REVIEW ARTICLE

Sampling and analysis frameworks for inference in ecology

Byron K. Williams [ID] | Eleanor D. Brown [ID]

U.S. Geological Survey, Science and
Decisions Center, Reston, VA, USA

Correspondence
Eleanor D. Brown
Email: ebrown@usgs.gov

Abstract

1. Reliable statistical inference is central to ecological research, much of which seeks to estimate population attributes and their interactions. The issue of sampling design and its relationship to inference has become increasingly important due to rapid proliferation of modelling methodology (line transect modelling, capture-recapture, estimation of occurrence, model selection procedures, hierarchical modelling) and new sampling approaches (adaptive sampling, other specialized designs). It is important for ecologists using these advanced methods to be aware of how the linkages between sample selection and data analysis can potentially affect inference.

2. We examine design-based and model-based inference frameworks for ecological data collected randomly, purposively or opportunistically. We elucidate differences in the probability structures for data arising from these frameworks, clarify the assumptions that underlie them, and demonstrate their differences.

3. Design based inference builds on a probability structure inherited from randomized data collection, whereas model-based inference relies on an assumed stochastic model of the data. By itself, a design-based approach is of limited value for inferences about causal hypotheses. In contrast, model-based inference is dependent on a conditionality principle that can seldom be shown to be met for an ecological system. We describe the conditions under which one can safely ignore sampling design in model-based analysis, along with inferential implications if these conditions are not met. The special case of opportunistic sampling is discussed.

4. We present a combined framework that takes advantage of both approaches to inference, and provides a robust methodology that can deal with the modelling of sampling problems such as non-detection and misclassification, as well as the exploration of causal hypotheses. The combined framework can be useful for identifying optimal sampling strategies.

5. Each approach to inference has its strengths and weaknesses, and practitioners should be aware of these in order to tailor designs and analyses to specific questions. We use the approaches and their underlying rationales to provide guidelines for choosing designs and estimators for reliable inference.

KEYWORDS
design-based inference, model-based analysis, sampling, statistical reliability, surveys

1 | INTRODUCTION

Reliable statistical inference is central to ecological research, much of which seeks to estimate population attributes such as size or vital rates, their interrelationships, and the influence of environment and management. In recent years new modelling approaches have proliferated (Buckland, Goudie, & Borchers, 2018; Williams, Nichols, & Conroy, 2002), along with new sampling approaches such as adaptive sampling (Thompson, 2012). Due to these advances, as well as the development of new technology for collecting and displaying complex data and large databases, the issue of sampling design and its relationship to statistical inference has become increasingly important. Ecologists using these advanced methods should be aware of how statistical inference is affected by sampling and its linkage to analysis, since ignoring sampling design can lead to biased estimation, and hence to ineffective decision making. An example is size-stratified fish sampling that can lead to erroneous estimation of parameters such as natural mortality rate (Goodyear, 2019), and in turn can bias the estimates of stock size and yield that factor into harvest rates (Clark, 1999).

Field studies in ecology involve collecting data about a population, followed by analysis based on hypothesized models. Both data collection and analysis can involve random elements, associated with sample selection and statistical estimation of model parameters. Although it seems intuitively clear that model-based inference may depend on the sampling protocol that determines what data are available, the links between sampling design and model analysis are rarely addressed (though see, e.g. studies by Edwards, Cutler, Zimmermann, Geiser, & Moisen, 2006 on sampling design effects on classification tree models for predicting species distributions; by Conn, Thorson, & Johnson, 2017 on testing and correcting for biasing effects of preferential sampling in population surveys; and by Irvine, Rodhouse, Wright, & Olsen, 2018 on erroneous model-based conclusions resulting from non-ignorable sampling designs).

Our objective in this paper was to explore the interplay between sampling design and statistical inference. We describe design-based and model-based frameworks for inference, and contrast them for data collected randomly, purposively or opportunistically. We provide an integrated framework combining both design-based and model-based factors, which can be useful for identifying effective sampling strategies. We examine the conditions under which the sampling design can be safely ignored in statistical inference, and discuss the special case of opportunistic sampling and its inferential limitations.

2 | TRADITIONAL INFERENCE PATHWAYS

Statistical analysis in ecology typically follows one of two well-known inferential tracks, depending on randomization in data collection and modelling as a basis for inference. Inference from data collected by means of random sampling is said to be *design-based*, in contrast to *model-based* inference, which relies on a hypothesized model that is assumed to describe the observed data. Though these two approaches both address the structure and function of ecological systems, they treat randomness in distinctive ways, often focusing on different ecological attributes and using different conceptual frameworks (Gregoire, 1998; Sterba, 2009).

The distinction between the two approaches to inference builds on the differing views of R.A. Fisher and Jerzy Neyman (Lenhard, 2006), who played prominent roles in the development of modern statistics. The modelling approach based on Fisher's work recognizes that empirical random sampling is often not feasible, particularly in observational studies, so its inferential framework relies on modelling—including distributional assumptions about observations—to mimic random sampling even when it is absent (Fisher, 1955, 1958). Observations inherit their randomness from model-based assumptions about observation probabilities, rather than from any empirical randomness associated with sampling. This leads to a focus on populations described by hypothesized models, and to sampling protocols that may be non-random.

In contrast, the design-based approach has grown out of Neyman's work, recognizing that hypothetical populations and models to fit them are fallible and subjective, and to be avoided when possible in making inferences from sampling data (Neyman, 1957; Neyman & Pearson, 1933). The framework for design-based inference focuses instead on finite populations that are randomly sampled. Samples inherit their randomness from a sampling design rather than from model-based distributional assumptions (Neyman, 1934).

Statistical practice in ecology includes both design- and model-based inference, as well as combined approaches that incorporate both random sampling and stochastic values in a single statistical assessment. Thus:

Design-based inference accounts only for sampling randomization, as in an evaluation of fixed values for units collected randomly. For example an analyst might focus on the random selection of units from a population, assuming that realized unit values were previously generated by a (possibly unrecognized) stochastic process and now are fixed. Thus, the analysis need not account for randomness in the unit values. However, sampling is considered to be random. Examples include well-known sampling designs such as cluster sampling, stratified random sampling, and systematic sampling.

