BIOMETRICS INFORMATION
(You're 95% likely to need this information)

PAMPHLET NO. 63 **DATE:** January 26, 2007

SUBJECT: Comparing Design-based and Model-based Inference: an Introduction

1. Introduction

The goal in survey sampling is to estimate an unknown parameter (such as the mean, total, etc.) of a population – e.g., the volume of timber or the number of moose within a particular area. Sample information is used to generalize to the parameters of the typically larger unknown population. The notion of generalizing to the unknown population, and determining the associated level of uncertainty, is the focus of statistical inference. *Model-based* and *design-based* inference are two rival views for making inference. Each approach to inference can lead to different outcomes, and it is sometimes difficult to decide which approach to use.

The purpose of this note is to introduce and compare model-based and design-based inference from the standpoint of survey sampling. This article will be most relevant to those forest researchers that use sampling to estimate abundance, density, basal area, volume, etc. We may delve into the concepts introduced here in more depth in future pamphlets.

1.1. Design-based Inference

Pure design-based inference is the backbone of traditional sampling theory. Well-known sampling texts such as Cochran (1977) or Scheaffer et al (1996) advocate this type of inference. Here, the population of interest is considered as a finite collection of elements at a particular moment in time. For example, the population might be the trees in a woodlot, the coarse woody debris in a forest, etc.

Design-based inference assumes that the population is fixed (i.e. unchanging). Since almost all natural populations are subject to change, this implies that we are interested in a characteristic or parameter of the population at the instance that our sample is drawn.

Each sample is viewed as a realization of a random process, so a different sample may have chosen a different set of units. The probabilistic nature of the sample is the only source of randomness that plays a part when making inference to the population. Figure 1 is a schematic representation, showing the myriad of possible samples for a given sampling design.

Inference in the design-based setting is generalized beyond the observed sample to all possible samples that could have been selected. This is achieved by considering hypothetical repeated applications of the sampling design. Each of these realizations would produce a different estimate[1] of the population parameter. The distribution of estimates from all possible samples provides a reference distribution, which allows inferences to the unknown population to be made. The variability of this reference distribution is described by the variance or standard error of an estimator. Fortunately, with ordinary sampling designs, the variance of an estimator can be expressed in terms of the sampled elements.

[1] The term *estimator* refers to a function of the sample data used to approximate a population parameter, such as the sample mean $\bar{y}$, which is an estimator of the population mean $\bar{Y}$. An *estimate* is the value of the estimator for a particular case (e.g. the estimate of $\bar{y}$ equals 7.5).

