Statistical Journal of the IAOS 31 (2015) 555–563
DOI 10.3233/SJI-150944
IOS Press

Calibrated Bayes, an inferential paradigm for official statistics in the era of big data

Roderick J. Little
Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48105, USA
E-mail: rlittle@umich.edu

Abstract. Official statistics is dipping its toe in the ocean of big data, and leaders are emphasizing the need for a major paradigm change. One aspect is the increased volume of data that are not collected on probability samples of the target population. Making full use of these data requires a fundamental change, not only in data collection and dissemination, but also in the methods of statistical inference. The classical "design-based" approach to survey inference, developed from the seminal work of Neyman [41], is simply not applicable to these data. Rather, statistical models are needed that potentially reflect selection bias from the lack of random sampling. I suggest that "Calibrated Bayes" is the appropriate statistical paradigm for addressing the analysis. Under this paradigm, inferences for a particular data set are Bayesian, but models are sought that yield inferences with robust repeated sampling properties. Probability sampling remains a powerful tool under this paradigm, since by ensuring that the selection mechanism is ignorable it enhances robust modeling, but it is not essential for the inference. I outline two applications of Calibrated Bayes to data collected by the U.S. Census Bureau.

Keywords: Bayesian statistics, big data, frequentist statistics, robust models, statistical inference

1. Introduction

The winds of change can be felt in the field of official statistics right now, and with the increasing proliferation of "big data", leaders in official statistics are calling for a paradigm change [43]. Robert Groves, a recent Director of the U.S. Census Bureau, put it this way:

"For decades, the Census Bureau has created "designed data" in contrast to "organic data." The questions we ask of businesses and households create data with a pre-specified purpose, with a use in mind. Indeed, designed data through surveys and censuses are often created by the users. This means that the ratio of information to data (for those uses) is very high, relative to much organic data. . . What has changed is that the volume of organic data produced as auxiliary to the Internet and other systems now swamps the volume of designed data. In 2004 the monthly traffic on the internet exceeded 1 exabyte or 1 billion gigabytes. The risk of confusing data with information has grown exponentially. . . The challenge to the Census Bureau is to discover how to combine designed data with organic data, to produce resources with the most efficient information-to-data ratio. This means we need to learn how surveys and censuses can be designed to incorporate transaction data continuously produced by the internet and other systems in useful ways. Combining data sources to produce new information not contained in any single source is the future. I suspect that the biggest payoff will lie in new combinations of designed data and organic data, not in one type alone." US Bureau of the Census Director's Blog, September 2011.

Surveys and censuses are expensive and challenging to mount, nonresponse is increasing, and (as Groves notes), non-probability sample sources of data are increasingly available. Combining information from surveys and other sources is reasonable and attractive in principle, but difficult in practice. Disseminating information for small areas is subject to the dangers from disclosure of confidential information from respondents.

The standard design-based approach of taking a random sample of the target population and weighting the

results up to the population is inadequate for this environment. Combining data from traditional probability surveys with administrative records and other information gleaned from cyberspace requires modern statistical modeling tools. Specifically, robust models are needed that correct for measurement errors and selection biases, yielding estimates with reliable statistical properties – for example confidence intervals that have close to nominal coverage.

With this backdrop, I discuss here limitations of the current prevailing philosophy of inference, which I call the Design-Model Compromise (DMC), and outline an alternative inferential paradigm, Calibrated Bayes (CB), which I think is a useful conceptual framework for melding probability sample data with various forms of "big data". These ideas are discussed in some detail elsewhere [39,40], so my goal here is to provide an overview.

2. The prevailing philosophies of statistical inference in official statistics

The classical randomization or design-based approach (e.g. [9,26,31]) treats the values of survey variables as fixed, and bases statistical inference on the distribution of estimates repeated sampling from the population. The inclusion indicators (say $\mathbf{I}$) for whether units are included or excluded from the sample are the random variables, and inferences are then generally based on normal large-sample approximations. For example, a 95% confidence interval for Q is $\hat{q} \pm 1.96\sqrt{\hat{v}}$, where $\hat{q}$ is an estimate of Q, $\hat{v}$ is an estimate of variance, and 1.96 is the 97.5^{th} percentile of the standard normal distribution.

Models play a role in determining the choice of estimator in this approach. Specifically, regression or ratio estimates are based on implicit models, and model-assisted methods such as generalized regression [53] incorporate model predictions. However, these methods remain fundamentally design-based, since the distribution of sample inclusion remains the basis for inference.

In contrast, the model-based approach derives inferences from a model for the distribution of the survey variables (say $\mathbf{Y}$), perhaps combined with a distribution of $\mathbf{I}$. Initial model formulations did not overtly assign a distribution for $\mathbf{I}$, but modeling both $\mathbf{Y}$ and $\mathbf{I}$ allows assumptions about the method of selection to be formalized, and clarifies the value of probability sampling. The model is used to predict the non-sampled values of the population, and hence finite population quantities Q. There are two major variants: superpopulation modeling and Bayesian modeling.