Model-based inference accounts only for stochasticity in unit values, as in an evaluation of a previously identified sample of units. For example the analyst might focus on the stochastic values of units that have already been selected. The analysis need not account for sampling randomization, because the sample is considered fixed. However, the unit values are considered random.

Combined inference incorporates both random sources, as in collection of random samples from a population and observation of stochastic unit values. Sampling and unit values are each considered to be random, and the approach accounts for both sources of randomness in estimation and reliability assessment.

2.1 | Design-based framework

A design-based framework for inference involves three key elements:

- a population of finitely many potential population units;
- a sampling design that describes a random or probability-based selection of units (i.e. the assignment of a probability of selection to potential samples); and
- a sampling scheme describing the mechanism for implementing the design.

The population is defined operationally, by the assignment of non-zero selection probabilities to subsets of units. By implication, units with no probability of selection are not considered part of the population, and no inference to them is possible. The targets of inference typically are straightforward population attributes such as population totals, means and ratios.

The statistical properties in a design-based framework derive from randomly sampling population units and recording unit values, denoted here by y_k. The values themselves are held to be fixed, whereas the sampling process is random. The sampling frame for design-based inference ideally consists of a list of units from which a sample can be selected. The units often are physical entities, for example individual organisms or clusters of organisms, plots of land or landscape patches of multiple plots, etc. The sampling design consists of assigning a probability of selection to potential samples (with, e.g. simple random sampling, stratified random sampling, cluster sampling). Finally, a sampling scheme describes the actual selection of units and observation of values on them.

A standard for probability sampling is the simple random sampling design, in which all same-size subsets of sampling units are equally likely to be selected. Sampling designs with varying selection probabilities are often referred to as 'complex designs' (Skinner & Wakefield, 2017). A common motivation for deviating from simple random sampling is efficiency, as measured by estimator variance. In many cases, estimator precision can be improved with stratification, clustering, and sampling based on unequal probabilities of unit selection, depending on the population structure. For example randomization within recognized strata of known stratum sizes can take advantage of systematic differences among strata to produce unbiased population estimators with smaller variance than simple random sampling. Other complex designs frequently arise in spatial sampling: transect sampling and other applications involve cluster and systematic sampling (Thompson, 2012), in which random selection of primary units is followed by the complete enumeration of all secondary units in each primary unit. If known, variable cluster sizes can be used in selection with unequal probabilities, or in forming ratio estimators. Multi-stage sampling is also possible, with random sampling of secondary units and complete enumeration of tertiary units. Ecological applications use a very large number of sampling strategies involving various combinations of stratification, multi-stage clustering, unequal probability sampling with auxiliary variables, and many other design factors.

The possibility that any unit in a population can be selected establishes the inferential linkage between units that are selected and those that are not. The design thereby allows statistical inference to be made to the whole population, including unselected units. A probability-based sampling design imparts stochasticity to samples, with variation among samples that declines to zero as sample size approaches a population census. Simple random sampling, and many other well-known designs such as stratified sampling, cluster sampling, systematic sampling, ratio and regression sampling, are probability-based and designed to control or reduce sampling variability.

A key strength of a design-based approach is the avoidance of challenges to inferential results, which can otherwise occur if results depend on models inadequately representing the structure of population values. Its main limitations are an inability to address analytical or causal hypotheses (given the absence of a process model by which to express them); a need for models to handle such factors as partial detection and non-response even though the framework seeks to avoid the use of models; and an inability to account for non-sampling errors such as measurement error.

2.2 | Model-based framework

A second generic approach to statistical inference uses a model-based framework, with different elements from those in a design-based framework. They typically include:

- a statistical model describing how observations on population units are thought to have been generated from a super-population with potentially infinitely many observations for each unit;
- an assumed stochastic structure that allows the unit values themselves to be seen as random variables; and
- a 'conditioning principle' by which any particular set of observations becomes statistically comparable to any other set of observations after hypothesized conditions (e.g. strata, clustering, disproportionate sampling effort) are accounted for.

The latter point is especially problematic in ecological investigations, because important structural features of ecological systems are often unknown. Many ecological studies focus on identifying relevant conditioning variables.

In a model-based approach, statistical inference is dependent on the assumed stochastic structure of the model, rather than the sample selection process (which may or may not be random). Unit values are treated as random variables from the super-population, and denoted here by Y_k to distinguish them from the fixed y_k values in the design-based framework. Targets of inference typically are the model parameters and causal or analytic relationships among parameters and the conditioning variables. Examples include mark–recapture and band-recovery models; occupancy models; models for distance sampling, survival and nest success; and many others that fit ecological data to an assumed model for purposes of estimating parameters and identifying model structures (Williams et al., 2002).

Model-based inference relies on modelling and model assumptions to impute stochasticity in an analysis. Because the model structure is held to apply to all potential observations, presumably any sample of observations will suffice; that is, sampling can be non-random. With non-random sampling there is no sampling distribution, and therefore no opportunity to use sampling probabilities for generalizing from sampled to unsampled units. Instead, inference must depend on distributions identified in the assumed model. The assumed applicability of the model to all potential sampling units allows inference to be extended beyond the sample to the population. Because the unit values are tied to a stochastic process describing the population, stochastic variation in the values remains, irrespective of the sample size. This contrasts with design-based inference, in which variation asymptotically vanishes as sample size approaches a census (Gregoire, 1998).

Key strengths of a model-based approach include the ability to make inferences when sample selection is non-random; the ability to investigate causal and analytic hypotheses; the ability to investigate responses to many factors with complex interactions; and sometimes the potential for improvements in estimation beyond what is possible with random sampling. Its main limitations include the potential inability to meet Fisher's conditionality principle because selection factors and strata or cluster indicators fail to be identified or observed; the potential to omit relevant factors unknowingly in sample selection; and the unbounded complexity of alternative model specifications, which can leave the analytic results suspect for any particular model.

