

Review

Design-Based Approach for Analysing Survey Data in Veterinary Research

D. Aaron Yang ^{1,*}  and Richard A. Laven ² 

¹ Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong

² School of Veterinary Science, Massey University, Palmerston North 4442, New Zealand; r.laven@massey.ac.nz

* Correspondence: aaron.yang@cityu.edu.hk

Abstract: Sample surveys are an essential approach used in veterinary research and investigation. A sample obtained from a well-designed sampling process along with robust data analysis can provide valuable insight into the attributes of the target population. Two approaches, design-based or model-based, can be used as inferential frameworks for analysing survey data. Compared to the model-based approach, the design-based approach is usually more straightforward and directly makes inferences about the finite target population (such as the dairy cows in a herd or dogs in a region) rather than an infinite superpopulation. In this paper, the concept of probability sampling and the design-based approach is briefly reviewed, followed by a discussion of the estimations and their justifications in the context of several different elementary sampling methods, including simple random sampling, stratified random sampling, and one-stage cluster sampling. Finally, a concrete example of a complex survey design (involving multistage sampling and stratification) is demonstrated, illustrating how finding unbiased estimators and their corresponding variance formulas for a complex survey builds on the techniques used in elementary sampling methods.

Keywords: sampling; survey methodology; design-based approach; unbiasedness; variance estimation



Citation: Yang, D.A.; Laven, R.A. Design-Based Approach for Analysing Survey Data in Veterinary Research. *Vet. Sci.* **2021**, *8*, 105. <https://doi.org/10.3390/vetsci8060105>

Academic Editor: Fabrizio Bertelloni

Received: 8 April 2021

Accepted: 3 June 2021

Published: 8 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Sample surveys, where data from a subset, or sample, of a population are used to make inferences about that population, are a traditional research methodology which has been widely used in veterinary research and investigation [1,2]. However, in this era of “big data”, with modern techniques such as machine learning, bioinformatics, or other computer-based technologies being increasingly used in veterinary research [3] across areas such as animal behaviour [4] and disease detection [5] and prediction [6], the sample survey is in danger of appearing “old fashioned” and “out-dated”.

However, the “old-fashioned” sample survey still has some advantages over cutting-edge big data methodologies. Firstly, in a sample survey, information or data can be collected actively in order to answer a specific research question, whereas the research question that can be answered by “big data” techniques is dependent on what information is available in the big data source. Secondly, in a well-planned sample survey the target population can be framed in advance and followed by a well-designed sampling process so that the samples are representative of the population [7]. This representativeness is often not achieved during the passive “big data” collection process, with data often being collected only from a particular subset of the target population—e.g., Revilla, et al. [8] analysed more than 10.5 million measurements from ~13,000 pigs obtained using automatic feeding systems. However, this dataset was collected from only one boar testing station, making generalisation to the wider population potentially difficult. Finally, it is not economically feasible to collect “big data” when novel information is required for some specific research topics.

Once the specific data for a research topic are collected, a rigorous and robust data analysis is essential to gain insight from the sample survey. As per Alexis de Tocqueville: “when statistics are not based on strictly accurate calculations, they mislead instead of guide” [9]. Generally, there are two approaches for analysing survey data: the model-based approach and the design-based approach [10]. The former is possibly better understood by many veterinary researchers who have undertaken standard quantitative research methodology training, as mainstream statistical training usually treats the observed data—e.g., production data or diagnostic test outcomes—as realisations of some relevant random variables. However, one important assumption in this method is usually overlooked—that the underlying population is treated as a “superpopulation” which contains infinitely many animals [11]. Strictly speaking, an estimated model parameter therefore refers to a property of the hypothetical “superpopulation” rather than a characteristic of the finite population which is of actual interest [7].

For example, suppose that simple random sampling had been implemented on a dairy farm to study the prevalence of bovine digital dermatitis using a large sampling fraction (~70% of the herd size). The analyst then fitted an intercept-only logistic regression to estimate the intercept. As the intercept represents the logit of the prevalence, using a suitable back-transformation, the analyst was then able to report the estimated prevalence of digital dermatitis on the farm. However, given a large sampling fraction, this estimate actually represents the prevalence of a hypothetical superpopulation from which the sample was drawn rather than the prevalence on the farm of interest [12]. To make an inference to the actual finite population, a superpopulation approach can be used [12]; however, although mathematically correct, explaining the approach is likely to create difficulties in communications with other rural professionals or companion animal practitioners.

The design-based approach for analysing survey data avoids the complexity in analysis and communication seen with the model-based approach. One key advantage of the design-based approach is that it focuses on inferences related to the target finite population(s) without introducing extra assumptions about the parametric form of the outcome variable. In addition, the analysis steps are consistent with the sampling steps, so the process of checking for potential mistakes during the analysis process is clearer [13]. Thirdly, the design-based approach has no requirement of an “assumed” probability distribution dictated by the design itself [14]. The aim of this review article is to provide a comprehensive introduction to the design-based approach for analysing survey data by (1) describing the analytical methods for elementary probability sampling methods, including simple random sampling, stratified random sampling, and cluster sampling, and (2) to demonstrate the key ideas necessary to understand and interpret those analytical methods, as well as how those ideas can be used to develop methods for any specific complex survey design.

2. Overview of Probability Sampling

First, we define a set U as a target population including M animals in a finite population (e.g., animals on a farm or all >2-year-old Jersey cows in a region). There are various ways to obtain samples which are just some subsets of U . Let us denote S for a particular sample chosen from U , then $S \subset U$. With the proper subset notation “ \subset ”, we restrict the sample size to being smaller than the population size. Suppose we want to obtain a random sample with m animals, we could have $\binom{M}{m} = L$ different samples: S_1, S_2, S_3, \dots . We can then define a set “sample space” (denoted as Ω) that contains all these samples. With probability sampling, a probability can be explicitly assigned for each of the samples, with the constraint that $\sum_{i=1}^L P(S_i) = 1$, as the axiom states that the probability of a sample space is 1 and the union of all the samples forms the sample space. The probability of obtaining each of the L samples does not have to be constant—i.e., $P(S_1) \neq P(S_2)$ is absolutely acceptable—and we can also restrict the probability of a particular sample to 0 if some animals within the sample are considered inappropriate as study units. The other feature of these samples is that two samples can include the same animals, and the probability of an animal k being selected (π_k) is calculated by summing the probabilities of all samples

including this animal—i.e., $\pi_k = \sum_{S:k \in S} P(S)$. An intuitive numeric example is displayed in Figure 1. Eventually, we define the sampling weights w_k as the reciprocal of the inclusion probability π_k for any type of sampling method [15]. Generally, it is recommended that the veterinary researcher interprets the sampling weight of the animal k as the number of animals in the target population represented by this animal (a deeper treatment of sampling weights can be found in Gelman [16]; however, non-response adjustments are beyond the scope of this article).

A small farm with 5 animals: $U = \{1,2,3,4,5\}$.

All possible samples with 3 animals are listed as below. There are in total $\binom{5}{3} = 10$ possible choices.

$$S_1 = \{1,2,3\} \quad S_2 = \{1,2,4\} \quad S_3 = \{1,2,5\} \quad S_4 = \{1,3,4\} \quad S_5 = \{1,3,5\}$$

$$S_6 = \{1,4,5\} \quad S_7 = \{2,3,4\} \quad S_8 = \{2,3,5\} \quad S_9 = \{2,4,5\} \quad S_{10} = \{3,4,5\}$$

If we assign uniform probability to S_i , such as $P(S_i) = \frac{1}{10}$, then $\pi_2 = P(S_1) + P(S_2) + P(S_3) + P(S_7) + P(S_8) + P(S_9)$.

The probability that animal 2 is selected is $\frac{3}{5}$.

If animal 5 is not suitable for the study, we may assign $P(S_3) = P(S_5) = P(S_6) = P(S_8) = P(S_9) = P(S_{10}) = 0$.

And we can customise the probabilities for the other possible samples: $P(S_1) = P(S_2) = \frac{1}{3}$, $P(S_4) = \frac{1}{12}$ and $P(S_7) = \frac{1}{4}$,

as long as $\sum_{i=1}^{10} P(S_i) = 1$. In this scenario, the probability that animal 2 is selected is $\frac{1}{3} + \frac{1}{3} + 0 + \frac{1}{4} + 0 + 0 = \frac{11}{12}$.

Figure 1. An intuitive explanation of the probability of a sample selection $P(S)$ and the probability of an animal selection π .

3. Design-Based and Model-Based Approaches

3.1. Overview of Design-Based Approach

In the design-based approach, the observed value (production record or test outcome) y_k is not considered to be a realisation generated from some data generation mechanism (or “population”); instead, it is regarded as a fixed constant, with the randomness arising solely from the sample selection [17]. In plainer language, although y_k is fixed, it remains unknown unless the animal is selected in the sample [18]. If all animals are tested from a given population, then the test results for all the animals are known without any uncertainty. In contrast, if we only test a sample of these animals, the only outcomes that we know are the animals included in the sample. If we randomly draw samples of a fixed size repeatedly from a target population, a particular animal may not be sampled repeatedly due to the randomness in the sample selection; hence, the sample statistics can vary across the samples. That is the only source of randomness. Therefore, it is natural to define a Bernoulli random variable to indicate whether an animal in the target population is also in the sample. For example, if the k th animal is included in the sample, the Bernoulli random variable $Z_k = 1$; otherwise, it is $Z_k = 0$. This random variable maps the (hypothetical) animal ID in the population into numeric values for selection status. This idea is essential when studying the properties of the estimators (unbiasedness) and deriving the variances of these estimators [19].

