Journal of Official Statistics, Vol. 28, No. 3, 2012, pp. 367–372

Rejoinder

Roderick J. Little

I thank the editors of JOS for the kind consideration they have given to my article, including the soliciting and editing of a stimulating set of discussions. The discussants include distinguished and experienced official statisticians, and I thank them for their positive and encouraging comments, while focusing here on points of disagreement.

Jean-François Beaumont's discussion is wide-ranging, and I will concentrate on a few major points. His main criticism of Bayesian inference is that it relies on specification of a full parametric model and a prior distribution, whereas the frequentist approach "only requires correct specification of the first two moments of a model". I suggest that counting the number of explicit specifications is a superficial way of comparing the approaches. The form of the estimates in the design-based approach has implicit assumptions, and the theory relies on asymptotic arguments. With large samples, inferences under the Bayesian approach are generally insensitive to violations of the additional specifications that are needed – specifically, the prior distribution and Gaussian distributional assumptions; in small samples, the additional specifications lead to improved frequentist inferences, such as t inferences for normal models. I agree with Beaumont and Michael Cohen that relatively nonparametric models have value when samples are large, and thank them for including additional references on this class of models.

Beaumont suggests that Bayes' inference requires that the model is correct – if this were true, the approach would never apply, since all models are wrong! Bayes inference requires a model that yields good inferences, and for this, the model does not need to be correct, only to capture key features of the population for the problem at hand. The sample design influences what these features are. It may seem that model-assisted estimation relies less on correctness of the model than Bayes estimates, but to the contrary I have found that robust Bayesian models give superior design-based inferences, if design features are incorporated.

The impact of model misspecification depends on the target for inference, so it is context-specific. In Beaumont's specific case of a normal linear regression model with simple random sampling, the inference – Bayes or non-Bayes – for the overall mean is robust to nonlinearity of the regression relationship. Design-based standard errors are (from the perspective of ancillarity) less satisfactory for a sample that happens by chance to be unbalanced in the distribution of the covariate (Cumberland and Royall 1988). Inference about predictions at particular values of the covariate – Bayes or non-Bayes – depends on the assumed form of the relationship, to a degree that depends on how far the covariate is from the center of the covariate distribution. Beaumont believes that the Bayes inference is more dependent on assumptions than the design-based inference. To me, the argument seems rather artificial, particularly when the confidence or credibility intervals

from the two approaches are identical. In practice, I want a conditional inference for the sample in hand, not the set of repeated samples that are never taken.

Concerning multiple imputation (MI), imputations and MI combining rules under a specific model are based on simulation approximations of the posterior distribution, and as such are Bayesian when the overall inference is Bayesian. Complex design features should be included in the MI model, as discussed in the article. The alleged problems of MI inferences arise when the imputer and analyst models differ, or the analyst adopts a non-Bayesian inferential perspective. I still feel that MI under a sensible model yields good frequentist inferences – tending to be conservative in terms of confidence coverage. I find the principles of MI – imputing draws, averaging over MI data sets to reduce the loss in efficiency resulting from draws, and MI combining rules to propagate imputation uncertainty – very useful in practical applications.

Philippe Brion discusses an interesting problem involving combining information on businesses from administrative and survey sources, focusing specifically on how to handle differences in industry codes, which are updated for the relatively small set of businesses in the sample. To me, this is inherently a modeling problem, with the goal to predict the updated codes for nonsampled businesses. I reject the conclusion that mass model-based imputation cannot work here – Bayesian mass imputation (even better, multiple imputation to propagate imputation error) was successfully applied *over twenty years ago* to a much larger problem, namely updating industry and occupation codes for the U.S. Census public use files (Clogg et al. 1991); another pertinent reference that applies multiple imputation to subset editing is Czajka et al. (1992).

The fact that the so-called "model-based" approach yielded biased estimates is for me a diagnostic that the model needs refinement, not that the model-based approach doesn't work. Finally, the "difference estimator" in Eq. (2) implies a model, and the assumptions of this model can be assessed, along with any other model.

I agree with Brion that a modeling perspective is useful for handling problems with editing – which to me is just a part of estimation – and thank him for mentioning my early work on this with Phil Smith. Readers should explore later work on this topic by Joe Schafer and his colleagues (e.g. Ghosh-Dastidar and Schafer 2003), which is more Bayesian and hence (naturally) even better!