3 | UNBIASED ESTIMATION IN DESIGN-BASED AND MODEL-BASED INFERENCE

In this section we consider how randomness in data collection can influence estimator performance. We address the two sources of stochastic variation mentioned above, involving random selection of samples and unit values that are generated by a stochastic process. Because design-based and model-based approaches focus on different statistical features, they treat the issue of estimator performance, and particularly estimator bias, somewhat differently.

Again, we consider a finite population of N discrete units, with realized values $\underline{y} = (y_1, \ldots, y_N)$ of a random vector $Y = (Y_1, \ldots, Y_N)$ from a super-population. An example might be 100 forest stands in a region, with individual stand values consisting of the proportions of stand area covered by the forest canopy. For each stand, a realized value of canopy cover can be observed, and repeated observations can yield variation in a stand's cover values.

To contrast the two inference scenarios, we address the issue of how well an estimator represents the population. Thus, consider a population attribute $Y_0 = h(Y)$ derived from Y (e.g. mean proportion of stand area covered by canopy for the population of stands), and an estimator $\hat{Y}_0 = g(Y_s)$ of Y_0 based on a sample s of the Y values. For any particular realization $\underline{y}$, we denote the value of the population attribute by $y_0 = h(Y = \underline{y})$ and the estimator value by $\hat{y}_0 = g(Y_s = \underline{y}_s)$.

Because $Y_0 = h(Y)$ is a function of the vector Y, it inherits stochasticity from the Y values. The estimator $\hat{Y}_0 = g(Y_s)$ does as well, but it is also influenced by sampling. By conditioning $\hat{Y}_0$ on one or the other of these factors, two conditional estimators can be identified, one associated with design-based inference and the other associated with model-based inference.

3.1 | Design-based inference

In design-based inference, the source of random variation involves the selection of a sample $s = \{s_1, \ldots, s_n\}$ of realized values $\{y_{s_1}, \ldots, y_{s_n}\}$, according to a sampling design that assigns probabilities $P(s)$ to samples. In our forest example, sampling might involve the random selection of 10 of the 100 stands, with a selection probability for any given stand that is proportional to its area. A large number of different samples of 10 stands can be selected from the population of 100 stands, each with its own selection probability $P(s)$. A scenario that accounts only for random sampling takes the canopy cover values as realized, that is treats them as fixed quantities $\{y_1, \ldots, y_N\}$.

By conditioning $\hat{Y}_0$ on the population values Y, the estimator $\left(\hat{Y}_0 | Y_s = \underline{y}_s\right) = \hat{y}_0$ ceases to be subject to the stochastic effect of Y, leaving only the effect of random sampling. Thus, sample values in the estimator are treated as fixed and non-random, so that statistical inference is based solely on probability sampling of the units. The values $\hat{y}_0 = g(\underline{y}_s)$ are averaged over the sample probabilities $P(s)$, where the expectation with respect to these probabilities is denoted by $E_p(\cdot)$. If $E_p(\hat{y}_0) = \sum_s P(s)g(\underline{y}_s)$ coincides with y_0, the estimator is said to be *design-unbiased* for y_0 (Gerow & McCulloch, 2000) (see Appendix S1).

A design-unbiased estimator $\hat{y}_0 = g(\underline{y}_s)$ will on average yield the population value $\underline{y}_0$ under the design probabilities $P(s)$, irrespective of any particular array of unit values $\underline{y}$. In that sense the distribution model $f(Y)$, and any process producing that distribution, are irrelevant. An obvious implication for design-unbiased estimation is that the sampling design must be probability-based, because only then is there a probability distribution $P(s)$ with which to determine $E_p(\hat{y}_0)$.

3.2 | Model-based inference

Alternatively, in model-based inference the source of variation involves unit-specific random values $\{Y_1, \ldots, Y_N\}$ that are assumed to have been generated by a stochastic process. Thus, Y has a joint distribution $f(Y)$, and any subset $Y_s = \{Y_{s_1}, \ldots, Y_{s_n}\}$ of values in Y has a marginal distribution $f_s(Y_s)$. In the forest example, the observed proportion of stand area covered by the canopy may vary with daily conditions (cloud cover, ambient light, and other factors), so the stand proportions are modelled as random variables with their own means and variances. A scenario that accounts only for stochastic unit values takes the sample of stands as given.

In general, conditioning $\hat{Y}_0$ on a sample s means the estimator $\left(\hat{Y}_0|s\right)=g(Y_s)$ ceases to be subject to the influence of random sampling, leaving only the effect of stochasticity in the vector Y. In this case the sample s is treated as fixed, so that statistical inference is based solely on the stochasticity of unit values. Letting $E_m(\cdot)$ denote expectation with respect to model stochasticity, if $E_m(\hat{Y}_0|s)=\int f_s(y_s)g\left(y_s\right)dy_s$ coincides with $E_m(Y_0)$ for every sample s, the estimator is said to be *model-unbiased* for Y_0 (Gerow & McCulloch 2000) (see Appendix S1).

A model-unbiased estimator will on average yield a value of $E_m(Y_0)$ under the model distribution $f(Y)$, irrespective of the particular sample s of units. In that sense the sample distribution $P(s)$ is irrelevant. A practical implication for model-unbiased estimation is that patterns in the expected values must uniformly reflect the population structure across samples.

With model-based inference one must be confident that the sample appropriately represents the population from which it was obtained, for otherwise any model-based inferences about that population can be biased and misleading. This contrasts with design-based inference, which requires accurate portrayal of the probabilities of the sampling design used in sample selection, for otherwise design-based inferences can be biased and misleading.

Both design-unbiased and model-unbiased estimators can be shown to be (unconditionally) unbiased for Y_0 (Thompson, 2012). Thus, there is no basis for preference for either approach in terms of bias alone. Instead, preference must be based on one of the other factors mentioned earlier (estimation objectives, estimator precision, capacity for hypothesis testing, treatment of non-sampling errors, vulnerability to challenge of inferential results).