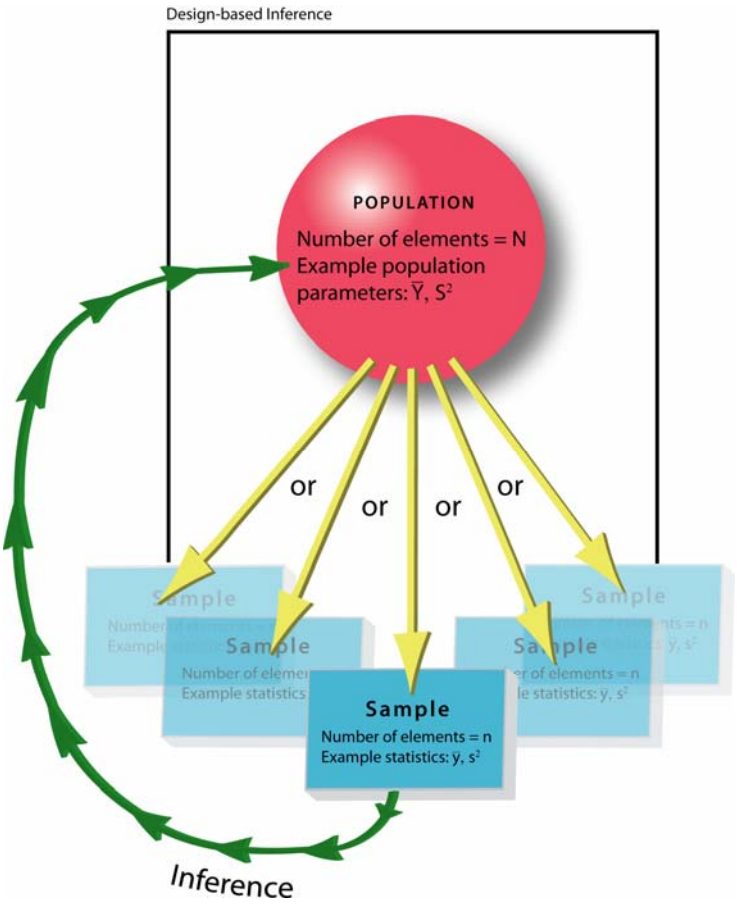

FIGURE 1. Schematic representation of design-based inference. Statistics from an observed sample are used to infer unknown parameters of a fixed population. Here, inference acknowledges that the observed sample is merely one of several possibilities.

With design-based inference, it's important that the estimator matches the sampling design to achieve an unbiased estimator(s). This will be discussed further in §2.

1.2. Model-based Inference

Unlike design-based inference, pure model-based inference does not regard the population of interest as fixed. We assume that an infinite *superpopulation* or *superpopulation model*, which includes a random component, is responsible for creating the elements in the population. One way to think of the superpopulation model is as the process, template or causal system used to create the elements in the population[2]. The specific form of the proposed infinite superpopulation model is often borrowed from classical and well-understood statistical methods such as regression.

As before, we are usually interested in making inference on a characteristic or parameter of the population (e.g. total volume) at the time our sample is drawn. We assume that the superpopulation model that created the population is tied to a specific instant in time, since the population will generally be different (even if imperceptibly) moments after the sample is chosen.

[2] Another view is that there exists an infinite population of finite populations similar but slightly different to ours, and we are observing merely one of them.

In model-based inference, the sampling design (i.e. how the sample was selected) plays little role, and inference stems entirely from the superpopulation model. We acknowledge that the sample is a subset of population elements, and therefore need to predict those population elements not sampled. However, inference is conditional on the observed sample, or any sample for that matter. That is, we make a leap of faith that our sample is a faithful representation of the population and ignore the random nature of the sampling design when making inference. The reference distribution is defined by the countless realizations of the population as governed by the superpopulation model. Figure 2 is a graphical description.

