensitivity and sample size than design-based inference (Eq. 3.2).
- When there is clustering, we can use Eq. 3.9 to compute sample size. This requires estimating or fixing the intra-cluster correlation coefficient (ICC) of the pathway. The ICC can be estimated using a hierarchical logistic Beta-Binomial model (Eq. 3.10) implemented in probabilistic software. When the risk cutoff is 0.5%, the

ICC can reliably be estimated from a pathway with at least 30 consignments, each with 600 samples (30 crates, each with 20 samples per crate) (Fig. 3.6). However, when the risk cutoff is 0.01% and the infestation rate is very low, even large pathway size (> 100 consignments) do not contain enough information to estimate the ICC (Fig. 3.7).

- Allows limited use of external information (for example, to estimate the ICC of the pathway from past data, section 3.2.2.2).

Bayesian inference

- In Bayesian inference, we postulate the potential values that the parameter of interest might take (prior information before seeing the data) as well as a model that might have generated the inspection data. We then combine the prior and the model with the inspection data to make our inference on the parameter of interest (typically the infestation rate of the consignment being inspected).

- When we use a non-informative uniform prior on the infestation rate of the consignment being inspected, Bayesian inference gives the same sample size as design-based and model-based inference for simple random sampling data (section 3.1.3) and as model-based inference for clustered sampling (section 3.2.3).

- The strength of Bayesian inference however is that it allows combining external information (informative prior) with inspection data (likelihood) to draw conclusion about the infestation rate of a consignment. Using informative prior (for example calibrated from past data on the pathway) allows to reduce sample size in both the simple random sampling and the clustered sampling cases (sections 3.1.3 and 3.2.3). In the case of a potential 0.5% pathway, the sample size can be reduced by a factor of around three compared to design-based inference. In the case of a potential 0.01% pathway, the sample size can be reduced by a factor of around 11.

- One issue that arises when using an informative prior is the assumption of stationarity (past data are representative of future data). We suggest monitoring and re-estimating the distribution of infestation rate among different consignments of the pathway regularly (perhaps every year). We can also use mixture priors to 'robustify' our prior.

- Another issue with Bayesian inference is that we do not always have analytical solutions for our estimates or our decision criteria. In the case of simple random sampling, we have an analytical distribution for the posterior p_j but we have to compute the sample size numerically. In the case of clustered sampling, both the posterior distribution of p_j and the sample size have to be computed numerically (by fitting the hierarchical model to a clean inspection data of different sizes and observing the effect on the posterior).

Dempster-Shafer theory of evidence

- Dempster-Shafer theory of evidence works directly on the decision scale (probability of compliance) rather than the infestation rate of the population. Dempster-Shafer theory is typically used to combine different lines of evidence when making inference.

Each line of evidence can arise from a model and inspection data (*e.g.*, a Binomial model generated the observed inspection data) or can be completely subjective (*e.g.*, experts think that the proportion of compliant consignments in this specific pathway that used a systems approach is 90%).

- With only one source of evidence and in the simple random sampling case, the sample sizes are similar to those given by Bayesian inference with non-informative prior.
- The framework might be difficult to extend to support clustered sampling.
- The Dempster-Shafer framework allows combining external information when making inference. There are several ways to do so and perhaps not much to decide between them (see for example Rathman et al., 2018).

Imprecise probability theory

- Imprecise probability theory is a specific type of Bayesian analysis that was created to avoid having to fix a specific non-informative prior when we are ignorant about the value of the parameter of interest.
- The sample sizes are similar to Bayesian inference with a uniform prior in the case of simple random sampling.
- The framework might be difficult to extend to support clustered sampling.
- Does not allow using external information when making inference (Imprecise probability theory is all about non-informative priors).

Adaptive Inspection Schemes

- Adaptive inspection schemes provide a light-touch approach for implementing risk-based intervention.
- The sample sizes depend on recent inspection history.
- Reasonably easy to implement.
- Does not explicitly allow using external information when making inference but work-arounds are possible.