In superpopulation modeling (e.g. [47,62,63]), the population values of $\mathbf{Y}$ are assumed to be a random sample from a "superpopulation", and assigned a probability distribution $p(\mathbf{Y}|\mathbf{Z}, \theta)$ indexed by fixed parameters θ, and conditioned on known design variables Z. Inference are based on predictions of non-sampled values from this model, with the parameter estimated by maximum likelihood or some such principle.

Bayesian survey inference [2,3,17–19,25,35,36,50, 52,59,60] requires in general the specification of a prior distribution $p(\mathbf{Y}, \mathbf{I}|\mathbf{Z})$ for the population values $\mathbf{Y}$ and inclusion indicators $\mathbf{I}$. Inferences for finite population quantities $Q(\mathbf{Y})$ are then based on the posterior predictive distribution of Q, given the data. The prior distribution $p(\mathbf{Y}, \mathbf{I}|\mathbf{Z})$ is often specified with a parametric model $p(\mathbf{Y}, \mathbf{I}|\mathbf{Z}, \theta)$ indexed by parameters θ, combined with a prior distribution $p(\theta|\mathbf{Z})$ for θ, that is:

$$p(\mathbf{Y}, \mathbf{I}|\mathbf{Z}) = \int p(\mathbf{Y}, \mathbf{I}|\mathbf{Z}, \theta) p(\theta|\mathbf{Z}) d\theta.$$

This general formulation includes a model for the inclusion indicators $\mathbf{I}$, which is potentially necessary in "big data" settings where data are not randomly sampled. The inclusion mechanism is generally difficult to model, and vulnerable to model misspecification. If, however, the inclusion mechanism is *ignorable* [49], then the distribution of the sample inclusion indicator $\boldsymbol{I}$ is not needed in this model, simplifying the modeling task to specification of the prior distribution $p(\mathbf{Y}|\mathbf{Z})$. An ignorable sampling mechanism is highly desirable for robust inference.

By an application of Rubin's [49] theory for missing data [35], a sufficient condition for ignoring the selection mechanism for Bayesian inference is that:

$$p(\mathbf{I}|\mathbf{Y}, \mathbf{Z}) = \mathbf{p}(\mathbf{I}|\mathbf{Y}_{\text{obs}}, \mathbf{Z}) \text{ for all } \mathbf{Y}_{\text{mis}}, \tag{1}$$

where $\mathbf{Y}_{\text{obs}}$ is the observed part of $\mathbf{Y}$ and $\mathbf{Y}_{\text{mis}}$ is the missing part of $\mathbf{Y}$. In particular, under *random sampling*,

$$p(\mathbf{I}|\mathbf{Y}, \mathbf{U}, \mathbf{Z}) = p(\mathbf{I}|\mathbf{Z}) \text{ for all } \mathbf{Y} \tag{2}$$

where $\mathbf{U}$ represents unobserved variables in the population, and $p(\mathbf{I}|\mathbf{Z})$ is known and determined by the probability sampling design. Note that Eq. (2) implies Eq. (1). The value of randomization for Bayesian sur-

vey inference is that it provides a practical way to ensure that the selection mechanism is *ignorable* for inference [23,48,61, Chapter 7] [35].

I emphasize the following differences between the sufficient condition for ignorability, Eq. (1), and the random sampling condition, Eq. (2):

1. Equation (1) is weaker than Eq. (2), that is Eq. (2) implies Eq. (1) but Eq. (1) does not imply Eq. (2).
2. Equation (1) is a modeling assumption, whereas Eq. (2) is not an assumption, providing units were randomly sampled according to the sampling design.
3. Equation (1) is outcome-specific, in that it might hold for some survey outcomes, but not for others. Equation (2) applies for any outcome, including variables not included in the survey.
4. Absent probability sampling, Rubin [49] is the appropriate theoretical framework, and a key issue is whether the selection is ignorable – is Eq. (1) a reasonable assumption? The answer to this question is a matter of degree and is survey variable-specific, since inferences for some variables may be more seriously biased by the sample selection than inferences for others.

Assessing ignorability is one of the central challenges of inference from non-probability samples, and is not straightforward. On the other hand, making randomization the basis for inference, as with the design-based approach, is too restrictive, since it does not provide a framework for handling deviations from randomization, or other non-sampling errors.