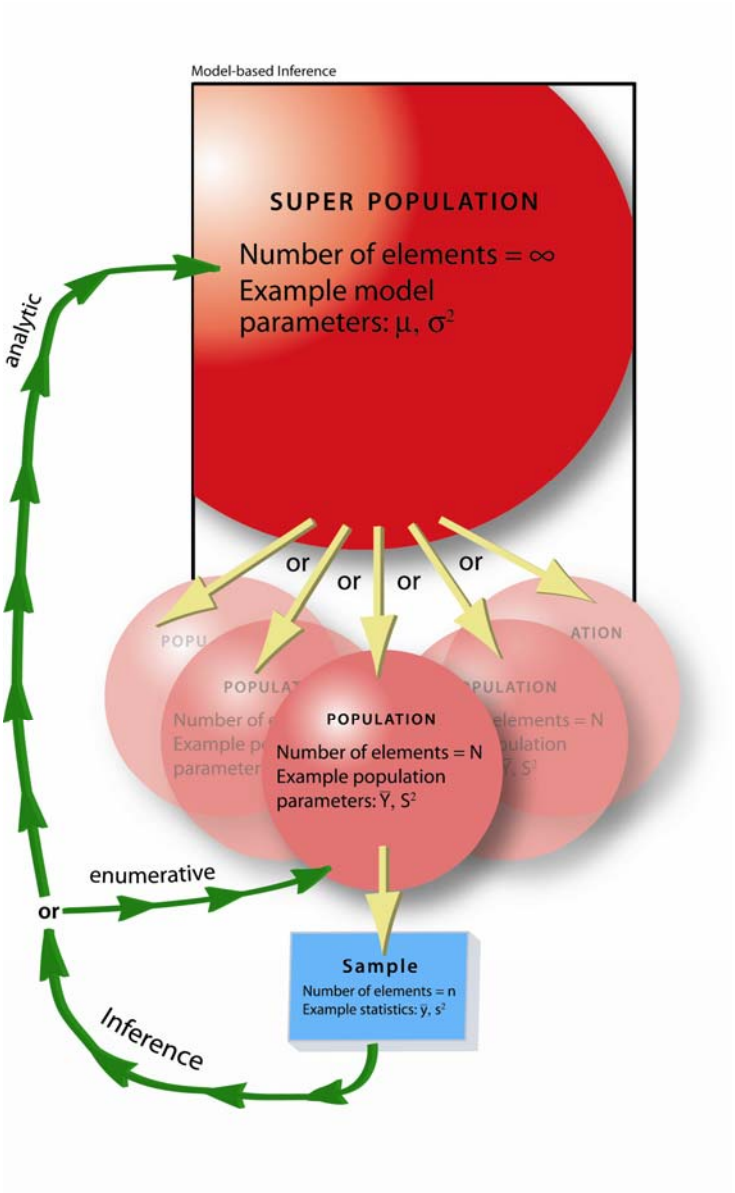

FIGURE 2. Schematic representation of model-based inference. Statistics from an observed sample are used to infer unknown parameters from the finite population, or the infinite superpopulation. We recognize that the sample is subset of elements, but the countless realizations of the population (as imposed by the superpopulation model) are the sole basis for inference.

You can see that it is important to check whether the sample data is adequately described by the proposed model. It is also prudent to sample in a way that is objective, rational and noninformative (Särndal 1978). The formal definition of a noninformative sampling design is fairly mathematical (Binder and Roberts 2001), but the basic idea is to sample elements without regard for the key variable(s) being studied. For example, if the objective is to estimate the average height of trees in a stand, one would not select the 10 tallest trees per plot because the selection criteria would clearly be related to parameter of interest.

Since it is postulated that the population characteristics are random variables generated by the superpopulation model, any function of the data such as a total or mean, will also be a random variable. For this reason uncertainty will exist in our estimate of a population parameter, unless we sample the entire finite population (i.e. conduct a complete census). As in design-based inference, a complete census would identify the population parameter at that moment in time, and its variance would be zero.

Deming (1953) distinguishes between studies where interest lies in a particular population characteristic, from those where interest lies in a specific superpopulation model parameter. The former are called *enumerative*, and the latter *analytic*. So far, we have been describing enumerative studies. In analytic studies, the desire is to extrapolate results to other populations (e.g. different locations), and interest lies in the process or causal system that created the population. In an analytic study, uncertainty around the estimated superpopulation model parameter would remain even after a census.

The key point here is that in the model-based setting, inference can be made to either a population parameter (e.g. total volume) or a superpopulation model parameter. This subtlety is depicted in Figure 2, and will be discussed in more detail near the end of §2.2.

2. Bias and Precision

Sampling theorists use the concepts of bias and precision to describe the quality of an estimator. We will now formally introduce these concepts, since they also help differentiate design-based and model-based inference. The material that follows is somewhat technical but hopefully the general ideas are apparent.

Recall that the customary goal in survey sampling is to estimate a population parameter after selecting a sample. We will first define y_i as a key value or characteristic such as height, volume, etc. associated with each unit in the population. Also, auxiliary information for each unit, such as weight, approximated volume, etc., is denoted as x_i. After taking, say, a simple random sample without replacement (SRS) of n units from the population of N units, we may wish to estimate the population mean per unit $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ with the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Note that uppercase ($\bar{Y}$) is used to symbolize the population parameter, and lower case ($\bar{y}$) is used to symbolize the estimator.

2.1. Design-based Inference

The bias measures how, on average, the sample mean differs from the true population mean. To begin with, let $p(t)$ represent the probability of observing each sample t[3]. For SRS, there

[3] We have used "*t*" rather than the more obvious choice "*s*" so that *s* can be reserved to represent the standard deviation.

are $\binom{N}{n}$ ways to chose *n* units from *N*, and each is equally likely. Therefore, the probability of observing each sample (consisting of *n* elements) is $p(t) = 1 \Big/ \binom{N}{n}$. For instance, the chance of matching all 6 of your numbers in the "Lotto 649" is $1 \Big/ \binom{49}{6} = 1/13{,}983{,}816$.

