

# DISCUSSION OF

## Why Bayes?

... for statistics in general?

... for missing data in particular?

By R.J.A. Little

**Nathaniel Schenker**

**Senior Scientist for Research and Methodology**

**National Center for Health Statistics**

**[nschenker@cdc.gov](mailto:nschenker@cdc.gov)**

**Presented at the Workshop on Bayesian Methods that  
Frequentists Should Know**

**University of Maryland**

**May 1, 2008**

## **CONTENTS**

- 1. “Unleash the power of B computation”: Examples**
- 2. Diffuse (but hopefully informative) comments on Rod’s general points**
- 3. Comments on Rod’s illustrations of Bayesian methods**
- 4. Conclusion**

# 1. “Unleash the power of B computation”: Examples

## A. Examining how survival of lung-cancer patients relates to post-operative smoking behavior (Faucett, Schenker, & Elashoff 1998)

- ▶ **Clinical trial data after surgery for lung cancer**
  - **Survival**
  - **Current smoking status at follow-up visits**
    - ◆ **Intermittently observed**
- ▶ **Markov chain model for current smoking status and proportional hazards model for survival given smoking behavior (and other covariates)**
- ▶ **Gibbs sampling to approximate posterior distributions under diffuse priors**

- ▶ **Analyzed effects of cumulative smoking as well as current smoking**
- ▶ **Example of joint modeling of longitudinal and survival data**

## **B. Multiple imputation of missing income data in the National Health Interview Survey (Schenker *et al.* 2006)**

- ▶ **NHIS: Rich source of data for studying relationships between health and other characteristics (e.g., income)**
- ▶ **About 30% nonresponse on “exact” family income**
  - **Have reported income categories for about 2/3 of the nonrespondents**
  - **Missingness appears related to several other characteristics**
- ▶ **Missing income data multiply imputed beginning with 1997 survey year ([www.cdc.gov/nchs/nhis.htm](http://www.cdc.gov/nchs/nhis.htm))**

- ▶ **Used IVEware to implement adaptation of Sequential Regression Multivariate Imputation (Raghunathan *et al.* 2001)**
  
- ▶ **Some complications handled during imputation**
  - **Structural dependencies between variables (e.g., employment status and personal earnings)**
  - **Imputation within bounds (e.g., based on categorical responses for income)**
  - **Several variables of different types (continuous, categorical, count) used as predictors**
    - ◆ **Small amounts of missingness (mostly < 2%)**
  
- ▶ **Another recent application of multiple imputation: Missing dual energy x-ray absorptiometry data in the National Health and Nutrition Examination Survey ([www.cdc.gov/nchs/nhanes.htm](http://www.cdc.gov/nchs/nhanes.htm))**

## **C. Combining Information from two health surveys for small-area estimation (Raghunathan *et al.* 2007)**

- ▶ Interest in small-area (e.g., county-level) estimates of the prevalence of cancer risk factors and screening**
  
- ▶ Behavioral Risk Factor Surveillance System**
  - + Large; almost all counties in sample**
  - Telephone survey**
    - ⇒ Non-coverage of non-telephone households; high nonresponse rates**
  
- ▶ National Health Interview Survey**
  - + Face-to-face survey**
    - ⇒ Includes non-telephone households, which can be identified; higher response rates**
  - Smaller; only about 25% of counties in sample**

- ▶ **Project developed Bayesian methods to combine information from the two surveys**
  - **Used Fay-Herriot type of model**
  - **Approximate posterior distributions obtained via Gibbs sampling**
  
- ▶ **National Cancer Institute aims to release small-area estimates on-line**

## 2. Diffuse (but hopefully informative) comments on Rod's general points

### A. Modes of inference within and outside of survey sampling

#### Simplified, non-exhaustive 2-way table

	Outside of survey sampling	Within survey sampling
<b>F</b>	<ul style="list-style-type: none"> <li>• Formulate <math>p(y   \theta)</math> (<math>y = \text{data}</math>, <math>\theta = \text{parameters}</math>).</li> <li>• Base inferences for <math>\theta</math> on <math>p(\hat{\theta}(y)   \theta)</math> (and large-sample theory).</li> </ul>	<ul style="list-style-type: none"> <li>• Estimate <math>Q(Y)</math> by <math>\hat{Q}(Y_{inc}, I)</math> (<math>Y = \text{population values}</math>, <math>Y_{inc} = \text{sampled values}</math>, and <math>I = \text{sample indicators}</math>).</li> <li>• Base inferences for <math>Q(Y)</math> on <math>p(\hat{Q}(Y_{inc}, I)   Y)</math> (and large-sample theory).</li> </ul>
<b>B</b>	<ul style="list-style-type: none"> <li>• Formulate <math>p(y   \theta)</math> and <math>p(\theta)</math>.</li> <li>• Base inferences for <math>\theta</math> on <math>p(\theta   y)</math>.</li> </ul>	<ul style="list-style-type: none"> <li>• Formulate <math>p(Y   \theta)</math> and <math>p(\theta)</math>.</li> <li>• Base inferences for <math>Q(Y)</math> on <math>p(Q(Y)   Y_{inc}, I)</math>.</li> </ul>

- ▶ **Choice between Bayesian and frequentist approaches to inference often seems more controversial within survey sampling than outside of it**
  - **Main issue outside of survey sampling:  
Use of prior/posterior versus use of sampling distribution, *both under a parametric model***
  - **Main issue within survey sampling:  
To model versus not to model**