Under ignorable selection and in large samples, the effect of the prior distribution $p(\theta|\mathbf{Z})$ on the parameters goes away, providing its support assigns positive prior probability to non-sampled values in $\mathbf{Y}$. Bayesian inferences under ignorable selection are then similar to inferences from superpopulation models with the same choice of $p(\mathbf{Y}|\mathbf{Z},\theta)$. However, for small sample problems Bayes is arguably superior, since uncertainty about unknown parameters is reflected in the inference [36,39].

Design-based and model-based systems of statistical inference both have strengths and weaknesses, and the key is to combine them in a way that capitalizes on their strengths. The "status quo" in current official statistics, which might be termed the "design/model compromise" (DMC [39]), favors design-based inference for descriptive statistics like means and totals based on large probability samples, and model-based inference for questions that are not well addressed by the design-based approach, such as small area estimation, survey nonresponse and response errors (e.g. [30,45,46]). The approach is pragmatic, but entails a degree of "inferential schizophrenia". This can give rise to inconsistencies, as discussed in Section 4 below for small-area estimation. The conflict between design-based and model-based inference creates confusion when statisticians impose design-based standards on substantive modelers who are not familiar with the design-based paradigm. See Little [39] for more discussion of these issues.

In the world of "big data", the design-based part of DMC no longer works, since the design-based approach makes no sense without probability sampling. An alternative compromise between randomization and modeling, Calibrated Bayes (CB), avoids "inferential schizophrenia" by assigning distinct roles to models (for the inference) and frequentist methods (for formulating and assessing the model). I now review this inferential paradigm.

3. Calibrated Bayesian (CB) inference

In CB, *all* inferences are explicitly Bayesian and hence model-based, but models are sought to yield inferences that are well calibrated in a frequentist sense; specifically, models are sought that yield posterior credibility intervals with (approximately) their nominal frequentist coverage in repeated sampling. Seminal references are Box [5] and Rubin [51]. Here I summarize some of the arguments for CB in the context of survey sample inference, presented in more detail elsewhere [3,38,39].

Frequentist inference is in essence a set of concepts, like unbiasedness, consistency, confidence coverage, efficiency, and robustness, for assessing *properties* of inference procedures. It is not a prescriptive system leading to a clear choice of estimator and inference. Of the many frequentist tools, such as least squares, method of moments, generalized weighting equations or maximum likelihood (ML), asymptotic inferences based on ML seem the closest to being prescriptive, but ML is not satisfactory for small sample inference. Exact small-sample inferences have been developed for some problems, but in many others there is no exact frequentist method, in the sense of yielding a confidence interval that has exact nominal confidence coverage for all values of the unknown parameters.

Design-based survey inferences are not only asymptotic, they fail for probability sampling schemes where the number of distinct repeated samples is limited. For

example, consider systematic sampling of units with a sampling interval of five, from a random start. The design-based standard error exists, but design-based estimates of standard error are not available, and since there are only five possible repeated samples and hence five possible estimates, design-based 90% or 95% confidence intervals do not exist. Models are needed to create and provide meaning to interval estimates.

Frequentist inference violates the likelihood principle [4], and is ambiguous about whether to condition on ancillary or approximately ancillary statistics when performing repeated sampling calculations [10, 11]. In the sample survey context, this issue arises in the question of whether the sampling distribution of post-stratified means should condition on post-stratum counts [27,34].

The Bayesian approach avoids these problems with frequentist inference. Once a model and prior distribution are specified, there is a clear path to inferences based on the posterior distribution, or optimal estimates for a given choice of loss function. Problems of inference under a model become purely computational, and a rich array of Bayesian computational tools are now available, even for complex high-dimensional problems. The likelihood principle is satisfied, issues about conditioning on ancillary statistics do not arise, and uncertainty about nuisance parameters is propagated by integrating them over their posterior distribution, an approach that (with noninformative prior distributions) leads to better small-sample inferences than ML.

The problem with Bayesian inference in practice is that it generally requires full specification of a likelihood and prior, and we never know the true model [15]. All models are wrong, and bad models lead to bad answers: under the frequentist paradigm, the search for procedures with good frequentist properties provides a degree of protection against model misspecification; there is no such built-in protection under a strict Bayesian paradigm where frequentist properties are not entertained.

We want model-based inferences with good frequentist properties, such as 95% credibility intervals that cover the unknown parameter approximately 95% of the time if the procedure was applied to repeated samples. In the absence of probability sampling, these repeated samples can be conceptualized as data drawn from a superpopulation according to some model, and robustness concerns retaining good properties in the face of alternative assumptions about that model. The Bayesian has some important tools for model development and checking, like Bayes factors and model averaging, but in my view frequentist ideas are essential when it comes to model development and assessment.