The expectation of the estimator $\bar{y}$ is defined as the average value of the estimator over all possible samples. Say each hypothetical sample produces an estimator of the mean $\bar{y}_t$, then the expectation is the weighted sum:

$$E(\bar{y}) = \sum_{t}^{\text{all possible samples}} \bar{y}_t \cdot p(t)$$

The estimator $\bar{y}$ will be design-unbiased if its expectation over all possible samples equals $\bar{Y}$. The formal definition is:

$$Bias(\bar{y}) = E(\bar{y} - \bar{Y}) = E(\bar{y}) - \bar{Y} .$$

Keep in mind that the population (including $\bar{Y}$) is fixed and we need not make any distributional assumptions on how it was generated, or the spatial distribution of the units in the population.

Accuracy of an estimator is measured using the mean squared error (*MSE*), which captures both precision and bias (see Appendix 1 for complete derivation):

$$MSE(\bar{y}) = E\left[(\bar{y} - \bar{Y})^2\right] = Var(\bar{y}) + [Bias(\bar{y})]^2 .$$

The above relationship is often shown pictorially using the well-known "dartboard" analogy (e.g. Figure 2.2 of Lohr (1999))[4].

The variance of $\bar{y}$ conveys how the estimated mean varies among hypothetical samples:

$$Var(\bar{y}) = E\left[(\bar{y} - E(\bar{y}))^2\right] = \sum_{t}^{\text{all possible samples}} [\bar{y}_t - E(\bar{y})]^2 \cdot p(t) .$$

The variance occasionally simplifies to something that looks more familiar, as in the case of simple random sampling[5]:

$Var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$, where S^2 is the squared population standard deviation:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2 .$$

[4] The formulae that define Bias and *MSE* are applicable to any estimator, not just $\bar{y}$.

[5] These steps have been omitted since the algebra is quite lengthy. See § 2.9 of Cochran (1977) if curious.

The above variance and standard deviation require information from the population that is unknown – remember we merely have a sample – so we normally work with an estimate of $Var(\bar{y})$. We maintain the convention of using lower case since it is an estimator:

$var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$, where s^2 is the squared sample standard deviation

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

For some sampling designs, it is also possible to assess the bias of our estimated variance using:

$Bias[var(\bar{y})] = E[var(\bar{y})] - Var(\bar{y})$, where $E[var(\bar{y})]$ represents the expected sample variance overall all possible samples: $E[var(\bar{y})] = \sum_{t}^{\text{all possible samples}} var(\bar{y}_t) \cdot p(t)$.

For the above SRS example, the squared sample standard deviation s^2 allows the estimated variance $var(\bar{y})$ to be unbiased, but some designs do not have an unbiased estimator of precision available (e.g. a single systematic sample with random start).

2.2. Model-based Inference

Now we will presume that an infinite superpopulation or superpopulation model is responsible for creating the elements in the finite population of interest. An estimator is model-unbiased if the expected discrepancy between the estimator and the finite population value is zero over repeated realizations of the population.

Expectation of the estimator is conditional on the observed sample, so the sampling design that produced the sample is irrelevant in our assessment. In other words, we recognize that the sample represents only a fraction of the population, but we don't pay attention to how it was selected. In pure model-based inference, we are content that our sample is a valid representation of the population and do not bother worrying about the matter further. This does not imply that sampling is unimportant in model-based inference. In fact, the opposite is true. A poorly chosen sample will always lead to poor (i.e. misleading) results because inference is limited to the observed sample.