I also agree with Brion and Paul Smith that objectivity is an important characteristic for official statistics, and is a motivation for design-based models. However, direct design-based estimates with or without GREG adjustments are inadequate for current needs, which require model-based estimates for small areas. Furthermore, essentially all so-called "design-based" estimates involve subjective elements, specifically in the choice of estimator and treatment of missing data. I believe that Bayesian model-based inferences that are well calibrated (in the sense of the article) can meet the bar of objectivity, and have the advantage that model assumptions are explicit, so alleged lapses in objectivity are correctable.

Brion and Smith both raise the need for consistency of estimates at different levels of aggregation. I agree that this is desirable, although it reflects an underlying accounting perspective that fails sufficiently to recognize that these are error-prone estimates. In principle, inference under a suitable calibrated Bayes model for the population implicitly predicts all the non-sampled values, and combining them yields estimates that are mutually consistent at all levels of aggregation. Current inconsistencies arise from trying

to reconcile estimates from different models or even different philosophies (design-based for large samples, model-based for small samples). This is another manifestation of "inferential schizophrenia". I concede, however, that although more work is needed to develop models that mesh well at different levels of aggregation.

I enjoyed Alan Dorfman's forthright discussion, which takes the position of a superpopulation modeler in the Richard Royall tradition. As a modeler, I am probably closer to Dorfman's position than the other discussants, but his points of disagreement are interesting.

I plead guilty of not paying sufficient attention to diagnostics in my article. I was focused on the official statistics world, where many statistics are generated routinely in a production setting. The danger of emphasizing model diagnostics is that working official statisticians will reject modeling as impractical in their setting; including design features in the model is one way of avoiding major traps in model misspecification. Valliant et al. (2000) argue that diagnostics can uncover the deficiencies of the estimators in Hansen et al. (1983) that ignore the design weights. They may be right in this instance, but more generally, I am not convinced that models that ignore design features can be rescued by judicious model checks. To take another example, would we entertain a model that ignores cluster effects when the sample design involves cluster sampling? I think not – relatively small within-cluster correlations can translate into substantial design effects for means. Furthermore, I fear that many diagnostic checks in the busy official statistics setting are what Dorfman would characterize as "half-baked".

On the other hand, it is hard to argue with the utility of model diagnostics, and I note that one advantage of the modeling perspective is that assumptions are explicit and hence subject to criticism. I think design-based statisticians need to check the models that underlie their estimates, for example the ratio model that (as discussed below) is pervasive for business surveys. Model diagnostics in the presence of a complex sample design is an area in need of more attention.

Dorfman interprets calibrated Bayes "in large samples, as a particular version of model-assisted sample estimation," where the adjusting sum is algebraically zero. Technically this is correct, but the statement minimizes important differences. I would say "choose a model that avoids the need for assistance, and make the inference Bayesian!" The small-sample aspect is important to me, as is the principle of inference.

Concerning Dorfman's alternatives to random sampling, my article is firmly within the tradition of sampling designs that incorporate randomization, to the extent feasible. Complex designs can incorporate prior information effectively, and the process of randomization has many benefits in terms of objectivity. I think failure to acknowledge the benefits of randomization has set back the Bayesian modeling cause in official statistics.

Risto Lehtonen notes that official statistics in Scandinavian countries are increasingly based on information from administrative registers. My lack of attention to this aspect reflects my ignorance rather than a conscious decision not to address it – the brevity of my discussion is perhaps well calibrated with the depth of my knowledge! Calibrated Bayes (like any other statistical inference approach) may have relatively little to offer when the whole population is included in the registry, sampling error is not important, and no information is available to adjust for errors. On the other hand, statistical models are often important when registry data are combined with data from other sources. For example, if

the administrative variables are proxies for "true" variables that are measured for a sample, then models are useful in order to (effectively) impute the true values in non-sampled cases (Zanutto and Zaslavsky 2001). Models also play a useful role in accounting for matching error between different data sources.

Lehtonen sees design bias as an important issue for official statistics production, and notes that model-based EBLUP estimators can lead to estimators where bias dominates the mean squared error, leading to invalid design-based confidence intervals. In other words, not all models (including hierarchical Bayes models) are well calibrated in the calibrated Bayesian sense. In particular, I note that the underperforming mixed models in Lehtonen et al. (2003) do not appear to include the sampling weight as a covariate, as I am suggesting they should. I suggest that Bayesian inference for a model that includes the sampling weight as a covariate, perhaps as a penalized spline as in Zheng and Little (2004), might yield better calibrated results.

Concerning his remarks on innovation in official statistics, Lehtonen is himself contributing to this by maintaining an active publication profile. It is a particularly interesting period for official statistics, and actively promoting publication is important to allow competing ideas to be broadcast, debated and refined.