4 | EXAMPLES COMPARING DESIGN-BASED AND MODEL-BASED FRAMEWORKS

The following examples show that a population estimator can have different statistical behaviours in model-based and design-based approaches.

4.1 | Simple random sampling without replacement

To contrast design-based and model-based inference, we use a simple example involving estimation of the mean of a population consisting of N sampling units, using a simple random sample of size n drawn without replacement. In our forest example, sampling might consist of the selection of 10 stands without replacement from a population of 100 stands.

A design-based estimator of the finite population total $T_y=N\mu_y=\sum_{i=1}^{N}y_i$ of realized unit values is given by

$$\hat{T}_y=N\hat{\mu}_y$$
$$=\frac{N}{n}\sum_{k\in s}y_k, \tag{1}$$

with statistical properties that depend exclusively on the sampling design, absent any consideration of the process generating the unit values themselves. The expected value and variance of $\hat{T}_y$ are $E_p(\hat{T}_y)=T_y$ and

$$\mathrm{var}_p(\hat{T}_y)=N^2\frac{S_y^2}{n}\left(1-\frac{n}{N}\right)$$
$$=N^2\left[\frac{\sum_{k=1}^{N}(y_k-\mu_y)^2}{N-1}\right]\left(\frac{1}{n}-\frac{1}{N}\right) \tag{2}$$

respectively (Cochran, 1977), where the subscript p refers to probability sampling.

The situation is somewhat different for model-based inference. Assume a model $Y_k=\mu+\varepsilon_k$ of independent unit values Y_k for the population units, with constant model variance across the units: $\varepsilon_k\sim N(0,\sigma^2)$. For a group of n units an optimal model-based estimator of μ is given by

$$\hat{T}_Y=N\hat{\mu}$$
$$=\frac{N}{n}\sum_{k\in s}Y_k \tag{3}$$

(Graybill, 1976), with statistical properties that depend exclusively on the underlying model absent consideration of the mechanism for sample selection. The expected value and variance of $\hat{T}_Y$ are $E_m(\hat{T}_Y)=N\mu$ and

$$\mathrm{var}_m(\hat{T}_Y)=\frac{N^2}{n}\sigma^2 \tag{4}$$

respectively, where the subscript m refers to model stochasticity.

Both estimators of the population total in Equations 1 and 3 seem plausible under the sampling and model assumptions. Indeed, both produce essentially the same estimates for the population total, namely the sample mean weighted by population size. However, the estimator variances in Equations 2 and 4 differ, a consequence of different stochastic sources in design-based and model-based inference. Other meaningful differences are that T_y, $\mu_y=T_y/N$ and σ_y^2 are population parameters for design-based inference, whereas T_Y is a random variable and μ and σ are model parameters for model-based inference.

That said, with large N one would expect the mean μ_y and variance σ_y^2 of the realized population to be close in value to the mean μ and variance σ^2 of the stochastic process generating the population values. Whether such concordance is observed in the data analysis depends mainly on whether the model that is assumed accurately describes the underlying stochastic process. When there is a mismatch between the actual process and the model used to represent it, there is no reason to expect consistency between estimators from the two approaches.

4.2 | Sampling with unequal inclusion probabilities

Here we replace the simple random sampling design in the previous example by a design in which unit inclusion probabilities π_k are unequal. The Horwitz–Thompson estimate (Horvitz & Thompson, 1952) of T_y for a sample s is

$$\hat{T}_y=\sum_{k\in s}\frac{y_k}{\pi_k},$$

which can be shown to be design unbiased (see Appendix S1).

Again, the situation differs for model-based inference. Assume the same model as in Example 1, $Y_k = \mu + \varepsilon_k$ for independent unit values Y_k with constant model variance across the units: $\varepsilon_k \sim N(0, \sigma^2)$. In this situation the model bias of the Horwitz–Thompson (H–T) estimator $\hat{T}_Y$ is

$$
\begin{aligned}
E_m\left(\hat{T}_Y - T_Y\right) &= E_m\left(\sum_{k \in s} \pi_k^{-1} Y_k - T_Y\right) \\
&= \sum_{k \in s} \pi_k^{-1} E_m\left(\mu + \varepsilon_k\right) - N\mu \\
&= \sum_{k \in s} \pi_k^{-1} \mu - N\mu \\
&= \mu\left(\sum_{k \in s} \pi_k^{-1} - N\right),
\end{aligned}
$$

which is non-zero unless $\sum_{k \in s} \pi_k^{-1} = N$. Thus, the estimator $\hat{T}_Y$ is generally model-biased for unequal probability sampling designs. The bias is likely to be slight for large values of N, since $\sum_{k \in s} \pi_k^{-1}$ is design-unbiased for N. Nevertheless, the H–T estimator, which is necessarily unbiased in design-based inference, is for the most part biased in model-based inference.

4.3 | Generalizations

The examples just described address estimators of the population total $T_y = \sum_{i=1}^{N} y_i$ and the expected value $E\left(T_Y\right) = N\mu$ based on a sample of population units. It is of course possible to consider other estimators of other population attributes. Whether to use design-based or model-based inference comes down to the focus of the estimation (on the mean of the realized unit values, or the mean of the process thought to generate them) and the acceptability of the model assumptions. If the assumptions are held to be acceptable, a model-based approach can produce reliable and efficient estimation; otherwise, without model assumptions, a design-based approach can yield reliable inference.

5 | INFORMATIVE AND NON-INFORMATIVE SAMPLING

In this section we consider sampling conditions that allow for population-level inference, and the inferential consequences of violating those conditions. At issue is whether one can extend statistical inference from the sample to the entire population of interest. Many ecological investigations simply overlook the linkage between sampling and analysis, by treating the sample as if it 'represents' the population irrespective of how it is collected, and proceeding directly to data analysis and inference. If a sample is not representative, population-level inferences, including model-based inferences, can be badly misleading, often unrecognizably so.