For example, let $E_m(\bar{y} \mid \text{sample})$ denote the expectation of the sample mean $\bar{y}$ conditional on the observed sample. Here the subscript *m* is used as a reminder that expectation is with respect to a model. Also let $E_m(\bar{Y})$ denote the expectation of the population mean $\bar{Y}$. Keep in mind that unlike the design-based setting, where $\bar{Y}$ is a fixed quantity, both $\bar{y}$ and $\bar{Y}$ are now random variables and we must work with the expected values for both. The formal definition for the model-bias of the sample mean $\bar{y}$ is then:

$$Bias_m(\bar{y}) = E_m(\bar{y} - \bar{Y} \mid \text{sample}) = E_m(\bar{y} \mid \text{sample}) - E_m(\bar{Y}).$$

For example, we might postulate the following superpopulation model: all *N* elements in the population are produced as independent realizations from a normal distribution having mean μ and variance σ^2. It's easy to show that under this scenario, $E_m(\bar{y} \mid \text{sample}) = E_m(\bar{Y}) = \mu$, so $\bar{y}$ is a model-unbiased estimate of $\bar{Y}$.

Model-based estimators determined via maximum likelihood are asymptotically unbiased (i.e. as the sample size *n* approaches ∞)[6]. This is one reason why maximum likelihood estimators (MLEs) are so popular.

In survey sampling, the accuracy of any point estimator, such as $\bar{y}$, can be evaluated using the *MSE*, but it has a slightly different interpretation than its design-based cousin. Here it is defined as the expected squared deviation between the estimated mean and the population mean (over repeated possible realizations of the population), conditional on the observed sample (see Appendix 2 for complete derivation):

$$MSE_m(\bar{y}) = E_m\left[(\bar{y}-\bar{Y})^2 \mid \text{sample}\right] = Var_m(\bar{y}-\bar{Y} \mid \text{sample}) + [Bias_m(\bar{y})]^2$$

Above, $Var_m(\bar{y}-\bar{Y} \mid \text{sample})$ represents the variance of the difference between two random variables $\bar{y}-\bar{Y}$. This distinctive variance occurs because the population characteristics are not fixed, and the model-based *MSE* must capture variability due to units from the sample, and all population units (including those that have not been sampled) that could have been generated from the superpopulation model. Also, $Var_m(\bar{y}-\bar{Y} \mid \text{sample})$ is generally different from $Var_m(\bar{y} \mid \text{sample})$.

You can also see that $MSE_m(\bar{y})$ equals $Var_m(\bar{y}-\bar{Y} \mid \text{sample})$ only when the bias is zero. Statisticians often concentrate on the variance rather than the *MSE*, so that precision of the estimators can be studied. However, the *MSE* reveals the subtle difference between the design-based and model-based concepts of variance. We have shown that seemingly identical definitions of *MSE* lead to different notions of variance between the two modes of inference. The design-based variance describes the variability in the estimator alone (due to sampling), whereas the model-based variance describes the variability in both the estimator and the population parameter (both due to the superpopulation model).

To save space, the remaining discussion will drop the conditioning on "sample" from all formulae pertaining to model-based inference.

Under the simple Normal superpopulation model described above, it turns out[7] that

$$MSE_m(\bar{y}) = Var_m(\bar{y}-\bar{Y}) = \left(1-\frac{n}{N}\right)\frac{\sigma^2}{n},$$

which appears almost identical to the design-based version. However, we cannot assume this equivalence will hold with other models.

The typical estimator for σ^2 is the squared sample standard deviation $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$, although another possibility is the maximum likelihood estimator (MLE) : $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$.

To determine whether an estimator of the variance is model-unbiased we would examine:

$$Bias_m\left[var_m(\bar{y}-\bar{Y})\right] = E_m\left[var_m(\bar{y}-\bar{Y})\right] - Var_m(\bar{y}-\bar{Y}),$$

---

[6] One potential drawback is that an estimator might still be biased if *N* is moderately small, even when sampling a sizeable fraction of the population (i.e. $n/N \cong 1$).

[7] See § 2.8 of Lohr (1999) for more detail.

where $E_m\left[var_m\left(\bar{y}-\bar{Y}\right)\right]$ is the expected (i.e. average) estimated variance taken over all possible realizations of the population. Usually, we can only assess this via simulation. In the above case, s^2 leads to an unbiased estimate of $Var_m\left(\bar{y}-\bar{Y}\right)$, whereas $\hat{\sigma}^2$ leads to an estimate that is slightly biased for small samples.

