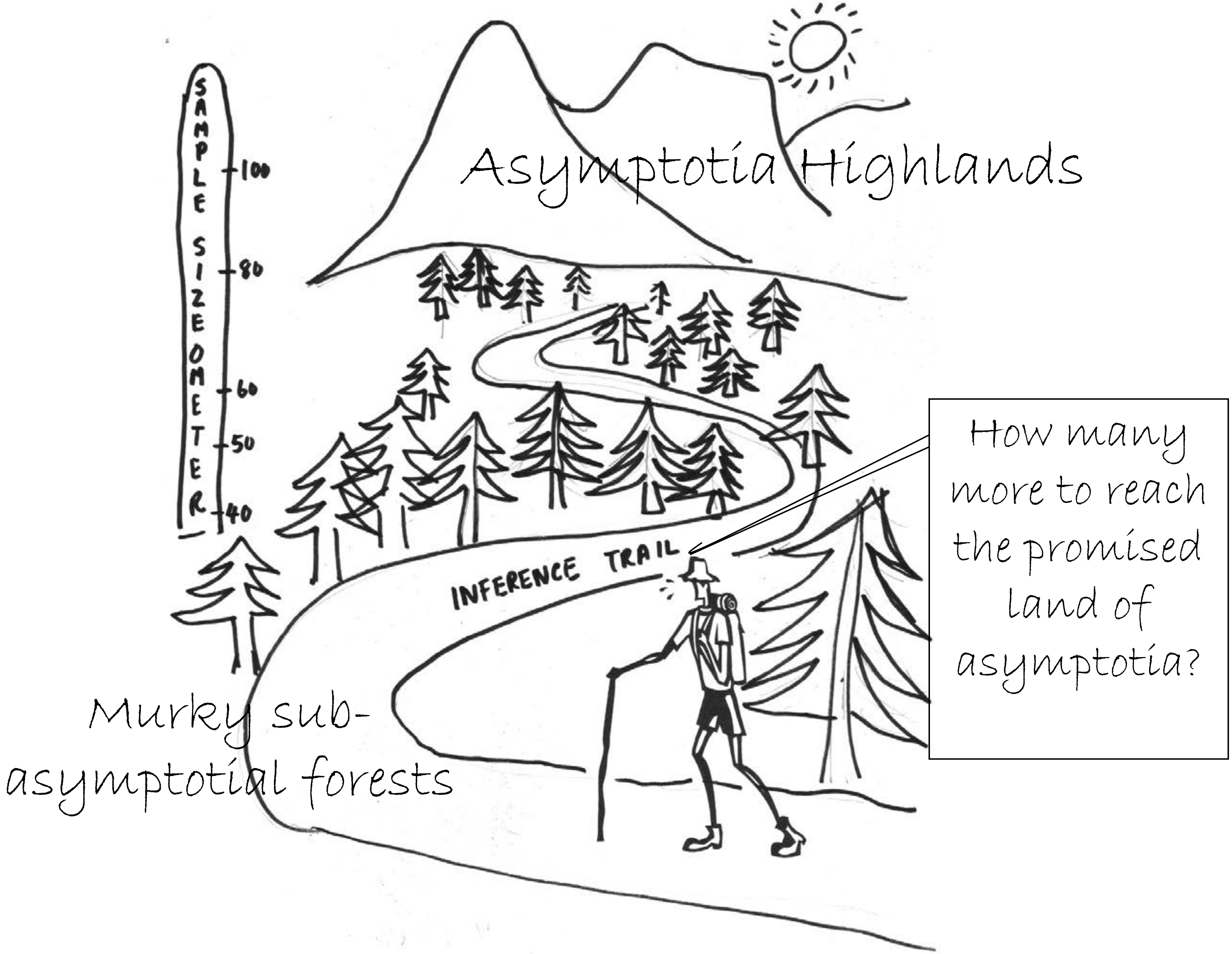
Bayesian Inference for Sample Surveys Module 12: Small Area Estimation

Roderick Little and T. Rathunathan

University of Michigan

Limitations of design-based approach

- Inference is based on probability sampling, but true probability samples are harder and harder to come by:
  - Noncontact, nonresponse is increasing
  - Face-to-face interviews increasingly expensive
  - Can't do "big data" (e.g. internet, administrative data) from the design-based perspective
- Theory is basically asymptotic -- limited tools for small samples, e.g. small area estimation