The design-based approach is of particular value when the finite population characteristics are of interest, as when a design-based approach is used, the researcher can direct inferences about the finite target population, even if the sample size to population size ratio is not small (i.e., when the finite population correction must be considered). For example, the prevalence in a finite population is interpreted as the proportion of diseased animals in that population. Assuming that 70% of the population is sampled, a design-based

approach gives a direct estimate of this population proportion, which is often a key target of veterinary investigations. In addition, the estimators (i.e., the rules or formulas) for estimating the finite population characteristics are consistent with the sampling method. Therefore, the estimation process is naturally understandable and easier to communicate with non-statistically inclined veterinarians and researchers [20]. Finally, with the design-based approach, the analyst does not need to decide which potential model generated the data, as the observed values are treated as fixed constants. For example, if the average milk production in a herd is of interest, one does not need to assume that milk yield from a cow is generated from a normal random variable, particularly when it is not. One just needs to calculate the sample mean as an estimate of the average milk production in the herd. Finally, whether an estimator is unbiased (i.e., whether its expected value and the true value of the parameter are effectively equivalent) is not dependent on the parametric form of the observed value.

3.2. Overview of Model-Based Approach

Although this approach may be more familiar to researchers, we do not advocate this approach in this paper from a practical point of view. As with most mainstream statistical methods, this approach treats the observed value as a realisation of an underlying random variable. For example, the test result y_k of the k th animal is generated from a random variable Y_k , whose parametric form must be decided. If Y_k is a Bernoulli random variable, then with the model-based approach a likelihood-based method is the most common approach for estimating the probability that a random animal will test positive. However, this probability is not a finite population prevalence; it is more correctly interpreted as a hypothetical infinite superpopulation prevalence. Although extra steps can help to make an inference back to the finite population, this significantly adds complexity to aspects of analysis and in communicating the results to stakeholders without a statistical background [12]. The other major disadvantage of the model-based approach is that the estimated parameters may be biased if the model is mis-specified [20].

4. Sampling Methods

4.1. Simple Random Sampling

Simple random sampling (SRS) is the most basic form of probability sampling. In this process, all possible samples of a given size have the same probability of being selected—i.e., $P(S)$ is constant for every possible sample. As a result, all the animals in the population have an equal probability of being included in the sample—i.e., the inclusion probability π is the same for every animal [21]. This sampling process has been applied to many veterinary studies, including recent investigations of lumpy skin disease [22], bovine mastitis [23], and foot-and-mouth disease [24]. Despite its simplicity, in the right situation it can be a powerful sampling method and provide the theoretical basis for more complicated sampling methods. There are two forms of SRS—with and without replacement. In this article, we will limit the discussion to SRS without replacement (the sample contains no duplicated animals,) as this is by far the most common practice in veterinary research.

The statistics in which a researcher is usually interested are the properties of the population, e.g., the average milk production of the herd or the prevalence of a disease within the herd. We denote this finite population mean as μ and prevalence p is just a special case of the finite population mean when the individual outcome value y_k can only be 1 or 0. In the SRS setting, estimating the mean is straightforward. However, for other sampling processes this is not always the case; hence, it is easier to start to estimate the finite population total before moving on to the mean (which is a linear function of the total). To ensure a consistent methodology is used in this review, we will stick with the two-step process—estimating the total first and then the mean or prevalence.

Suppose we have a herd with M animals, of which a sample of m animals has been obtained using SRS. The Horvitz–Thompson (HT) estimator of the finite population total is [25]:

$$\hat{\tau} = \sum_{k=1}^m w_k y_k = \frac{M}{m} \sum_{k=1}^m y_k \tag{1}$$

In the SRS setting, the sampling weight w_k is a constant as the inclusion probability is the same for every animal, such that $\pi_k \equiv P(Z_k = 1) = \frac{m}{M}$ (see Appendix A for technical details), where Z_k is the Bernoulli random variable for selection and $Z_k = 1$ if the animal k is selected; otherwise, $Z_k = 0$. The HT estimator is, by design, unbiased—i.e., its expected value is equal to the true value of the finite population characteristic [7]:

$$\mathbb{E}[\hat{\tau}] = \mathbb{E}\left[\sum_{k=1}^m w_k y_k\right] = w_k \mathbb{E}\left[\sum_{k=1}^M y_k Z_k\right] = \frac{1}{\pi_k} \sum_{k=1}^M y_k \mathbb{E}[Z_k] = \frac{M}{m} \sum_{k=1}^M y_k \frac{m}{M} = \sum_{k=1}^M y_k = \tau,$$

where $\mathbb{E}[\cdot]$ is the expectation operator which takes all the possible values generated by the random variable and returns the weighted average value, so $\mathbb{E}[Z_k] = 1 \times P(Z_k = 1) + 0 \times P(Z_k = 0) = P(Z_k = 1) = \frac{m}{M}$.

The unbiased estimator for the mean \bar{y} or the prevalence \hat{p} is, therefore:

$$\bar{y} \equiv \hat{p} = \frac{\hat{\tau}}{M} = \frac{1}{m} \sum_{k=1}^m y_k \tag{2}$$

The proof is trivial. By observing Equation (2), we see that, in SRS, the sample mean or sample proportion is the unbiased estimator for the population mean or prevalence. This means that, for other sampling strategies, building up the sample mean from SRS will also result in an unbiased estimator if done correctly.

To derive the variance of the estimator for the mean or the prevalence, it is also easier to start with the variance of the total. The detailed derivation can be seen in Appendix B; here, we only provide the formulas for the variances. First, the variance for the estimated population total is:

$$Var(\hat{\tau}) = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \sigma^2, \tag{3}$$

where $\sigma^2 = \frac{\sum_{k=1}^M (y_k - \mu)^2}{M-1}$ is the variance of the finite population. In the special case where we estimate prevalence, we can replace μ with p with some algebra (see Appendix B), resulting in $\sigma^2 = \frac{Mp(1-p)}{M-1}$. Therefore, the variances for \bar{y} and \hat{p} are given as follows:

$$Var(\bar{y}) = Var\left(\frac{\hat{\tau}}{M}\right) = \frac{\sigma^2}{m} \left(1 - \frac{m}{M}\right), \tag{4}$$

$$Var(\hat{p}) = \left(\frac{M-m}{M-1}\right) \frac{p(1-p)}{m}. \tag{5}$$

However, the finite population variance depends on an unknown quantity μ or p , which we are attempting to estimate; in practice, we often replace σ^2 with $s^2 = \frac{\sum_{k=1}^m (y_k - \bar{y})^2}{m-1}$, which is the sample variance (or p with \hat{p}). Therefore, the estimated variance for \bar{y} and \hat{p} is:

$$\widehat{Var}(\bar{y}) = \frac{s^2}{m} \left(1 - \frac{m}{M}\right), \tag{6}$$

$$\widehat{Var}(\hat{p}) = \left(1 - \frac{m}{M}\right) \frac{\hat{p}(1-\hat{p})}{m-1}, \tag{7}$$

where $1 - \frac{m}{M}$ is usually referred to as the finite population correction factor [26].

To illustrate this process, consider an investigator who wants to estimate the prevalence of digital dermatitis in lactating cows in a dairy herd. A random sample of 100 cows is obtained from a herd of 300 cows, of which 35 sampled cows are diagnosed as diseased. These 35 cows have records $y_k = 1$ and the remaining 65 sampled cows have records $y_k = 0$. The estimated prevalence is calculated using Equation (2), thus it is 0.35. The variance of this estimate is calculated using Equation (7). As the actual prevalence is

unknown, we need to use the estimated prevalence to calculate the estimated variance:

$$\left(1 - \frac{100}{300}\right) \frac{0.35 \times (1 - 0.35)}{100 - 1} = 0.0015.$$

4.2. Stratified Random Sampling

In the stratified random sampling procedure (STRRS), the target finite population (e.g., the total number of animals within a herd) is partitioned into non-overlapping groups based on some pre-defined attributes and each of the groups is referred to as a stratum. These strata constitute the entire population; therefore, each animal belongs to a specific stratum. Within each stratum, SRS is commonly used to sample animals, and the sampling processes in the different strata are independent [27]. There is no requirement to select all strata within a population. If only some strata are of interest (e.g., only those which include lactating cows), these can be selected and strata that are not of interest can be excluded. If this approach is used, it needs to be made clear that the target population is no longer the entire finite population, but rather the population represented by the selected strata.

The finite population mean or prevalence is then estimated by pooling the information from all the strata. Like SRS, STRRS is commonly used in veterinary research, for example stratification by area. This allows the researcher to investigate prevalences and associations across a country or a region—e.g., Heayns and Baugh [28] investigated the opinions of veterinarians across the UK about serological testing to assess revaccination requirements in dogs. In this study, each county of Great Britain was considered as a stratum and 10% of the small animal veterinary practices within each stratum were randomly selected (if there were fewer than 10 practices in a county, one practice was randomly chosen to represent the county). Similarly, Atuman et al. [29] investigated dog ecology, dog bites, and rabies vaccination rates in Bauchi, Nigeria, using STRRS. They stratified Bauchi into five areas, and within each area randomly selected 10% of the streets for direct street counts and the administration of a questionnaire. However, other sources of strata are also used—e.g., as part of a randomised clinical trial of footrot treatments in Kashmir, India, Kaler, et al. [30] allocated sheep with acute footrot to one of three treatments using STRRS, with the strata being based on each sheep's maximum footrot score. Stratification is useful to ensure that the sample includes individuals which could otherwise be missed by chance in SRS due to the limited number of individuals in their stratum. For example, at a certain period a pig farm in Hong Kong may keep few finisher pigs, but many piglets and sows are present on the farm. With SRS, it is likely that none of the finisher pigs is included in the sample, therefore one can argue that there is error in the representation of the population which could potentially diminish the accuracy of the estimate. For this reason, it is also common to sample a fixed number of individuals in each stratum. Compared to SRS, however, extra information such as the variable used for stratification (membership) must be obtained for all sampling units.