To address the potential linkage between population sampling and analysis we highlight the concept of informative and non-informative sampling, relating to the probability distributions for sampled and unsampled units. We focus again on the outcome of two random events: (a) generation of random values for a finite population, and (b) random selection of units from the population. A correlation between selection probabilities and unit values (after accounting for any environmental covariates [see Appendix S1]) is definitive of informative sampling, and

implies that to avoid bias the selection process must be taken into account when making inferences from survey data (Sarndal, 1978).

To illustrate, assume that unit values Y_k for a population are independent random variables with a population probability density function $f(y_k|z_k)$, where the z_k is a conditioning variable, as in a regression model. By sampling the population, groups of sampled and unsampled population units can be identified. In our example of forest stands, the population of 100 stands is divided by sampling into one class of 10 stands selected in the sample, and a second class of the remaining 90 stands that are not selected.

Sampling allows one to recognize a distribution $f_s(y_k|z_k)$ for the values of units in the selected sample, as indicated by the subscript s. A common practice is to base inferences about the population on the sample distribution $f_s(y_k|z_k)$ (e.g. the 10 stands selected in the sample), on the assumption that the latter adequately represents the distribution of values across the whole population, including the 90 stands that are not selected. That is, $f_s(y_k|z_k)$ is assumed to coincide with, or closely approximate, the population distribution $f(y_k|z_k)$. At issue is whether that assumption is met, because otherwise inference based on $f_s(y_k|z_k)$ misrepresents the population as a whole.

To see how such a misinterpretation can occur, consider the relationship between the sample and population distributions, as expressed by Bayes' theorem:

$$
f_s(y_k|z_k) = \frac{\Pr\left(k \in s|y_k, z_k\right) f(y_k|z_k)}{\Pr\left(k \in s|z_k\right)}. \tag{5}
$$

when $\Pr\left(k \in s|y_k, z_k\right) \neq \Pr\left(k \in s|z_k\right)$, the sample distribution $f_s(y_k|z_k)$ in Equation 5 differs from the population distribution $f(y_k|z_k)$. Sampling then is said to be *informative*, in that the sampling probabilities $\Pr\left(k \in s|y_k, z_k\right)$ are related to the population values y_k (Little, 2004; Sugden & Smith, 1984).

On the other hand, when $\Pr\left(k \in s|y_k, z_k\right) = \Pr\left(k \in s|z_k\right)$, $f(y_k|z_k)$ coincides with $f_s(y_k|z_k)$. In that case sampling is said to be *non-informative*, in that sampling probabilities are unrelated to the population values. Inference about the population then can be based on the sample distribution $f_s(y_k|z_k)$, without accounting for the conditional sampling probabilities $\Pr\left(k \in s|z_k\right)$.

Clearly, the informative or non-informative nature of a sampling plan is key to the use of sample data for population inference. Sampling probabilities for population units must be unrelated to the corresponding values (non-informative sampling) in order to support reliable inference without further considerations. If sampling probabilities are related to unit values, (informative sampling), the differences in Equation 5 between sample and population distributions complicates inference for both design-based and model-based approaches.

6 | COMBINING SAMPLING DESIGN AND MODELLING

In this section we describe a framework that combines both sources of stochasticity, and use it to discuss sampling informability and the ignorability of sampling design. These attributes are important in

recognizing when a sampling design is (or is not) relevant to population-level inference.

A general framework incorporates both the randomized selection of units as well as model-based stochasticity of unit values Y in a joint distribution. This allows us to consider both stochastic elements in the statistical treatment of data, and account for both the sampling and analysis components of field ecology in a fully integrated way.

The framework explicitly includes stochastic unit values, here represented by the distribution $f(y|z)$, and sampling randomization that is captured with a vector $I = (I_1, \ldots, I_N)$ of binary indicators denoting inclusion (or exclusion) of the population units. A joint distribution that includes both factors can be written as

$$f(\underline{y}, i|\underline{z}) = f(\underline{y}|\underline{z}) \Pr(i|\underline{y}, \underline{z}), \qquad (6)$$

where $f(\underline{y}, i|\underline{z})$ is the conditional distribution of $(Y, I)|Z = \underline{z}$ at $(\underline{y}, i)$, and similarly for $\Pr(i|\underline{y}, \underline{z})$. Letting $\bar{s}$ represent population units not in sample s, inference with sample data is based on the joint distribution

$$f(\underline{y}_s, i|\underline{z}) = \int \Pr(i|\underline{y}_s, \underline{y}_{\bar{s}}, \underline{z}) f(\underline{y}_s, \underline{y}_{\bar{s}}|\underline{z}) d\underline{y}_{\bar{s}} \qquad (7)$$

(see Appendix S1).

Ignoring the sampling mechanism means that $\Pr(i|\underline{y}_s, \underline{y}_{\bar{s}}, \underline{z})$ is omitted in Equation 7, so that inference is based on

$$f(\underline{y}_s|\underline{z}) = \int f(\underline{y}_s, \underline{y}_{\bar{s}}|\underline{z}) d\underline{y}_{\bar{s}}. \qquad (8)$$

The question here is under what conditions the sampling probabilities $\Pr(i|\underline{y}_s, \underline{y}_{\bar{s}}, \underline{z})$ can be safely ignored. It is argued in the Appendix S1 that if sampling is non-informative, that is,

$$\Pr(i|\underline{y}_s, \underline{y}_{\bar{s}}, \underline{z}) = \Pr(i|\underline{z}), \qquad (9)$$

the distribution $f(\underline{y}_s|i, \underline{z})$ that accounts for the effect of sampling is identical to the distribution $f(\underline{y}_s|\underline{z})$ that omits it. That is, sampling is ignorable for statistical inference when sampling is non-informative.

Though informative sampling and ignorability are related concepts, each is associated with one of the two random events for sampling and analysis mentioned earlier:

- the terms informative and non-informative apply to sampling (i.e. non-informative sampling means that sampling probabilities are not informed by population unit values); and
- ignorability applies to population inference (i.e. ignorability means that population inference can ignore the sampling mechanism)

(Sugden & Smith, 1984). An example of a sampling design that is informative involves sample selection targeting large Y values and avoiding small Y values (e.g. retaining only the units with large Y values for data analysis). Thus, one might retain and analyse only those forest stands with large measures of canopy cover, because they include a larger proportion of the forest under study. Restricting the sampled units to those with larger values violates condition (9), leading to potential estimator bias.