Design-based methods live in the land of asymptotia

The current "status quo" -- design-model compromise

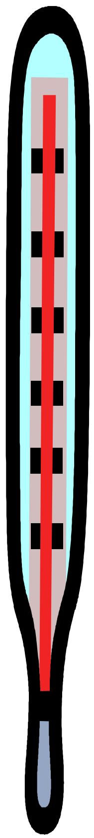
- Design-based for large samples, descriptive statistics
  - But may be *model assisted*, e.g. regression calibration:

$$\hat{T}_{\mathrm{GREG}} = \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{N} I_i (y_i - \hat{y}_i) / \pi_i, \hat{y}_i = \text{ model prediction}$$

  - model estimates adjusted to protect against misspecification, (e.g. Särndal, Swensson and Wretman 1992).
  - Can incorporate auxiliary information, but does not borrow strength for small areas
- Model-based for small area estimation, nonresponse, time series,…
- Attempts to capitalize on best features of both paradigms… but … at the expense of "inferential schizophrenia" (Little 2012)?

Example: when is an area “small”?

n
-
o
m
e
t
e
r



Design-based inference

--- ← n_0 = “Point of inferential schizophrenia”

Model-based inference

How do I choose n_0?

If $n_0 = 35$, should my entire statistical philosophy and inference be different when n=34 and n=36?

n=36, CI: [ ] (wider since based on direct estimate)

n=34, CI: [ ] (narrower since based on model)

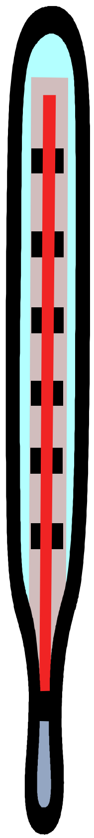
Multilevel (hierarchical Bayes) models

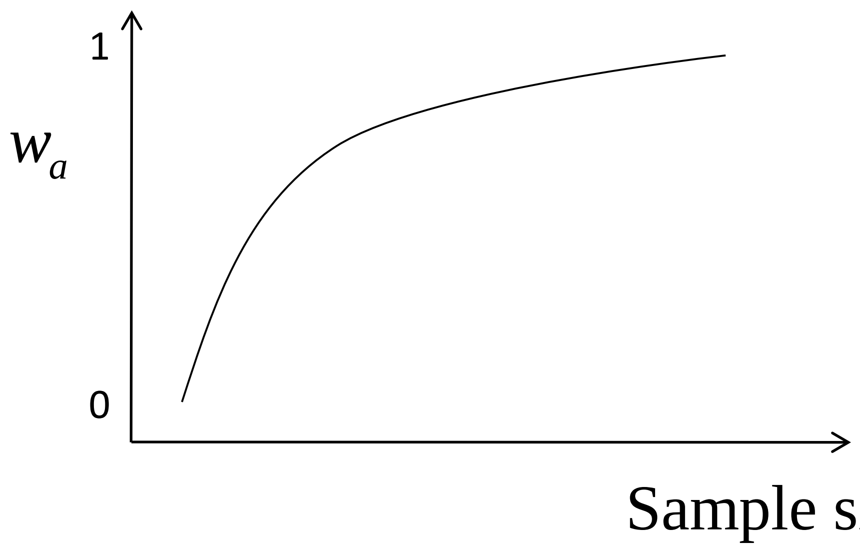
Model estimate

$$\tilde{\mu}_a = w_a \bar{y}_{\pi a} + (1 - w_a)\hat{\mu}_a$$

Direct estimate

n-ometer





Bayesian multilevel model estimates borrow strength increasingly from model as *n* decreases

A hierarchical Bayes model for small areas

- Fixed-effects models have distinct parameters (means, variances) for small areas, e.g.

$$y_{ai} \mid \mu_a, \sigma_a^2 \sim N(\mu_a, \sigma_a^2), \text{ for unit } i \text{ in area } a$$

- Hierarchical Bayes models assign distributions to the parameters for each area, e.g.

