COMBINING INFORMATION FROM A CENSUS, A COVERAGE MEASUREMENT SURVEY, AND DEMOGRAPHIC ANALYSIS

Michael R. Elliott and R.J.A. Little
Department of Biostatistics
University of Michigan

- Capture-recapture methods and demographic analysis (DA) have established that subgroups of the US population are undercounted in the US Census (Robinson 1993).

- DA indicates a 1990 male-female ratio among African-Americans 30-49 of 0.91, whereas the 1990 Census determined a sex ratio of 0.86.

- DA provides national-level estimates only.

- Capture-recapture methods require modeling assumptions that are difficult to assess based on fit to the data.

Coverage Measurement Survey (CMS)

- The 1990 CMS was conducted in a sample of Census blocks during July-September 1990 (Hogan 1992).
- Imputations and erroneous enumerations were removed from Census counts (E-sample).
- Persons in the CMS (P-sample) were cross-checked against Census records and assigned to In- or Out-Census category.

P-sample counts were then inflated by the inverse of selection probability and 2 $\times$ 2 tables formed for males and females ($S = M, F$) in each of K post-strata (area $\times$ owner $\times$ age $\times$ race)

		Observed Data			Underlying Parameters		
		CMS			CMS		
		In	Out		In	Out	
Census	In	y^S_{k11}		$z^S_{k1.}$	ψ^S_{k11}	ψ^S_{k10}	$\psi^S_{k1.}$
	Out	y^S_{k01}			ψ^S_{k01}	ψ^S_{k00}	$\psi^S_{k0.}$
		$y^S_{k.1}$			$\psi^S_{k.1}$	$\psi^S_{k.0}$	$\psi^S_{k..}$

- Only 3 data elements present to estimate 4 parameters.
- Constraints must be placed on the parameters to obtain unique estimates.
- Assume independence of capture and recapture (ICR). If probabilities of capture and recapture are not independent or are unequal across individuals, result will be downwardly biased ("correlation bias"). Hence Sekar and Deming (1949) suggested dividing capture-recapture tables into K homogenous poststrata and estimating ψ_{k00}^S under the constraint:

$$\text{ICR: } \theta_k^S = 1 \text{ for all } k, S.$$

  where $\theta_k^S = (\psi_{k11}^S/\psi_{k10}^S)/(\psi_{k01}^S/\psi_{k00}^S)$.

- May still be difficult to achieve homogeneity in practice: estimated male-female ratio among African-Americans aged 30-49 is 0.84 under ICR, not the DA estimate of 0.91. Hence Wolter (1990) relaxes independence assumption for males and instead assumes a fixed OR for males across K strata:

$$\text{FOR: } \theta_k^M = \theta^M \text{ and } \theta_k^F = 1,$$

  Identifiability maintained by forcing parameter estimates to equal known nationwide sex ratio ($\rho = \hat{\psi}^M_{\cdots} / \hat{\psi}^F_{\cdots}$).

- Bell (1993) and Das Gupta (1996) extend Wolter's approach by assuming different constraints for males (fixed relative risk, etc.), again assuming independence for females and known nationwide sex ratios.

Generic Problem

- All models are saturated and provide equally good fit to data with non-trivial differences.
  - Propose qualitative principles for choosing between saturated models.

Principles for Choosing Between Saturated Models

1. PLAUSIBILITY: The model should imply a plausible description of behavior.

2. FIT: The model should minimize contradiction with available data.

3. PREDICTION: The model should provide plausible predictions of key unobserved quantities.

4. ICR INCLUSION: The model should include the ICR model, which assumes zero correlation bias within poststrata, as a particular case.

5. STABILITY: If alternative competing models are not distinguished on the basis of 1-4, models that yield more stable estimates of key estimands should be favored.

6. CONSERVATISM: If alternative competing models are not distinguished on the basis of 1-5, then models that are more conservative with respect to undercount adjustment should be favored.