A natural compromise is thus to use frequentist methods for model development and assessment, and Bayesian methods for inference under a model. This capitalizes on the strengths of both paradigms, and is the essence of Calibrated Bayes (CB) [5,12,14,42,51, 64]. Rubin [51] wrote that

> "The applied statistician should be Bayesian in principle and calibrated to the real world in practice – appropriate frequency calculations help to define such a tie... frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is the calibration of Bayesian probabilities to the frequencies of actual events."

3.1. Calibrated Bayes inference for sample surveys

The CB approach is prescriptive given the model, but there is no clear prescription on how to build models that are well calibrated – this remains somewhat of an art in the current state of knowledge. However, CB has certain implications for sample survey inference. The main features that distinguish survey sampling inference from other areas of statistics are (a) the focus on descriptive finite population quantities (though analytic parameters are also of interest) and (b) the emphasis on probability sample designs with features like stratification, weighting and clustering, which render simple "iid" assumptions invalid.

Concerning (a), Bayesian inference for finite population quantities is based on the posterior predictive distribution. The target population quantity does not need to be a parameter of the CB model used for inference; it could be the quantity obtained by applying a "target model" to the full population. CB inference is then based on the posterior predictive distribution of this finite population quantity, for an "analysis model", which captures key features of the sample design, and which may differ from the target model.

Concerning (b), the "calibrated" part of CB is enhanced by probability sampling, which make the selection mechanism ignorable and hence allows for simplified and robust inference. However, CB models need to incorporate explicitly design features like stratification, weighting and clustering, since models that ignore these features are vulnerable to model misspecification. Thus with disproportionate stratified sam-

pling, stratum effects need to be included in the model, leading to large sample Bayesian inferences that correspond to standard design-based weighted estimates, and small sample inference that are superior since they incorporate uncertainty in estimating variances [39]. Models that ignore stratum effects lead to unweighted estimates of means, which are potentially very biased if the stratum means are related to the stratum selection rates. More generally, the role of sampling design features like survey weights in multiple regression is discussed in Little [39]. Frequentist concepts like design consistency or asymptotic design unbiasedness [6,28] are useful in developing CB models for probability samples, since design-consistency tends to promote good confidence coverage, particularly in large samples; the class of Bayesian models that yield design consistent estimates is very broad [20], so design consistency is relatively easy to achieve under the CB paradigm.

With non-probability samples, issues of selection bias are not solved by the CB paradigm, but CB provides a useful theoretical framework for addressing them. Thus one strategy is to incorporate auxiliary population information into the model, using poststratification or other approaches [34] to reduce the selection bias explained by these factors.

Other features of CB models for surveys are (a) relatively weak prior distributions, so that the evidence in the data overshadows the evidence in the prior; and (b) model checks to ensure that models are not contradicted by the data. The latter point should not be controversial, since any statistical approach, frequentist or Bayesian, needs to evaluate assumptions. Diagnostic approaches include posterior predictive checks [24, 51], and cross-validation approaches [13,14].

4. Two applications of the CB perspective

The DMC philosophy suggests that when there are sufficient data to support "direct" estimates that do not borrow strength across subdomains, inferences are design-based, but when the data are too limited then model-based small area estimates are acceptable. This dichotomy implies, for any particular survey, the existence of a tipping point (say n_0), the "point of inferential schizophrenia", such that inferences are design-based when $n > n_0$ and model-based when $n < n_0$. For planned domain estimates, the choice of n_0 is sometimes based on a pre-set coefficient of variation; or it may be unspecified. In either case the choice is somewhat arbitrary, and it is troubling that one's entire philosophy of statistics, and the nature of the estimator, changes depending on where the sample size falls relative to this value. In particular, the (design-based) confidence intervals for the mean for sample sizes slightly more than n_0 will tend to be wider than the (model-based) confidence intervals for the mean for sample sizes slightly less than n_0, even though they are based on more data.

The CB philosophy avoids this inconsistency. Hierarchical Bayes models yield estimates close to "direct" estimates when sample sizes are large, and as the sample size decreases, move seamlessly towards predictions from a fixed-effects model. Thus, for a typical hierarchical Bayes model, the posterior mean of the population mean $\bar{Y}_a$ in area a, given covariate information X, has the form

$$E(\bar{Y}_a|\text{ data}) = w_a\bar{y}_a + (1-w_a)(\bar{y} + \hat{\beta}(\bar{x}_a - \bar{X}), \quad (3)$$

where $\bar{y}_a, \bar{x}_a, n_a$ are the sample means of Y and X and sample size in area a, $(\bar{y} + \hat{\beta}(\bar{x}_a - \bar{X})$ is the regression prediction for the mean of Y aggregated over all areas, and w_a assigns most of the weight to the sample mean when n_a is large, and most of the weight to the regression prediction over all areas when n_a is small.