I thank Paul Smith for articulating his practical concerns. As he implies, business surveys are one area where more serious attention to models is needed. Smith is probably right that the prevailing perspective is largely design-based, but many of the estimates are based on regression models with variance a function of a measure of size, such as ratio estimates. I feel there is much scope for model development here, using multivariate models that allow for general patterns of missing data and make efficient use of all available covariate information. Concerning computational issues for statistics requiring rapid turnaround, the issues seem to me more organizational than computational: Matt Wand recently presented a talk at the Australian Statistics Conference which displayed Bayes' inferences for a relatively complex semiparametric regression model, updated *in real time* as new data points are added (Wand 2012). If new estimates can be created in seconds, monthly estimates should not be a problem!

Smith's comments on the need to assess model sensitivity are on target – his remarks on potential complaints about the model when it is the basis for resource allocation are interesting, and my impression is that they are consistent with experience at the U.S. Census Bureau.

Like Michael Cohen, I'd like to honor the memory of David Binder, an influential pioneer in the application of nonparametric Bayesian methods to surveys. I have known David since we were both graduate students in the 1970s at Imperial College, London, and he was always good for a stimulating but good-natured argument about statistics. I am sure he would have had strong opinions about the topic of this discussion, and regret that we are not able to hear them[1].

Cohen distinguishes descriptive inferences for finite population characteristics and analytic inferences that are focused on causal mechanisms. In Bayesian inference the

[1]Short of that, the readers can find some of David Binder's views on the subject by consulting his discussion in JOS Vol. 24(4), 2008, pp. 513–516, where he in a positive context mentions Professor Little's calibrated Bayes and then goes on to list some applications where he thought that the Bayesian approach was necessary or worked well. – *Editors-in-chief*

distinction is reflected in the posterior distribution of the finite population quantity (e.g. $p(\bar{Y} \mid \text{data})$ for the finite population mean $\bar{Y}$) and the posterior distribution of the superpopulation parameter ($p(\mu \mid \text{data})$ for the superpopulation mean μ). The distinction matters, particularly for estimates of uncertainty. Bayes handles both types of problem without difficulty, and as noted in the article, "finite population corrections" are automatically incorporated in the posterior distributions of finite population quantities. A strength of Bayes is that this is true even for complex quantities where frequentist fpcs are not obvious, for example the ratio of two finite population regression coefficients.

Cohen believes that avoiding bias is paramount for descriptive statistics. Here I think it is important to distinguish between avoiding policy bias, seeking "an overall pattern of estimates with no tendency to favor one side over another in any controversy" and avoiding statistical bias, a technical property of statistics in repeated samples that is now broadly regarded as unhelpful in modern statistics. If Cohen is seriously advocating design-unbiased estimates of descriptive statistics, he is disqualifying small area estimates that smooth direct estimates towards model predictions, since these are technically design-biased. Such estimates are useful and broadly accepted, by design-based statisticians and Bayesians alike. Requiring design-unbiasedness restricts us to direct estimates, meaning that when the direct estimates lack sufficient precision, the small area problem cannot be addressed at all.

Dorfman closes his discussion with a related question: when is a model good enough, particularly for small areas? In small area estimation, one might classify estimates in terms of their degree of reliance on the model, based on how much weight is being given to the model rather than the direct estimate. Also, the quality of the model prediction depends on the extent to which key features that differentiate the areas with respect to the survey outcome are measured and included as predictors in the model. Thus, models with high R-squareds are more compelling. But, as Dorfman implies, good ways of answering this question are needed.

To me, calibrated Bayes is the right approach to inference, but, in the absence of major analytic errors, the quality of the results is determined more by the quality of the data than how it is analyzed.

References

Binder, D.A. (2008). Discussion. Journal of Official Statistics, 24, 513–516.

Clogg, C.C., Rubin, D.B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple Imputation of Industry and Occupation Codes in Census Public-use Samples Using Bayesian Logistic Regression. Journal of the American Statistical Association, 86, 68–78.

Cumberland, W.G. and Royall, R.M. (1988). Does Simple Random Sampling Provide Adequate Balance? Journal of Royal Statistical Society, Series B, 50, 118–124.

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A., and Rubin, D.B. (1992). Projecting from Advance Data Using Propensity Modeling; an Application to Income and Tax Statistics. Journal of Business and Economic Statistics, 10, 117–132.

Ghosh-Dastidar, B. and Schafer, J.L. (2003). Multiple Edit/Multiple Imputation for Multivariate Continuous Data. Journal of the American Statistical Association, 98, 807–817.

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

Wand, M. (2012). Real Time Semiparametric Regression. Keynote presentation, Australian Statistics Conference, Adelaide 2012.