Among five models suggested by Wolter, Bell, and Das Gupta, principles of ICR inclusion and stability suggest use of the fixed relative-risk (FRR) model, which assumes a constant relative risk for enumeration in the Census and CMS for males:

$$\text{FRR: } \gamma_k^M = \frac{\psi_{k11}^M / \psi_{k1.}^M}{\psi_{k01}^M / \psi_{k0.}^M} = \gamma^M \text{ for all } k.$$

Maximum Likelihood Estimation Problems

- Natural estimate of ψ_{k10}^S (Out CMS-In Census) given by $z_{k1.}^S - y_{k11}^S < 0$, which can lead to unstable poststratum estimates.
  - Use Bayesian model to introduce priors that have non-zero support.
- Unclear whether negative Out CMS-In Census cells due to variability or bias.
  - Obtain posterior predictive distributions of ψ_{k10}^S and compare with $z_{k1.}^S - y_{k11}^S$.
- How to incorporate uncertainty in sex ratio estimates?
  - Use prior distribution incorporating knowledge about variability of ρ.

Bayesian Fixed Relative Risk Model

Transform $8K$ poststratum parameters:

Parameter	Description	Prior
$\psi_{k..}$	Total population count	$\propto 1$
$\rho_k = \psi^M_{k..}/\psi^F_{k..}$	Sex ratio	$N(\rho, \sigma^2)$
$\delta^S_k = \psi^S_{k1.}/\psi^S_{k..}$	Census undercount proportion	$BETA(a^S, b^S)$
$\phi^S_k = \psi^S_{k11}/\psi^S_{k1.}$	Proportion of Census cases in CMS	$BETA(c^S, d^S)$
$\gamma^M_k = (\psi^S_{k11}/\psi^S_{k1.})/(\psi^S_{k01}/\psi^S_{k0.})$	Relative proportion of Census and non-Census cases in CMS	$GAMMA(\alpha, \beta)$

Also assume $\gamma^M_k = \gamma^M$, $\gamma^F_k = 1$, and y^S_{k11}, y^S_{k01}, and $z^S_{k1.}$ are independent and normally distributed with known variance.

Utilize Gibbs sampling to obtain posterior parameter estimates of $\psi_k^* = (\psi_{k..}, \rho_k, \delta_k^M, \delta_k^F, \phi_k^M, \phi_k^F, \gamma_k^M, \gamma_k^F)$:

- $\psi_{k..} \mid \text{Rest} \propto N(\eta_k^1, (\xi_k^1)^2)$

- Draw $M_k = \frac{\rho_k}{1+\rho_k}$ (the fraction of males): $M_k \mid \text{Rest} \propto N(\eta_k^2, (\xi_k^2)^2)$. Draws made from $N_K(\eta^2, (\xi^2)^2)$ conditional on $\overline{M} = \rho/(1+\rho)$ where $\eta = (\eta_1^2, \ldots, \eta_K^2)'$, $\xi = \text{diag}((\xi_1^2)^2, \ldots, (\xi_K^2)^2)$ via the SWEEP operator (Goodnight 1979; Little and Rubin 1987).

- $\delta_k^S \mid \text{Rest} \propto (\delta_k^S)^{a^S-1}(1-\delta_k^S)^{b^S-1} exp(\eta_k^3)$

- $\phi_k^S \mid \text{Rest} \propto (\phi_k^S)^{c^S-1}(1-\phi_k^S)^{d^S-1} exp(\eta_k^4)$

- $\gamma^M \mid \text{Rest} \propto (\gamma^M)^{\alpha-1} exp\left\{-(\gamma^M/\beta) + \sum_k \eta_k^5\right\}$