The weights in Eq. (3) depend on the between-area variance τ^2 and within-area variance σ^2 , which in practice need to be estimated. Superpopulation modeling approaches replace the variances by point estimates, typically computed by the method of moments or maximum likelihood. When these variance estimates go negative, they are replaced by a value 0 on the boundary of the parameter space, and uncertainty in the variance estimates is not reflected in inferences. Fully Bayes methods based on weak priors on the variance components propagate uncertainty and avoid estimates on the boundary of the parameter space, though care is needed with the choice of prior distribution for τ^2 [22]. As a result, Bayes estimates tend to be better calibrated, that is, yield credibility intervals with better confidence coverage.

Example 1: Language Provisions of the Voting Rights Act. The United States Voting Rights Act determines that certain counties and townships are required to provide language assistance at the polls. Let P denote the proportion of voting age citizens in a political district who (a) are members of a single language minority and (b) are limited English Proficient (LEP). A key (if not sole) criterion for requiring language as-

sistance is that $P > 0.05$. The U.S. Census Bureau is charged with determining which jurisdictions are covered under the Act, and until now have used direct estimates of $P$from Long Form Decennial Census Data. With the replacement of the long form, estimates are henceforward to be based on the smaller ACS, and some districts have small ACS samples and hence have direct estimates of P with unacceptably high variance. The 2011 determinations use a small-area model that combines information from the 2005–2009 ACS and 2010 Census data. The approach to the "more than 5%" provision was to:

(a) to build a district level regression model to predict P based on variables in the ACS, and Census 2010 counts of minority groups;
(b) classify districts into classes with similar predicted P based on the model – an approach known as predictive mean stratification;
(c) within classes, apply a hierarchical random-effects model that pulls the direct ACS estimate of P towards the average P for districts in that class; and
(d) compare the model estimate with 5% for this aspect of the determination.

Comparison of the Bayesian model estimates with the direct ACS estimates indicated large gains in precision, particularly for the small voting districts. The predictive mean stratification is used to reduce dependency on model assumptions, since the regression model is used to group similar jurisdictions rather than to create direct predictions. See Joyce et al. [29] for more details.

Example 2: Inferences for proportions in the American Community Survey. Official statistics often presents uncertainty in the form of standard errors or margins of error. In particular, users of the U.S. American Community Survey (ACS) have the ability to generate tables of estimated counts of individuals by race, age and gender, in small areas. Results are reported by an estimate and a margin of error, chosen so that the estimate plus or minus the margin of error is asymptotically a 90% confidence interval. However, in many instances the margin of error is larger than the estimate, yielding intervals containing negative counts of people! The ACS documentation suggests truncating the resulting intervals so that they are bounded below by zero, but the confidence interval based on the margin of error still fails to have the nominal coverage in small samples, since it is based on a large-sample approximation.

This exemplifies a general weakness of design-based inferences – that they are too focused on estimates and standard errors, assuming that we are in the "land of asymptotia" where an estimate plus or minus two standard errors is truly a 95% confidence interval. We learn in elementary statistics that this is false when the sample size is small, as when a t correction is applied to a normal test or confidence interval when the variance is not known. In simulation studies with realistic sample sizes, design-based confidence intervals often fail to achieve the nominal coverage (e.g. [8,65–68]). A comprehensive theory for finite samples should be able to deal with small sample sizes, and (as discussed below) the simplest general way to achieve this is to make the inference Bayesian. The concern is that the introduction of the prior distribution adds subjective information, but Bayes credibility intervals with noninformative priors tend to be more, not less, conservative than design-based confidence intervals.

In particular, it is well known that asymptotic Wald confidence intervals for proportions do not achieve nominal coverage when the sample size is small, particularly for proportions close to zero or one [7]. Simple fixes such as the Wilson estimate, which for a 95% interval adds 2 to the numerator and 4 to the denominator of the proportion [1], have a Bayesian interpretation. The Bayesian posterior credibility interval based on a noninformative Jeffreys' prior distribution is constrained to lie between 0 and 1, is appropriately asymmetric when the estimate is close to zero or one, and has better confidence coverage than the asymptotic Wald interval [7].

The ideal solution to estimating proportions from small ACS samples would be to create a posterior predictive distribution based on a full Bayesian hierarchical model. This, however, is a daunting proposition for a survey of the scale of the ACS, and my colleague and successor at the Census Bureau, Tom Louis, suggested the following simpler approximate approach:

A. Compute design-based estimates $\hat{p}$ of the proportion and s of the standard error using existing design-based methods;
B. Pretend the data are binomial with number of successes x^* and sample size n^* that lead to the estimates in A, that is, such that $x^*/n^* = \hat{p}$ and $\sqrt{(x^*/n^*)(1 - x^*/n^*)/n^*} = s$;
C. Compute Beta posterior distribution with noninformative prior (e.g. uniform or Jeffreys);
D. Compute 90% posterior credibility interval based on this Beta posterior – this interval reflects asymmetry, and is always between 0 and 1.

