Application of Bayesian Analysis for Sample Surveys

Danutė Krapavickaitė

Lithuania

18th of June, 2019

Baltic-Nordic Conference on Survey Statistics
June 16–20, Örebro, Sweden

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Outline

1. Introduction to Bayesian statistics
2. Comparison of frequentist and Bayesian approaches
3. Design-based and model-based inferece in sample surveys
4. Modelling the sample selection mechanism
5. Bayesian models for sample surveys
   - simple random sampling
   - stratified srs
6. Multiple imputation
7. References

Example

A coin is tossed 10 times and 7 heads occur. Is the coin regular?

Frequentist approach

Let $A = \{\text{a head occur after one toss}\}$,

$p = P(A)$, $n = 10$, $m = 7$

$\hat{p} = m/n = 0.7$,

$\hat{p} \sim$ *Normal* under a very large number of identical repeated experiments

$CI_{0.95}(p) = \hat{p} \pm z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n} = (0.43; 0.97)$,

$p = 0.5 \in (0.43; 0.97) \Rightarrow$ with probability 0.95 it is no contradiction that the coin is regular.

$P_{10}(m > 5) \approx 0.377$.

Bayesian approach

Assume $\pi = P(A) \in \{\pi_1, \pi_2, \pi_3\} = (0.4; 0.5; 0.6)$

$$H_i = \{\pi = \pi_i\},\ i = 1, 2, 3.$$

$$P(H_1) = 0.2;\ P(H_2) = 0.6;\ P(H_3) = 0.2.$$

$B = \{7 \text{ heads occur after } 10 \text{ tosses}\}$

$P(B|H_i) = C_{10}^7 \pi_i^7 (1 - \pi_i)^3,\ i = 1, 2, 3$

$P(B) = \sum_{i=1}^3 P(B|H_i)P(H_i) = 0.218$

Bayes theorem:

$$P(H_i|B) = \frac{P(B|H_i)P(H_i)}{P(B)}, \tag{1}$$

$i = 1, 2, 3.\ P(H_i|B) \in (0.07;\ 0.577;\ 0.353)$

$P(\pi > 0.5|B) = P(H_3|B) = 0.353$

$\pi \in\ (0.4; 0.5; 0.6)$ prior distribution
$P(B|H_i),\ i = 1, 2, 3$ likelihood
$P(H_i|B),\ i = 1, 2, 3$ posterior distribution

Classical/frequentist approach to parameter estimation

Let Y be a random variable with the distribution $F(y, \theta)$,
θ be a vector of the distribution parameters.

$Y_1, Y_2, ..., Y_n \sim$ i. i. d. (iid) Y (sample)

The aim is to estimate the parameter θ.

Population parameter θ is *fixed but unknown constant*,
no distribution associated to it.

The sample is random. The only probability distribution is distribution of the random sample of size n given the parameter θ
$\Rightarrow \hat{\theta}$ is random

Bayesian approach

The *data* $y_1, y_2, ..., y_n$ (iid sample realization) is available and it *is fixed*.
It can be obtained under various values of the parameter θ, $\theta \in \Theta$.
Therefore *parameter* θ *is random*.

The aim: to find distribution of θ given data $y = (y_1, y_2, ..., y_n)$

Let f distribution of data; g distribution of parameter
Let $f(y|\theta)$ be a density/probability of a distribution F. Then

$$f(\theta, y) = f(y|\theta)g(\theta) = g(\theta|y)f(y)$$

$$g(\theta|y) = \frac{f(y|\theta)g(\theta)}{f(y)} = \frac{f(y|\theta)g(\theta)}{\int_\Theta f(y|\theta)g(\theta)d\theta} \tag{2}$$

$f(y|\theta) = \prod_{i=1}^n f(y_i|\theta)$ – *likelihood* of the data
$g(\theta)$ – *prior* (subjective!) distribution of θ
$g(\theta|y)$ – *posterior* distribution of θ given data y.