The critical assumption that sample unit selection does not depend on Y is met with random sampling, and sometimes (though not necessarily) with non-random sampling. More generally, it is met with any sampling scheme for which unit selection is based solely on the auxiliary z values (e.g. balanced sampling [Royall & Pfeffermann, 1982; Yates, 1960]).

Non-informative sampling, and therefore ignorability, apply most conveniently to simple random sampling, because that plan is designed to represent the sampled population. Other plans such as stratified random sampling can also represent the population, while accounting for population structure. In addition, some purposive plans such as balanced sampling can be non-informative. But because sampling designs are not automatically ignorable for reliable inference, an analyst needs to determine whether the design actually used is non-informative, in order to produce reliable and convincing inferences.

These results are especially germane to field studies involving sample selection followed by inference about ecological parameters. With non-informative sampling the process of generating a sample can effectively be 'ignored' in the subsequent inference process, which allows one to use any sample of a reasonable size to make inferences about population parameters. It is the ability to extend parametric inference from a specific set of sampled units to the whole population that underlies model-based inference. The importance of this feature is frequently overlooked by ecologists, who often assume or even assert a broad inferential range for their work, though they often fail to justify it.

7 | SAMPLING STRATEGIES

In this section we describe how the combined framework can be used for formulating sampling strategies that account for both design and estimation. Many ecological studies involve selection of a sampling design as well as estimators of targeted population parameters. Because multiple sampling plans can be used to provide data for many different estimators, an important question is how best to sample the population and aggregate the data into an appropriate estimator T, so as to account for the influence of sampling and stochasticity in unit values.

To illustrate, assume that the target of the sampling strategy is the population mean, to be estimated with counts of organisms on sample units. Here we seek estimators and sampling designs to minimize the average over the sample probabilities of the mean square error, that is,

$$E_p E_m \left(T - \bar{Y}\right)^2 = \sum_s P(s) E_m \left(T - \bar{Y}\right)^2,$$

where the subscripts p and m refer, respectively, to probability sampling and model stochasticity.

With this formulation one can condition on a particular sample s, identify a predictor T that minimizes $E_m \left(T - \bar{Y}\right)^2$, and then search for sampling plans to reduce the expected mean squared error. The estimator T for this problem can usefully be expressed in terms of

units in the sample for which values are available, and units not in the sample for which values must be model-based predictions:

$$T = \sum_s Y_k/N + \sum_{\bar{s}} \hat{Y}_k/N$$
$$= \sum_s Y_k/N + U/N.$$

Under reasonable assumptions $E_m\left(T - \bar{Y}\right)^2$ can be written in terms of the component variances,

$$E_m\left(T - \bar{Y}\right)^2 = \left[V_m(U) + \sum_{\bar{s}} V_m(Y_k)\right]/N^2$$

(see Appendix S1), so that the task of finding an optimal predictor T^* reduces to a search for the predictor U^* that minimizes $V_m(U)$. The resulting predictor T^* then can be used to look for optimal sampling designs.

For example we consider a population with independent unit values that is modelled with

$$E_m(Y_k) = \beta z_k$$
$$V_m(Y_k) = v_k \sigma^2. \tag{10}$$

In our forest example the auxiliary variable z_k might be aspect, slope, or elevation for stand k, which is thought to influence the amount of the stand canopy cover Y_k. In addition, variation in the Y values is thought to be influenced by a unit-specific attribute v_k, as in $v_k \sigma^2$. If $v_k = z_k$ the optimal predictor T^* for this model reduces to

$$T^* = \sum_s Y_k/N + U^*/N$$
$$= \frac{\bar{Y}_s}{\bar{z}_s} \bar{z},$$

so that

$$E_m E_p \left(T^* - \bar{Y}\right)^2 = \bar{z} E_p \left(N\bar{z}/\sum_s z_k - 1\right) \sigma^2/N \tag{11}$$

(see Appendix S1).

Perhaps surprisingly, $E_m E_p \left(T^* - \bar{Y}\right)^2$ in Equation 11 is minimized by the purposive selection of a sample with the largest z values, rather than by randomized sampling. When the model is correct, such a design can give striking improvements over simple random sampling. However, its precision is sensitive to the model in Equation 10, and it performs poorly under different model assumptions (Cassel, Sarndal, & Wretman, 2017). Basically, the extreme design with samples based on the largest z values is not robust. When there is uncertainty in the model, as is so often the case in ecology, simple random sampling or balanced sampling with $\bar{z}_s = \bar{z}$ can produce more robust results. In essence, sampling that is designed to be representative of the population at large is more likely to lead to robust estimation when structural features are omitted or inadequately represented.

Of course, other models than Equation 10 can be envisioned, along with other estimation targets than the population mean. And other criteria than $E_m E_p \left(T - \bar{Y}\right)^2$ can be formulated, which might include constraints to avoid different kinds of bias (Sarndal, 1978). Each model/parameter/criterion combination can be expected to produce its own optimal strategy, with its own robustness considerations.

The bottom line is that a framework incorporating both the randomized selection of units and model-based stochasticity for the unit values can be used to identify strategies that include both sampling designs and estimators. Under certain conditions it is possible to optimize sampling strategy as above, such that an optimal predictor can be identified conditional on a sampling plan, and then used to explore the performance of different sampling plans. There is, however, a trade-off between optimality and strategy robustness. Optimal performance based on an assumed population model may be quite suboptimal if the model is incorrect. On the other hand, performance that is suboptimal for a particular model may turn out to be robust over a broad range of different population conditions.

8 | OPPORTUNISTIC SAMPLING

Many recent papers in the ecological literature describe analysis of data collected opportunistically in the field, without a specific sampling design (Brown & Williams, 2018). Thus, observers record chance observations of some phenomenon (wild-life species presence, visible damage from flooding) over a general area. Such opportunistic sampling is neither probability-based nor guided by a model-assisted design, and is substantively different from sampling based on randomization, or purposive sampling based on assumed environmental features. It can be subject to selection bias, non-detection, observer bias, recording errors, and other factors influencing the observations (Isaac, Strien, August, Zeeuw, & Roy, 2014).

