

# Bayesian inference for sample surveys

Roderick Little and Trivellore Raghunathan

Module 9: Bayesian models for  
stratified sample designs



# Modeling sample selection

- Role of sample design in model-based (Bayesian) inference
- Key to understanding the role is to include the sample selection process as part of the model
- Modeling the sample selection process
  - Simple and stratified random sampling
  - Cluster sampling, other mechanisms
  - See Chapter 7 of *Bayesian Data Analysis* (Gelman, Carlin, Stern and Rubin 1995)

# General set-up: models that include data collection

$Y = (y_1, \dots, y_N)$  = population data;  $y_i$  may be a vector

$Z$  = fully-observed covariates, design variables

$Q = Q(Y, Z)$  = finite population quantity

$I = (I_1, \dots, I_N)$  = Sample Inclusion Indicators

$$I_i = \begin{cases} 1, & y_i \text{ observed} \\ 0, & \text{otherwise} \end{cases}$$

$Y = (Y_{\text{inc}}, Y_{\text{exc}})$

$Y_{\text{inc}}$  = included part of  $Y$ ,  $Y_{\text{exc}}$  = excluded part of  $Y$

# Full model for $Y$ and $I$

$$p(Y, I | Z, \theta, \phi) = p(Y | Z, \theta) p(I | Y, Z, \phi)$$

Model for  
Population

Model for  
Inclusion

- Observed data:  $(Y_{\text{inc}}, Z, I)$  (No missing values)

- Observed-data likelihood:

$$L(\theta, \phi | Y_{\text{inc}}, Z, I) \propto p(Y_{\text{inc}}, I | Z, \theta, \phi) = \int p(Y, I | Z, \theta, \phi) dY_{\text{exc}}$$

- Posterior distribution of parameters:

$$p(\theta, \phi | Y_{\text{inc}}, Z, I) \propto p(\theta, \phi | Z) L(\theta, \phi | Y_{\text{inc}}, Z, I)$$

# Ignoring the data collection process

- The likelihood *ignoring the data-collection process* is based on the model for  $Y$  alone with likelihood:

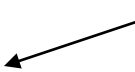
$$L(\theta | Y_{\text{inc}}, Z) \propto p(Y_{\text{inc}} | Z, \theta) = \int p(Y | Z, \theta) dY_{\text{exc}}$$

- The corresponding posteriors for  $\theta$  and  $Y_{\text{exc}}$  are:

$$p(\theta | Y_{\text{inc}}, Z) \propto p(\theta | Z) L(\theta | Y_{\text{inc}}, Z)$$

$$p(Y_{\text{exc}} | Y_{\text{inc}}, Z) \propto \int p(Y_{\text{exc}} | Y_{\text{inc}}, Z, \theta) p(\theta | Y_{\text{inc}}, Z) d\theta$$

Posterior predictive  
distribution of  $Y_{\text{exc}}$



- When the full posterior reduces to this simpler posterior, the data collection mechanism is called *ignorable* for Bayesian inference about  $\theta, Y_{\text{exc}}$ .

# Conditions when data collection mechanism can be ignored

- Two general and simple sufficient conditions for ignoring the data-collection mechanism are:

Selection at Random (SAR):

$$p(I | Y, Z, \phi) = p(I | Y_{\text{inc}}, Z, \phi) \text{ for all } Y_{\text{exc}}.$$

Bayesian Distinctness:

$$p(\theta, \phi | Z) = p(\theta | Z) p(\phi | Z)$$

- It is easy to show that these conditions together imply that:

$$p(\theta, Y_{\text{exc}} | Y_{\text{inc}}, Z) = p(\theta, Y_{\text{exc}} | Y_{\text{inc}}, Z, I)$$

so the model for the data-collection mechanism does not affect inferences about the parameter  $\theta$  or finite population quantities  $Q$ .

# Bayes inference for probability samples

- In probability sampling designs, selection does not depend on values of  $Y$  and the mechanism is known, that is:

$$p(I | Y, Z, \phi) = p(I | Z) \text{ for all } Y.$$

- This means that the data-collection mechanism is ignorable for Bayesian inference (with complete data)
- But the model needs to appropriately account for relationship of survey variables  $Y$  with the design variables  $Z$ .

