A simulation of nonresponse and imputation (or, a framework for narrative modeling)

Ben Klemens

Center for Statistical Research and Methodology
U.S. Census Bureau
ben.klemens@census.gov

20 August 2013

United States™ Census Bureau

Disclaimer

This paper is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

United States Census Bureau

The Outline

- The theory part, defining a statistical model.
  - Using a random number generator as a statistical model.
- The simulation
  - Interviewer agents seek out respondent agents.
  - This is a work in progress.

United States Census Bureau

My storyline

- I had agent-based models that needed statistical analysis. My options:
  - Write the ABM in the language with all the stats functions.
    - It was too slow.

United States Census Bureau

My storyline

- I had agent-based models that needed statistical analysis. My options:
  - Write the ABM in the language with all the stats functions.
    - It was too slow.
  - Write stats functions in a language fast enough for ABM.
    - I wound up actually doing this

United States Census Bureau

My storyline

- I had agent-based models that needed statistical analysis. My options:
  - Write the ABM in the language with all the stats functions.
    - It was too slow.
  - Write stats functions in a language fast enough for ABM.
    - I wound up actually doing this
- In natural programmer form, I started writing up a model object.
  - Slots for parameters, a pointer to the input data, functions describing the estimation, the score, &c.
  - Functions that take in a model object (see below)

United States Census Bureau

My storyline

- I had agent-based models that needed statistical analysis. My options:
  - Write the ABM in the language with all the stats functions.
    - It was too slow.
  - Write stats functions in a language fast enough for ABM.
    - I wound up actually doing this
- In natural programmer form, I started writing up a model object.
  - Slots for parameters, a pointer to the input data, functions describing the estimation, the score, &c.
  - Functions that take in a model object (see below)
- The epiphany: The same model object that describes regressions and distributions also works for ABMs.

United States Census Bureau

Apophenia

- A library of stats functions written around a model object as described here.

United States Census Bureau

Apophenia

- A library of stats functions written around a model object as described here.
- Pretty mature:
  - $\sim$ 250 functions for data shunting and modeling
  - Documentation is 70,000 words long (plus a book)
  - Testing suite about as large as the code base
  - $\sim 13,500$ nontrivial lines of code
  - Slipped into Census production here and there

United States Census Bureau

Apophenia

- A library of stats functions written around a model object as described here.
- Pretty mature:
  - $\sim$ 250 functions for data shunting and modeling
  - Documentation is 70,000 words long (plus a book)
  - Testing suite about as large as the code base
  - $\sim 13,500$ nontrivial lines of code
  - Slipped into Census production here and there
  - Raking (IPF) for sparse tables

United States Census Bureau

Apophenia

- A library of stats functions written around a model object as described here.
- Pretty mature:
  - $\sim$ 250 functions for data shunting and modeling
  - Documentation is 70,000 words long (plus a book)
  - Testing suite about as large as the code base
  - $\sim 13,500$ nontrivial lines of code
  - Slipped into Census production here and there
  - Raking (IPF) for sparse tables
- v1 coming soon
- Interested? Contribute!

United States Census Bureau

Apophenia

- A library of stats functions written around a model object as described here.
- Pretty mature:
  - $\sim$ 250 functions for data shunting and modeling
  - Documentation is 70,000 words long (plus a book)
  - Testing suite about as large as the code base
  - $\sim 13,500$ nontrivial lines of code
  - Slipped into Census production here and there
  - Raking (IPF) for sparse tables
- v1 coming soon
- Interested? Contribute!
- Not interested? This talk will be language independent—implement it in your favorite computing platform!

United States Census Bureau

A bundle of functions

United States Census Bureau

Notation

- $\mathbb{D}$: Data space. Anything required by the model; 'private' to the model unless otherwise noted. $\leq$ is defined. [sample space]
- $\mathbb{P}$: Parameter space. Similarly model-specific. [state space]
- $\mathbb{M}$: Model space. The set of categories of $\{\mathbb{D}$, $\mathbb{P}$, ML-consistent functions as per the next slide$\}$.