If STRRS has been used, care is required when pooling the information from the strata in order to obtain an unbiased estimator for the finite population mean or prevalence. A “natural” estimator for the mean/prevalence might involve summing up all the observed values in the sample and dividing by the sample size (equivalent to the process of the SRS). However, this estimator is unbiased if the sample size in each of the strata is proportional to the actual size of the stratum—i.e., there has been proportional allocation (this is demonstrated in more detail in Appendix C). The more general common approach to obtain an unbiased estimator for the finite population mean or prevalence follows the two principles we have mentioned: (1) following the actual sampling process and (2) starting with the finite population total. Consider a farm with M animals. A researcher has created J strata based on the ages of the animals. For the j th stratum, there are M_j animals, and clearly $M = \sum_{j=1}^J M_j$. Suppose that m_j animals are sampled using SRS independently from each of the strata and that the value of the variable of interest is denoted as y_{jk} for the k th animal in the j th stratum.

The unbiased estimator (using weight notation) for the finite population total:

$$\hat{\tau} = \sum_{j=1}^J \sum_{k=1}^{m_j} y_{jk} w_{jk}, \tag{8}$$

where w_{jk} is the sampling weight which is the reciprocal of the inclusion probability π_{jk} . For STRRS, this is the probability of the k th animal in the j th stratum being selected. However, writing the estimator in this form is not very intuitive, and it can be rewritten into a different formula in order to provide a more intuitive and meaningful picture for veterinary researchers. As SRS has been implemented within each of the strata, the inclusion probability π_{jk} for the k th animal in the j th stratum is simply the sample size m_j divided by the stratum size M_j , which leads to $w_{jk} = \frac{1}{\pi_{jk}} = \frac{M_j}{m_j}$. Now, Equation (8) can be rewritten as:

$$\hat{\tau} = \sum_{j=1}^J M_j \frac{\sum_{k=1}^{m_j} y_{jk}}{m_j}. \tag{9}$$

This formula says that in order to estimate the finite population total, we need to first compute the mean/prevalence for each of the strata $\bar{y}_j \equiv \hat{p}_j = \frac{\sum_{k=1}^{m_j} y_{jk}}{m_j}$ using the estimator we have seen in SRS and then multiply it by the stratum size M_j to obtain the estimated total for each stratum. We then sum up all these estimated stratum totals to obtain the estimated finite population total. This is consistent with and follows the actual sampling process, as well as producing an unbiased estimator:

$$\mathbb{E}[\hat{\tau}] = \sum_{j=1}^J \mathbb{E} \left[\frac{M_j}{m_j} \sum_{k=1}^{M_j} y_{jk} Z_{jk} \right] = \sum_{j=1}^J \frac{M_j}{m_j} \sum_{k=1}^{M_j} y_{jk} \mathbb{E}[Z_{jk}] = \sum_{j=1}^J \frac{M_j}{m_j} \sum_{k=1}^{M_j} y_{jk} \frac{m_j}{M_j} = \sum_{j=1}^J \sum_{k=1}^{M_j} y_{jk} = \tau,$$

where Z_{jk} is the Bernoulli random variable for selection, representing whether the k th animal in the j th stratum is selected with an inclusion probability π_{jk} , and $\mathbb{E}[Z_{jk}] = \pi_{jk} = \frac{m_j}{M_j}$ due to SRS. Once the estimated total is found, the estimated finite population mean or prevalence is just the total divided by the population size:

$$\bar{y} \equiv \hat{p} = \frac{\sum_{j=1}^J \frac{M_j}{m_j} \sum_{k=1}^{m_j} y_{jk}}{M}. \tag{10}$$

Since each stratum is independently sampled, building on the SRS, the variances for \bar{y} and \hat{p} using STRRS are also straightforward:

$$Var(\bar{y}) = \frac{1}{M^2} Var(\hat{\tau}) = \sum_{j=1}^J \frac{M_j^2}{M^2} \left(1 - \frac{m_j}{M_j} \right) \frac{\sigma_j^2}{m_j}, \tag{11}$$

$$Var(\hat{p}) = \frac{1}{M^2} Var(\hat{\tau}) = \sum_{j=1}^J \frac{M_j^2}{M^2} \left(\frac{M_j - m_j}{M_j - 1} \right) \frac{p_j(1 - p_j)}{m_j}, \tag{12}$$

where both σ_j^2 and p_j are unknown quantities representing the population variance and prevalence in the j th stratum. Similar to the SRS, the estimated variances are obtained by substituting estimated quantities into the unknowns, such as:

$$\widehat{Var}(\bar{y}) = \sum_{j=1}^J \frac{M_j^2}{M^2} \left(1 - \frac{m_j}{M_j} \right) \frac{s_j^2}{m_j}, \tag{13}$$

$$\widehat{Var}(\hat{p}) = \sum_{j=1}^J \frac{M_j^2}{M^2} \left(1 - \frac{m_j}{M_j} \right) \frac{\hat{p}_j(1 - \hat{p}_j)}{m_j - 1}, \tag{14}$$

where s_j^2 is the sample variance of the j th stratum and the formula is given in the SRS section.

To illustrate this, consider an investigation of the seroprevalence of pseudorabies on a farm where STRRS is used. First, pigs are divided into groups based on the five production stages (strata): piglets, weaners, growers, finishers, and sows (breeding herds). The total numbers of pigs in each stratum are 30, 30, 40, 20, and 60, respectively. Within each stratum, a fixed number of pigs (10) are sampled using SRS and the numbers of infected pigs are 5, 6, 3, 2, and 7. The estimated prevalence can then be calculated using Equation (10): $\frac{\frac{30}{10} \times 5 + \frac{30}{10} \times 6 + \frac{40}{10} \times 3 + \frac{20}{10} \times 2 + \frac{60}{10} \times 7}{30 + 30 + 40 + 20 + 60} = 0.506$. The variance of this prevalence estimate can then be estimated using Equation (14). This is carried out stratum by stratum; for example, for the piglets, $\frac{M_1^2}{M^2} \left(1 - \frac{m_1}{M_1}\right) \frac{\hat{p}_1(1-\hat{p}_1)}{m_1-1} = \frac{30^2}{(30+30+40+20+60)^2} \times \left(1 - \frac{10}{30}\right) \times \frac{5}{10} \times \left(1 - \frac{5}{10}\right) \div (10 - 1)$. This process is then repeated for all the strata, and the estimated variance is the sum of the quantities calculated for each stratum. In the example, the final estimated variance is 0.004.

4.3. Cluster Sampling

In this sampling method, the animals in a finite population (animals in a herd, region, or country) are aggregated into larger sampling units: clusters. A cluster is similar to a stratum; however, the sampling process is different. In a cluster sampling procedure, a set of (n) clusters is sampled using SRS from a population with N clusters. These clusters are usually referred to as primary sampling units, and the members within each cluster as secondary sampling units. Within the primary sampling units, all secondary sampling units may be measured or observed (one-stage cluster sampling) or the secondary sampling units may be sampled using SRS (two-stage cluster sampling). The selected individuals within the selected clusters then form a sample of the finite population [26]. In contrast, in STRRS all strata of interest must be included, and SRS is usually used to sample individuals within each stratum. These different sampling strategies mean that the sources of variability in cluster sampling are different from those in STRRS. In STRRS, the variability of the estimated population mean/prevalence arises only from individual variability within a stratum. For cluster sampling, the variability of the estimated population mean/prevalence comes from one or more sources [27]. In one-stage cluster sampling, where all individuals in a selected cluster are included, the variability of the estimated population characteristic or quantity is dependent on the variability between clusters. In two-stage cluster sampling, where only a sub-sample is collected from selected clusters, the variability of the estimated population characteristic comes from two sources: the within- and between-cluster variabilities [31]. One advantage of cluster sampling is that it overcomes some of the logistics issues associated with SRS or STRRS and therefore generally requires less spending on administration and travel expenses. However, the estimates provided by cluster sampling are usually less precise than those provided by SRS, given the same sample size [27].

Cluster sampling is possibly the most widely used approach in livestock research. Usually, a farm or a herd is regarded as a cluster and a number of farms/herds are selected. This was the approach adopted by Getahun, et al. [32], who studied mastitis and antibiotic resistance patterns in dairy cows in central Ethiopia. This design treated a farm as a cluster and a number of farms were chosen using SRS; within each farm, all the dairy cows were sampled. A similar approach was later used to estimate the prevalence of bovine tuberculosis in southern Ethiopia [33]. In this study, the target population was only cows above 6 months of age, and all cows above 6 months old were included on the selected dairy farms (clusters). We list here three examples of two-stage cluster sampling in veterinary research for interested readers [34–36]. In the rest of this section, we will first provide insights into the estimation process for one-stage cluster sampling and do the same for a two-stage cluster sampling where STRRS instead of SRS is used at the second stage (essentially a complex sampling) with details.