Conclusion and recommendations.

Of the five frameworks reviewed, Bayesian inference seems to be the most promising to allow incorporating sources of data other than the current inspection sample when making a decision. Bayesian inference is also compatible with current methods used in biosecurity: when using non-informative priors (*i.e.*, representing our ignorance of the infestation rate of the consignment before inspection), Bayesian answers are similar to design and model-based inference (*e.g.*, after a clean '600 samples' inspection with a uniform prior, Bayesian methods infer that there are 95% chances that the infestation rate of the inspected consignment is below 0.5%). If available, Bayesian inference allows using information from external sources of data, which reduces the sample size required

to make a decision on the compliance of a consignment. However, this comes at a cost: if future data are different from past data, we are no longer guaranteed to detect a given prevalence with a given sensitivity (as with the design-based inference procedure). There are different ways to penalize an informative prior. The most promising approach is to use a mixture prior that combines the informative prior with a uniform prior. This approach allows for the possibility that some of the future consignments might have an infestation rate higher than what we have seen in past data.

Data collected from a clustered population can result in noticeably reduced sensitivity for an inspection scheme. Keeping the sensitivity constant (with respect to simple random sample inspection) requires sampling more units. How many more units to sample will depends on the intra-cluster correlation coefficient ρ (*i.e.*, the degree of similarity among units sampled from the same cluster) and the number of units sampled per cluster n_k (the higher ρ and n_k, the higher we will need to increase the sample size to be to keep the sensitivity constant) (see Fig. 3.3). When the infestation rate is relatively high (in the 0.5–2% range), it is possible to reliably estimate ρ for a pathway using model-based or Bayesian inference. However, when the infestation rate is very low (*e.g.*, in the case of NZ data, with a typical mean infestation rate of 0.003%), it is difficult to reliably estimate ρ from a pathway, even for large pathways (100 consignments).

Adaptive inspection regimes might be a useful first step if early action is valuable and when we do not have enough data to apply Bayesian approaches.

Bibliography

Aldrich, J. (1977). "R.A. Fisher and the Making of Maximum Likelihood 1912–1922". In: *Statistical Science* 12, pp. 162–176.

Altham, P.M.E. (1978). "Two generalizations of the binomial distribution". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, pp. 162–167.

Augustin, T., F.P.A. Coolen, G. de Cooman, and M.C.M. Troffaes, eds. (2014). *Introduction to Imprecise Probabilities*. New York: Wiley.

Barron, M.C. (2006). "Effects of aggregation on the probability of detecting infestations in fresh produce consignments". In: *New Zealand Plant Protection* 59, pp. 103–108.

Bayes, T. (1763). "An essay towards solving a problem in the doctrine of chances". In: *Philosophical Transactions of the Royal Society of London* 53, pp. 370–418.

Beale, R., J. Fairbrother, A. Inglis, and D. Trebeck (2008). *One Biosecurity: a Working Partnership*. Commonwealth of Australia.

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag.

– (1990). "Robust Bayesian analysis: sensitivity to the prior". In: *Journal of Statistical Planning and Inference* 25, pp. 303–328.

Berger, J.O., J.M. Bernardo, and D. Sun (2015). "Overall objective priors (with discussion)". In: *Bayesian Analysis* 10, pp. 189–246.

Bernardo, J.M. and A.F.M. Smith (1994). *Bayesian Theory*. New York: Wiley.

Box, G.E.P. and G.C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Brewer, K.R.W. (1963). "Ratio estimation and finite populations: some results deducible from the assumption of an underlying stochastic process". In: *Australian Journal of Statistics* 5, pp. 93–105.

Brown, L.D., T.T. Cai, and A. DasGupta (2001). "Interval estimation for a binomial proportion (with discussion)." In: *Statistical Science* 16, pp. 101–133.