# Stratified and PPS samples

- For **stratified samples**,  $Z$  consists of stratum indicators, so models for  $Y$  need to include stratum indicators as covariates
  - Same selection fraction across strata yields epsem design
  - Different selection fractions across strata yields unequal probability design – sampling weights are the inverse of selection fractions
- For **PPS sampling**,  $Z$  is the size variable, and models for  $Y$  need to include size as a covariate
  - Sampling weight is then proportional to *inverse* of  $Z$
- In either case, other auxiliary variables can be included, but correctly modeling the relationship between  $Y$  and  $Z$  is particularly important to avoid bad inferences because of model misspecification



# Design-based weighting

- A pure form of **design-based** estimation is to **weight** sampled units by inverse of inclusion probabilities
  - Sampled unit  $i$  “represents”  $w_i = 1 / \pi_i$  units in the  $\pi_i$  population
- More generally, a common approach is:
$$w_i = w_{is} \times w_{in} \times w_{ip}$$
$$w_{is} = \text{sampling weight}$$
$$w_{in} = \text{nonresponse weight}$$
$$w_{ip} = \text{post-stratification weight}$$
- We’ll compare Bayesian (predictive) inference with weighting

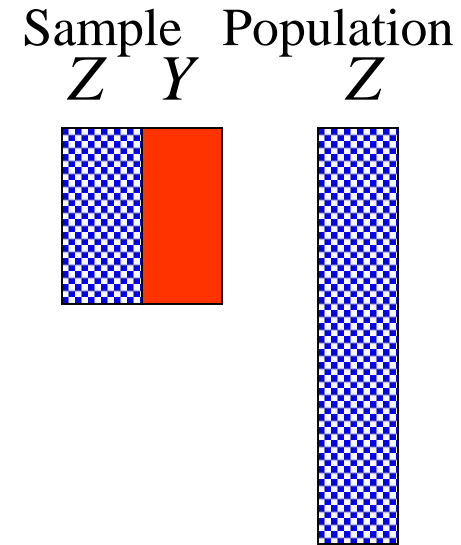
# Weighting and models

- The weights can't generally be ignored from a modeling perspective
  - Ignores different selection effects that bias estimates
- Weights are auxiliary covariates from a modeling perspective
- Design: weight the respondents
  - Simple: same weights for all  $Y$  variables, but:
  - Weighting adds noise for  $Y$ 's unrelated to weights
- Model: use weights as covariates to predict non-sampled and non-responding values
  - More flexible, but need a good model

# Ex 1: stratified random sampling

- Population is divided into  $J$  strata
- Simple random sample of  $n_j$  units selected from population of  $N_j$  units in stratum  $j$ .
- $Z$  is a variable indicating stratum:

$$z_i = j, \text{ if unit } i \text{ is in stratum } j \ (j = 1, \dots, J)$$



- In a regression model,  $Z$  is represented by a set of  $J - 1$  binary indicators for stratum, the stratum left out being the reference stratum (dummy variable regression)
- This design is ignorable *providing* model for survey variable  $Y$  conditions on the stratum indicators  $Z$ .

# Inference for a mean from a stratified sample

- A normal model that includes stratum effects is:

$$[y_i | z_i = j] \sim_{\text{ind}} N(\theta_j, \sigma_j^2)$$

- For simplicity assume  $\sigma_j^2$  is known and the flat prior:

$$p(\theta_j | Z) \propto \text{const.}$$

- Standard Bayesian calculations lead to

$$[\bar{Y} | Y_{\text{inc}}, Z, \{\sigma_j^2\}] \sim N(\bar{y}_{\text{st}}, \sigma_{\text{st}}^2)$$

where:

$$\bar{y}_{\text{st}} = \sum_{j=1}^J P_j \bar{y}_j, P_j = N_j / N, \bar{y}_j = \text{sample mean in stratum } j,$$

$$\sigma_{\text{st}}^2 = \sum_{j=1}^J P_j^2 (1 - f_j) \sigma_j^2 / n_j, f_j = n_j / N_j$$

# Bayes for stratified normal model

- Bayes inference for this model is equivalent to standard classical inference for the population mean from a stratified random sample
- The posterior mean weights case by inverse of inclusion probability:

$$\bar{y}_{\text{st}} = N^{-1} \sum_{j=1}^J N_j \bar{y}_j = N^{-1} \sum_{j=1}^J \sum_{i: x_i=j} y_i / \pi_j,$$

where  $\pi_j = n_j / N_j =$  selection probability in stratum  $j$ .

- With unknown variances, Bayes' for this model with flat prior on log(variances) yields useful t-like corrections for small samples (See module 7)

# Suppose we ignore stratum effects?

- Suppose we assume instead that:

$$[y_i | z_i = j] \sim_{ind} N(\theta, \sigma^2),$$

the previous model with no stratum effects.