4.3.1. One-Stage Cluster Sampling

In one-stage cluster sampling, all animals within a farm are sampled; therefore, the farm total $\tau_i = \sum_{k=1}^{M_i} y_{ik}$ is directly measured, where y_{ik} is the value of the variable of interest measured for the k th animal on the i th farm given the herd size of M_i . Common research tasks might be to estimate the farm-level and animal-level averages, such as the average milk production or average number of positive animals per farm and average milk production per cow or overall prevalence at the animal level. Suppose n farms are sampled from N farms in a region using SRS. As before, to estimate the population mean or prevalence it is always recommended to start by estimating the total. Since SRS is used for sampling clusters, the unbiased estimator for the finite population total (e.g., the number of all diseased dairy cows in a region) is straightforward and therefore given without proof:

$$\hat{\tau} = \sum_{i=1}^n w_i \tau_i = \frac{N}{n} \sum_{i=1}^n \tau_i. \tag{15}$$

The variance and estimated variance for this estimator can also be straightforwardly determined by applying the theory introduced in the SRS section: $Var(\hat{\tau}) = N^2(1 - \frac{n}{N}) \frac{\sigma_\tau^2}{n}$ and $\widehat{Var}(\hat{\tau}) = N^2(1 - \frac{n}{N}) \frac{s_\tau^2}{n}$, where σ_τ^2 and s_τ^2 are the finite population variance and sample variance (at the farm level), such that $\sigma_\tau^2 = \frac{1}{N-1} \sum_{i=1}^N (\tau_i - \frac{\tau}{N})^2$ and $s_\tau^2 = \frac{1}{n-1} \sum_{i=1}^n (\tau_i - \frac{\hat{\tau}}{N})^2$. The estimated farm-level average and its corresponding variance and estimated variance are straightforward:

$$\bar{y}_F = \frac{\hat{\tau}}{N} = \frac{1}{n} \sum_{i=1}^n \tau_i, \tag{16}$$

$$Var(\bar{y}_F) = \left(1 - \frac{n}{N}\right) \frac{\sigma_\tau^2}{n}, \tag{17}$$

$$\widehat{Var}(\bar{y}_F) = \left(1 - \frac{n}{N}\right) \frac{s_\tau^2}{n}. \tag{18}$$

The total number of animals in the region is $M = \sum_{i=1}^N M_i$. Hence, the estimated average at the cow level is given by:

$$\bar{y} \equiv \hat{p} = \frac{\hat{\tau}}{M}. \tag{19}$$

The variances and estimated variances for the cow-level average or overall prevalence are given as:

$$Var(\bar{y}) \equiv Var(\hat{p}) = \frac{N^2}{M^2} \left(1 - \frac{n}{N}\right) \frac{\sigma_\tau^2}{n}, \tag{20}$$

$$\widehat{Var}(\bar{y}) \equiv \widehat{Var}(\hat{p}) = \frac{N^2}{M^2} \left(1 - \frac{n}{N}\right) \frac{s_\tau^2}{n}. \tag{21}$$

Note that at the farm level, we work on counts of positive animals instead of binary values even if we are estimating prevalence, therefore the variance formulas for \bar{y} and \hat{p} are indistinguishable.

4.3.2. Two-Stage Cluster Sampling

The main purpose of this section is to illustrate the estimation process for a complex survey—i.e., how to obtain the unbiased estimators and derive their corresponding variances. Suppose there are M dairy cows in a region with N dairy herds. The herd size for herd i is M_i . The cows are separately managed based on a certain criterion; that is, within the i th herd there are J groups, and within each of the groups there are M_{ij} cows. The groups can be treated as strata, as they are not overlapping and constitute the entire herd. A research team is interested in knowing the prevalence of a disease among cows in this region. Based on the demographic information, a two-stage cluster sampling is decided. First, n herds will be selected using SRS. Within each of the sampled herds, STRRS will

be used to sample cows from each of the strata in each of the herds. Before going to the estimation process, we shall define some notations (Table 1).

Table 1. Quantities used in a two-stage cluster sampling design, where stratified random sampling is implemented in the second stage.

N	The number of dairy herds in the region.
n	The number of dairy herds in the sample.
M_{ij}	The number of cows in the j th stratum in the i th herd.
m_{ij}	The sample size in the j th stratum in the i th herd.
M_i	The number of cows in the i th herd (herd size for herd i), $M_i = \sum_{j=1}^J M_{ij}$.
m_i	The sample size for herd i .
M	The total number of cows in the region, $M = \sum_{i=1}^N M_i$.
y_{ijk}	The disease outcome (1/0) of the k th cow in the the j th stratum in the i th herd.
τ_i	The total number of diseased cows in the i th herd, $\tau_i = \sum_{j=1}^J \sum_{k=1}^{M_{ij}} y_{ijk}$.
p_i	The herd prevalence for the i th herd, $p_i = \frac{\tau_i}{M_i}$.
τ	The total number of diseased cows in the region, $\tau = \sum_{i=1}^N \sum_{j=1}^J \sum_{k=1}^{M_{ij}} y_{ijk}$.
p	The overall prevalence in the region, $p = \frac{\tau}{M}$.

The ultimate goal for this sample survey is to estimate p ; however, as in the previous examples it is the best to start by estimating the total τ . Additionally, the computation process needs to be consistent with the actual sampling steps. Thus, we start by estimating the total diseased animals in the j th stratum in the i th herd. Within each stratum, SRS is used, therefore the estimated total can be computed based on Equation (1). The second step is to estimate the total diseased animals in the i th herd. Because we used STRRS, this can be achieved by adopting Equation (9). Finally, we can estimate the total number of diseased animals in the region by using Equation (15), as SRS is used to select herds. Hence, the unbiased estimated region total is computed in the following way:

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^n \sum_{j=1}^J \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} y_{ijk} \tag{22}$$

To prove that the outcome of this process is unbiased, we simplify the notation, letting $\hat{\tau}_i = \sum_{j=1}^J \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} y_{ijk}$. We know that $\hat{\tau}_i$ is unbiased (namely, $E[\hat{\tau}_i] = \tau_i$), because we have used STRRS. Secondly, we specify a binary indicator variable Z_i that = 1 if herd i is selected or 0 if it is not. Let π_i denote the probability that herd i is selected (inclusion probability of a herd); we then have $\pi_i \equiv P(Z_i = 1) = E[Z_i] = \frac{n}{N}$, since SRS is used for the first stage of selection (i.e., the selection of herds). Given that sampling within any herd is independent of the sampling in any other herd and that $\hat{\tau}_i$ is independent of $\mathbf{Z} = (Z_1, Z_2, \dots, Z_N)$, we have:

$$\begin{aligned} E[\hat{\tau}] &= E[E[\hat{\tau}|\mathbf{Z}]] = E\left[E\left[\sum_{i=1}^n \frac{N}{n} \hat{\tau}_i \mid \mathbf{Z}\right]\right] = E\left[E\left[\sum_{i=1}^N \frac{N}{n} Z_i \hat{\tau}_i \mid \mathbf{Z}\right]\right] \text{ (partition theorem for expectations)} \\ &= E\left[\sum_{i=1}^N \frac{N}{n} E\left[\frac{N}{n} Z_i \hat{\tau}_i \mid \mathbf{Z}\right]\right] \text{ (the conditional expectation of a sum is the sum of the conditional expectations)} \\ &= E\left[\sum_{i=1}^N \frac{N}{n} E[Z_i \hat{\tau}_i \mid \mathbf{Z}]\right] \text{ (expectation is a linear operator and } \frac{N}{n} \text{ is a constant)} \\ &= E\left[\sum_{i=1}^N \frac{N}{n} Z_i E[\hat{\tau}_i \mid \mathbf{Z}]\right] \text{ (knowing a vector means the same as knowing every element of the vector; conditional on the selection status of every herd means knowing the selection status of any herd)} \end{aligned}$$

$$\begin{aligned}
 &= \mathbb{E} \left[\sum_{i=1}^N \frac{N}{n} Z_i \mathbb{E}[\hat{\tau}_i] \right] \text{ (}\hat{\tau}_i \text{ and } \mathbf{Z} \text{ are independent)} \\
 &= \mathbb{E} \left[\sum_{i=1}^N \frac{N}{n} Z_i \tau_i \right] \text{ (unbiased estimator for stratified random sampling for each herd)} \\
 &= \sum_{i=1}^N \frac{N}{n} \tau_i \mathbb{E}[Z_i] = \sum_{i=1}^N \frac{N}{n} \tau_i \frac{n}{N} = \sum_{i=1}^N \tau_i = \tau \text{ (linear property of expectation).}
 \end{aligned}$$

Therefore, the unbiased estimator for the overall prevalence is simply:

$$\hat{p} = \frac{\hat{\tau}}{M} = \frac{N}{nM} \sum_{i=1}^N \sum_{j=1}^J \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} y_{ijk}. \tag{23}$$

In order to find the variance formula for \hat{p} , it is easier to start with $\hat{\tau}$. It is necessary to first identify the sources of variability. In this two-stage cluster sampling process, we have between- and within-herd variances. The variance partition formula can thus decompose the total variance into two parts: $Var(\hat{\tau}) = Var(\mathbb{E}[\hat{\tau}|\mathbf{Z}]) + \mathbb{E}[Var(\hat{\tau}|\mathbf{Z})]$, where $Var(\mathbb{E}[\hat{\tau}|\mathbf{Z}])$ measures the variability between herds and $\mathbb{E}[Var(\hat{\tau}|\mathbf{Z})]$ measures the variability within a herd. Since SRS is implemented at the herd level, according to Equation (3), we have:

$$Var(\mathbb{E}[\hat{\tau}|\mathbf{Z}]) = Var \left(\mathbb{E} \left[\sum_{i=1}^N \frac{N}{n} Z_i \hat{\tau}_i \mid \mathbf{Z} \right] \right) = Var \left(\sum_{i=1}^N \frac{N}{n} Z_i \tau_i \right) = N^2 \left(1 - \frac{n}{N} \right) \frac{\sigma_\tau^2}{n},$$

where $\sigma_\tau^2 = \frac{1}{N-1} \sum_{i=1}^N (\tau_i - \frac{\tau}{N})^2$. This part of the variance is the same as that of one-stage cluster sampling, since the herd sampling procedures are exactly the same. The detailed derivation is essentially the same as the derivation of variance in SRS (see Appendix B).