Cai, T.T. (2005). "One-sided confidence intervals in discrete distributions". In: *Journal of Statistical Planning and Inference* 131, pp. 63–88.

Camac, J., A. Dodd, N Bloomfield, and A. P. Robinson (2020). *Sampling to support claims of area freedom*. Tech. rep. Centre of Excellence for Biosecurity Risk Analysis, pp. 1–28.

Caselton, W.F. and W. Luo (1992). "Decision making with imprecise probabilities: Dempster-Shafer theory and application". In: *Water Resources Research* 28, pp. 3071–3083.

Chambers, R.L. and R.G. Clark (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford, UK: Oxford University Press.

Chatfield, C. and G.J. Goodhardt (1970). "The beta-binomial model for consumer purchasing behaviour". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 19, pp. 240–250.

Chen, Cuicui, Rebecca S. Epanchin-Niell, and Robert G. Haight (2017). "Optimal Inspection of Imports to Prevent Invasive Pest Introduction". In: *Risk Analysis* 38.3, pp. 603–619. ISSN: 0272-4332. DOI: 10.1111/risa.12880.

Cochran, W.G. (1977). *Sampling Techniques*. 3rd. New York: John Wiley and Sons.

Colautti, R.I., S.A. Bailey, C.D.A. Van Overdijk, K. Amundsen, and H.J MacIsaac (2006). "Characterised and projected costs of nonindigenous species in Canada". In: *Biological Invasions* 8, pp. 45–59.

Coolen, F.P.A. and M.A. Elsaeiti (2009). "Nonparametric Predictive Methods for Acceptance Sampling". In: *Journal of Statistical Theory and Practice* 3, pp. 907–921.

Dalal, S. R. and W. J. Hall (1983). "Approximating Priors by Mixtures of Natural Conjugate Priors". In: 45.2, pp. 278–286. ISSN: 00359246.

de Finetti, B. (1937). "La Prévision: Ses Lois Logiques, Ses Sources Subjectives". In: *Annales de l'Institut Henri Poincaré* 7, pp. 1–68.

– (1974). *Theory of Probability*. Vol. I. New York: Wiley.

– (1975). *Theory of Probability*. Vol. II. New York: Wiley.

Defense, Department of (1996). *Test Method Standard: DOD Preferred Methods For Acceptance Of Product*. Tech. rep. MIL-STD-1916. Washington, D.C.: Department of Defense (DOD), p. 28.

– (1999). *Department of Defense Handbook: Companion Document to MIL-STD-1916*. Tech. rep. MIL-HDBK-1916. Washington, D.C.: Department of Defense (DOD), p. 127.

Dempster, A.P. (1966). "New methods for reasoning towards posterior distributions based on sample data". In: *Annals of Mathematical Statistics* 37, pp. 355–374.

– (1967a). "Upper and lower probabilities induced by a multivalued mapping". In: *Annals of Mathematical Statistics* 38, pp. 325–339.

– (1967b). "Upper and lower probability inference based on a sample from a finite univariate population". In: *Biometrika* 54, pp. 515–528.

– (1968a). "A Generalization of Bayesian inference". In: *Journal of the Royal Statistical Society, Series B* 30, pp. 205–247.

– (1968b). "Upper and lower probabilities generalized by a random closed interval". In: *Annals of Mathematical Statistics* 39, pp. 957–966.

Dempster, A.P. and A. Kong (1987). "Probabilistic expert systems in medicine: practical issues in handling uncertainty — comment." In: *Statistical Science* 2, pp. 32–36.

Denoeux, T. (2014). "Likelihood-based belief function: justification and some extensions to low-quality data". In: *International Journal of Approximate Reasoning* 55, pp. 1535–1547.

– (2016). "40 Years of Dempster-Shafer theory". In: *International Journal of Approximate Reasoning* 79, pp. 1–6.