Recall that μ and σ^2 are the unknown model parameters of the infinite superpopulation model. We have already seen that σ^2 is required to determine $Var_m\left(\bar{y}-\bar{Y}\right)$. But what if we conducted an analytic study, and wanted to make inference on the superpopulation mean μ rather than finite population mean $\bar{Y}$? Perhaps we were interested in the process that generated data, so we could generalize to several populations. For instance, we might have a sample from one of many comparable populations, all believed to be generated by the same superpopulation model.

The MLE of μ turns out to be identical to the sample mean $\bar{y}$. However, this is not the case in general, especially if a more complicated model is assumed (e.g. one where the y_i are not independent). The variance of the MLE, say $Var_m(\hat{\mu})$, is $\frac{\sigma^2}{n}$, which differs from $MSE_m(\bar{y})$ or $Var_m\left(\bar{y}-\bar{Y}\right)$ by not having the finite population correction factor (fpc) $\left(1-\frac{n}{N}\right)$. The increase in variance is the cost of making inference to a broader space (i.e. beyond the finite population to the infinite superpopulation).

3. Confidence Intervals

While both design-based and model-based approaches may lead to the same confidence interval (CI) when they produce the same variance for an unknown parameter, the interpretations are different.

The design-based CI has a repeated sampling interpretation. To begin with, let α denote the level of significance (e.g. $\alpha = 0.05$). If we take all possible simple random samples[8] of size *n* from *N* fixed population units and construct a $100\cdot(1-\alpha)\%$ CI for each sample, then $100\cdot(1-\alpha)\%$ of these CI's will include the true population mean $\bar{Y}$. Thanks to the central limit theorem, we can then use a parametric distribution such as the Normal or Student's *t* to build design-based CI's[9].

The model-based CI can be interpreted by considering repeated realizations of the population due to the superpopulation model. If we repeatedly (a) generate individual units within the population using the model, (b) select a sample, and (c) construct a $100\cdot(1-\alpha)\%$ CI from each resulting sample, then $100\cdot(1-\alpha)\%$ of these CI's will include the true superpopulation parameter μ[10]. The parametric distribution required to build a model-based CI follows from the model itself. For example, in our model-based example, it follows that $\bar{y}\sim N\left(\mu,\sigma^2/n\right)$, so the normal distribution provides the basis for the CI.

---

[8] Sampling designs different from SRS are obviously also permissible.

[9] This is a bit paradoxical, since the assumed distribution is in essence, a type of model.

[10] Another way to think of this is with the notion of an infinite population of finite populations. If we could sample all of the finite populations and construct a CI for each one, $100\cdot(1-\alpha)\%$ of these would contain the true mean (of the infinitely many populations).

4. Conclusion

In an attempt to unify both types of inference, one could regard design-based inference as the case where inference is conditioned on (i.e. limited to) a particular realization from the superpopulation model, and model-based inference as the case where inference is conditioned on the observed sample. A combined view is also possible, where both sources of randomness are incorporated into inference; this idea is developed further in §12.2 of Särndal et. al. (1992).

Thompson (1992) states that "even the best model is something one not so much believes as tentatively entertains." For pure model-based inference, it's important to check that the assumed model appropriately describes the sample data. If you cannot unearth an appropriate model, then the design-based approach may be best. Also, we have not discussed missing data, non-response or measurement error, but the usual way to deal with these issues is with some sort of model.

Ibid. lists some advantages of the two approaches to inference. He points out that design-based inference is good at:

- obtaining unbiased point estimators (and estimators of variance) that do not depend upon on assumptions about the population - a sort of nonparametric approach,
- obtaining estimators acceptable to users with differing interests, and
- avoiding ordinary human biases in sample selection.

He also indicates that model-based inference excels at**:**

- assessing the efficiency of standard sampling designs and estimators under different assumptions about the population,
- deriving estimators that make the most efficient use of the sample data, including auxiliary information, and
- dealing with observational data obtained without a proper sampling design.