A bundle of functions

- Likelihood: $(\mathbb{D}, \mathbb{P}) \rightarrow \mathbb{R}^+$.
  - Integrates to a finite value; always nonnegative.
  - In some cases, better described as an 'objective function'. Doesn't have to integrate to one.
- Estimation: $\mathbb{D} \rightarrow \mathbb{P}$
  - ML-consistency: L(**d**, **p**) is maximized by $\mathbf{p} = \textsc{Est}(\mathbf{d})$.
- RNG: $\mathbb{P}$ (and uniform prng) $\rightarrow \mathbb{D}$.
  - Likelihood of draw $\mathbf{d} = \mathrm{RNG}(\mathbf{p}) \propto L(\mathbf{d}, \mathbf{p})$.
- CDF: $(\mathbb{D}, \mathbb{P}) \rightarrow [0, 1]$.
  - Proportion of random draws $\mathrm{RNG}(\mathbf{p}) \leq \mathbf{d} \rightarrow \mathrm{CDF}(\mathbf{d}, \mathbf{p})$.

United States Census Bureau

The Normal example

- Likelihood: $\mathcal{N}(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-(x-\mu)^2/2\sigma^2)$ or
- Estimation: $\hat{\mu} =$ mean of D; $\hat{\sigma} = \sum(d - \hat{\mu})^2/n$.
- RNG: See Devroye (1986).
- CDF: `gsl_cdf_gaussian_P(d-mu, sd)` (or see `erf`).

United States Census Bureau

Just a likelihood

Given only a likelihood function, $P(\mathbf{D}, \mathbf{P})$.

- Score (dlog likelihood): numeric deltas.
- Estimation: Use Maximum likelihood estimation.
  - All MLE algorithms repeatedly sample from the likelihood. Some use the score.
- RNG: ARMS (Gilks 1995)
- CDF: make random draws, count the percent up to a given point

United States Census Bureau

Just an RNG

Given only $\mathrm{RNG}(\mathbf{P})$.

- CDF: make random draws, count the percent up to a given point
- Likelihood: make a million draws, write down a PMF using those draws. Optional: smooth the PMF.
  - The problem has been reduced to the prior slide

United States Census Bureau

The category of $\{\mathbb{M}$, functions$\}$

- Transformations are of the form $t : \mathbb{M} \rightarrow \mathbb{M}$
  - Truncation
  - Jacobian: e.g., map $d \rightarrow \sqrt{d}$
  - Constrain the parameters.
- Transformations that join models: $j : (\mathbb{M}, \mathbb{M}) \rightarrow \mathbb{M}$
  - Stack it
  - Mix it
  - Bayesian-update it $(\mathbb{D}_{\text{out}} = \mathbb{P}_{\text{in}})$
  - Data-compose it $(\mathbb{D}_{\text{out}} = \mathbb{D}_{\text{in}})$

United States Census Bureau

So we can build complex models via transformations

- The square of x is Normally distributed; has uncertain μ, σ.

United States Census Bureau

So we can build complex models via transformations

- The square of x is Normally distributed; has uncertain μ, σ.
  - $\mathcal{N}$ = Normal distribution
  - InvWish = Inverse Wishart distribution

United States Census Bureau

So we can build complex models via transformations

- The square of x is Normally distributed; has uncertain μ, σ.
  - $\mathcal{N}$ = Normal distribution
  - InvWish = Inverse Wishart distribution
  - M_{pri} = Stack($\mathcal{N}$, InvWishart)

United States Census Bureau

So we can build complex models via transformations

- The square of x is Normally distributed; has uncertain μ, σ.
  - $\mathcal{N}$ = Normal distribution
  - InvWish = Inverse Wishart distribution
  - M_{pri} = Stack($\mathcal{N}$, InvWishart)
  - M_{L} = Truncate$_{x>0}(\mathcal{N})$

United States Census Bureau

So we can build complex models via transformations

- The square of x is Normally distributed; has uncertain μ, σ.
  - $\mathcal{N}$ = Normal distribution
  - InvWish = Inverse Wishart distribution
  - M_{pri} = Stack($\mathcal{N}$, InvWishart)
  - M_{L} = Truncate$_{x>0}(\mathcal{N})$
  - M_{sq} = Update(M_{pri}, M_{L})

United States Census Bureau

So we can build complex models via transformations

- The square of x is Normally distributed; has uncertain μ, σ.
  - $\mathcal{N}$ = Normal distribution
  - InvWish = Inverse Wishart distribution
  - M_{pri} = Stack($\mathcal{N}$, InvWishart)
  - M_{L} = Truncate$_{x>0}(\mathcal{N})$
  - M_{sq} = Update(M_{pri}, M_{L})
  - M = Jacobian$_{\sqrt{x}}(M_{sq})$