Bayesian inference

A proportional form of (2)

$$g(\theta|y) \propto f(y|\theta)g(\theta)$$

Marginal / *prior predictive distribution* of the data:

$$f(y) = \int f(y,\theta)d\theta = \int f(y|\theta)g(\theta)d\theta$$

We can predict unknown observable $\tilde{y}$ from the same process

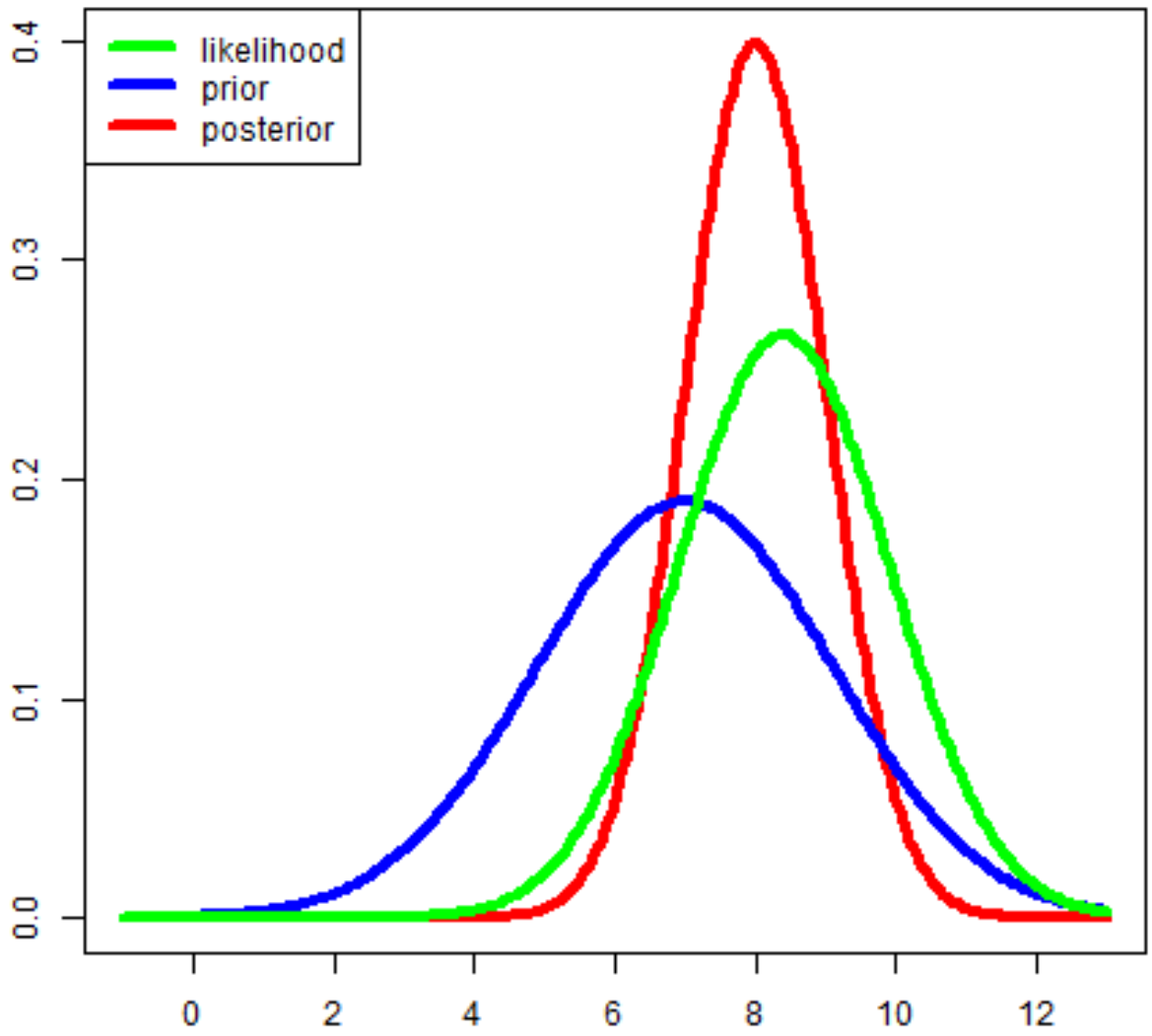
$$f(\tilde{y}|y) = \int f(\tilde{y},\theta|y)d\theta = \int f(\tilde{y}|y,\theta)g(\theta|y)d\theta = \int f(\tilde{y}|\theta)g(\theta|y)d\theta$$

due to conditional independence of y and $\tilde{y}$ given θ.
This is the *posterior predictive distribution*
If we are interested in the function $Q = Q(y)$, it is based on its posterior predictive distribution given the data:

$$f(Q(y)|y) = f(Q(y)|y_1, y_2, ..., y_n) = \int f(Q(y)|\theta)g(\theta|y)d\theta$$

Bayesian inference – grafical representation

Posterior $\propto$ prior $\times$ likelihood



Frequentist and Bayesian approach: comparison

Frequentist and Bayesian point of view is different in

- point estimation;
- confidence interval/credibility interval (high probability region estimation);
- hypothesis testing

Frequentist inference about the parameter requires probabilities calculated from the sampling distribution of the data given fixed but unknown parameter.