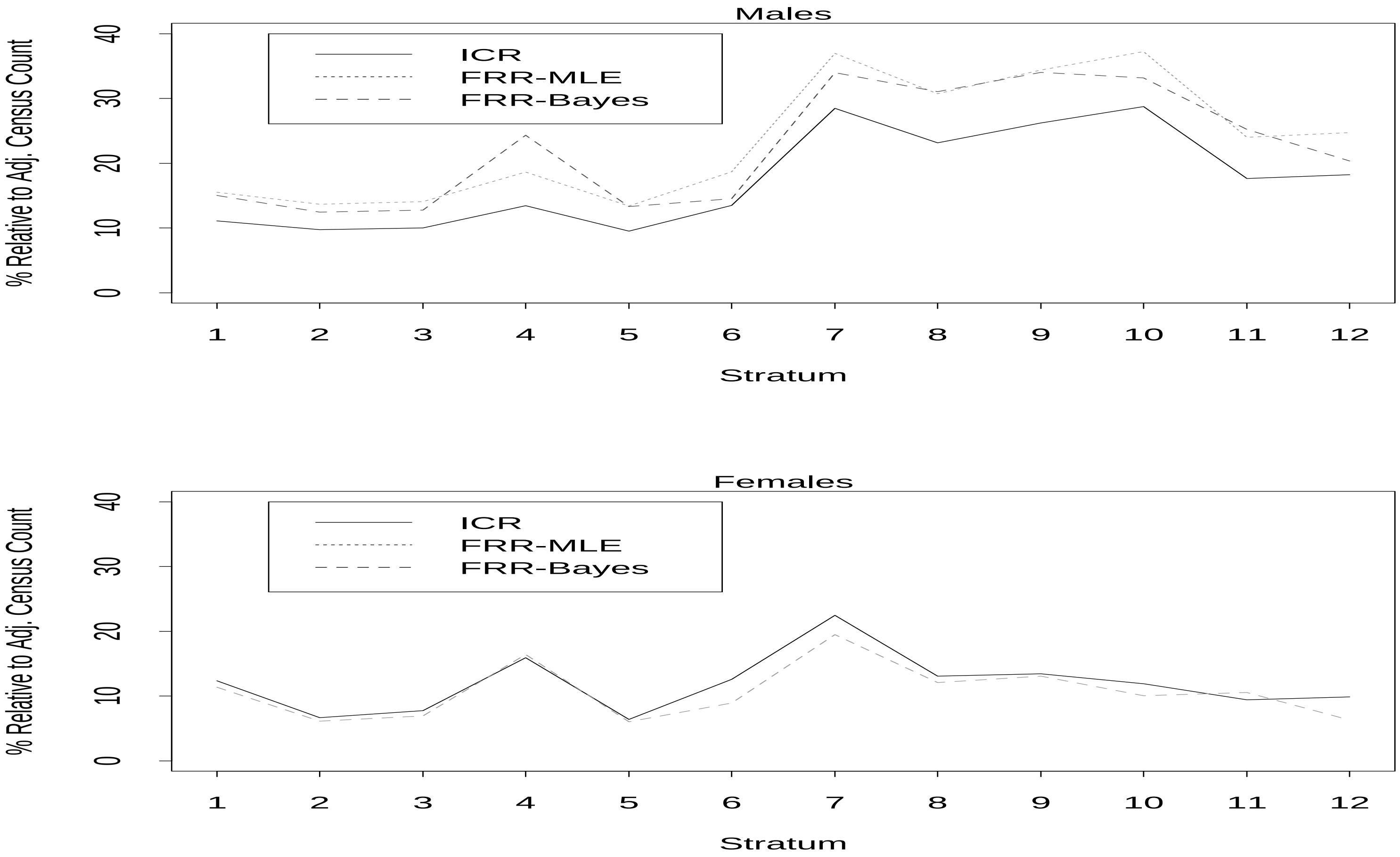
where η_k, ξ_k are linear functions of $\psi^*_{k\{-\theta_j\}}$ and data

Application: 1990 US Census, African-Americans 30-49

- $K = 24$: Gender $\times$ owner-rental residence $\times$ 6 geographic regions.
- Ran 5 chains of 1,000 draws from Gibbs sampler, discarded first 200.
- Gelman-Rubin statistic $\hat{R}$ (ratio of marginal posterior variance for all chains to mean within-chain posterior variance) lies between 1.00 and 1.07 (Gelman et al. 1995).

Undercount when compared against published 1990 US Census counts of African-Americans 30-49:

			Posterior Means
	Independence	MLE FRR	Bayes FRR (95% PPI)
Total	0.4%	4.1%	2.9% (0.6% - 5.2%)
Females	1.7%	1.7%	0.5% (-1.8% - 2.6%)
Males	-1.1%	6.7%	5.4% (3.2% - 7.5%)



Percent of Total Population Relative to Adjusted Census Counts ($z_{k1.}$) under the ICR, FRR MLE, and FRR Bayes Models, by Gender and Stratum.