United States Census Bureau

So we can build complex models via transformations

- The square of x is Normally distributed; has uncertain μ, σ.
  - $\mathcal{N}$ = Normal distribution
  - InvWish = Inverse Wishart distribution
  - M_{pri} = Stack($\mathcal{N}$, InvWishart)
  - M_{L} = Truncate$_{x>0}$($\mathcal{N}$)
  - M_{sq} = Update(M_{pri}, M_{L})
  - M = Jacobian$_{\sqrt{x}}$(M_{sq})

$$M = \text{Jacobian}_{\sqrt{x}}(\text{Update}(\text{Stack}(\mathcal{N}, \text{InvWishart}), \text{Trunc}_{x>0}(\mathcal{N})))$$

United States Census Bureau

$$\mathbb{D}_{\text{out}} = \mathbb{D}_{\text{in}}$$

- Given $\mathbb{P}_1$, M_1 produces via RNG.
- Given $\mathbb{P}_2$, M_2 gives likelihood of d.
- $M_3 = Dcompose(M_1, M_2)$ has params $\mathbb{P}_1 \times \mathbb{P}_2$.

United States Census Bureau

$\mathbb{D}_{\text{out}} = \mathbb{D}_{\text{in}}$

- Given $\mathbb{P}_1$, M_1 produces via RNG.
- Given $\mathbb{P}_2$, M_2 gives likelihood of d.
- $M_3 = Dcompose(M_1, M_2)$ has params $\mathbb{P}_1 \times \mathbb{P}_2$.
- Find the most likely value of λ for a Poisson or Exponential based on data generated via a simulation.
- Likelihood for M_2 is the parameterless distance function between d and fixed $d_g \Rightarrow$ a traditional model calibration.

United States Census Bureau

$\mathbb{D}_{\text{out}} = \mathbb{D}_{\text{in}}$

- Given $\mathbb{P}_1$, M_1 produces via RNG.
- Given $\mathbb{P}_2$, M_2 gives likelihood of d.
- $M_3 = Dcompose(M_1, M_2)$ has params $\mathbb{P}_1 \times \mathbb{P}_2$.
- Find the most likely value of λ for a Poisson or Exponential based on data generated via a simulation.
- Likelihood for M_2 is the parameterless distance function between d and fixed $d_g \Rightarrow$ a traditional model calibration.

United States Census Bureau

More computational epistemology

- Frequentist linear models: $g(Y) = f(X\beta) + \epsilon$
  - Deterministic narrative, with uncertainty added at the end

United States Census Bureau

More computational epistemology

- Frequentist linear models: $g(Y) = f(X\beta) + \epsilon$
  - Deterministic narrative, with uncertainty added at the end
- Bayesian hierarchies:

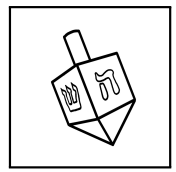
$$\begin{aligned} \mu_L &\sim \mathcal{N}(\mu_p, \sigma_p) \\ \sigma_L &\sim \text{InvWish}(\Sigma) \\ x &\sim \mathcal{N}(\mu_L, \sigma_L) \end{aligned}$$

  - Uncertainty starts at the outset; goes throughout the model
  - Textbook distributions expressing a narrative of how the data was generated

United States Census Bureau

More computational epistemology

- Agent-based model:

$$
\begin{aligned}
\text{ה} &\sim \mathcal{N}(\mu_p, \sigma_p) \\
\text{ש} &\sim \text{InvWish}(\Sigma) \\
x &\sim \text{  }
\end{aligned}
$$

  - Black box model expressing a narrative of how the data was generated
  - Uncertainty also expressed throughout the model

United States Census Bureau

Simulating a survey

United States Census Bureau

The survey and impute process

- $\mathcal{D}$ is the true population distribution
- In N dimensions, where N is large: $\mathcal{D}$(age, sex, race, block, income, commute time, gov't attitude, work hours, . . .).

United States Census Bureau

The survey and impute process

- $\mathcal{D}$ is the true population distribution
- In N dimensions, where N is large: $\mathcal{D}$(age, sex, race, block, income, commute time, gov't attitude, work hours, . . .).
- Surveying process is a transformation: $\mathbb{M} \rightarrow \mathbb{M}$ that induces distortions.
- Imputation: $Imp : \mathbb{M} \rightarrow \mathbb{M}$
- Our job: select Imp such that $Imp(Survey(\mathcal{D})) \sim \mathcal{D}$.

United States Census Bureau

Why an ABM

- We don't actually know $Survey(\cdot)$.
  - When are people home? What is their inclination toward gov't surveys? Where/when did the Census send interviewers?
- It's easier to write $Survey(\mathcal{D})$ as its own model, with an explicit description of agents.