These probabilities are based on all possible unknown samples that could have occur.

The *Bayesian inference* use probabilities calculated from the posterior distribution.
That makes them conditional on the samples that actually did occur.

Conjugate distributions

Proportional form of the Bayesian inference

$$g(\theta|y) \propto f(y|\theta)g(\theta)$$

If the data $y:\ y_1, y_2, ..., y_n$ is obtained observing the iid random variables then the likelihood can be expressed

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Definition

If for a distribution $f(y|\theta) \in \mathcal{F}$ (family of the distributions)
can be chosen a distribution $g(\theta) \in \mathcal{P}$ (family of the distributions)
so that $g(\theta|y) = f(y|\theta)g(\theta) \in \mathcal{P}$
then $\mathcal{P}$ is said to be *conjugate* to $\mathcal{F}$.

The classes of the conjugate distributions are known.

Example. Left-handed students

n_i	Students	20	15	9	8	5	25	18	24	3
m_i	Left-handed	2	1	3	2	1	2	3	3	0

What is a proportion λ of left-handed?

Frequentist: $\hat{\lambda} = \frac{\sum m_i}{\sum n_i} = 0.1338,\ sd = \sqrt{\lambda(1-\lambda)/n} = 0.1135$

Bayesian. Left-handed $\sim\ Pois(\lambda)$.

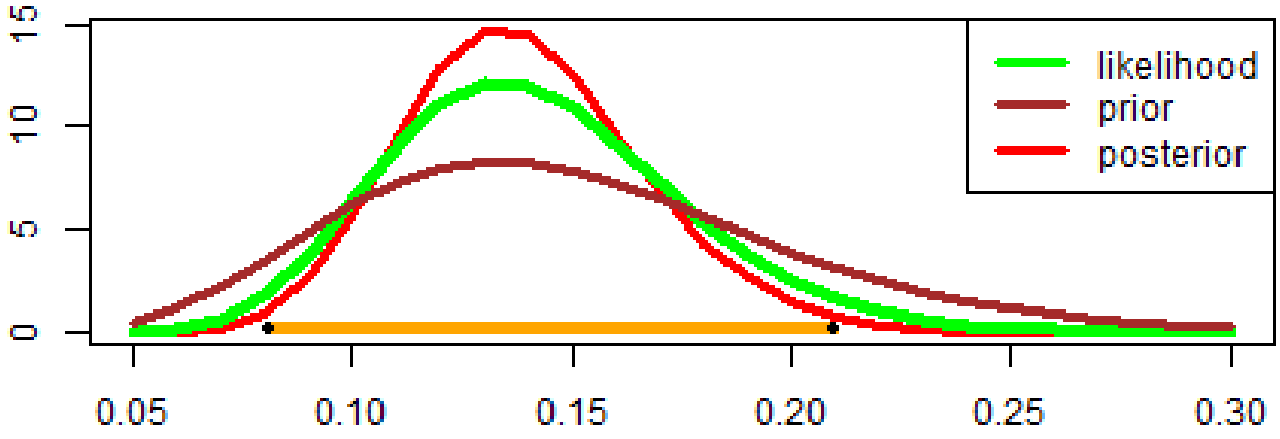
Likelihood $f(y|\lambda) = \prod_{i=1}^{n} f(y_i|\lambda) \propto gamma(r', v')$,