For the within-herd component of the variance, $\mathbb{E}[Var(\hat{\tau}|\mathbf{Z})]$, the formula inside the expectation operator, $Var(\hat{\tau}|\mathbf{Z}) = \mathbb{E}[\hat{\tau}^2|\mathbf{Z}] - \mathbb{E}[\hat{\tau}|\mathbf{Z}]^2$ according to the conditional variance formula. The detailed mathematical derivation is available in Appendices D and E provides the statistical theorems required in this paper. Here, we only give an essential intermediate result: $\mathbb{E}[Var(\hat{\tau}|\mathbf{Z})] = \frac{N}{n} \sum_{i=1}^N Var(\hat{\tau}_i)$. Since STRRS is implemented within each herd, $Var(\hat{\tau}_i)$ can be easily obtained from Equation (11) or Equation (12) depending on the nature of y_{ijk} . In our particular example, where y_{ijk} takes a binary value (either 1 or 0), we have $Var(\hat{\tau}_i) = \sum_{j=1}^J \frac{M_{ij}^2}{m_{ij}} \left(1 - \frac{m_{ij}}{M_{ij}} \right) \frac{M_{ij} p_{ij} (1 - p_{ij})}{M_{ij} - 1}$. Generally, $Var(\hat{\tau}_i) = \sum_{j=1}^J \frac{M_{ij}^2}{m_{ij}} \left(1 - \frac{m_{ij}}{M_{ij}} \right) \sigma_{ij}^2$.

Therefore, the general formula for the variance of $\hat{\tau}$ is given as:

$$Var(\hat{\tau}) = \frac{N^2}{n} \left(1 - \frac{n}{N} \right) \sigma_\tau^2 + \frac{N}{n} \sum_{i=1}^N \sum_{j=1}^J \frac{M_{ij}^2}{m_{ij}} \left(1 - \frac{m_{ij}}{M_{ij}} \right) \sigma_{ij}^2, \tag{24}$$

where $\sigma_{ij}^2 = \left(\frac{1}{M_{ij}-1} \right) \sum_{k=1}^{M_{ij}} (y_{ijk} - \mu_{ij})^2$, μ_{ij} is the unknown mean of the j th stratum in the i th herd. When y_{ijk} takes a binary value, the special form is given by applying the method introduced in the SRS section (see Equation (5)):

$$Var(\hat{\tau}) = \frac{N^2}{n} \left(1 - \frac{n}{N} \right) \frac{1}{N-1} \sum_{i=1}^N \left(\tau_i - \frac{\tau}{N} \right)^2 + \frac{N}{n} \sum_{i=1}^N \sum_{j=1}^J \frac{M_{ij}^2}{m_{ij}} \left(1 - \frac{m_{ij}}{M_{ij}} \right) \frac{M_{ij} p_{ij} (1 - p_{ij})}{M_{ij} - 1}. \tag{25}$$

Again, this variance depends on some unknown quantities which we have estimated. These estimates can then be used to replace these unknown quantities, as we have done previously. Thus, the estimated variance (general form) will be:

$$\widehat{Var}(\hat{\tau}) = \frac{N^2}{n} \left(1 - \frac{n}{N} \right) \hat{\sigma}_\tau^2 + \frac{N}{n} \sum_{i=1}^N \sum_{j=1}^J \frac{M_{ij}^2}{m_{ij}} \left(1 - \frac{m_{ij}}{M_{ij}} \right) s_{ij}^2, \tag{26}$$

where $\hat{\sigma}_\tau^2 = \frac{1}{n-1} \sum_{i=1}^n \left(\hat{\tau}_i - \frac{\hat{\tau}}{N} \right)^2$. Note the difference between $\hat{\sigma}_\tau^2$ and s_τ^2 in the one-stage cluster sampling; $s_{ij}^2 = \left(\frac{1}{m_{ij}-1} \right) \sum_{k=1}^{m_{ij}} (y_{ijk} - \bar{y}_{ij})^2$ is the sample variance within the j th

stratum in the i th herd, with $\bar{y}_{ij} = \frac{1}{m_{ij}} \sum_{k=1}^{m_{ij}} y_{ijk}$ being the estimated sample mean. When y_{ijk} takes a binary value, the special form is given as:

$$\widehat{Var}(\hat{\tau}) = \frac{N^2}{n} \left(1 - \frac{n}{N}\right) \left(\frac{1}{n-1}\right) \sum_{i=1}^n \left(\hat{\tau}_i - \frac{\hat{\tau}}{N}\right)^2 + \frac{N}{n} \sum_{i=1}^n \sum_{j=1}^J M_{ij}^2 \left(1 - \frac{m_{ij}}{M_{ij}}\right) \frac{\hat{p}_{ij}(1 - \hat{p}_{ij})}{m_{ij} - 1}. \tag{27}$$

Finally, the variance and estimated variance for \hat{p} are found simply by multiplying the results of Equations (25) and (27) by a constant $\frac{1}{M^2}$. The same process can be applied to find the variance and estimated variance for \bar{y} when y_{ijk} is not limited to binary values. A numerical illustration example in this design would be tedious to present manually; we have therefore provided the Python code for computation (see the Supplementary Materials: Python code for the two-stage cluster sampling where stratification is implemented within the clusters).

5. Sample Size Consideration

Although the paper has focused principally on the estimation of outcomes of interest, sample size calculations are also a critical part of the study design process. For ready-to-use sample size calculation formulas, readers are directed to Stevenson 2021 [37]. However, for a complex survey where the formula needs to be derived on a case-by-case basis, it is of value to briefly introduce the principles behind the sample size calculation. The formulas for sample size calculations are closely related to the sampling distributions of the estimators. The investigator needs to come up with an expected value for the finite population characteristic of interest and then think about how precise the estimate needs to be. The narrower the sampling distribution of an estimator, the more precise the estimate needs to be (and thus the larger the sample size). Therefore, it is natural to think what the sampling distribution is and which quantity defines the spread of the distribution.

Here, we use the SRS as an example, as the SRS serves as the theoretical basis for other more complicated sampling methods. Suppose that in the SRS for estimating prevalence in a finite herd, the sampling distribution for \hat{p} is approximately normal, with mean = p and variance = $\left(\frac{M-m}{M-1}\right) \frac{p(1-p)}{m}$ (see Equation (5) of this paper for the variance formula) [27]. Note that for sample size calculations as opposed to calculating the variance of an estimator from the empirical data, we use the theoretical variance formula instead of the estimated variance formula. Clearly, the variance determines the spread of a distribution and it is a function of the sample size m ; thus, this is the equation we are targeting.

The investigator then needs to specify the expected prevalence $p = p_0$ and think about the quantiles ($q_{\alpha/2}$ and $q_{1-\alpha/2}$) of this sampling distribution at the $1 - \alpha$ confidence level (usually, we set $1 - \alpha = 0.95$). These quantiles can be interpreted as the farthest acceptable estimates from the expected prevalence, and $q_{\alpha/2} = -q_{1-\alpha/2}$ due to the symmetry of the normal density curve. This suggests that $P(-q_{1-\alpha/2} \leq \hat{p} \leq q_{1-\alpha/2}) = 1 - \alpha$. Let SE denote the square root of the variance in this specific example. Standardisation gives $\frac{\hat{p}-p_0}{SE} \sim N(0, 1)$; thus, $\pm \frac{q_{1-\alpha/2}-p_0}{SE} = \pm z_{1-\alpha/2}$, where $\pm z_{1-\alpha/2}$ are the quantiles of standard normal distribution that we know—for example, $\pm z_{1-\alpha/2} \approx \pm 1.96$ for 95% confidence. We thus need to find a sample size value which makes $z_{1-\alpha/2}SE$ equal to the absolute difference (d) between the farthest acceptable estimate and the expected prevalence. Now, it is just a matter of solving this equation. With a little algebra, the sample size is:

$$m = \frac{z_{1-\alpha/2}^2 M p_0 (1 - p_0)}{(M - 1) d^2 + z_{1-\alpha/2}^2 p_0 (1 - p_0)}. \tag{28}$$

6. Conclusions

This paper has provided a brief overview of the principles of probability sampling and the design-based approach for estimating finite population characteristics. In addition, we summarised the analytical methods for various commonly used sampling methods in detail. Instead of feeding the formulas to the readers, we have attempted to introduce and illustrate the ideas to help the readers understand, interpret, derive, and prove the unbiased

estimators of the design-based samples and their corresponding variance formulas. We hope the ideas and methods presented in this paper can inspire the readers, so that the readers are encouraged to find the proper estimators and corresponding variances in their own sample surveys.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/vetsci8060105/s1>, File S1: Python code for the two-stage cluster sampling where stratification is implemented within the clusters.

Author Contributions: Writing—original draft preparation, D.A.Y.; writing—review and editing, R.A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analysed in this study. Data sharing is not applicable to this article.

Acknowledgments: We acknowledge the technical supports offered by Geoff Jones at Massey University, New Zealand and Ciprian Giurcaneanu at The University of Auckland, New Zealand.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

First: we consider the simple random sampling scenario. In a population with M animals, let a Bernoulli random variable Z_k denote whether animal k is selected. $Z_k = 1$ if the individual k is selected; otherwise, $Z_k = 0$. Let $\pi_k \equiv P(Z_k = 1)$ be the inclusion probability that animal k is selected. Claim that $\pi_k \equiv P(Z_k = 1) = \frac{m}{M}$ for a sample with a fixed sample size m . We shall apply the principle introduced in the section “Overview of probability sampling” to prove. For a sample with a fixed sample size m , there are in total $\binom{m}{M}$ ways to obtain a sample. In the simple random sampling, all possible samples of a given size have the same probability of being selected—that is, $P(S) = \frac{1}{\binom{m}{M}} = \frac{m!(M-m)!}{M!}$.