United States Census Bureau

Why an ABM

- We don't actually know $Survey(\cdot)$.
  - When are people home? What is their inclination toward gov't surveys? Where/when did the Census send interviewers?
- It's easier to write $Survey(\mathcal{D})$ as its own model, with an explicit description of agents.
- Information sponge: you give me a relevant fact about how respondents and interviewers interact, and I can fit it into the narrative.

United States Census Bureau

The population

- Two types of agent: respondent households and interviewers.
- Draw households from DC PUMS (urban, square)
  - The geography in the simulation is terrible so far.
- Census Barriers, Attitudes, and Motivators Survey (CBAMS, Bates et al 2009):
  - Marketing survey about survey attitudes
  - Five categories (see next slide)
  - Mulry & Olson: a Logit model of type given race, income, urban/rural, U.S. born, Primary English.
  - Use their coefficients to calculate the odds that a person is of any given type; draw a type.

United States Census Bureau

The mindsets

- Leading edge (26%): pro-Census, "say that they will inform family and friends about the Census"
- Cynical Fifth (19%), "anti-government and anti-institution"
- Head nodders (41%), largely positive, but not well informed. "might be vulnerable to negative publicity"
- Insulated (7%), have heard of the Census, but are unfamiliar
- Unacquainted (6%) Have never heard of the Census

United States Census Bureau

The interviewers

- Based on a simulation by Bor-Chung Chen.
- Explicitly models drive time and drive costs.
- Uniform characteristics and abilities.
- Everything is one mode. With more data, one could model mail/telephone.

United States Census Bureau

A day

- Interviewers are allocated
  - Current strategy: in proportion to demand
  - Uniform, no-reallocation, or other strategies are possible

United States Census Bureau

A day

- Interviewers are allocated
  - Current strategy: in proportion to demand
  - Uniform, no-reallocation, or other strategies are possible
- Knock on a door:
  - Odds of respondent being at home is a three-hump density function.
  - If a householder is home, answers with probability determined by type.
- If respondent answers, all data is gathered with 100% accuracy.

United States Census Bureau

A day

- Interviewers are allocated
  - Current strategy: in proportion to demand
  - Uniform, no-reallocation, or other strategies are possible
- Knock on a door:
  - Odds of respondent being at home is a three-hump density function.
  - If a householder is home, answers with probability determined by type.
- If respondent answers, all data is gathered with 100% accuracy.
- At end of day, interviewees tell their neighbors.

United States Census Bureau

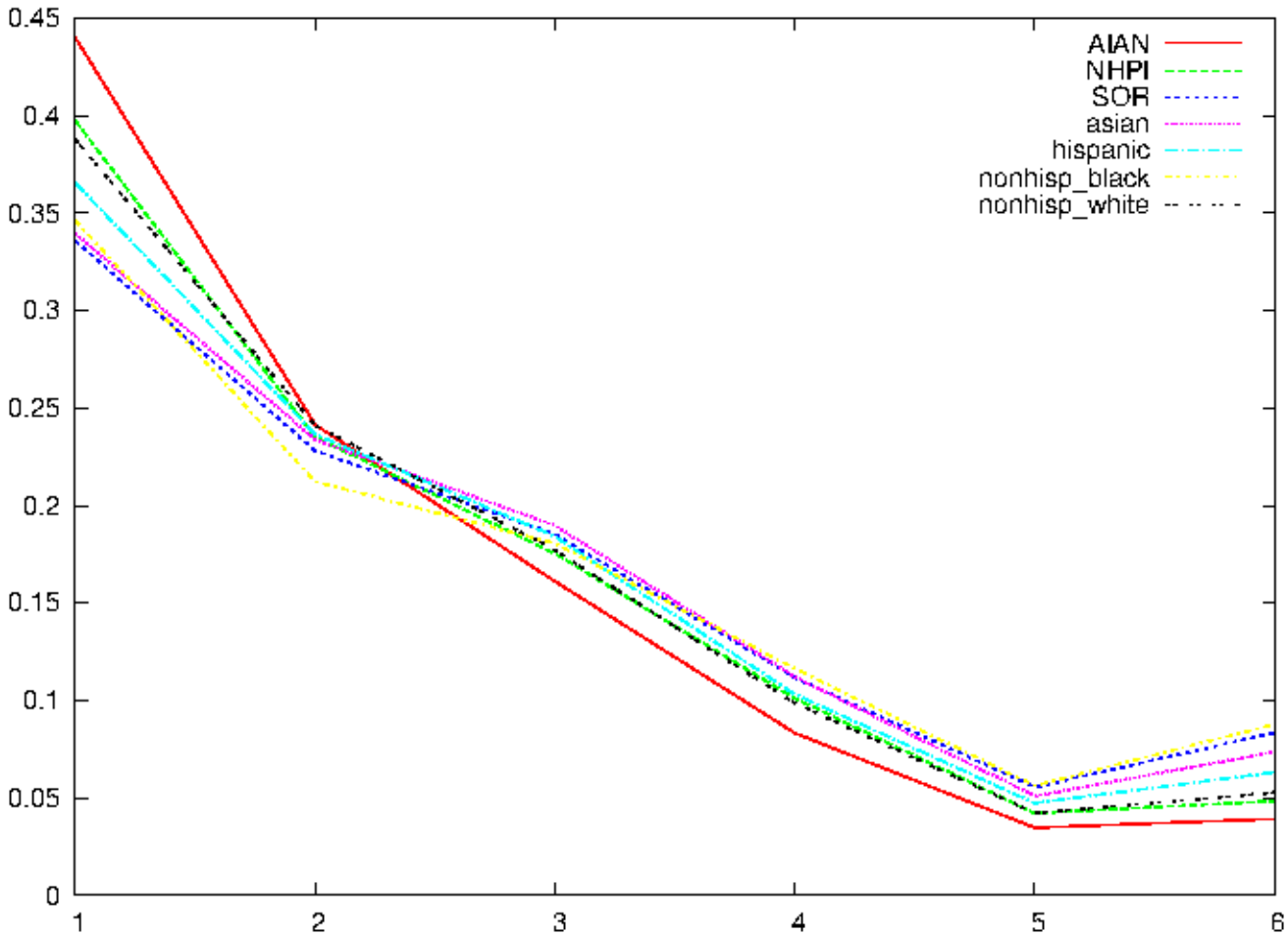
Calibrating the model–parameters and settings