Diaconis, P. and D. Ylvisaker (1985). "Bayesian Statistics 2". In: ed. by Bernardo J.M., M.H. DeGroot, Lindleyand D.V., and A.F.M. Smith. North Holland: Elsevier. Chap. Quantifying prior opinion, pp. 133–156.

Dodge, Harold F (1943). "A sampling inspection plan for continuous production". In: *The Annals of mathematical statistics* 14.3, pp. 264–279.

– (1955). "Skip-lot sampling plan". In: *Industrial Quality Control* 11.5, pp. 3–5.

Dodge, Harold F and Mary N Torrey (1951). "Additional continuous sampling inspection plans". In: *Industrial Quality Control* 7.5, pp. 7–12.

Ducey, M.J. (2001). "Representing uncertainty in silvicultural decisions: an application of the Dempster-Shafer theory of evidence". In: *Forest Ecology and Management* 150, pp. 199–211.

Ellenberg, D. and D. Mueller-Dombois (1974). *Aims and methods of vegetation ecology*. New York: Wiley.

Ericson, W.A. (1969). "Subjective Bayesian models in sampling finite populations". In: *Journal of the Royal Statistical Society, Series B* 31, pp. 195–233.

Everson, P.J. and E.T. Bradlow (2002). "Bayesian inference for the beta-binomial distribution via polynomial expansions". In: *Journal of Computational and Graphical Statistics* 11, pp. 202–207.

Fienberg, S.E. (2006). "When did Bayesian inference become "Bayesian"?" In: *Bayesian analysis* 1, pp. 1–40.

Fisher, R.A. (1922). "On the mathematical foundations of theoretical statistics". In: *Philosophical Transactions of the Royal Society of London Series A* 222, pp. 309–368.

Fisher, R.A (1925). *Statistical Methods for Research Workers*. Edinburgh and London, UK: Oliver and Boyd.

Fuller, W.A. (2009). *Sampling Statistics*. Wiley.

Geisser, S. (1984). "On Prior Distributions for Binary Trials". In: *American Statistician* 38, pp. 244–247.

Gelman, A., J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D. Rubin (2014). *Bayesian Data Analysis. Third edition*. Chapman and Hall.

Giera, N. and B. Bell (2009). *Economic Costs of Pests to New Zealand*. MAF Biosecurity New Zealand Technical Paper 2009/31. Wellington: Biosecurity New Zealand, Ministry of Agriculture and Forestry.

Good, I. J. (1976). "The Bayesian Influence, or How to Sweep Subjectivism under the Carpet". In: *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science: Proceedings of an International Research Colloquium held at the University of Western Ontario, London, Canada, 10–13 May 1973. Volume II, Foundations and Philosophy of Statistical Inference*. Ed. by W.L. Harper and C.A. Hooker. Dordrecht: Springer Netherlands, pp. 125–174.

Gregoire, T.G. (1998). "Design–based and model–based inference in survey sampling: appreciating the difference". In: *Canadian Journal of Forest Research* 28, pp. 1429–1447.

Gregoire, T.G. and H.T. Valentine (2008). *Sampling Techniques for Natural and Environmental Resources*. Boca Raton, FL: Chapman and Hall.

Griffiths, D.A. (1973). "Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease". In: *Biometrics*, pp. 637–648.

Halpern, J.Y. and R. Fagin (1992). "Two views of belief: belief as generalized probability and belief as evidence". In: *Artificial intelligence* 54, pp. 275–317.

Hansen, M.H., W.N. Hurwitz, and W.G. Madow (1953). *Sampling Survey Methods and Theory*. Vol. I and II. New York: Wiley.

Hemming, Victoria, Mark A. Burgman, Anca M. Hanea, Marissa F. McBride, and Bonnie C. Wintle (2018). "A practical guide to structured expert elicitation using the IDEA protocol". In: *Methods Ecol Evol* 9.1, pp. 169–180. ISSN: 2041-210X. DOI: 10.1111/2041-210X.12857.