Model fit: Sex Ratios

- MLE estimates of between-stratum sex ratio variability appear excessive.
- Prior distribution $\rho_k \sim N(\rho, \sigma^2)$ allows analyst to include prior information about variability of poststratum sex ratios through σ.
- $\sigma = 1$: nearly flat prior relative to likelihood
- $\sigma = .033$ ($P(\rho_k > 1)$=.002)

	Adj. Census Estimate	ICR MLE	FRR MLE	($\sigma = 1$) Bayes MLE	($\sigma = .033$) Bayes FRR
PS 1	0.826	0.815	0.858	0.863	0.890
PS 2	0.842	0.871	0.911	0.904	0.905
PS 3	0.845	0.866	0.907	0.901	0.902
PS 4	0.929	0.902	0.960	1.042	0.912
PS 5	0.874	0.904	0.945	0.948	0.931
PS 6	0.874	0.833	0.940	0.934	0.911
PS 7	0.668	0.725	0.822	0.817	0.900
PS 8	0.746	0.844	0.937	0.952	0.916
PS 9	0.623	0.731	0.822	0.822	0.890
PS 10	0.833	1.030	1.170	1.123	0.918
PS 11	0.716	0.788	0.853	0.857	0.893
PS 12	0.769	0.846	0.919	0.904	0.906
Total	0.782	0.836	0.906	0.906	0.906
$s(\times 10^{-2})$	9.18	8.24	9.20	8.49	1.23

Estimates of the ratio of males to females within each poststratum.

Model Fit: Posterior Predictive Distributions

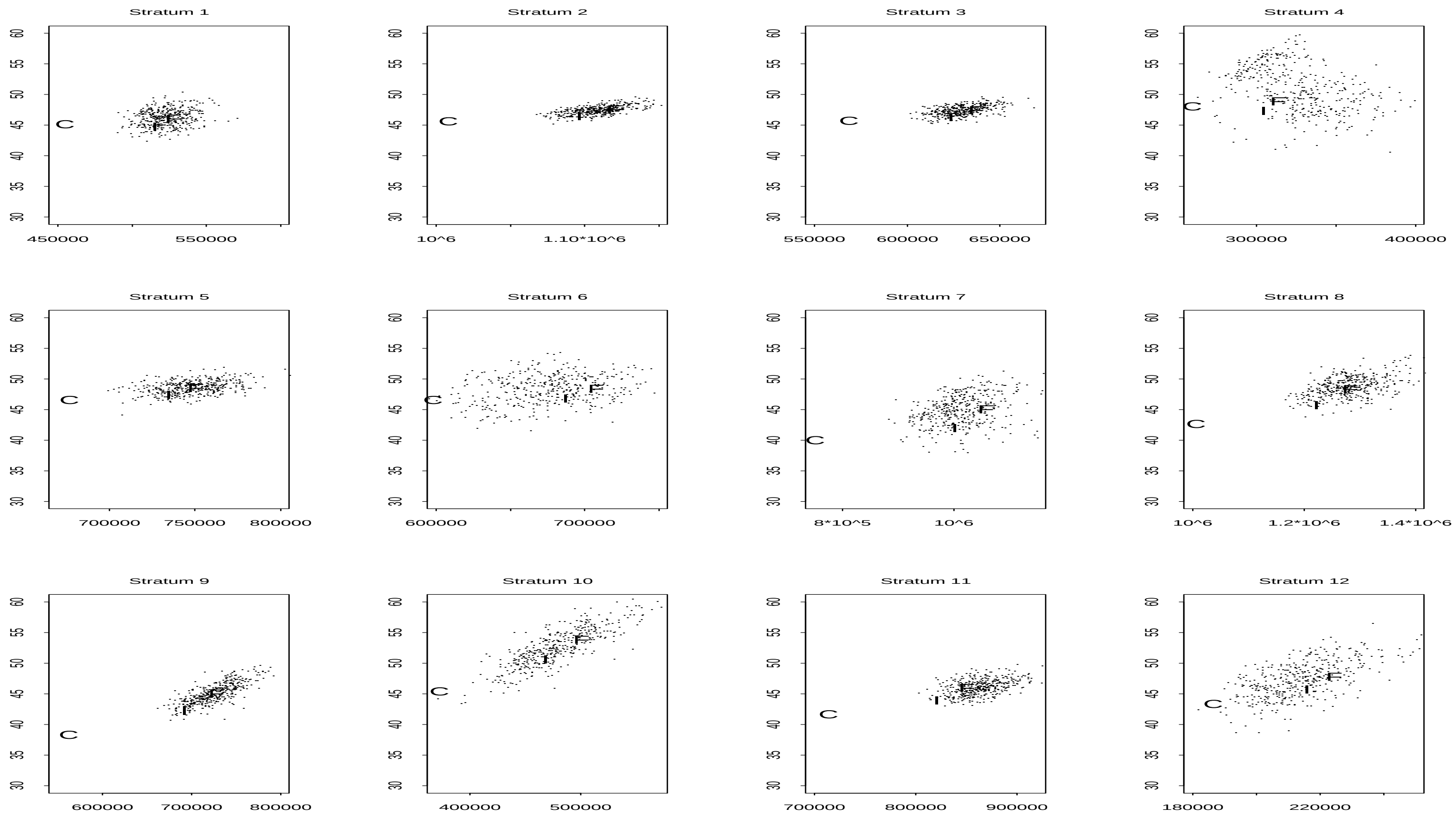
- By definition $\psi^S_{k11} < \psi^S_{k10}$.
- Persistent overstating of matches from the CMS or of erroneous enumerations from the Census could cause bias, so $E(y^S_{k11}) \neq \psi^S_{k11}$, $E(z^S_{k1.}) \neq \psi^S_{k1.}$.
- Large negative values of $y^S_{k11} - z^S_{k1.}$ could give evidence of this.
- Utilize posterior predictive values (Gelman et al. 1995; Gelman, Meng, and Stern 1996).