- Settings: details that are fixed and aren't worth optimizing:
  - population size
  - driving costs
  - for this run: staff size
- Parameters: the optimizer will search for a best value
  - P(response|unacquainted)
  - P(response|insulated)
  - P(response|head nodder)
  - P(response|leading edge)
  - P(response|cynical fifth)

United States Census Bureau

Response rate by race

Odds of responding on NRFU 1, 2, . . . , 6+ broken down by race:



Calibrating the model

- We could do it nonparametrically by comparing K-L divergence between observed and simulated distributions; I parameterized it.
- Recall the dcompose transformation.

United States Census Bureau

Calibrating the model

- We could do it nonparametrically by comparing K-L divergence between observed and simulated distributions; I parameterized it.
- Recall the dcompose transformation.
- Distributions are parameterized by λ_D^{hisp}, $\lambda_D^{\text{nonhisp black}}$, $\lambda_D^{\text{nonhisp white}} \ldots$
- Data generates distributions of NRFUs with parameters λ_S^{hisp}, $\lambda_S^{\text{nonhisp black}}$, $\lambda_S^{\text{nonhisp white}} \ldots$
- Loss function is

$$|\lambda_S^{\text{hisp}} - \lambda_D^{\text{hisp}}| + |\lambda_S^{\text{nonhisp black}} - \lambda_D^{\text{nonhisp black}}| + \ldots$$

United States Census Bureau

Optimization

- The optimization search happens here, via simulated annealing:
  - SimAn is ideal for optimizing a stochastic function.
  - Nelder-Mead Simplex also tends to work OK in practice.

```
estimated_model = apop_estimate(data, survey_sim)
```

United States Census Bureau

Variation

Given the optimum from the last slide, I calculate variances using the local inverse Hessian.

- Inverse Hessian calculation easily takes a black-box model
- Comparative statics are valid: if σ_i is $3\sigma_j$, then in expectation whatever causes a 1-unit shift in β_i will cause, in expectation, a 3-unit shift in β_j.
- Parameter sensitivity is inversely proportional to Hessian^{-1}: as perturbances happen, does β_i fluctuate or hold steady?

United States Census Bureau

Variation

Given the optimum from the last slide, I calculate variances using the local inverse Hessian.

- Inverse Hessian calculation easily takes a black-box model
- Comparative statics are valid: if σ_i is $3\sigma_j$, then in expectation whatever causes a 1-unit shift in β_i will cause, in expectation, a 3-unit shift in β_j.
- Parameter sensitivity is inversely proportional to Hessian^{-1}: as perturbances happen, does β_i fluctuate or hold steady?
- Asking whether zero or β_j is $2\sigma_i$ from β_i has limited utility—how much do we believe $\boldsymbol{\beta} \sim$ Multivariate Normal?