Hoffman, B.D. and L.M. Broadhurst (2016). "The Economic Cost of Managing Invasive Species in Australia". In: *Neobiota* 31, pp. 1–18.

Horvitz, D.G. and D.J. Thompson (1952). "A generalization of sampling without replacement from a finite universe". In: *Journal of the American Statistical Association* 47, pp. 663–685.

Hughes, G., L.V. Madden, and G.P. Munkvold (1996). "Cluster sampling for disease incidence data". In: *Phytopathology* 86, pp. 132–137.

Hulme, P.E., S. Bacher, M. Kenis, S. Klotz, I. Kühn, D. Minchin, W. Nentwig, S. Olenin, V. Panov, J. Pergl, P. Pyšek, A. Roques, D. Sol, A. Solarz, and M. Vilà (2008). "Grasp-

ing at the routes of biological invasions: a framework for integrating pathways into policy". In: *Journal of Applied Ecology* 45.2, pp. 403–414.

International Plant Protection Convention (2008). *International Standard for Phytosanitary Measures 31: Methodologies for Sampling of Consignments*.

– (2016). *International Standard for Phytosanitary Measures 6: Guidelines for Surveillance*.

IPPC (2002). *International standard for phytosanitary measures. ISPM No. 14: The use of integrated measures in a systems approach for pest risk management*. Tech. rep. Secretariat of the International Plant Protection Convention.

– (2008). *International standard for phytosanitary measures. ISPM No. 31: methodology for sampling of consignment*. Tech. rep. Secretariat of the International Plant Protection Convention.

Jarrad, F. C., S. Barrett, J. Murray, J. Parkes, R. Stoklosa, K. Mengersen, and P. Whittle (2011). "Improved design method for biosecurity surveillance and early detection of non-indigenous rats". In: *New Zealand Journal of Ecology* 35.2, pp. 132–144.

Jeffreys, H. (1939). *Theory of Probability*. Oxford: Clarendon Press.

Kemp, C.D. and A.W. Kemp (1956). "The analysis of point quadrat data". In: *Australian Journal of Botany* 4, pp. 167–174.

Keynes, J.M. (1921). *A Treatise on Probability*. London: MacMillan.

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

– (1995). "The Hundred Years' Wars of Survey Sampling". In: *Statistics in Transition* 2, pp. 813–830.

Kruskal, W. and F. Mosteller (1980). "Representative Sampling, IV: The History of the Concept in Statistics, 1895–1939". In: *International Statistical Review / Revue Internationale de Statistique* 48, pp. 169–195.

Kupper, L.L. and J.K. Haseman (1978). "The use of a correlated binomial model for the analysis of certain toxicological experiments". In: *Biometrics*, pp. 69–76.

Kyburg Jr., H.E. (1987). "Bayesian and non-Bayesian evidential updating". In: *Artificial intelligence* 31, pp. 271–293.

Lane, S.E., R.M. Cannon, A.D. Arthur, and A.P. Robinson (2018a). "Sample-Based Regulatory Intervention for Managing Risk within Heterogeneous Populations, with Phytosanitary Inspection of Mixed Consignments as a Case Study". In: *bioRxiv*, p. 441543.

Lane, S.E., R. Gao, M. Chisholm, and A.P. Robinson (2017). "Statistical profiling to predict the biosecurity risk presented by non-compliant international passengers". In: *arXiv preprint arXiv:1702.04044*.

Lane, S.E., R. Souza Richards, C. McDonald, and A.P. Robinson (2018b). *Sample size calculations for phytosanitary testing of small lots of seed*. Tech. rep. Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne.

Laplace, P.-S. (Marquis de) (1774). *Théorie analytique des probabilités*. Paris: Courcier.

– (1812). "Mémoire sur la probabilité des causes par les événements". In: *Mémoires de Mathématiqe et de Physique presentés á l'Académie Royale des Sciences, par Divers Savans, & Lûs dans ses Assemblées* 6, pp. 621–656.