The inclusion probability π_k is calculated by summing the probabilities of all samples including the animal k —i.e., $\pi_k = \sum_{S:k \in S} P(S)$. There are $\binom{m-1}{M-1}$ scenarios that a sample contains the animal k . As once the animal k is fixed, we need to choose another $m-1$ animal from the rest $M-1$ animals to form a sample. We then just need to add $\binom{m-1}{M-1} P(S)$ s up to obtain π_k . Therefore, $\pi_k = \binom{m-1}{M-1} \frac{m!(M-m)!}{M!} = \frac{(M-1)!}{(m-1)!(M-m)!} \frac{m!(M-m)!}{M!} = \frac{m}{M}$.

Building on the idea of the simple random sampling, the inclusion probability for stratified random sampling is easy to derive. Let $Z_{jk} \sim \text{Bern}(\pi_{jk})$ represent whether the k th animal in the j th stratum is selected, where π_{jk} is the inclusion probability. For the j th stratum, $\pi_{jk} = \frac{m_j}{M_j}$, as simple random sampling is implemented within the stratum.

For the cluster sampling, let $Z_{ik} \sim \text{Bern}(\pi_{ik})$ represent whether the k th animal on the i th farm is selected. Let random variable $Z_i \sim \text{Bern}(\pi_i)$ represent whether the i th farm is selected. Given a one-stage cluster sampling scenario, if the farm is selected, all animals within the farm must be selected—namely, $P(Z_{ik} = 1 | Z_i = 1) = 1$. As simple random sampling is used to select farms, $\pi_i \equiv P(Z_i = 1) = \frac{n_i}{N}$. By the law of total probability, $\pi_{ik} \equiv P(Z_{ik} = 1) = P(Z_{ik} = 1 | Z_i = 1)P(Z_i = 1) + P(Z_{ik} = 1 | Z_i = 0)P(Z_i = 0)$. If the farm is not selected, the probability of the animal on this farm being selected is 0, namely, $P(Z_{ik} = 1 | Z_i = 0) = 0$, therefore $\pi_{ik} = \frac{n_i}{N}$. The inclusion probability for the two-stage

cluster sampling is similar; the difference is $P(Z_{ik} = 1|Z_i = 1) \neq 1$, but this depends on the sampling method used within a farm.

Appendix B

The well-known results for the variance formulas in the simple random sampling will be derived here. For a finite population with size M , we have Z_1, Z_2, \dots, Z_M Bernoulli random variables representing the selection of the animals. These are identically distributed with the same marginal distribution, such that $Z_k \sim \text{Bern}(\pi_k)$; however, they are not independent, as $P(Z_k = 1|Z_l = 1) \neq P(Z_k = 1)$, where $k \neq l$. The probability of the second animal being selected depends on the first animal's selection result. For a Bernoulli random variable, we know that $\mathbb{E}[Z_k] = P(Z_k = 1) = \pi_k = \frac{m}{M}$ if the sample size is m ; $\text{Var}(Z_k) = \frac{m}{M}(1 - \frac{m}{M})$. Therefore, $\mathbb{E}[Z_k^2] = \text{Var}(Z_k) + \mathbb{E}[Z_k]^2 = \frac{m}{M}$. If $k \neq l$, the probability that both animals are selected is $P(Z_k = 1, Z_l = 1) = P(Z_l = 1)P(Z_k = 1|Z_l = 1) = \frac{m}{M} \frac{m-1}{M-1}$, and $\mathbb{E}[Z_k Z_l] = \sum_{Z_k} \sum_{Z_l} z_k z_l P(Z_k = z_k, Z_l = z_l) = P(Z_k = 1, Z_l = 1) = \frac{m}{M} \frac{m-1}{M-1}$. Therefore, the covariance of Z_k and Z_l is computed as:

$$\text{Cov}(Z_k, Z_l) = \mathbb{E}[Z_k Z_l] - \mathbb{E}[Z_k]\mathbb{E}[Z_l] = \frac{m}{M} \frac{m-1}{M-1} - \frac{m}{M} \frac{m}{M} = -\left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \frac{m}{M}.$$

Now the variance of the estimated finite population total $\hat{\tau}$ shall be derived as:

$$\begin{aligned} \text{Var}(\hat{\tau}) &= \text{Var}\left(\frac{M}{m} \sum_{k=1}^m y_k\right) = \frac{M^2}{m^2} \text{Var}\left(\sum_{k=1}^m y_k Z_k\right) = \frac{M^2}{m^2} \text{Cov}\left(\sum_{k=1}^m y_k Z_k, \sum_{l=1}^m y_l Z_l\right) = \frac{M^2}{m^2} \sum_{k=1}^m \sum_{l=1}^m y_k y_l \text{Cov}(Z_k, Z_l) \\ &= \frac{M^2}{m^2} \left\{ \sum_{k=1}^m y_k^2 \text{Var}(Z_k) + \sum_{k=1}^m \sum_{l \neq k}^m y_k y_l \text{Cov}(Z_k, Z_l) \right\} \\ &= \frac{M^2}{m^2} \left\{ \sum_{k=1}^m y_k^2 \frac{m}{M} \left(1 - \frac{m}{M}\right) - \sum_{k=1}^m \sum_{l \neq k}^m y_k y_l \frac{m}{M} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \right\} \\ &= \frac{M^2}{m^2} \frac{m}{M} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ (M-1) \sum_{k=1}^m y_k^2 - \sum_{k=1}^m \sum_{l \neq k}^m y_k y_l \right\} \\ &= \frac{M}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ (M-1) \sum_{k=1}^m y_k^2 - \left[\left(\sum_{k=1}^m y_k\right)^2 - \sum_{k=1}^m y_k^2 \right] \right\} = \frac{M}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ M \sum_{k=1}^m y_k^2 - \left(\sum_{k=1}^m y_k\right)^2 \right\} \\ &= \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m y_k^2 - \frac{\left(\sum_{k=1}^m y_k\right)^2}{M} \right\} \\ &= \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m y_k^2 - M \left(\frac{\sum_{k=1}^m y_k}{M}\right)^2 \right\} = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m y_k^2 - M\mu^2 \right\} = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m y_k^2 - 2M\mu^2 + M\mu^2 \right\} \\ &= \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m y_k^2 - 2\mu M\mu + M\mu^2 \right\} = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m y_k^2 - 2\mu \sum_{k=1}^m y_k + \sum_{k=1}^m \mu^2 \right\} \\ &= \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m y_k^2 - 2\mu y_k + \mu^2 \right\} = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \left(\frac{1}{M-1}\right) \left\{ \sum_{k=1}^m (y_k - \mu)^2 \right\}. \end{aligned}$$

Observe that $\frac{\sum_{k=1}^m (y_k - \mu)^2}{M-1}$ is the finite population variance σ^2 , therefore:

$$\text{Var}(\hat{\tau}) = \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \sigma^2,$$

which is the Equation (3) shown in the main body of the text. In our particular case, y_k is binary valued, we shall let $\mu \equiv p$ and $\bar{y} \equiv \hat{p}$, therefore:

$$\begin{aligned} \sigma^2 &= \frac{1}{M-1} \sum_{k=1}^M (y_k - p)^2 = \frac{1}{M-1} \sum_{k=1}^M (y_k^2 - 2py_k + p^2) = \frac{1}{M-1} \sum_{k=1}^M y_k^2 - 2p \sum_{k=1}^M y_k + \sum_{k=1}^M p^2 = \frac{Mp - 2pMp + Mp^2}{M-1} \\ &= \frac{M(p-p^2)}{M-1} = \frac{Mp(1-p)}{M-1}. \end{aligned}$$

and the variance of the estimated finite population prevalence are:

$$\text{Var}(\hat{p}) \equiv \text{Var}(\bar{y}) = \frac{1}{M^2} \text{Var}(\hat{\tau}) = \frac{1}{M^2} \frac{M^2}{m} \left(1 - \frac{m}{M}\right) \sigma^2 = \left(1 - \frac{m}{M}\right) \frac{1}{m} \frac{Mp(1-p)}{M-1} = \left(\frac{M-m}{M}\right) \frac{1}{m} \frac{Mp(1-p)}{M-1} = \left(\frac{M-m}{M-1}\right) \frac{p(1-p)}{m},$$

which is Equation (5) shown in the main body of the text. Since the variance depends on the unknown quantity p , we shall replace the finite population variance σ^2 with the sample variance $s^2 = \frac{\sum_{k=1}^m (y_k - \hat{p})^2}{m-1}$. With the same algebra, $s^2 = \frac{m\hat{p}(1-\hat{p})}{m-1}$. Therefore, the estimated variance is:

$$\widehat{Var}(\hat{p}) = \frac{1}{M^2} \widehat{Var}(\hat{\tau}) = \frac{1}{M^2} \frac{M^2}{m} \left(1 - \frac{m}{M}\right) s^2 = \frac{1}{m} \left(1 - \frac{m}{M}\right) \frac{m\hat{p}(1-\hat{p})}{m-1} = \left(1 - \frac{m}{M}\right) \frac{\hat{p}(1-\hat{p})}{m-1},$$

which is the Equation (7) shown in the main body of the text.