United States Census Bureau

Odds of responding

head nodder	80% (7.2%)
leading edge	76% (1.3%)
unacquainted	74% (6.3%)
insulated	57% (8.5%)
cynical fifth	53% (1.7%)

- This was a test (and vindication) of CBAMS: loss function improves when the ABM uses type labels; labels have meaningful interpretation.

United States Census Bureau

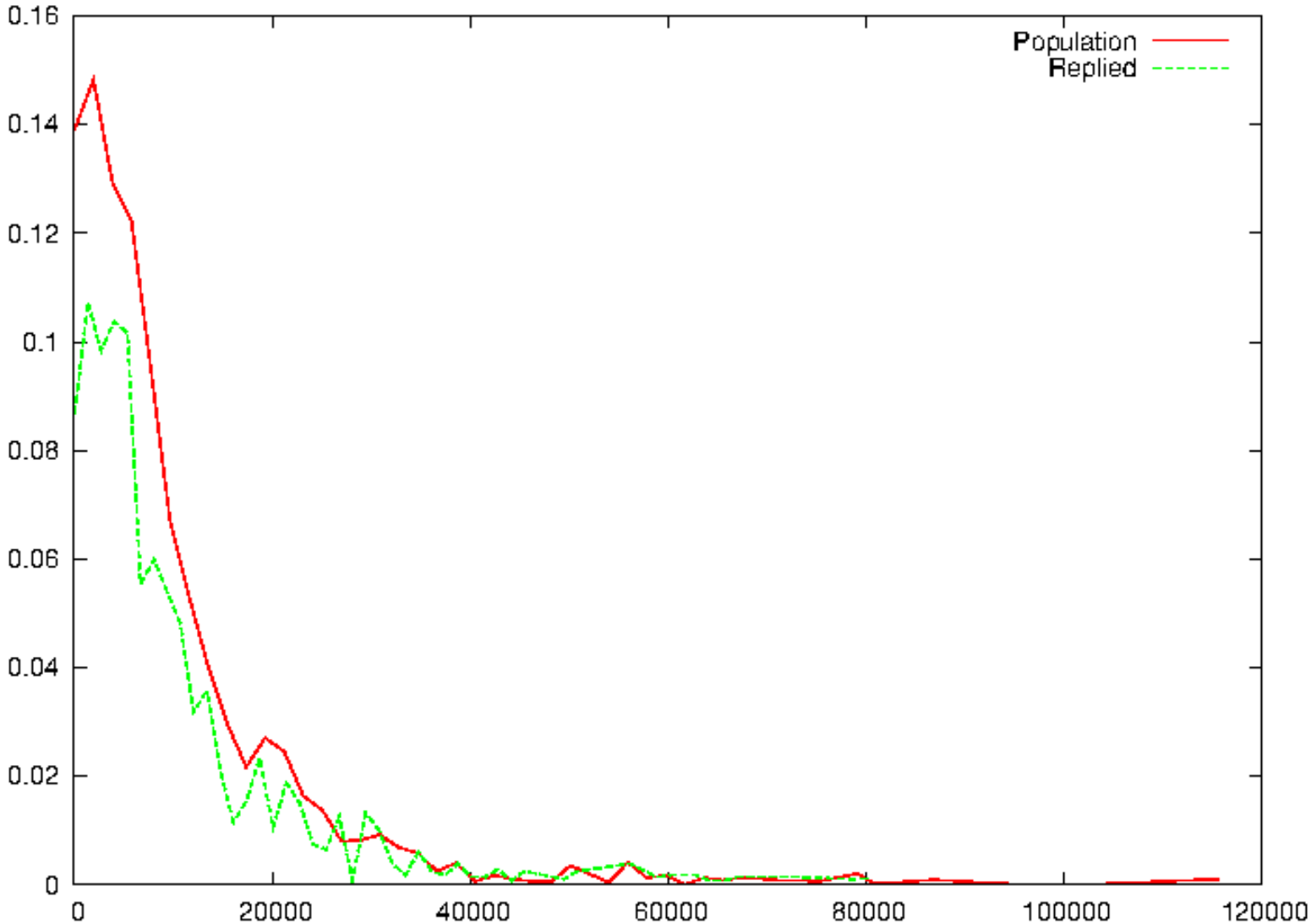
A few imputations

Imputing with the sim

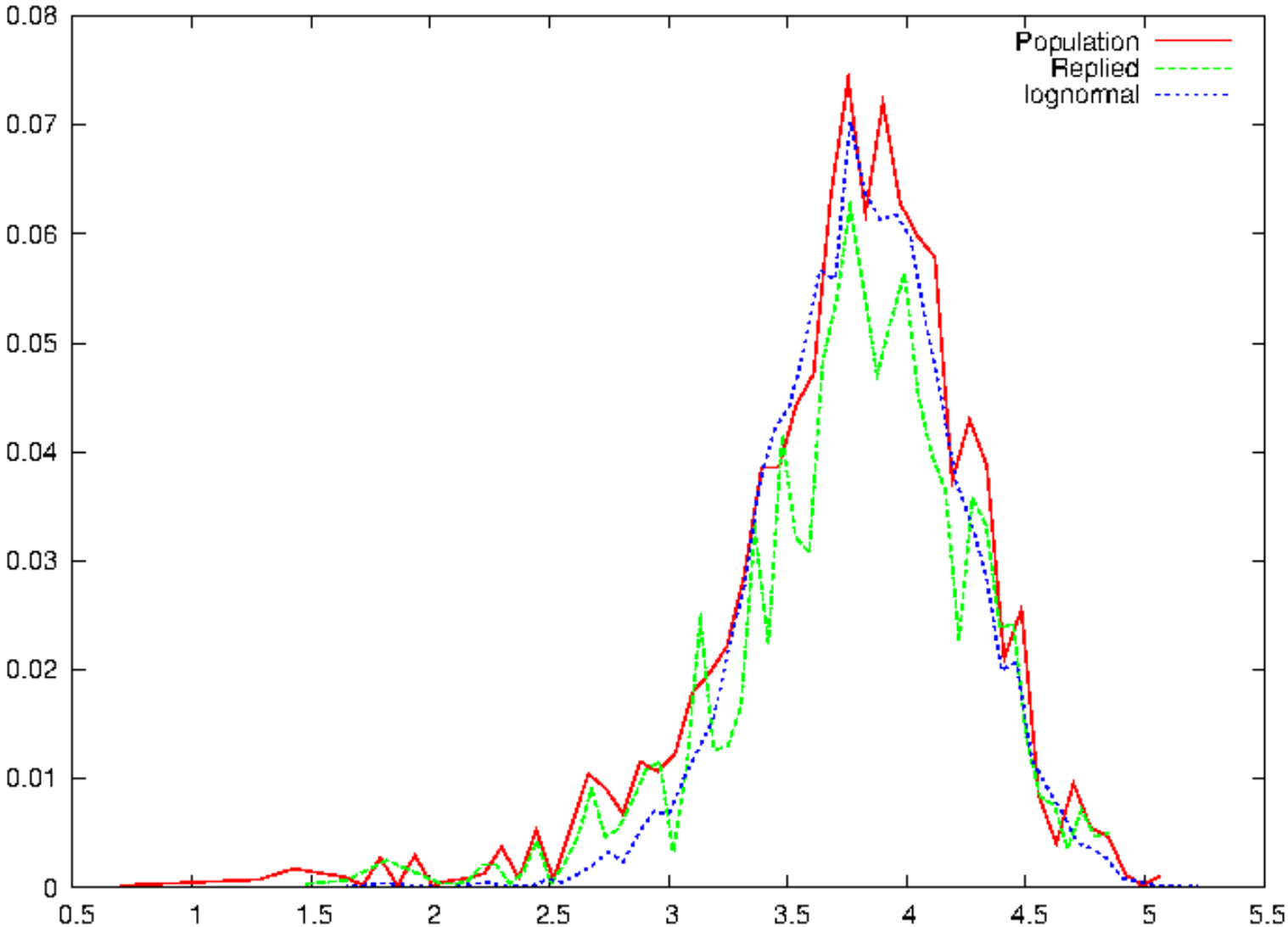
- Distinct analysis from the above
- Takes parameters as fixed. Assigning a prior to them is left as an exercise for the reader