Although there is no role for design-based inference in opportunity sampling (given the lack of a priori assignment of selection probabilities to potential sample units), model-based inference may be applicable (e.g. Kéry et al., 2010) if the factors influencing observations are assumed to be known, data on them are collected, and sampling is repeated. A key question is whether the models actually represent the population over the area of interest, that is, whether the conditioning principle in model-based inference is operative. Misrepresentation of population structure and unreliable estimation result from failing to satisfy the conditioning principle. For example, inflated variances can result from failing to include strata. Variances that are too small can result from failing to account for clustering. And bias in the estimators of model parameters and their interactions can result from disproportionate sampling effort that is not accounted for. If model-based inference is to be valid in a context of opportunistic sampling, design features such as stratification and clustering need to be reflected in the model analysis, and the model itself must be applicable across all population units. (Skinner & Wakefield, 2017).

Even if the critical factors influencing observations are specified (or weights corresponding to them are identified), reliable inferences can be undermined by failing to collect the necessary data on those factors. For example data on an important covariate

may not be collected at all, or the data range may be too truncated to be useful for inference. Even if all the necessary response variables and covariates are observed where the data are collected (e.g. in disturbed areas), statistical inference will be compromised for areas where data are not collected (e.g. in undisturbed areas). The latter situation is an example of informative sampling, where differences between sampled and unsampled areas lead to estimation bias.

The many environmental and observer factors influencing observations in a broad-scale field study make it unlikely that opportunistically collected data can produce reliable inferences, without special assumptions or auxiliary information. One approach might be to treat the opportunistically collected information as auxiliary to other more rigorously collected data, if possible. Without the other information or restrictive assumptions, opportunistically collected data are better suited for exploratory data techniques (Tukey, 1977, 1980), in which patterns that emerge from the data point to hypotheses that can be investigated by follow-up studies designed specifically for reliable inference (Lenhard, 2006).

9 | DISCUSSION

We have described and compared design-based and model-based approaches to the collection and analysis of survey data. Each approach has strengths and limitations. A design-based approach is robust for descriptive population parameters, but does not permit inference about causal hypotheses. Model-based inference accommodates non-random sampling and causal hypotheses, but is dependent on the conditionality principle, which cannot be shown to be met for any studied system. A combined framework can incorporate the assumed distribution of observations in a model-based approach and the sampling probabilities in a design-based approach. The combined framework takes advantage of both approaches and provides a robust methodology to deal with the modelling of sampling problems such as non-detection and misclassification, as well as with the investigation of causal hypotheses.

These results do not support a uniform preference for design-based inference, with its randomized sampling, over model-based inference with its assumed stochastic structures, or vice-versa. For example, claims that randomized sampling is always preferred in ecological investigations are unjustified. In fact, balanced sampling and other purposive designs are often used to good effect in estimating ecological parameters. On the other hand, indifference or lack of attention to the potential consequences of complex and non-random sampling is also unjustified in model-based inference, because informative sampling designs can produce badly misleading bias.

Limitations of model-based inference are often due to non-independence of spatially adjacent sample units, and an inability to identify important ecological factors affecting population values. If the unit values themselves can be assumed to be independent and identically distributed across the population, many standard statistical results apply regardless of how the sample is selected.

This conclusion follows from the factorization of Equation 6 into a part that involves the observation values and a part that does not, which points to the conditions in which response and design variables are independent of one another. Unfortunately, the assumption that the unit values are independent and identically distributed can be problematic in ecological investigations, due to the tendency of values for nearby population units to exhibit correlation. One way around this difficulty is to use randomized sampling designs.

A related challenge for model-based inference is the use of survey results for different purposes. For example the same dataset may be used to set hunting quotas, to allocate program resources, or to assess recreational preferences. Because no single combination of model and sampling design is likely to suffice for such diverse purposes, there is a positive incentive to use random sampling and design-based inference to eliminate any appearance of bias in sample selection (Hansen, Madow, & Tepping, 1983).

However, if there is convincing evidence of patterns in unit values, they can be used to identify sampling plans and models. For example, it may be well documented that on average the number of organisms in a habitat patch is proportional to patch area. Modelling such patterns underlies many recent methodological advances in ecology (abundance and distribution analysis, hierarchical modelling, Bayesian approaches to inference, reinforcement learning). Modelling in some form is also used to treat non-sampling errors such as differential responses, measurement errors, and imperfect detectability (Thompson, 2012).

Promising developments in the partial integration of design-based and model-based approaches may overcome some of their respective limitations. A hybrid partially integrated framework would apply to finite and infinite populations and incorporate measurement error, while producing analytic statistics without the need to condition on all sampling features during model specification. Advances include disproportionate sample selection (Kish & Frankel, 1974); use of model weights to accommodate stratification and clustering (Binder, 2018; Fuller, 1975); use of model estimates for finite population parameters (Godambe & Thompson, 1986); incorporation of measurement error (Muthen & Satorra, 1995; Stapleton, 2008); and use of a mixture of model specification and estimation from both approaches (Korn & Graubard, 2003; Rabe-Hesketh & Skrondal, 2006).

Both design-based and model-based approaches to inference have always been a part of ecological investigation, and will continue to be useful for the foreseeable future. Which approach, or combination of approaches, is most appropriate depends on the focus of the investigation and the need for efficiency, accuracy and accountability. Ecologists must understand the strengths and limitations of each approach in order to tailor designs and analyses to specific questions and produce unbiased inferences from survey data. Failing to do so, or failing to conduct the necessary follow-up assessment after exploration of the data for patterns, can undermine reliable inference as well as its practical application, which so often motivates the investigation in the first place.

AUTHORS' CONTRIBUTIONS

B.K.W. and E.D.B. conceived the concept. B.K.W. led the writing of the manuscript. Both authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

We used no data, simulated data or code in preparing this manuscript.