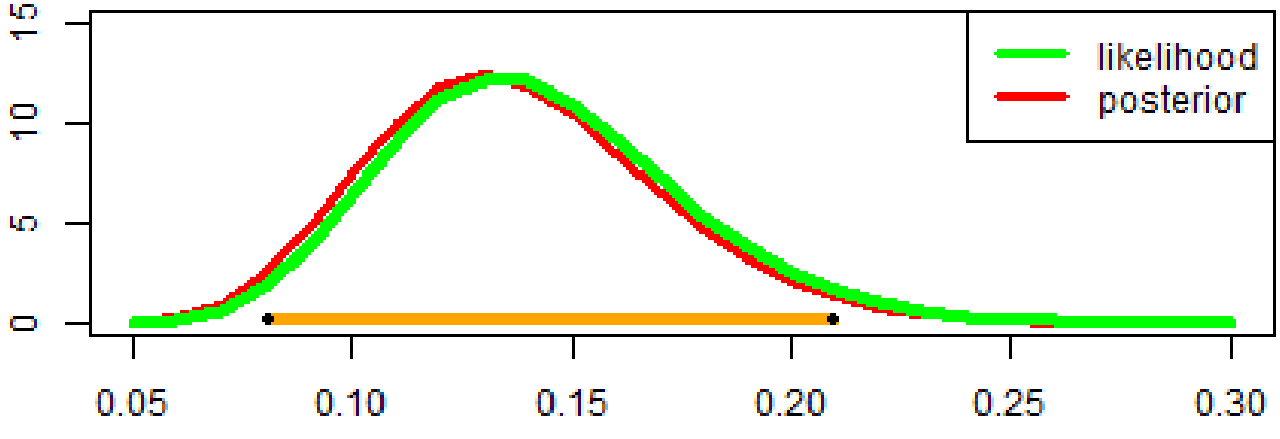
$r' = \sum y_i + 1 = 17,\ v' = n = 127.$

Conjugate priors.

$g(\lambda)$	$g(\lambda\|y)$	$\hat{\lambda}$	$sd(\hat{\lambda})$	$q_{0.975} - q_{0.025}$
Frequentist		0.1339	0.1135	0.3562
$gamma(9, 60)$	$gamma(26, 187)$	0.1390	0.0330	0.1065
$g(\lambda) = 1/\sqrt{\lambda}$	$gamma(17.5; 127)$	0.1378	0.0323	0.1284

Example. Bayesian analysis





Problems

- The conjugate posterior distribution $g(\theta|y)$ can be found easily and used further. But even in this case the values of the distribution $g(\theta|y)$ can be simulated many times, and empirical characteristics of this distribution can be used for inference.
- Conjugate distribution for the likelihood is not available $\Rightarrow$ The only way to make inference about the posterior distribution $g(\theta|y)$ is simulation from the posterior.
- "...easy in theory hard in practice" (Bolstadt). Approximate calculation of the integrals
  - Complicated hierarchical models can be used for data
  - Likelihood includes many parameters, main and additional (nuisance).
  - Checking how the model estimated fits the data.

Approaches to Survey Inference

- Design-based (randomization) inference
- Superpopulation modeling
  - Frequentist
    - Superpopulation model with fixed parameters is specified
    - Frequentist estimators of parameters are based on the repeated samples from the superpopulation and finite population
  - Bayesian modeling
    - Full probability model of the data and parameters (including prior distribution on the parameters) is specified
    - Bayesian inference is based on the *posterior distributions* of the finite population quantities

Bayesian analysis for finite population statistics

Main principles of Bayesian statistics: sufficiency and likelihood

Definition

The *likelihood principle* asserts that given

- a fixed model (including the prior distribution for underlying data),
- fixed observed values of the data

Bayesian inference is determined regardless of the sampling design.

Method of the data collection is important in Bayesian analysis. The information of a sampling design should be included in a model and conclusions should be made conditional on the variable describing sampling design.

As more explanatory variables are included into the model, the inferential conclusions become more valid conditionally but possibly more sensitive to the model specifications.

Notations

$y = (y_1, y_2, ..., y_N)$ – matrix of potential data,
each y_i may be a vector y_{ij}, $j = 1, 2, ..., J$.
$I = (I_1, I_2, ..., I_N)$ – matrix of the same dimension as y of indicators for observations y:

$$I_{ij} = \begin{cases} = 1, & y_{ij} \text{ is observed,} \\ = 0, & \text{otherwise.} \end{cases} \tag{3}$$

$obs = inc = \{(i, j) : \ I_{ij} = 1\}$,
$exc = \{(i, j) : \ I_{ij} = 0\}$
Assumption 1: no y measurement errors,
no missing y values

Modelling sample selection mechanism

The joined probability model for y and I is broken into two parts:

- the model for complete data y, observed and unobserved components,
- the model for inclusion indicator I:

$$f(y, I|\theta, \phi)) = f(y|\theta)f(I|y, \phi).$$

By θ, ϕ are denoted parameters of the distributions of the complete data y and the inclusion matrix I, respectively.