- With a flat prior on the mean, the posterior mean of  $\bar{Y}$  is then the unweighted mean

$$E(\bar{Y} | Y_{inc}, Z, \sigma^2) = \bar{y} \equiv \sum_{j=1}^J p_j \bar{y}_j, p_j = n_j / n$$

- This is potentially a very biased estimator if the selection rates  $\pi_j = n_j / N_j$  vary across the strata
  - The problem is that results from this model are highly sensitive to violations of the assumption of no stratum effects ... and stratum effects are likely in most realistic settings.
  - Hence prudence dictates a model that allows for stratum effects, such as the model in the previous slide.

# Design consistency

- Loosely speaking, an estimator is *design-consistent* if (irrespective of the truth of the model) it converges to the true population quantity as the sample size increases, holding design features constant.
- For stratified sampling, the posterior mean  $\bar{y}_{st}$  based on the stratified normal model converges to  $\bar{Y}$ , and hence is design-consistent
- For the normal model that ignores stratum effects, the posterior mean  $\bar{y}$  converges to

$$\bar{Y}_{\pi} = \sum_{j=1}^J \pi_j N_j \bar{Y}_j / \sum_{j=1}^J \pi_j N_j$$

and hence is not design consistent unless  $\pi_j = \text{const.}$

- We generally advocate Bayesian models that yield design-consistent estimates, to limit effects of model misspecification

## Ex 2: PPS sampling

- In certain applications, it is efficient to sample “large” units (firms, tax returns, transactions in an audit...) with higher probability than “small” units – in particular when variability of outcome increases with size (as with variables like total sales, number of employees, ...)
- For a continuous stratifying size variable  $Z$ , this is conveniently achieved by probability proportional to size (pps) sampling
- Units in the population are first ordered, either randomly or by values of  $Z$ . Then:



# PPS sampling

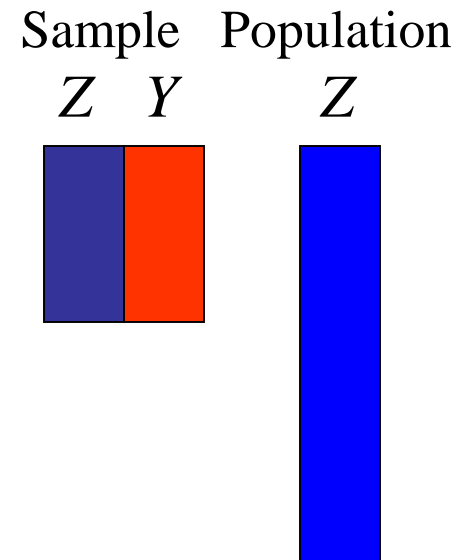
- Associate unit  $i$  with interval  $(c_{i-1}, c_i)$ , where  $c_0 = 0$ ,  $c_i = z_1 + \dots + z_i$  are cumulated sizes up to  $i$ ,  $i = 1, \dots, n$ .
- Choose a sampling interval  $I = z_n/n$ .
- Choose a random start between 0 and  $I$ , say  $x$
- Units corresponding to the intervals that contain the values  $x, x+I, x+2I, \dots, x+(n-1)I$  are sampled
- Notes:
  - Units with size greater than  $I$  are selected with probability 1. They are pre-selected and removed from the list prior to sampling from the list
  - With units randomly ordered, creates a pps sample with no implicit stratification
  - With units sorted by size, creates a pps sample with implicit stratification on size, and  $n$  implicit strata of size 1. More efficient, but sampling variance requires models

## Ex 2. PPS sampling, $Z = \text{size}$

Consider PPS sampling,  $Z = \text{measure of size}$

Standard design-based estimator is weighted Horvitz-Thompson estimate

$$\bar{y}_{\text{HT}} = \frac{1}{N} \left( \sum_{i=1}^n y_i / \pi_i \right); \pi_i = \text{selection prob (HT)}$$



Question: is there a model for  $Y$  for which the predictions yield the HT estimate of the mean?

An alternative to HT: Hajek:  $\bar{y}_{\text{HK}} = \frac{1}{\hat{N}} \sum_{i=1}^n (y_i / \pi_i), \hat{N} = \sum_{i=1}^n (1 / \pi_i)$

( $\bar{y}_{\text{HK}} = \bar{y}_{\text{HT}}$  when  $\hat{N} = N$ )

Question: when  $\bar{Y}_{\text{HK}} \neq \bar{Y}_{\text{HT}}$ , which is better? More on this later...