Lee, J. and Y.L. Lio (1999). "A note on Bayesian estimation and prediction for the beta-binomial model". In: *Journal of Statistical Computation and Simulation* 63, pp. 73–91.

Lee, J.C. and D.J. Sabavala (1987). "Bayesian estimation and prediction for the beta-binomial model". In: *Journal of Business & Economic Statistics* 5.3, pp. 357–367.

Liebhold, A.M., T.T. Work, D.G. McCullough, and J.F. Cavey (2006). "Airline baggage as a pathway for alien insect species invading the United States". In: *American Entomologist* 52.1, pp. 48–54.

Lindley, D.V. (1978). "The Bayesian Approach [with Discussion and Reply]". In: *Scandinavian Journal of Statistics*, pp. 1–26.

Little, R.J. (2004). "To model or not to model? Competing modes of inference for finite population sampling". In: *Journal of the American Statistical Association* 99, pp. 546–556.

Low-Choy, S. (2012). "Priors: Silent or active partners of Bayesian inference?" In: *Case studies in Bayesian statistical modelling and analysis*. Ed. by C.L. Alston, K.L. Mengersen, and A.N. Pettitt. New York: John Wiley & Sons. Chap. 3, pp. 30–65.

– (2015a). "Getting the Story Straight: Laying the Foundations for Statistical Evaluation of the Performance of Surveillance". In: *Biosecurity Surveillance: Quantitative Approaches*. Ed. by F. Jarrad, S. Low-Choy, and K. Mengersen. Wallingford, UK: CAB International. Chap. 3, pp. 43–74.

– (2015b). "Hierarchical Models for Evaluating Surveillance Strategies: Diversity Within a Common Modular Structure". In: *Biosecurity Surveillance: Quantitative Approaches*. Ed. by F. Jarrad, S. Low-Choy, and K. Mengersen. Wallingford, UK: CAB International. Chap. 4, pp. 75–108.

Lui, K.-J., W.G. Cumberland, and L. Kuo (1996). "An interval estimate for the intraclass correlation in beta-binomial sampling". In: *Biometrics*, pp. 412–425.

Madden, L.V. and G. Hughes (1995). "Plant disease incidence: distributions, heterogeneity, and temporal analysis". In: *Annual Review of Phytopathology* 33, pp. 529–564.

Magnussen, S. (2015). "Arguments for a model-dependent inference?" In: *Forestry* 88, pp. 317–325.

Mak, T.K. (1988). "Analysing intraclass correlation for dichotomous variables". In: *Applied Statistics*, pp. 344–352.

Martin, T.G., M.A. Burgman, F. Fidler, P.M. Kuhnert, S. Low-Choy, M. McBride, and K. Mengersen (2012). "Eliciting expert knowledge in conservation science". In: *Conservation Biology* 26, pp. 29–38.

Martz, H.F. and M.G. Lian (1974). "Empirical Bayes estimation of the binomial parameter". In: *Biometrika* 61, pp. 517–523.

Mátern, B. (1960). "Spatial Variation". In: *Medd. Statens Skogforskningsinstitut* 49.5.

McBride, G.B. and P. Johnstone (2011). "Calculating the probability of absence using the Credible Interval Value". In: *New Zealand Journal of Ecology* 35, pp. 189–190.

Montgomery, Douglas C. (2009). *Introduction to Statistical Quality Control*. 6-th. John Wiley & Sons, Inc.

New Zealand Ministry for Primary Industries (2016). *Standard 152.02, Importation and Clearance of Fresh Fruit and Vegetables into New Zealand*.

Neyman, J. (1934). "On the two different aspects of the representative method: the method of stratified sampling and the purposive method". In: *Journal of the Royal Statistical Society* 97, pp. 558–625.

– (1938). "Contribution to the Theory of Sampling Human Populations". In: *Journal of the American Statistical Association* 33, pp. 101–116.