This approach is very easy to implement, involving a minor amount of post-processing over current methods. It easily beats standard Wald-type confidence intervals in simulations [21].

These two examples both involve proportions that are potentially close to zero or one, where the symmetric large sample confidence interval is inappropriate – the corresponding posterior distribution of the proportion is right-skewed when the proportion is low, or left-skewed when the proportion is close to one. The Bayesian approach yields an appropriately skewed posterior distribution, and there are different ways of summarizing the center of this distribution – for example the posterior mean, median or mode, yield different point estimates. The choice requires a consideration of the loss function, an important element that is missing in the design-based approach which assumes large samples. This is another advantage of the Bayesian paradigm.

5. Conclusions

Colleagues trained in the classical survey tradition are understandably skeptical of an overtly model-based, even worse Bayesian, approach to official statistics. Models are mistrusted, and should be avoided at all costs! However, what is the alternative? Classical design-based methods cannot handle the complex problems that increasingly arise in official statistics in the era of "big data". Judicious choices of well-calibrated models are needed to tackle problems in small area estimation, nonresponse and response errors, file linkage and combining information across probabilistic and non-probabilistic sources. The power of Bayesian methods for combining data sources is shown in a number of applications sponsored by the U.S. National Center for Health Statistics [44,54–57]. For a Census application, see Elliott and Little [16]. Attention to design features and objective priors can yield Bayesian inferences that avoid subjectivity, and modeling assumptions are explicit, and hence capable of criticism and refinement.

The move to a more overt modeling approach means that government agencies need to recruit and train statisticians who are adept in modeling (and yes, Bayesian) methods, as well as being familiar with survey sampling design. Survey sampling needs to be considered a part of mainstream statistics, in which Bayesian models that incorporate complex design features play a central role. A CB philosophy would improve statistical output, and provide a common philosophy for statisticians and researchers in substantive disciplines such as economics and demography. A strong research program within government statistical agencies, including cooperative ties with statistics departments in academic institutions, would also foster examination and development of the viewpoints advanced in this article [32,33].

Change is also needed before statisticians are recruited into government agencies. Currently Bayesian statistics is absent or "optional" in many programs for training masters' level statisticians, and even Ph.D. statisticians are often trained with very little exposure to Bayesian ideas, beyond a few lectures in a theory sequence dominated by frequentist ideas. This is clearly incompatible with the rising prominence of Bayes in science, as evidenced by the strong representation of modern-day Bayesians in science citations [58].

Formulating useful statistical models for real problems is not simple, and students need more instruction on how to fit models to complicated data sets, and to check the important features of these models. We need to elucidate the subtleties of model development. Issues include the following: (a) models with better fits can yield worse predictions; (b) all model assumptions are not equal, for example in regression lack of normality of errors is secondary to misspecification of the error variance, which is in turn secondary to misspecification of the mean structure; (c) If inferences are to be Bayesian, more attention needs to be paid to the difficulties of picking prior distributions in high-dimensional complex models, objective or subjective. So there will be many challenging but interesting problems to keep official statisticians engaged in the big data era.

Acknowledgements

I thank the referee and associate editor for useful comments.

References

[1] A. Agresti and B.A. Coull, Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions, *The American Statistician* **52** (1998), 119–126.

[2] D. Basu, An Essay on the Logical Foundations of Survey Sampling, Part 1. In Foundations of Statistical Inference, Toronto: Holt, Rinehart and Winston, 1971, pp. 203–242.

[3] D.A. Binder, Non-parametric Bayesian Models for Samples from Finite Populations, *Journal of the Royal Statistical Society* **44**(3) (1982), 388–393.