$$(y_{ai} \mid \mu_a, \sigma_a^2, \beta, \tau^2) \sim N(\mu_a, \sigma_a^2);$$

$$\Rightarrow (\bar{y}_a \mid \mu_a, \sigma_a^2, \beta, \tau^2) \sim N(\mu_a, \sigma_a^2 / n_a)$$

$$(\mu_a \mid \mu_a, \sigma_a^2, \beta, \tau^2) \sim N(\beta z_a, \tau^2)$$

z_a is a vector of covariates for area a, including constant term

Hierarchical Bayes Models for small areas

$$(\bar{y}_a \mid \mu_a) \sim N(\mu_a, \sigma_a^2 / n_a)$$

$$(\mu_a) \sim N(\beta z_a, \tau^2)$$

$$p(\bar{y}_a, \mu_a) \propto \exp -\tfrac{1}{2}\left[A(\bar{y}_a - \mu_a)^2 + B(\mu_a - \beta z_a)^2 \right]$$

$$\left(A = n_a / \sigma_a^2, B = 1/\tau^2 \right)$$

$$p(\bar{y}_a, \mu_a) \propto \exp -\tfrac{1}{2}(A+B)\left[(\mu_a - \tfrac{A\bar{y}_a + B\beta z_a}{A+B})^2 \right]$$

Hence $E(\mu_a \mid \bar{y}_a, \beta) = \frac{A\bar{y}_a + B\beta z_a}{A+B} = w_a \bar{y}_a + (1 - w_a)\beta z_a$

where weight on $\bar{y}_a$ is $w_a = A/(A+B) = \dfrac{n_a / \sigma_a^2}{n_a / \sigma_a^2 + 1/\tau^2}$

Proper prior ($\tau^2 < \infty$) on μ_a moves the direct area estimate $\bar{y}_a$ towards the model prediction βz_a; w_a increases with n_a

Integrating over posterior of β replaces β by its posterior mean

Hierarchical Bayes Models for small areas

Treatment of variances σ^2, τ^2:

Empirical Bayes: replaces them by estimates $\hat{\sigma}^2, \hat{\tau}^2$

(Maximum likelihood, or the simpler method of moments)

Bayes: assigns them dispersed prior distributions and integrates over their posterior distribution

(note that τ^2 cannot be assigned the Jeffreys' prior $p(\tau^2) \propto 1/\tau^2$)

Bayes approach is better, particularly if $\hat{\tau}^2 = 0$

A Basic Beta/Binomial small area model for binary outcomes

$n_a =$ count in area a

$m_a =$ count with $y = 1$ in area a

$m_a \mid p_a \sim \text{Bin}\left(n_a ; p_a\right)$

$p_a \sim \text{Beta}\left(\alpha, \beta\right)$ (conjugate prior distribution)

Prior mean of p_a: $E(p_a) = \dfrac{\alpha}{\alpha + \beta}$

A Basic Beta/Binomial small area model for binary outcomes

Posterior distribution of p_a is Beta:

$(p_a \mid n_a, m_a) \sim \text{Beta}\left(\alpha + m_a, \beta + n_a - m_a\right)$

Posterior mean of p_a :

$$E(p_a \mid n_a, m_a) = \frac{\alpha + m_a}{\alpha + \beta + n_a} = w_a \frac{m_a}{n_a} + (1 - w_a)\frac{\alpha}{\alpha + \beta}$$

where weight on sample proportion is $w_a = \dfrac{n_a}{n_a + \alpha + \beta}$

"Prior adds α successes and β failures to data"

Applications

- U.S. Census Bureau SAIPE (Small Area Income and Poverty Estimates) Program
- Voting Rights Act special tabulation
- The American Community Survey (ACS) and the “standard error error”

Example 1: SAIPE project

- Objective: Estimates of poverty for various age groups and median household income for all *states*, *counties*, and *school districts* in the U.S.

- Problem: Direct survey estimates (from Current Population Survey (CPS) or, later, American Community Survey (ACS) are too unreliable for many areas
  - CPS sample small for most states; no sample in ≈ 2/3 counties
  - ACS (single year) sample small for many counties and most school districts.

- Solution: Use small area model to integrate survey data with data from admin records (IRS, SNAP program) and previous census long form.