Classic p-values: Probability under the model that the observed statistics $T(y; \theta = \hat{\theta})$ will be less than (or greater than) the values of the statistic that would be seen in repeated observations: $P(T(y) \leq T(y^{rep}) \mid \hat{\theta})$.

PPD p-value: Probability that the observed statistic $T(y, \theta)$ is more extreme than replicated statistic, conditional on the observed data:
$P(T(y, \theta) \leq T(y^{rep}, \theta) \mid y)$.

1. Construct $(\psi^S_{k11})^{rep}$ and $(\psi^S_{k1.})^{rep}$ from the draws of $\rho_k, \delta^S_k, \phi^S_k$, and $\psi_{k..}$.
2. Draw $(y^S_{k11})^{rep}$ from $N((\psi^S_{k11})^{rep}, (\tau^S_{k11})^2)$ and $(z^S_{k1.})^{rep}$ from $N((\psi^S_{k1.})^{rep}, (u^S_{k1.})^2)$.
3. Compare $(z^S_{k1.})^{rep} - (y^S_{k11})^{rep}$ with $z^S_{k1.} - y^S_{k11}$.

- The only strata for which the observed data appear in the tail of the predictive distributions are poststratum 4 for both males and females: the PPD p-value for males is 0.070 and for females is 0.054.
- Large negative Out Census-In CMS cells are not necessarily evidence of bias.



Inference about Populations: Gibbs Sampling Draws of the Total Population Estimates with each Poststratum: Percent Male (y-axis) vs. Total (x-axis). Adjusted Census counts and ML estimates for the ICR and FRR models given by "C", "I", and "F" respectively.

Conclusions

I have proposed six general principles – PLAUSIBILITY, FIT, PREDICTION, INDEPENDENCE MODEL INCLUSION, STABILITY, and CONSERVATISM – to choose among saturated models that have been used to combine Census, CMS, and demographic information. I also proposed a Bayesian form of a fixed relative risk model that

- addresses the negative In-Census-Out CMS cell problem;
- provides smoothing of estimates across strata;
- provides inference about estimates;
- tests whether negative cells can be plausibly caused by sampling variability or whether they are due to other reasons; and
- can be extended to incorporate uncertainty in sex ratio estimates.