Ormsby, M. (2017). "International Developments in Determining Levels of Intervention in Risk Pathways". In: *Proceedings, International Symposium on Risk-Based Sampling, Baltimore, Maryland, June 26–30, 2017*. North American Pest Protection Organization, pp. 31–36.

Pearl, J. (1990). "Reasoning with belief functions: An analysis of compatibility". In: *International Journal of Approximate Reasoning* 4, pp. 363–389.
Perry, Robert L (1973). "Skip-lot sampling plans". In: *Journal of Quality Technology* 5.3.
Pimentel, D. (2011). *Biological Invasions: Economic and Environmental Costs of Alien Plant, Animal, and Microbe Species*. 2nd. Hoboken, NJ: CRC Press.
Pimentel, D., R. Zuniga, and D. Morrison (2005). "Update on the environmental and economic costs associated with alien-invasive species in the United States". In: *Ecological Economics* 52, pp. 273–288.
Quinlan, M., M. Stanaway, and K. Mengersen (2015). "Biosecurity surveillance in agriculture and environment: a review". In: *Biosecurity Surveillance: Quantitative Approaches*. Ed. by F. Jarrad, S. Low-Choy, and K. Mengersen. Wallingford, UK: CAB International. Chap. 2, pp. 9–42.
Ransom, L. (2017). "Australia's Experience with Risk-Based Sampling". In: *Proceedings, International Symposium on Risk-Based Sampling, Baltimore, Maryland, June 26–30, 2017*. North American Pest Protection Organization, pp. 12–17.
Rathman, James F., Chihae Yang, and Haojin Zhou (2018). "Dempster-Shafer theory for combining in silico evidence and estimating uncertainty in chemical risk assessment". In: *Computational Toxicology* 6, pp. 16–31. ISSN: 2468-1113.
Robinson, A.P., M.A. Burgman, and R. Cannon (2011). "Allocating surveillance resources to reduce ecological invasions: maximizing detections and information about the threat". In: *Ecological Applications* 21, pp. 1410–1417.
Rossiter, A. and S. Hester (2017). "Designing biosecurity inspection regimes to account for stakeholder incentives: An inspection game approach". In: *Economic Record* 93.301, pp. 277–301.
Rossiter, Anthony, Andreas Leibbrandt, Bo Wang, Felicity Woodhams, and Susie Hester (2018). *Testing Compliance-Based Inspection Protocols*. Tech. rep. 1404c. Melbourne, Australia: Centre of Excellence for Biosecurity Risk Analysis (CEBRA), p. 63.
Royall, R.M. (1970). "On Finite Population Sampling Under Certain Linear Regression Models". In: *Biometrika* 57, pp. 377–387.
Saha, K.K. and S. Wang (2018). "Confidence intervals for the common intraclass correlation in the analysis of clustered binary responses". In: *Journal of Biopharmaceutical Statistics* 28, pp. 682–697.
Särndal, C.E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer.
Savage, L.J. (1954). *The Foundations of Statistical Inference*. New York: Wiley.
Schmidli, Heinz, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander (2014). "Robust meta-analytic-predictive priors in clinical trials with historical control information". In: *Biom* 70.4, pp. 1023–1032. ISSN: 0006-341X. DOI: 10.1111/biom.12242.
Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton, New Jersey: Princeton University Press.
– (2016). "Dempster's rule of combination". In: *International Journal of Approximate Reasoning* 79, pp. 26–40.
Shafer, G. and J. Pearl, eds. (1990). *Readings in Uncertain Reasoning*. San Mateo: Morgan Kaufman.
Skellam, J.G. (1948). "A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 10, pp. 257–261.

Stephens, K.N. (2001). *Handbook of Applied Acceptance Sampling: Plans, Procedures, and Principles*. Milwaukee, WI: American Society for Quality Press.