Appendix C

In the stratified random sampling section, we mentioned a “natural” estimator for the prevalence on a farm—that is:

$$\check{p} = \frac{1}{m} \sum_{j=1}^J \sum_{k=1}^{m_j} y_{jk}.$$

This sums up all the observed values in the sample and divides by the sample size. This estimator is generally biased; to see why, let us suppose that we sampled the entire herd (herd size = M):

$$\begin{aligned} p &= \frac{1}{M} \sum_{j=1}^J \sum_{k=1}^{M_j} y_{jk} = \frac{1}{M} \sum_{j=1}^J M_j p_j \\ \mathbb{E}[\check{p}] &= \frac{1}{m} \sum_{j=1}^J \mathbb{E} \left[\sum_{k=1}^{M_j} Z_{ij} y_{jk} \right] = \frac{1}{m} \sum_{j=1}^J \sum_{k=1}^{M_j} y_{jk} \mathbb{E} [Z_{jk}] \\ &= \frac{1}{m} \sum_{j=1}^J M_j p_j \frac{m_j}{M_j} = \frac{1}{m} \sum_{j=1}^J m_j p_j \neq \frac{1}{M} \sum_{j=1}^J M_j p_j = p \\ \therefore \mathbb{E}[\check{p}] - p &\neq 0 \implies \check{p} \text{ is a biased estimator.} \end{aligned}$$

However, under the proportional allocation, the sample size in each of the strata is taken in proportion to the actual size of the corresponding stratum, that is $m_j = CM_j$, then $\sum_{j=1}^J m_j = C \sum_{j=1}^J M_j \implies m = CM$. With a little algebra, we can see that $\frac{m_j}{m} = \frac{M_j}{M}$; therefore, $\mathbb{E}[\check{p}] = p$.

Appendix D

In the two-stage cluster sampling, the variance of the estimated total comes from two sources. Based on the variance partition formula, $Var(\hat{\tau}) = Var(\mathbb{E}[\hat{\tau}|\mathbf{Z}]) + \mathbb{E}[Var(\hat{\tau}|\mathbf{Z})]$, where $Var(\mathbb{E}[\hat{\tau}|\mathbf{Z}])$ measures the variability between herds and $\mathbb{E}[Var(\hat{\tau}|\mathbf{Z})]$ measures the variability within a herd. Since simple random sampling is implemented at the herd level, $Var(\mathbb{E}[\hat{\tau}|\mathbf{Z}])$ is relatively easier to derive by following the method shown in the simple random sampling scenario.

Here, we focus on deriving the variance formula for the within-herd component $\mathbb{E}[Var(\hat{\tau}|\mathbf{Z})]$. We apply the conditional variance formula inside the expectation operator: $\mathbb{E}[Var(\hat{\tau}|\mathbf{Z})] = \mathbb{E} \left[\mathbb{E}[\hat{\tau}^2|\mathbf{Z}] - \mathbb{E}[\hat{\tau}|\mathbf{Z}]^2 \right]$. We will work on the two terms inside the expectation operator, then apply the expectation. A gentle courtesy here is to explain the notations again. A Bernoulli random variable $Z_i \sim Bern(\pi_i)$ represents whether the herd i is selected, and we have $\pi_i \equiv P(Z_i = 1) = \mathbb{E}[Z_i] = \mathbb{E}[Z_i^2] = \frac{n}{N}$. Sampling within any herd is independent of the sampling in any other herd and $\hat{\tau}_i$ is independent of $\mathbf{Z} = (Z_1, Z_2, \dots, Z_N)$.

$$\begin{aligned} Var(\hat{\tau}|\mathbf{Z}) &= \mathbb{E}[\hat{\tau}^2|\mathbf{Z}] - \mathbb{E}[\hat{\tau}|\mathbf{Z}]^2 = \mathbb{E} \left[\left(\sum_{i=1}^N \frac{N}{n} Z_i \hat{\tau}_i \right)^2 \middle| \mathbf{Z} \right] - \left(\sum_{i=1}^N \frac{N}{n} Z_i \tau_i \right)^2 \\ &= \mathbb{E} \left[\sum_{i=1}^N \frac{N^2}{n^2} Z_i^2 \hat{\tau}_i^2 + \sum_{i=1}^N \sum_{h \neq i}^N \frac{N^2}{n^2} \hat{\tau}_i \hat{\tau}_h Z_i Z_h \middle| \mathbf{Z} \right] - \left(\sum_{i=1}^N \frac{N}{n} Z_i \tau_i \right)^2 \\ &= \frac{N^2}{n^2} \sum_{i=1}^N \mathbb{E} [Z_i^2 \hat{\tau}_i^2 | \mathbf{Z}] + \frac{N^2}{n^2} \sum_{i=1}^N \sum_{h \neq i}^N \mathbb{E} [\hat{\tau}_i \hat{\tau}_h Z_i Z_h | \mathbf{Z}] - \left(\sum_{i=1}^N \frac{N}{n} Z_i \tau_i \right)^2 \\ &= \frac{N^2}{n^2} \left(\sum_{i=1}^N Z_i^2 \mathbb{E} [\hat{\tau}_i^2 | \mathbf{Z}] + \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \mathbb{E} [\hat{\tau}_i \hat{\tau}_h | \mathbf{Z}] - \sum_{i=1}^N Z_i^2 \tau_i^2 - \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \tau_i \tau_h \right) \\ &= \frac{N^2}{n^2} \left(\sum_{i=1}^N Z_i^2 \mathbb{E} [\hat{\tau}_i^2] + \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \mathbb{E} [\hat{\tau}_i \hat{\tau}_h] - \sum_{i=1}^N Z_i^2 \tau_i^2 - \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \tau_i \tau_h \right) \end{aligned}$$

($\hat{\tau}_i$ and \mathbf{Z} are independent random variables, real valued functions f and g defined for $\hat{\tau}_i$ and \mathbf{Z} are also independent random variables; this justifies $\mathbb{E}[\hat{\tau}_i^2|\mathbf{Z}] = \mathbb{E}[\hat{\tau}_i^2]$ as f and g are a quadratic and an identity functions, respectively; more generally $\hat{\tau}_i$, $\hat{\tau}_h$ and \mathbf{Z} are

independent, thus $f(\hat{\tau}_i, \hat{\tau}_h)$ and $g(\mathbf{Z})$ are independent; this justifies $\mathbb{E}[\hat{\tau}_i \hat{\tau}_h | \mathbf{Z}] = \mathbb{E}[\hat{\tau}_i \hat{\tau}_h]$ as $f(\hat{\tau}_i, \hat{\tau}_h) = \hat{\tau}_i \hat{\tau}_h$ and $g(\mathbf{Z}) = \mathbf{Z}$

$$= \frac{N^2}{n^2} \left(\sum_{i=1}^N Z_i^2 \mathbb{E}[\hat{\tau}_i^2] + \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \mathbb{E}[\hat{\tau}_i] \mathbb{E}[\hat{\tau}_h] - \sum_{i=1}^N Z_i^2 \tau_i^2 - \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \tau_i \tau_h \right)$$

(sampling independently from the farms implies independence between $\hat{\tau}_i$ and $\hat{\tau}_h$, for two independent random variables, $\mathbb{E}[\hat{\tau}_i \hat{\tau}_h] = \mathbb{E}[\hat{\tau}_i] \mathbb{E}[\hat{\tau}_h]$)

$$= \frac{N^2}{n^2} \left(\sum_{i=1}^N Z_i^2 \left\{ \mathbb{E}[\hat{\tau}_i^2] + \text{Var}(\hat{\tau}_i) \right\} + \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \tau_i \tau_h - \sum_{i=1}^N Z_i^2 \tau_i^2 - \sum_{i=1}^N \sum_{h \neq i}^N Z_i Z_h \tau_i \tau_h \right)$$

$$= \frac{N^2}{n^2} \left(\sum_{i=1}^N Z_i^2 \tau_i^2 + \sum_{i=1}^N Z_i^2 \text{Var}(\hat{\tau}_i) - \sum_{i=1}^N Z_i^2 \tau_i^2 \right) = \frac{N^2}{n^2} \sum_{i=1}^N Z_i^2 \text{Var}(\hat{\tau}_i).$$

$$\therefore \mathbb{E}[\text{Var}(\hat{\tau} | \mathbf{Z})] = \mathbb{E} \left[\frac{N^2}{n^2} \sum_{i=1}^N Z_i^2 \text{Var}(\hat{\tau}_i) \right] = \frac{N^2}{n^2} \sum_{i=1}^N \mathbb{E}[Z_i^2] \text{Var}(\hat{\tau}_i) = \frac{N}{n} \sum_{i=1}^N \text{Var}(\hat{\tau}_i)$$

(sampling independently from the farms).

Appendix E

1. Joint probability: $P(A, B) = P(A|B)P(B)$,
2. Independence of two events: $P(A|B) = P(A)$,
3. Law of total probability: $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$,
4. Expectation of a discrete random variable: $\mathbb{E}[X] = \sum_x x f(x)$,
5. Property of expectation: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$,
6. Variance of a random variable: $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
7. Property of variance:
 - $\text{Var}(aX + b) = a^2 \text{Var}(X)$,
 - If X_1, X_2, \dots, X_n are mutually independent, then $\text{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \text{Var}(X_i)$,
8. Bias: $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$, unbiasedness implies $\text{Bias}(\hat{\theta}) = 0$,
9. Expectation for a function of two random variables: $\mathbb{E}[g(X, Y)] = \sum_x \sum_y g(x, y) f(x, y)$,
10. Covariance: $\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$,
11. Properties of covariance:
 - Independence between X and $Y \implies \text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$,
 - $\text{Var}(X) = \text{Cov}(X, X)$,
 - $\text{Cov}(aX + b, cX + d) = ac \text{Cov}(X, Y)$,
 - $\text{Cov}(\sum_{i=1}^n X_i, \sum_{j=1}^n X_j) = \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) = \sum_{i=1}^n \text{Var}(X_i) + \sum_{i=1}^n \sum_{j \neq i}^n \text{Cov}(X_i, X_j)$,
12. Conditional expectation: $\mathbb{E}[X|Y]$ is a random variable subject to the variation of X ,
13. Properties of conditional expectation:
 - Independence between X and $Y \implies \mathbb{E}[X|Y] = \mathbb{E}[X]$ and $\mathbb{E}[g(X)|Y] = \mathbb{E}[g(X)]$,
 - $\mathbb{E}[g(Y)|Y] = g(Y)$ and $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$,
 - $\mathbb{E}[aX + bY|Z] = a\mathbb{E}[X|Z] + b\mathbb{E}[Y|Z]$,
14. Partition theorem for expectations: $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$,
15. Variance partition formula: $\text{Var}(X) = \text{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\text{Var}(X|Y)]$,
16. Conditional variance formula: $\text{Var}(X|Y) = \mathbb{E}[X^2|Y] - \mathbb{E}[X|Y]^2$.