The actual information available is y_{inc} and I, so appropriate likelihood for Bayesian inference is

$$f(y_{inc}, I|\theta, \phi) = \int f(y, I|\theta, \phi)dy_{exc} \tag{4}$$

Complete data likelihood

Let z means the variables that are fully observed (auxiliary).
If they are available, all the previous expressions are conditional on z. The complete data likelihood:

$$f(y, I|z, \theta, \phi) = f(y|z, \theta) f(I|z, y, \phi)$$

The joint posterior of the model parameters θ and ϕ given the observed information (z, y_{inc}, I) is

$$\begin{aligned} g(\theta, \phi|z, y_{inc}, I) &\propto g(\theta, \phi|z) f(y_{inc}, I|z, \theta, \phi) \\ &= g(\theta, \phi|z) \int f(y, I|z, \theta, \phi) dy_{exc} \\ &= g(\theta, \phi|z) \int f(y|z, \theta) f(I|z, y, \phi) dy_{exc} \end{aligned}$$

Evaluating integrals is avoiding by drawing *posterior simulations* of the joint vector of unknowns (y_{exc}, θ, ϕ) and processing on the estimands of interest.

Finite population and superpopulation inference

There are two kinds of estimands:
$g(\theta, \phi | z, y_{inc}, I)$ – superpopulation quantities

$f(y_{exc} | z, y_{inc}, I, \theta, \phi)$ – finite population quantities (non-observed), descriptive statistics.

The parameters ϕ are characteristics of data collection and are not of scientific interest. Quite often they are absent at all.

y_{exc} are obtained by first drawing (θ, ϕ) from the joint posterior distribution and then drawing y_{exc} from the conditional distribution given (θ, ϕ).

Ignorability

Definition

Sampling design is called *ignorable* if

$$g(\theta|z, y_{inc}, I) = g(\theta|z, y_{inc})$$

If the *data collection process is ignored* then the posterior distribution of θ can be computed by conditioning only on y_{inc}, but not on I:

$$g(\theta|z, y_{inc}) \propto g(\theta|z)g(y_{inc}|z, \theta)) = g(\theta|z) \int f(y|z, \theta))dy_{exc}. \quad (5)$$

Example. Sampling design is ignorable and known:
SRS without covariates,
stratified SRS with stratification covariates given.

Exchangeability

de Finetti (1991) introduced exchangeability as a weaker condition than independency.

Definition

Observations are exchangeable if the *conditional density* of the sample $y_1, y_2, ..., y_n$ is unchanged for any permutation of the subscripts.

He proved: exchangeable observations can be treated as independent, given θ.

Likelihood of the exchangeable observations is expressed:

$$f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Ignorability, exchangeability $\Rightarrow$ likelihood principle $\Rightarrow$ Bayesian inference is valid

Simple random sampling

We consider finite population of N units with a study variable $y = (y_1, ..., y_N)$ (for example, weight of cargo carried by the cargo vehicles last month in the country). Parameter of interest is the finite population average $\bar{y}$.
n-size simple random sampling is defined as

$$f(I|y,\phi) = f(I) = \begin{cases} \left(C_N^n\right)^{-1} & \text{if } \sum_{i=1}^N I_i = n, \\ 0 & \text{otherwise.} \end{cases}$$

It is ignorable because does not depend on y or unknown parameters $\Rightarrow$ straightforward to deal with inferentiality.
Inference for superpopulation. The posterior density from (5):

$$g(\theta|y_{inc}) \propto g(\theta) f(y_{inc}|\theta)$$

The finite population average

$$\bar{y} = \frac{n}{N}\bar{y}_{inc} + \frac{N-n}{N}\bar{y}_{exc}, \tag{6}$$

$\bar{y}_{inc}$ and $\bar{y}_{exc}$ – averages for included and excluded y_is.

Simulation from the posterior

$\bar{y}$ can be determined using simulations of $\bar{y}_{exc}$ from its posterior predictive distribution as follows.