Fay-Herriot (1979) Model

(Hierarchical Bayesian Formulation)

$$y_i \mid \theta_i, v_i \sim N(\theta_i, v_i)$$

$$\theta_i \mid \beta, \sigma^2 \sim N(x_i'\beta, \sigma^2)$$

- y_i = direct survey estimate of population quantity θ_i for area i
- v_i = sampling variance of y_i (assumed known)
- x_i = vector of regression variables for area i
- β = vector of regression parameters
- σ^2 = variance of small area random effects

Example: State 5-17 poverty rate model

- Direct survey estimates y_i originally from CPS, but since 2005 from ACS
- Regression variables in x_i include a constant term and, for each state
  - Pseudo-poverty rate for children from tax return data
  - Tax “nonfiler rate”
  - SNAP (food stamp) participation rate
  - Previous census estimated state 5-17 poverty rate, or residuals from regressing previous census estimates on other elements of x_i for the census year.
  - Recent work explicitly Bayesian to propagate error in variance components

Posterior Variances from State Model for 2004 CPS 5-17 Poverty Rates

Results for four states

State	n_i	v_i	$\mathrm{Var}(Y_i \mid \text{data})$	approx. wt. on y_i in $E(Y_i \mid \text{data})$
CA	5,834	1.1	0.8	.61
NC	1,274	4.6	2.0	.28
IN	904	8.1	2.0	.18
MS	755	12.0	3.9	.13

Example 2: Voting Rights Tabulations

- Section 203 Language Provisions of the Voting Rights Act
- Determines counties and townships required to provide language assistance at the polls
- Determinations are based <u>in part</u> on the following "more than 5%" provision:

  … More than 5 percent of voting age citizens of political district are members of a single language minority and are LEP.

Voting Rights Tabulations

- Previously used direct estimates from Long Form Decennial Census Data
- Use ACS 2005-2009 and 2010 Census data to produce a Federal Register Notice by mid-summer 2011
- Direct estimates for some districts are based on small ACS sample and hence have unacceptably high variance
- E.g. let P be proportion of voting age citizens in political district who are members of a single language minority and are LEP
- Suppose ACS was a simple random sample, a direct estimate of P is the sample proportion m/n
  - District A with n=105, m=5, $m/n < 0.05$
  - District B with n=105, m=6, $m/n > 0.05$
  - Direct ACS estimation is more complex, but same idea applies

Voting Rights Tabulations

- Alternative approach to the “more than 5%” provision:
- Build a district level regression model to predict *P* based on variables in the ACS
- Classify districts into classes with similar predicted *P* based on the model [predictive mean stratification]
- Within classes, apply a hierarchical random-effects model that pulls the direct ACS estimate of *P* towards the average *P* for districts in that class
- Compare HRE model estimate with 5% for this aspect of the determination
- Rationale: increased precision of HRE estimates in small samples increases the probability of getting the determination right, particularly in small districts

Example 3: ACS tabulations

- American Fact Finder gives users access to a dazzling array of ACS tables, for small areas
- The move to make more data available is highly commendable, but the methodology remains design-based and assumes large samples.
- Need for methods appropriate for small samples…

B01001A. SEX BY AGE (WHITE ALONE) -
Universe: WHITE ALONE POPULATION
Data Set: 2005-2009 American
Community Survey 5-Year Estimates
Survey: American Community Survey

90% CI = Estimate
+/- Margin of Error

Northfield township, Washtenaw County, Michigan		
	Estimate	Margin of Error
	8,062	+/-210
Male:	4,164	+/-239
Under 5 years	321	+/-108
5 to 9 years	239	+/-90
10 to 14 years	342	+/-171
15 to 17 years	151	+/-57
18 and 19 years	14	+/-21
20 to 24 years	332	+/-175
…	…	…

90% CI = Estimate
+/- Margin of Error

Oops!

Northfield township, Washtenaw County, Michigan		
	Estimate	Margin of Error
	8,062	+/-210
Male:	4,164	+/-239
Under 5 years	321	+/-108
5 to 9 years	239	+/-90
10 to 14 years	342	+/-171
15 to 17 years	151	+/-57
18 and 19 years	14	+/-21
20 to 24 years	332	+/-175
…	…	…

Example: ACS tabulations

B01001B. SEX BY AGE (BLACK OR AFRICAN AMERICAN ALONE) - Universe: BLACK OR AFRICAN AMERICAN ALONE POPULATION
Data Set: 2005-2009 American Community Survey 5-Year Estimates
Survey: American Community Survey

90% CI = Estimate
+/- Margin of Error

(1) Margin of Error gives a rough idea of uncertainty, but …
(2) Intervals often contains negative values
(3) Truncated to be positive, CI still does not have stated coverage
(4) Simple fixes for SRS, less simple for complex designs

Northfield township, Washtenaw County, Michigan		
	Estimate	Margin of Error
Total:	43	+/-41
Male:	28	+/-32
Under 5 years	0	+/-109
5 to 9 years	0	+/-109
10 to 14 years	0	+/-109
15 to 17 years	0	+/-109
…	…	…
35 to 44 years	0	+/-109
45 to 54 years	28	+/-32
55 to 64 years	0	+/-109

Imagine a better table…

90% CI =(LL, UL)
LL, UL based on Bayesian model

…but more complex… and billions of cells!