ORCID

Byron K. Williams https://orcid.org/0000-0001-7644-1396
Eleanor D. Brown https://orcid.org/0000-0001-7798-830X

REFERENCES

Binder, D. A. (2018). On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279–292. https://doi.org/10.2307/1402588

Brown, E. D., & Williams, B. K. (2018). The potential for citizen science to produce reliable and useful information in ecology. *Conservation Biology*, *33*, 561–569. https://doi.org/10.1111/cobi.13223

Buckland, S. T., Goudie, I. B. J., & Borchers, D. L. (2018). Wildlife population assessment: Past developments and future directions. *Biometrics*, *56*, 1–12. https://doi.org/10.1111/j.0006-341X.2000.00001.x

Cassel, C. M., Sarndal, C. E., & Wretman, J. H. (2017). *Foundations of inference in survey sampling*. New York: John Wiley.

Clark, W. G. (1999). Effect of an erroneous natural mortality rate on a simple age-structured stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences*, *56*, 1721–1731. https://doi.org/10.1139/f99-085

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley.

Conn, P. B., Thorson, J. T., & Johnson, D. S. (2017). Confronting preferential sampling and analyzing population distributions: Diagnosis and model-based triage. *Methods in Ecology and Evolution*, *8*, 1535–1546. https://doi.org/10.1111/2041-210x.12803

Edwards, T. C., Cutler, D. R., Zimmermann, N. E., Geiser, L., & Moisen, G. G. (2006). Effects of sample survey design on the accuracy of classification tree models in species distribution models. *Ecological Modelling*, *199*, 132–141. https://doi.org/10.1016/j.ecolmodel.2006.05.016

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B*, *17*, 69–78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Fisher, R. A. (1958). The nature of probability. *Centennial Review*, *2*, 261–274.

Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya Series C*, *37*, 117–132.

Gerow, K., & McCulloch, C. E. (2000). Simultaneously model-unbiased, design-unbiased estimation. *Biometrics*, *56*, 873–878. https://doi.org/10.1111/j.0006-341X.2000.00873.x

Godambe, V. P., & Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, *54*, 127–138. https://doi.org/10.2307/1403139

Goodyear, C. P. (2019). Modeling growth: Consequences from selecting samples by size. *Transactions of the American Fisheries Society*, *148*, 528–551. https://doi.org/10.1002/tafs.10152

Graybill, F. A. (1976). *Theory and application of the linear model*. Belmont, CA: Duxbury Press.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: Appreciating the difference. *Canadian Journal of Forest Research*, *28*, 1429–1447. https://doi.org/10.1139/x98-166

Hansen, M. H., Madow, W. G., & Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*, 776–793. https://doi.org/10.1080/01621459.1983.10477018

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*, 663–685. https://doi.org/10.1080/01621459.1952.10483446

Irvine, K. M., Rodhouse, T. J., Wright, W. J., & Olsen, A. R. (2018). Occupancy modeling species-environment relationships with non-ignorable designs. *Ecological Applications*, *28*, 1616–1625.

Isaac, N. J. B., van Strien, A. J., August, T. A., de Zeeuw, M. P., & Roy, D. B. (2014). Statistics for citizen science: Extracting signals of change from noisy ecological data. *Methods in Ecology and Evolution*, *5*, 1052–1060. https://doi.org/10.1111/2041-210X.12254

Kéry, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Haffinger, G., & Zbinden, N. (2010). Site-occupancy distribution modeling to correct population trend estimates derived from opportunistic observations. *Conservation Biology*, *24*, 1388–1397. https://doi.org/10.1111/j.1523-1739.2010.01479.x

Kish, L., & Frankel, M. R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B*, *36*, 1–37.

Korn, E. L., & Graubard, B. I. (2003). Estimating variance components by using survey data. *Journal of the Royal Statistical Society, Series B*, *65*, 175–190. https://doi.org/10.1111/1467-9868.00379

Lenhard, J. (2006). Models and statistical inference: The controversy between Fisher and Neyman-Pearson. *British Journal for the Philosophy of Science*, *57*, 69–91. https://doi.org/10.1093/bjps/axi152

Little, R. J. (2004). To model or not to model: Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, *99*, 546–556. https://doi.org/10.1198/016214504000000467

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. https://doi.org/10.2307/271070

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, Series A*, *97*, 558–606. https://doi.org/10.2307/2342192

Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *Revue de L'institut International de Statistique*, *23*, 7–22. https://doi.org/10.2307/1401671

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London, Series A*, *231*, 289–337. https://doi.org/10.1098/rsta.1933.0009

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series A, 169*, 805–827.

Royall, R. M., & Pfeffermann, D. (1982). Balanced samples and robust Bayesian inference in finite population sampling. *Biometrika, 69*, 401–410. https://doi.org/10.1093/biomet/69.2.401

Sarndal, C. E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics, 5*, 27–52.

Skinner, C., & Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science, 32*, 165–175. https://doi.org/10.1214/17-STS614

Stapleton, L. M. (2008). Variance replication using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling, 9*, 475–502.

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research, 44*, 711–740. https://doi.org/10.1080/00273170903333574

Sugden, R. A., & Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika, 71*, 495–506. https://doi.org/10.1093/biomet/71.3.495

Thompson, S. (2012). *Sampling* (3rd ed.). Hoboken, NJ: John Wiley.

Tukey, J. W. (1977). *Exploratory data analysis*. New York: Pearson.

Tukey, J. W. (1980). We need both exploratory and confirmatory. *The American Statistician, 34*, 23–25. https://doi.org/10.1080/00031305.1980.10482706

Williams, B. K., Nichols, J. D., & Conroy, M. J. (2002). *Analysis and management of animal populations*. San Diego, CA: Academic Press.

Yates, F. (1960). *Sampling methods for censuses and surveys* (3rd ed.). London: Griffin.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Williams BK, Brown ED. Sampling and analysis frameworks for inference in ecology. *Methods Ecol Evol.* 2019;10:1832–1842. https://doi.org/10.1111/2041-210X.13279