1. Simulate θ: θ^s, $s = 1, 2, ..., S$, from the posterior
2. For each θ^s draw the vector y^s_{exc} from

$$f(y_{exc}|\theta^s, y_{inc}) = f(y_{exc}|\theta^s) = \prod_{i:I_i=0} f(y_i|\theta^s).$$

3. Average y^s_{exc} over $s = 1, 2, ..., S$ and get $\bar{y}_{exc}$.
4. $\bar{y}_{inc}$ is known, compute $\bar{y}$ from (6).
5. Repeat 1-4 steps D times and get distribution of $\bar{y}$
6. Find mean, median, other descriptive statistics of this distribution

Any function $Q(y)$ can be used instead of $\bar{y}$.

Finite population mean, conjugate prior

A basic model for a single continuous survey outcome with simple random sampling is

$$\begin{aligned} [y_i|\theta,\sigma^2] &\sim \mathcal{N}(\theta,\sigma^2) \\ g(\theta) &\sim \text{const.} \end{aligned}$$

An improper uniform prior for the mean θ is assigned.
I) Assume the variance σ^2 to be known. Then the posterior

$$\begin{aligned} g(\theta|y) &\sim \mathcal{N}(\bar{y}_s,\sigma^2/N) \\ g(\bar{y}|y_{inc}) &\sim \mathcal{N}(\bar{y}_s,(1-n/N)\sigma^2/n) \end{aligned}$$

The Bayesian inference coincides with the design-based inference.
II) The variance σ^2 is unknown, prior $g(\theta,\sigma^2) \sim 1/\sigma^2$

$$g(\bar{y}|y_{inc}) \sim t_{n-1}(\bar{y}_s,(1-n/N)s^2/n)$$

Continuation

III) Normal prior $g(\theta) \sim \mathcal{N}(m, v^2)$ with known variance v^2. Then the posterior of the parameter

$$\begin{aligned} g(\theta|y) &\sim \mathcal{N}(m', v'^2), \\ m' &= \frac{\sigma^2}{\sigma^2 + v^2} \times m + \frac{v^2}{\sigma^2 + v^2} \times y, \\ \frac{1}{v'^2} &= \frac{1}{v^2} + \frac{1}{\sigma^2} \end{aligned}$$

IV) For non-normal prior the posterior for mean θ is not normal.

Stratified SRS and conjugate prior

N size population divided into H strata of size N_h. SRS of size n_h drawn independently in each stratum, $h = 1, 2, ..., H$. This design is ignorable given H vectors $z_1, z_2, ..., z_H$ with $z_h = (z_{1h}, ..., z_{nh})$ and

$$z_{ih} = \begin{cases} 1 & \text{if unit } i \text{ is in stratum } h, \\ 0 & \text{otherwise, } \ i = 1, ..., n. \end{cases}$$

A natural analysis:

- to model the distributions of the measurements y_i within each stratum h in terms of parameters θ_h
- to perform Bayesian inference on all the sets of parameters $\theta_1, ..., \theta_H$.

A hierarchical model can be assigned to the θ_h's.
The finite population inferences can be obtained by weighting the inferences from the separate strata.
For example, the population mean $\bar{y}$ is written in terms of individual stratum means $\bar{y}_h$, as $\bar{y} = \sum_{h=1}^{H} \bar{y}_h N_h / N$.
The finite population quantities $\bar{y}_h$ can be simulated given the simulated parameters θ_h for each stratum.

Stratified SRS and conjugate prior

The model/likelihood:

$$[y_i|z_{ih} = h,\ \{\theta_h, \sigma_h^2\}] \sim \mathcal{N}(\theta_h, \sigma_h^2)$$

Let σ_h^2 are known and flat prior on the stratum means is assigned:

$$g(\theta_h|z) \sim \text{const}$$

Bayesian analysis similar to SRS (**I**) leads to

$$[\bar{y}|z, y_{inc}, \{\sigma_h^2\}] \sim \mathcal{N}(\bar{y}_{st}, \sigma_{st}^2),$$

The finite population parameters

$$\begin{aligned}
\bar{y}_{st} &= \sum_{h=1}^{H} P_h \bar{y}_{sh}, \quad P_h = N_h/N,\ \bar{y}_{sh} = \text{sample mean in stratum } h, \\
\sigma_{st}^2 &= \sum_{h=1}^{H} P_h^2 (1 - n_h/N_h)\sigma_h^2/n_h.
\end{aligned}$$

Two-stage sampling

Suppose the population is divided into C clusters. Two-stage sampling:

- SRS of c from C clusters selected,
- n_c units from N_c units in each sampled cluster selected.