► **Conclusions of Hansen, Madow, and Tepping (1983)**

**1. For descriptive inference from large, well-designed sample surveys, use design-based inference**

- ◆ **Avoids errors due to model misspecification**
- ◆ **Little efficiency lost relative to model-based inference**

**2. Models can be useful and important**

- ◆ **Sample design**
- ◆ **Inference for small samples**
- ◆ **Inference in the presence of non-sampling errors**
- ◆ **When inference under model is of intrinsic interest**

- ▶ **Some issues regarding conclusions of Hansen, Madow, and Tepping (1983)**
  - **How large a sample is large enough?**
  - **Near efficiency requires that estimator be nearly optimal for population and chosen design**
  - **Does a pragmatic approach (sometimes design-based, sometimes model-based) make us look like we have a “split personality”?**

## B. Calibrated Bayes

- ▶ **Attractive idea, but in a specific situation, how do we ensure that we're well calibrated?**
- ▶ **Predictive checks of Box (1980) and Rubin (1984) use frequentist ideas, but do they ensure calibration in repeated experience?**
- ▶ **Suggestion from Little (1983) and Rod's discussion of Hansen, Madow, and Tepping (1983): For inference from sample surveys, choose Bayesian procedures that are design-consistent.**
  - **How easy is this to do in general?**

## C. Multiple imputation

- ▶ **Very helpful for public-use data for a number of reasons, BUT:**
  - **Difficult (impossible?) to make imputation models general enough to cover all possible analyses by secondary users**
    - ◆ **I tend to “throw the kitchen sink” into models**
  - **Need further work on incorporating features of sample designs into imputation models**
    - ◆ **I tend to include weights, stratum/PSU indicators, and/or other variables related to sample designs**
  
- ▶ **Issues above apply to *single* imputation as well**

### **3. Comments on Rod's illustrations of Bayesian methods**

#### **A. Penalized spline of propensity models**

- ▶ **Nice example of conditioning on probabilities of selection in model-based inference**
  - **See Rubin's discussion of Hansen, Madow, and Tepping (1983), and HMT's rejoinder**
- ▶ **Use of "modern nonparametric" methods, whereas design-based inference is reminiscent of "classical nonparametrics"**
- ▶ **As model becomes "weaker," how much efficiency is lost?**
- ▶ **Look forward to development for multi-stage sample surveys**

▶ **Simulation study**

- **Example of difficulty in assessing whether “asymptotia” has been reached**
- **How stable are the Yates-Grundy variance estimators for the HT and GR methods?**
  - ◆ **Would be interesting to compare to linearization or resampling under assumption of sampling with replacement**

▶ **Wald-type confidence intervals for proportions can perform poorly in general, especially for proportions near 0 or 1**

- **Rubin and Schenker (1987): Improvements via approximate Bayesian intervals in a non-survey context**

## B. Proxy pattern-mixture analysis

- ▶ **Nice example of sensitivity analysis when there are missing data**
  - **Especially important when the amount of missing information is large and/or missingness at random is questionable**
  - **Emphasis in early writings on multiple imputation**
  
- ▶ **Method implicitly assumes that relationship between  $Y$  and  $Y^*$  (i.e.,  $\rho$ ) is the same for respondents and nonrespondents?**
  - **Perhaps reasonable, but not directly verifiable**
  - **Rubin (1977) specified a subjective prior distribution for nonrespondents' parameters given respondents' parameters**

- ▶ **Is PPMA model able to incorporate features of complex sample designs (e.g., for NHANES III analysis)?**
  - **Is this an important issue?**

## 4. Conclusion

- ▶ **Second Rod's call for more work on model checks and software (!!)**
- ▶ **Also call for more demonstrations of usefulness of Bayesian methods in real applications**
- ▶ **THANKS TO:**
  - **Rod for a very interesting talk**
  - **Partha Lahiri and Eric Slud for organizing the workshop and inviting me to be a discussant**

**REFERENCES**

- Box, G.E.P. (1980), "Sampling and Bayes' Inference in Scientific Modelling and Robustness" (with discussion and rejoinder), *Journal of the Royal Statistical Society, Series A: General*, 143, 383-430.
- Faucett, C.L., Schenker, N., and Elashoff, R.M. (1998), "Analysis of Censored Survival Data with Intermittently Observed Time-Dependent Binary Covariates," *Journal of the American Statistical Association*, 93, 427-437.
- Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys" (with discussion and rejoinder), *Journal of the American Statistical Association*, 78, 776-807.
- Little, R.J.A. (1983), "Estimating a Finite Population Mean from Unequal Probability Samples," *Journal of the American Statistical Association*, 78, 596-604.
- Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, 27, 85-95.
- Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W., and Feuer, E.J. (2007), "Combining Information From Two Surveys to Estimate County-Level Prevalence Rates of Cancer Risk Factors and Screening," *Journal of the American Statistical Association*, 102, 474-486.
- Rubin, D.B. (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538-543.
- Rubin, D.B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151-1172.
- Rubin, D.B., and Schenker, N. (1987), "Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior," *Sociological Methodology*, 17, 131-144.
- Schenker, N., Raghunathan, T.E., Chiu, P.-L., Makuc, D.M., Zhang, G., and Cohen, A.J. (2006), "Multiple Imputation of Missing Income Data in the National Health Interview Survey," *Journal of the American Statistical Association*, 101, 924-933.