[4] A. Birnbaum, On the Foundations of Statistical Inference (with discussion), *Journal of the American Statistical Association* **57** (1962), 269–326.
[5] G.E.P. Box, Sampling and Bayes Inference in Scientific Modelling and Robustness (with discussion), *Journal of the Royal Statistical Society Series A* **143** (1980), 383–430.
[6] K.R.W. Brewer, A Class of Robust Sampling Designs for Large-Scale Surveys, *Journal of the American Statistical Association* **74** (1979), 911–915.
[7] L.D. Brown, T.T. Cai and A. DasGupta, Interval Estimation for a Binomial Proportion, *Statistical Science* **16**(2) (2001), 101–133.
[8] Q. Chen, M.R. Elliott and R.J.A. Little, Bayesian Penalized Spline Model-Based Inference for a Finite Population Proportion in Unequal Probability Sampling, *Survey Methodology* **36**(1) (2010), 23–34.
[9] W.G. Cochran, Sampling Techniques, 3rd Edition, New York: Wiley, 1977.
[10] D.R. Cox, The Choice between Alternative Ancillary Statistics, *Journal of the Royal Statistical Society Series B* **33** (1971), 251–255.
[11] D.R. Cox and D.V. Hinkley, Theoretical Statistics, London: Chapman Hall, 1974.
[12] A.P. Dawid, The Well-Calibrated Bayesian, *Journal of the American Statistical Association* **77** (1982), 605–610.
[13] D. Draper, Assessment and Propagation of Model Uncertainty (with discussion), *Journal of the Royal Statistical Society Series B* **57** (1995), 45–97.
[14] D. Draper and M. Krnjajic, Calibration Results for Bayesian Model Specification, *Bayesian Analysis* **1**(1) (2010), 1–43.
[15] B. Efron, Why Isn't Everyone a Bayesian? The American Statistician, 40, 1-11 (with discussion and rejoinder), 1986.
[16] M. Elliott and R.J. Little, A Bayesian Approach to Census 2000 Evaluation Using A.C.E. Survey Data and Demographic Analysis, *Journal of the American Statistical Association* **100** (2005), 380–388.
[17] W.A. Ericson, Subjective Bayesian Models in Sampling Finite Populations, *Journal of the Royal Statistical Society Series B* **31** (1969), 195–234.
[18] W.A. Ericson, Bayesian Inference in Finite Populations. In Handbook of Statistics 6, Amsterdam: North-Holland, 1988, 213–246.
[19] S.E. Fienberg, Bayesian Models and Methods in Public Policy and Government Settings, *Statistical Science* **26**(2) (2011), 212–226.
[20] D. Firth and K.E. Bennett, Robust Models in Probability Sampling, *Journal of the Royal Statistical Society, Series B* **60** (1998), 3–21.
[21] C. Franco, R. Little, T. Louis and E. Slud, Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys. Proceedings of the Survey Research Methods Section, Joint Statistical Meetings, 2014.
[22] A. Gelman, Prior Distributions for Variance Parameters in Hierarchical Models, *Bayesian Analysis* **1**(3) (2006), 515–533.
[23] A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin, Bayesian Data Analysis, 2nd. edition. New York: CRC Press, 2003.
[24] A. Gelman, X.-L. Meng and H. Stern, Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies. (with discussion), *Statistica Sinica* **6** (1996), 733–807.
[25] M. Ghosh and G. Meeden, Bayesian Methods for Finite Population Sampling. London: Chapman and Hall, 1997.
[26] M.H. Hansen, W.N. Hurwitz and W.G. Madow, Sampling Survey Methods and Theory, Vols. I and II, New York: Wiley, 1953.
[27] D. Holt and T.M.F. Smith, Poststratification, *Journal of the Royal Statistical Society Series A* **142** (1979), 33–46.
[28] C.T. Isaki and W.A. Fuller, Survey Design under the Regression Superpopulation Model, *Journal of the American Statistical Association* **77** (1982), 89–96.
[29] P.M. Joyce, D. Malec, R.J. Little, A. Gilary, A. Navarro and M.E. Asiala, Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations, *Journal of the American Statistical Association* **109** (2014), 36–47.
[30] G. Kalton, Models in the Practice of Survey Sampling (Revisited), *Journal of Official Statistics* **18** (2002), 129–154.
[31] L. Kish, Survey Sampling. New York: Wiley, 1965.
[32] R. Lehtonen, E. Pahkinen and C.-E. Sarndal, Research and Development in Official Statistics and Scientific Co-operation with Universities: An Empirical Investigation, *Journal of Official Statistics* **18** (2002), 87–110.
[33] R. Lehtonen and C.-E. Särndal, Research and Development in Official Statistics and Scientific Co-operation with Universities: A Follow-Up Study, *Journal of Official Statistics* **25**(4) (2009), 467–482.
[34] R.J. Little, Post-Stratification: a Modeler's Perspective, *Journal of the American Statistical Association* **88** (1993), 1001–1012.
[35] R.J. Little, The Bayesian Approach to Sample Survey Inference, in: *Analysis of Survey Data*, R.L. Chambers and C.J. Skinner, eds, Wiley: New York, 2003, pp. 49–57.
[36] R.J. Little, To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling, *Journal of the American Statistical Association* **99** (2004), 546–556.
[37] R.J. Little, Calibrated Bayes: A Bayes/Frequentist Roadmap, *The American Statistician* **60**(3) (2006), 213–223.
[38] R.J. Little, Calibrated Bayes, for Statistics in General, and Missing Data in Particular (with Discussion and Rejoinder), *Statistical Science* **26**(2) (2011), 162–186.
[39] R.J. Little, Calibrated Bayes: an Alternative Inferential Paradigm for Official Statistics (with discussion and rejoinder), *Journal of Official Statistics* **28**(3) (2012), 309–372.
[40] R.J. Little, Survey Sampling: Past Controversies, Current Orthodoxies, and Future Paradigms, in: *Past, Present and Future of Statistical Science*, COPSS 50th Anniversary Volume, X. Lin, D.L. Banks, C. Genest, G. Molenberghs, D.W. Scott and J.-L. Wang, eds, CRC Press, 2014.
[41] J. Neyman, 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* **97** (1934), 558–606.
[42] H.W. Peers, On Confidence Points and Bayesian Probability Points in the Case of Several Parameters, *Journal of the Royal Statistical Society Series B* **27** (1965), 9–16.
[43] W. Radermacher, Are we at the Edge of a New Era for Statistics? Keynote address, Eurostat Conference on New Techniques and Technologies for Statistics, Brussels, Belgium, 2015.
[44] T.E. Raghunathan, D. Xie, N. Schenker, V.L. Parsons, W.W. Davis, K.W. Dodd and E.J. Feuer, Combining Information from Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening, *Journal of the American Statistical Association* **102** (2007), 474–486.
[45] J.N.K. Rao, Small Area Estimation. Wiley: New York, 2003.
[46] J.N.K. Rao, Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal, *Statistical Science* **26**(2) (2011), 240–256.