References

1. Sano, H.; Barker, K.; Odom, T.; Lewis, K.; Giordano, P.; Walsh, V.; Chambers, J.P. A survey of dog and cat anaesthesia in a sample of veterinary practices in New Zealand. *N. Z. Vet. J.* **2018**, *66*, 85–92. [[CrossRef](#)]
2. Thomson, K.; Rantala, M.; Hautala, M.; Pyörälä, S.; Kaartinen, L. Cross-sectional prospective survey to study indication-based usage of antimicrobials in animals: Results of use in cattle. *BMC Vet. Res.* **2008**, *4*, 15. [[CrossRef](#)] [[PubMed](#)]
3. Ouyang, Z.; Sargeant, J.; Thomas, A.; Wycherley, K.; Ma, R.; Esmailbeigi, R.; Versluis, A.; Stacey, D.; Stone, E.; Poljak, Z.; et al. A scoping review of ‘big data’, ‘informatics’, and ‘bioinformatics’ in the animal health and veterinary medical literature. *Anim. Health Res. Rev.* **2019**, *20*, 1–18. [[CrossRef](#)] [[PubMed](#)]

4. Valletta, J.J.; Torney, C.; Kings, M.; Thornton, A.; Madden, J. Applications of machine learning in animal behaviour studies. *Anim. Behav.* **2017**, *124*, 203–220. [[CrossRef](#)]
5. Cernek, P.; Bollig, N.; Anklam, K.; Döpfer, D. Hot topic: Detecting digital dermatitis with computer vision. *J. Dairy Sci.* **2020**, *103*, 9110–9115. [[CrossRef](#)]
6. Astill, J.; Dara, R.A.; Fraser, E.D.G.; Sharif, S. Detecting and predicting emerging disease in poultry with the implementation of new technologies and big data: A focus on avian influenza virus. *Front. Vet. Sci.* **2018**, *5*. [[CrossRef](#)] [[PubMed](#)]
7. Skinner, C.; Wakefield, J. Introduction to the design and analysis of complex survey data. *Stat. Stat. Sci.* **2017**, *32*, 165–175. [[CrossRef](#)]
8. Revilla, M.; Lenoir, G.; Flatres-Grall, L.; Muñoz-Tamayo, R.; Friggens, N.C. Quantifying growth perturbations over the fattening period in swine via mathematical modelling. *bioRxiv* **2020**. [[CrossRef](#)]
9. Mansfield, H.C.; Winthrop, D. *Alexis de Tocqueville, Democracy in America*; University of Chicago Press: Chicago, IL, USA, 2000.
10. Gregoire, T.G. Design-based and model-based inference in survey sampling: Appreciating the difference. *Can. J. For. Res.* **1998**, *28*, 1429–1447. [[CrossRef](#)]
11. Jones, G.; Johnson, W.O. A Bayesian superpopulation approach to inference for finite populations based on imperfect diagnostic outcomes. *J. Agric. Biol. Environ. Stat.* **2016**, *21*, 314–327. [[CrossRef](#)]
12. Yang, D.A.; Johnson, W.O.; Müller, K.R.; Gates, M.C.; Laven, R.A. Estimating the herd and cow level prevalence of bovine digital dermatitis on New Zealand dairy farms: A Bayesian superpopulation approach. *Prev. Vet. Med.* **2019**, *165*, 76–84. [[CrossRef](#)] [[PubMed](#)]
13. Little, R.J. To model or not to model? Competing modes of inference for finite population sampling. *J. Am. Stat. Assoc.* **2004**, *99*, 546–556. [[CrossRef](#)]
14. Baffetta, F.; Fattorini, L.; Franceschi, S.; Corona, P. Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sens. Environ.* **2009**, *113*, 463–475. [[CrossRef](#)]
15. Pfeffermann, D. The use of sampling weights for survey data analysis. *Stat. Methods Med Res.* **1996**, *5*, 239–261. [[CrossRef](#)]
16. Gelman, A. Struggles with survey weighting and regression modeling. *Stat. Sci.* **2007**, *22*, 153–164. [[CrossRef](#)]
17. Chen, Q.; Elliott, M.R.; Haziza, D.; Yang, Y.; Ghosh, M.; Little, R.J.; Sedransk, J.; Thompson, M. Approaches to improving survey-weighted estimates. *Stat. Sci.* **2017**, *32*, 227–248. [[CrossRef](#)]
18. Stehman, S.V. Practical implications of design-based sampling inference for thematic map accuracy assessment. *Remote Sens. Environ.* **2000**, *72*, 35–45. [[CrossRef](#)]
19. Tate, J.E.; Hudgens, M.G. Estimating population size with two-and three-stage sampling designs. *Am. J. Epidemiol.* **2007**, *165*, 1314–1320. [[CrossRef](#)]
20. Dorazio, R.M. Design-based and model-based inference in surveys of freshwater mollusks. *J. N. Am. Benthol. Soc.* **1999**, *18*, 118–131. [[CrossRef](#)]
21. West, P.W. Simple random sampling of individual items in the absence of a sampling frame that lists the individuals. *N. Z. J. For. Sci.* **2016**, *46*, 15. [[CrossRef](#)]
22. Abera, Z.; Degefu, H.; Gari, G.; Kidane, M. Sero-prevalence of lumpy skin disease in selected districts of West Wollega zone, Ethiopia. *BMC Vet. Res.* **2015**, *11*, 135. [[CrossRef](#)]
23. Abebe, R.; Hatiya, H.; Abera, M.; Megersa, B.; Asmare, K. Bovine mastitis: Prevalence, risk factors and isolation of *Staphylococcus aureus* in dairy herds at Hawassa milk shed, South Ethiopia. *BMC Vet. Res.* **2016**, *12*, 270. [[CrossRef](#)]
24. Sulayeman, M.; Dawo, F.; Mammo, B.; Gizaw, D.; Shegu, D. Isolation, molecular characterization and sero-prevalence study of foot-and-mouth disease virus circulating in central Ethiopia. *BMC Vet. Res.* **2018**, *14*, 110. [[CrossRef](#)]
25. Horvitz, D.G.; Thompson, D.J. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* **1952**, *47*, 663–685. [[CrossRef](#)]
26. Cochran, W.G. *Sampling Techniques*, 3rd ed.; John Wiley & Sons, Inc.: New York, NY, USA, 1977.
27. Lohr, S.L. *Sampling: Design and Analysis*; Cengage Learning: Boston, MA, USA, 2010.
28. Heayns, B.; Baugh, S. Survey of veterinary surgeons on the introduction of serological testing to assess revaccination requirements. *Vet. Rec.* **2012**, *170*, 74. [[CrossRef](#)]
29. Atuman, Y.J.; Ogunkoya, A.B.; Adawa, D.A.Y.; Nok, A.J.; Biallah, M.B. Dog ecology, dog bites and rabies vaccination rates in Bauchi State, Nigeria. *Int. J. Vet. Sci. Med.* **2014**, *2*, 41–45. [[CrossRef](#)]
30. Kaler, J.; Wani, S.A.; Hussain, I.; Beg, S.A.; Makhdoomi, M.; Kabli, Z.A.; Green, L.E. A clinical trial comparing parenteral oxytetracycline and enrofloxacin on time to recovery in sheep lame with acute or chronic footrot in Kashmir, India. *BMC Vet. Res.* **2012**, *8*, 12. [[CrossRef](#)]
31. Wickham, J.D.; Stehman, S.V.; Smith, J.H.; Wade, T.G.; Yang, L. A priori evaluation of two-stage cluster sampling for accuracy assessment of large-area land-cover maps. *Int. J. Remote Sens.* **2004**, *25*, 1235–1252. [[CrossRef](#)]
32. Getahun, K.; Kelay, B.; Bekana, M.; Lobago, F. Bovine mastitis and antibiotic resistance patterns in Selalle smallholder dairy farms, central Ethiopia. *Trop. Anim. Health Prod.* **2008**, *40*, 261–268. [[CrossRef](#)]
33. Regassa, A.; Tassew, A.; Amenu, K.; Megersa, B.; Abunna, F.; Mekibib, B.; Macrotty, T.; Ameni, G. A cross-sectional study on bovine tuberculosis in Hawassa town and its surroundings, Southern Ethiopia. *Trop. Anim. Health Prod.* **2010**, *42*, 915–920. [[CrossRef](#)]

34. Solís-Calderón, J.J.; Segura-Correa, J.C.; Aguilar-Romero, F.; Segura-Correa, V.M. Detection of antibodies and risk factors for infection with bovine respiratory syncytial virus and parainfluenza virus-3 in beef cattle of Yucatan, Mexico. *Prev. Vet. Med.* **2007**, *82*, 102–110. [[CrossRef](#)]
35. Hotchkiss, J.W.; Reid, S.; Christley, R. A survey of horse owners in Great Britain regarding horses in their care. Part 1: Horse demographic characteristics and management. *Equine Vet. J.* **2007**, *39*, 294–300. [[CrossRef](#)]
36. Bisson, A.; Maley, S.; Rubaire-Akiiki, C.; Wastling, J. The seroprevalence of antibodies to *Toxoplasma gondii* in domestic goats in Uganda. *Acta Trop.* **2000**, *76*, 33–38. [[CrossRef](#)]
37. Stevenson, M.A. Sample size estimation in veterinary epidemiologic research. *Front. Vet. Sci.* **2021**, *7*. [[CrossRef](#)]