Northfield township, Washtenaw County, Michigan			
	LL	Estimate	UL
Total:	0	?	?
Male:	0	?	?
Under 5 years	0	?	?
5 to 9 years	0	?	?
10 to 14 years	0	?	?
15 to 17 years	0	?	?
…	0	?	?
35 to 44 years	0	?	?
45 to 54 years	0	?	?
55 to 64 years	0	?	?

American Community Survey

- US Census Bureau is making available thousands of ACS tables, with millions of cells
- A high fraction of these estimates are based on very little data, and hence are very noisy
  - Many people want information, not data, so ACS should produce information products, as well as data products
  - When noise swamps the signal, the information content is buried
  - Data products are highly constrained by confidentiality requirements, leading to incompleteness

The Statistical Problem

- The ACS philosophy is essentially to produce “direct” (“design-based”) estimates, together with margins of error
- This works fine with large samples, but most of the ACS estimates are based on small samples
  – The estimates are often too noisy to be useful
  – The confidence intervals derived from the estimates and margins of error are known to be of poor quality, violating statistical standards
    - Intervals include proportions outside the range (0,1)
    - Intervals do not have nominal coverage

The “standard error” error

- ACS reports estimates and margins of error that yield asymptotic 90% confidence intervals
- But in small samples, the implied confidence intervals do not have the stated coverage; so
- Calibrated Bayes: Seek to replaces estimates and margins of error by posterior means and 5% to 95% credibility intervals that have the approximately the nominal coverage
  - A non-Bayesian can interpret the posterior means as estimates, and the 90% credibility intervals as 90% confidence intervals.

Pragmatic “pseudo-Bayes” approach

A fully Bayesian hierarchical model for proportions is feasible but beyond current Bureau of Census capabilities

Tom Louis suggested a simple “Bayes-like” approach, which “gets the Calibrated Bayes foot in the door” (my words)

Pragmatic “pseudo-Bayes” approach

A. Compute design-based estimate of proportion and standard error using existing design-based methods

B. Pretend data are binomial with number of successes m_a* and sample size n_a* that lead to the estimates in A.

C. Compute Beta posterior distribution with noninformative prior (e.g. uniform or Jeffreys)

D. Compute 90% posterior credibility interval based on this Beta posterior (reflects asymmetry, always between 0 and 1)

Simple to implement and easily beats standard Wald-type confidence intervals in simulations (Franco, Little, Louis and Slud 2014)

Approximate Beta posterior distribution for complex designs

SRS: Posterior distribution of p_a is Beta:

$$(p_a \mid n_a, m_a) \sim \text{Beta}\left(\alpha + m_a, \beta + n_a - m_a\right)$$

Complex Design: estimate $\hat{p}_a$, SE $\hat{s}_a$ using a design-based approach

$$\hat{s}_a^2 = \frac{\hat{p}_a(1-\hat{p}_a)}{n_a^*},\ n_a^* = \text{effective sample size; } n_a / n_a^* = \text{design effect}$$

Hence define effective ss and count: $n_a^* = \dfrac{\hat{p}_a(1-\hat{p}_a)}{\hat{s}_a^2}$, $m_a^* = n_a^* \hat{p}_a$,

Approximate posterior distribution as

$$(p_a \mid n_a^*, m_a^*) \sim \text{Beta}\left(\alpha + m_a^*, \beta + n_a^* - m_a^*\right)$$

References

Franco, C., Little, R., Louis, T. and Slud, E. (2014). Coverage Properties of Confidence Intervals for Proportions in Complex Sample Surveys . *ASA Proc Survey Research Methods Section*.

Joyce, P.M., Malec, D., Little, R.J., Gilary, A., Navarro, A. and Asiala, M.E. (2014). Statistical Modeling Methodology for the Voting Rights Act Section 203 Language Assistance Determinations. *JASA,* 109, 36-47*.*

Little, R.J. (2012). Calibrated Bayes: an alternative inferential paradigm for official statistics (with discussion and rejoinder). *JOS*, 28, 3, 309-372*.*

Särndal, C.-E., Swensson, B. & Wretman, J.H. (1992), *Model Assisted Survey Sampling*, Springer Verlag: New York.