[47] R.M. Royall, On Finite Population Sampling Under Certain Linear Regression Models, *Biometrika* **57** (1970), 377–387.
[48] D.B. Rubin, Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology* **66**(5) (1974), 688–701.
[49] D.B. Rubin, Inference and missing data (with discussion), *Biometrika* **63** (1976), 581–592.
[50] D.B. Rubin, Comment on "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys" by M.H. Hansen, W.G. Madow and B.J. Tepping, *Journal of the American Statistical Association* **78** (1983), 803–805.
[51] D.B. Rubin, Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician, *Annals of Statistics* **12** (1984), 1151–1172.
[52] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987.
[53] C.-E. Särndal, B. Swensson and J.H. Wretman, Model Assisted Survey Sampling. Springer Verlag: New York, 1992.
[54] N. Schenker, Bridging Across Changes in Classification Systems, in: *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives*, A. Gelman and X.-L. Meng, eds, Chichester: Wiley, 2004, pp. 117–128.
[55] N. Schenker, J.F. Gentleman, D. Rose, E. Hing and I.M. Shimizu, Combining Estimates from Complementary Surveys: A Case Study Using Prevalence Estimates from National Health Surveys of Households and Nursing Homes, *Public Health Reports* **117** (2002), 393–407.
[56] N. Schenker and T.E. Raghunathan, Discussion of "Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities" by J. Sedransk, *Journal of Official Statistics* **24** (2008), 507–512.
[57] N. Schenker, T.E. Raghunathan and I. Bondarenko, Improving on Analyses of Self -Reported Data in a Large-Scale Health Survey by Using Information from an Examination-Based Survey, *Statistics in Medicine* **9** (2010), 533–545.
[58] Science Watch (2002). Vital Statistics on the Numbers Game: Highly Cited Authors in Mathematics, 1991–2001. Science Watch, 13, 3, 2.
[59] A.J. Scott, Large-Sample Posterior Distributions for Finite Populations, *Annals of Mathematical Statistics* **42** (1977), 1113–1117.
[60] J. Sedransk, Assessing the Value of Bayesian Methods for Inference about Finite Population Quantities, *Journal of Official Statistics* **24** (2008), 495–506.
[61] R.A. Sugden and T.M.F. Smith, Ignorable and Informative Designs in Survey Sampling Inference, *Biometrika* **71** (1984), 495–506.
[62] M.E. Thompson, Superpopulation Models, *Encyclopedia of Statistical Sciences* **1**(9) (1988), 93–99.
[63] R. Valliant, A.H. Dorfman and R.M. Royall, Finite Population Sampling and Inference: a Prediction Approach. New York: Wiley, 2000.
[64] B.L. Welch, On Comparisons between Confidence Point Procedures in the Case of a Single Parameter, *Journal of the Royal Statistical Society Series B* **27** (1965), 1–8.
[65] Y. Yuan and R.J. Little, Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Unit Nonresponse, *Journal of the Royal Statistical Society, Ser. C* **56** (2007), 79–97.
[66] Y. Yuan and R.J. Little, Model-Based Estimates of the Finite Population Mean for Two-Stage Cluster Samples with Item Nonresponse, *Journal of Official Statistics* **24** (2008), 193–211.
[67] H. Zheng and R.J. Little, Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples, *Survey Methodology* **30**(2) (2004), 209–218.
[68] H. Zheng and R.J. Little, Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model, *Journal of Official Statistics* **21** (2005), 1–20.