Sukhatme, P.V. and B.V. Sukhatme (1970). *Sampling Theory of Surveys with Applications*. Ames, IA: Iowa State University Press.

Thompson, S.K. (2012). *Sampling*. 3rd. New York: John Wiley and Sons.

Tuyl, F., R. Gerlach, and K. Mengersen (2008). "A comparison of Bayes–Laplace, Jeffreys, and other priors: the case of zero events". In: *The American Statistician* 62, pp. 40–44.

– (2009). "Posterior predictive arguments in favor of the Bayes-Laplace prior as the consensus prior for binomial and multinomial parameters". In: *Bayesian Analysis* 4, pp. 151–158.

Valliant, R., A.H. Dorfman, and R.M. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

vanKlinken, R. D., K. Fiedler, L. Kingham, K. Collins, and D. Barbourc (n.d.). "A risk framework for using systems approaches to manage horticultural biosecurity risks for market access". In: *In prep.*

Venette, R.C., R.D. Moon, and W.D. Hutchinson (2002a). "Strategies and statistics of sampling for rare individuals". In: *Annual Review of Entomology* 47, pp. 143–174.

Venette, Robert C., Roger D. Moon, and William D. Hutchison (2002b). "Strategies and Statistics of Sampling for Rare Individuals". In: *Annu. Rev. Entomol.* 47.1, pp. 143–174. ISSN: 0066-4170. DOI: 10.1146/annurev.ento.47.091201.145147.

Vijayaraghavan, R (2000). "Design and evaluation of skip-lot sampling plans of type SkSP-3". In: *Journal of Applied Statistics* 27.7, pp. 901–908.

von Neumann, J. (1951). "Various techniques used in connection with random digits". In: *U.S. National Bureau of Standards Applied Math. Series* 12, pp. 36–38.

Voorbraak, F. (1981). "On the justification of Dempster's rule of combination". In: *Artificial Intelligence* 48, pp. 171–197.

Walley, P. (1987). "Belief function representations of statistical evidence". In: *The Annals of Statistics* 15, pp. 1439–1465.

– (1991). *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

– (1996a). "Inferences from multinomial data: learning about a bag of marbles". In: *Journal of the Royal Statistical Society, Series B* 58, pp. 3–57.

– (1996b). "Measures of uncertainty in expert systems". In: *Artificial intelligence* 83, pp. 1–58.

– (2000). "Towards a unified theory of imprecise probability". In: *International Journal of Approximate Reasoning* 24, pp. 125–148.

– (2002). "Reconciling frequentist properties with the likelihood principle". In: *Journal of Statistical Planning and Inference* 105, pp. 35–65.

Walley, P. and S. Moral (1999). "Upper probabilities based only on the likelihood function". In: *Journal of the Royal Statistical Society, Series B, Part 4* 61, pp. 831–847.

Wan, Fang-Hao and Nian-Wan Yang (2016). "Invasion and Management of Agricultural Alien Insects in China". In: *Annu. Rev. Entomol.* 61.1, pp. 77–98. ISSN: 0066-4170. DOI: 10.1146/annurev-ento-010715-023916.

Ware, K. D. and T. Cunia (1962). "Continuous forest inventory with partial replacement of samples". In: *Forest Science monographs*.

Whattam, M., G. Clover, M. Firko, and T. Kalaris (2014). "The Biosecurity Continuum and Trade: Border Operations". In: *The Handbook of Plant Biosecurity*. Ed. by G. Gordh and S. McKirdy. Dordrecht: Springer. Chap. 6, pp. 149–188.

Williams, D.A. (1975). "The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity". In: *Biometrics* 31, pp. 949–952.

Yang, R. and J.O. Berger (1998). *A Catalog of Noninformative Priors*. West Lafayette, IN: Purdue University.

Zou, G. and A. Donner (2004). "Confidence interval estimation of the intraclass correlation coefficient for binary outcome data". In: *Biometrics* 60, pp. 807–811.