The inclusion mechanism is ignorable conditional on cluster information, but model needs to account for within cluster correlation in the population.
Let y_{ic} =outcome for unit i in cluster c, $i = 1, ..., N_c$; $c = 1, ..., C$.
A normal model is

$$[y_{ic}|\theta_c, \sigma^2] \sim \mathcal{N}(\theta_c, \sigma^2), \quad [\theta_c|\mu, \phi] \sim \mathcal{N}(\mu, \phi), \quad \mu \sim \ , \phi \sim \ .$$

A flat prior $g(\theta_c) = const.$ cannot be assigned to the cluster means, because only a subset of clusters is sampled;
the uniform prior does not allow information from sampled clusters to predict means for non-sampled clusters.

Unequal selection probabilities

Independent sampling with unequal inclusion probabilities for elements.

The design is ignorable conditionally on the variables z determining the sampling probabilities, which are known for the general population.

The *critical step* then in Bayesian modelling is formulating the conditional distribution of y given z.

Multiple imputation for item nonresponse

Assume missing at random (MAR) mechanism for a variable y.
In order to make multiple imputation in Bayesian way do the following:

- Impute *draws*, not means, from the posterior predictive distribution of the missing values;
- Create $D > 1$ different filled in data sets with different values imputed; Estimate the parameters needed.
- Average the estimates over multiply imputed data sets.

MCMC computations for multiple imputation

Let obs, $miss$ means observed and missed values, correspondingly.
$y = (y_{obs}, y_{miss})$
Ignoring the missingness model, the posterior distribution can be written:

$$\begin{aligned} g(\theta|y_{obs}) &\propto g(\theta) f(y_{obs}|\theta), \\ f(y_{obs}|\theta) &= \int f(y|\theta) dy_{miss} \end{aligned}$$

A general algorithm of data augmentation is used.
At the step $t+1$ the algorithm draws y_{miss} and θ by alternating the following steps:

$$\begin{aligned} [y_{miss}^{(d,t+1)}|y_{obs}, \theta^{(d,t)}] &\sim f(y_{miss}|y_{obs}, \theta^{(d,t)}) \\ [\theta^{(d,t+1)}|y_{miss}^{(d,t+1)}] &\sim g(\theta|y_{obs}, y_{miss}^{(d,t+1)}) \end{aligned}$$

As t tends to infinity, this sequence converges to a draw from the joint posterior distribution of (y_{miss}, θ), as required.
This is an application of the Gibbs sampler used for estimation of the multi-parameter models.

Survey statistics and Bayesian inference

- Design-based statistics works well for large samples
- Model-based Bayesian inference works well for small samples and large samples
- Bayesian inference is common for SAE, missing data imputation, editing, outlier adjustment
-

$$\text{Little}: \begin{cases} n \text{ large}: \text{ design} - \text{based inference} \\ n_0 = \text{"point of inferential schizophrenia"} \\ n \text{ small}: \text{ model} - \text{based inference} \end{cases}$$

Daniel Thorburn (Kyiv, 2009):
Always use Bayesian methods!

References

1. Bolstad William M., Curran James M. (2017). *Intoduction to Bayesian Statistics*, Third Edition. John Wiley & Sons.
2. Van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL.: Chapman & Hall/CRC Press.
3. Gelman Andrew, Carlin John B., Stern Hal S., Dunson David B., Vehtari Aki, Rubin Donald B. (2014). *Bayesian Data Analysis*, Third Edition. CRC Press, Taylor & Francis Group.
4. Rao J. N. K. (2011) Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal. *Statistical Science*, **26**(2), 240-256.
5. Little Roderick (2011). *Bayesian Inference for sample surveys*. https://www.scb.se/Grupp/Produkter_Tjanster/Kurser/_Dokument/JOS-2015/little-bayesmodule1intro.pdf
6. Statisticat, LLC. (2018). *LaplacesDemon*. R package version 16.1.1. [https://cran.r-project.org/]

A Tutorial with R, JAGS, and Stan

John K. Kruschke