Careful consideration of the last point is in order. Model-based inference is not an excuse for a poor sample design!

Design-based and model-based interference can lead to different point estimators and/or variances. It is also common to find an estimator that is model-unbiased but not design-unbiased or vice versa. In these situations, it's important to understand the underlying concepts that characterize the two approaches. The excellent paper by Gregoire (1998) may shed more light in this area. Future pamphlets may build on the foundation developed here, and examine in detail examples where the two modes of inference do lead to different estimators.

Prepared by:

Peter Ott. Phone: 250-387-7982; email: peter.ott@gov.bc.ca
B.C. Ministry of Forests and Range, Research Branch

References

Binder, D.A. and G.R. Roberts. 2001. Can informative designs be ignorable? Newsletter of the Survey Research Methods Section (ASA) 12: 1-3.

Cochran, W. G. 1977. Sampling Techniques, 3rd Edition. John Wiley & Sons, Inc. New York, NY.

Deming, W.E. 1953. On the distinction between enumerative and analytic surveys. JASA 48: 244-255.

Gregoire, T. G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. 28: 1429-1447.

Lohr, S. L. 1999. Sampling: Design and Analysis. Duxbury Press. Pacific Grove, CA.

Särndal, C. E. 1978. Design-based and model-based inference in survey sampling. Scandinavian Journal of Statistics 5: 27-52.

Särndal, C. E., Swensson, B. and J. Wretman. 1992. Model Assisted Survey Sampling. Springer-Verlag. New York, NY.

Scheaffer, R. L., Mendenhall, W. and L. Ott. 1986. Elementary Survey Sampling, Third Edition. Duxbury Press. Boston, MA.

Thompson, S. K. 1992. Sampling. John Wiley & Sons, Inc. New York, NY.

Appendix 1. Design-based *MSE*

$$
\begin{aligned}
MSE(\bar{y}) &= E\left[\left(\bar{y} - \bar{Y}\right)^2\right] \\
&= E\left[\left(\bar{y} - \bar{Y} - E(\bar{y}) + E(\bar{y})\right)^2\right] \\
&= E\left[\left[\left(\bar{y} - E(\bar{y})\right) + \left(E(\bar{y}) - \bar{Y}\right)\right]^2\right] \\
&= E\left[\left(\bar{y} - E(\bar{y})\right)^2\right] + E\left[\left(E(\bar{y}) - \bar{Y}\right)^2\right] + 2 \cdot E\left[\left(\bar{y} - E(\bar{y})\right)\left(E(\bar{y}) - \bar{Y}\right)\right] \\
&= E\left[\left(\bar{y} - E(\bar{y})\right)^2\right] + \left[E\left(\bar{y} - \bar{Y}\right)\right]^2 + 2 \cdot (0) \\
&= Var(\bar{y}) + \left[Bias(\bar{y})\right]^2
\end{aligned}
$$

Appendix 2. Model-based *MSE*

$$
\begin{aligned}
MSE_m(\bar{y}) &= E_m\left[\left(\bar{y} - \bar{Y}\right)^2 \mid \text{sample}\right] \\
&= E_m\left[\left(\bar{y} - \bar{Y}\right)^2 \mid \text{sample}\right] - \left[Bias_m(\bar{y})\right]^2 + \left[Bias_m(\bar{y})\right]^2 \\
&= E_m\left[\left(\bar{y} - \bar{Y}\right)^2 \mid \text{sample}\right] - \left[E_m\left(\bar{y} - \bar{Y} \mid \text{sample}\right)\right]^2 + \left[Bias_m(\bar{y})\right]^2 \\
&= E_m\left[\left[\left(\bar{y} - \bar{Y}\right) - E_m\left(\bar{y} - \bar{Y}\right)\right]^2 \mid \text{sample}\right] + \left[Bias_m(\bar{y})\right]^2 \\
&= Var_m\left(\bar{y} - \bar{Y} \mid \text{sample}\right) + \left[Bias_m(\bar{y})\right]^2
\end{aligned}
$$