United States Census Bureau

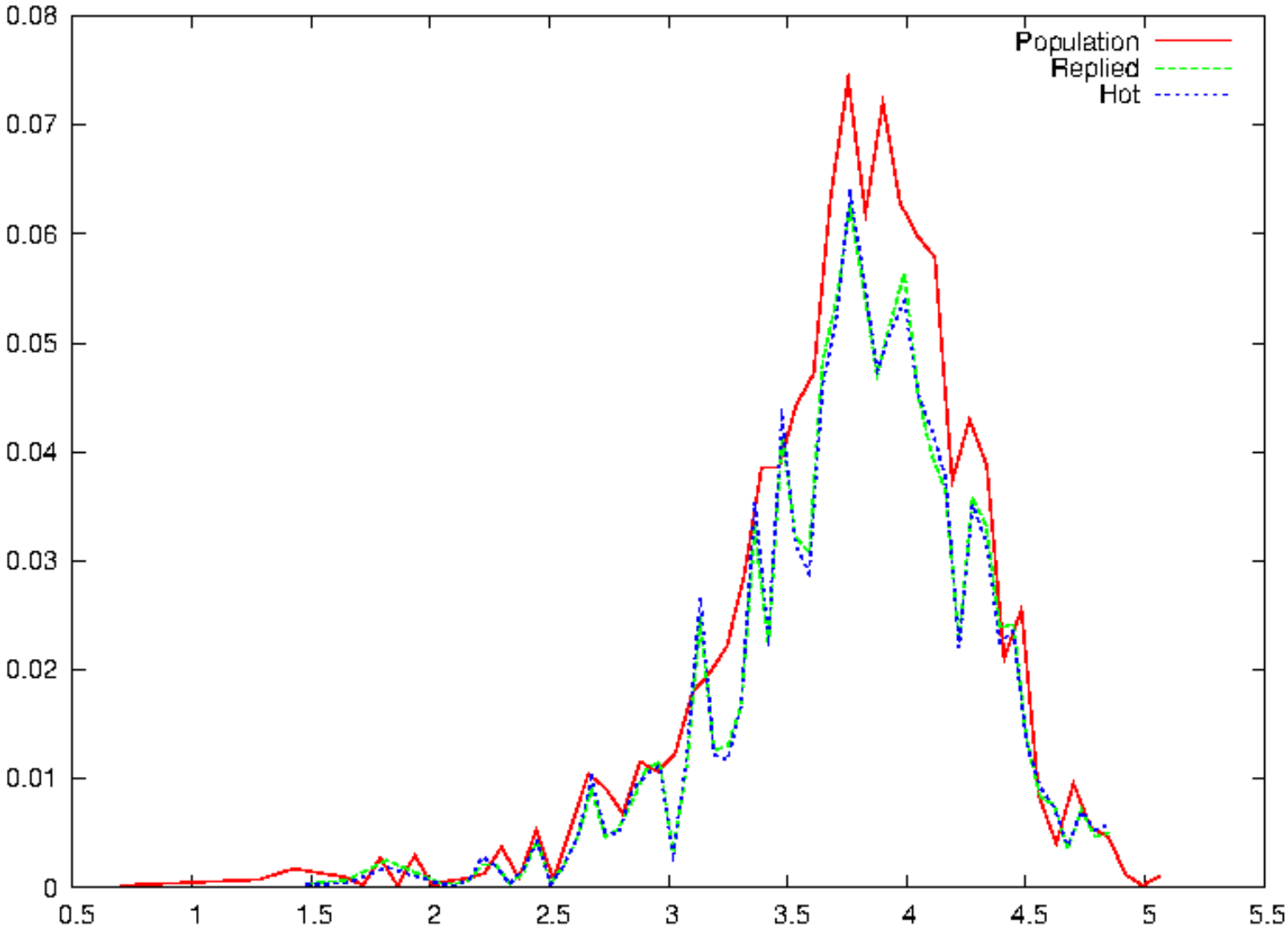
Our respondents



The imputations–lognormal



The imputations–hot deck



Log(income)–lognormal vs hot deck

	bias	MSE
lognormal distribution	0.035 ($\pm$0.005)	.166 ($\pm$0.002)
hot deck	-0.002 ($\pm$0.002)	.237 ($\pm$0.02)

Est (replication variance)

United States Census Bureau

Conclusion

United States Census Bureau

The conclusion slide

- A random number generator is sufficient to describe a statistical model.
- An ABM produces random outcomes from fixed input parameters.
- $\therefore$ An ABM is sufficient to describe a statistical model.

United States Census Bureau

The conclusion slide

- A random number generator is sufficient to describe a statistical model.
- An ABM produces random outcomes from fixed input parameters.
- $\therefore$ An ABM is sufficient to describe a statistical model.
- Things we do with a statistical model:
  - Calibrate it, by searching for parameters that produce data as close as possible to real-world data.
  - Evaluate the robustness of those parameters.
  - Transform it, such as applying an imputation.

United States Census Bureau