# Projection vs Prediction

Sample  $i = 1, \dots, n$ , non-sample  $i = n + 1, \dots, N$

$\bar{Y}$  = population mean  $= \sum_{i=1}^n y_i / N$

$\hat{y}_i$  = prediction of  $y_i$  from a model

prediction estimator:  $\bar{Y}_{\text{pred}} = \left( \sum_{i=1}^N y_i + \sum_{i=n+1}^N \hat{y}_i \right) / N$

projection estimator:  $\bar{Y}_{\text{proj}} = \sum_{i=1}^N \hat{y}_i / N = \bar{Y}_{\text{pred}} + \frac{n}{N} \sum_{i=1}^n (y_i - \hat{y}_i) / n$

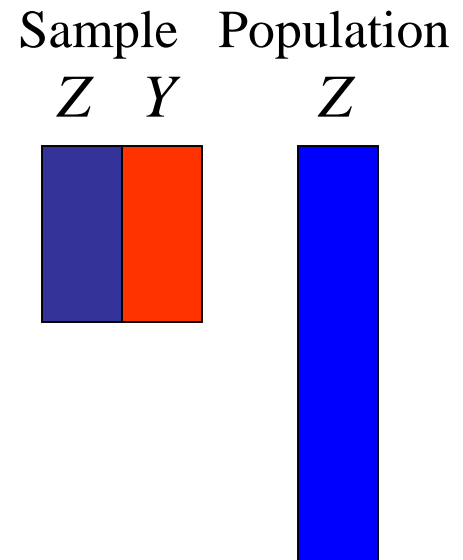
Similar, particularly if  $n \ll N$

## Ex 2. PPS sampling, $Z = \text{size}$

$y_i \sim \text{Nor}(\beta\pi_i, \sigma^2\pi_i^2)$  ("HT model")

$r_i = y_i / \pi_i \sim \text{Nor}(\beta, \sigma^2)$ , so

$\hat{\beta} = \bar{r} = \frac{1}{n} \sum_{i=1}^n (y_i / \pi_i)$ , yielding prediction  $\hat{y}_j = \hat{\beta}\pi_j$



$$\begin{aligned} \bar{Y}_{\text{proj}} &= \frac{1}{N} \sum_{j=1}^N \hat{y}_j = \frac{\hat{\beta}}{N} \sum_{j=1}^N \pi_j = \frac{1}{Nn} \sum_{i=1}^n (y_i / \pi_i) \sum_{j=1}^N \pi_j \\ &= \frac{1}{N} \sum_{i=1}^n (y_i / \pi_i), \left( \text{since } \sum_{j=1}^N \pi_j = n \right) = \bar{y}_{\text{HT}} \end{aligned}$$

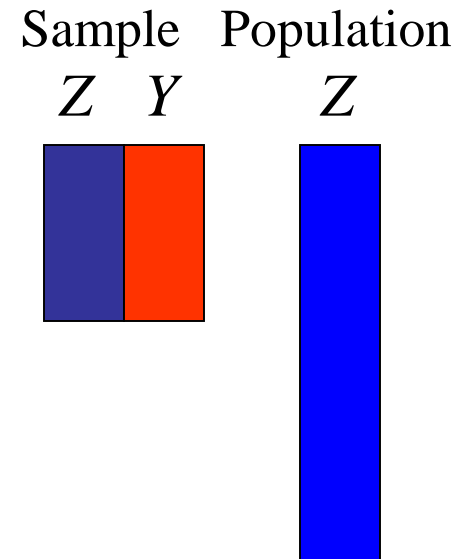
That is, the HT estimator is the projection estimator under the HT model.

## Ex 2. PPS sampling, $Z = \text{size}$

Implication:

When the relationship between  $Y$  and  $Z$  is well described by the HT model, the HT estimate performs well

When the relationship between  $Y$  and  $Z$  deviates a lot from the HT model, the HT estimate is inefficient and CI's can have poor coverage



# Ex. Basu's inefficient elephants

$(y_1, \dots, y_{50}) =$  weights of  $N = 50$  elephants

Objective:  $T = y_1 + y_2 + \dots + y_{50}$ . Only one elephant can be weighed!

- Circus trainer wants to choose “average” elephant (Sambo)
- Circus statistician requires “scientific” prob. sampling:

Select Sambo with probability 99/100

One of other elephants with probability 1/4900

Sambo gets selected! Trainer:  $\hat{t} = y_{(\text{Sambo})} \times 50$

Statistician requires unbiased Horvitz-Thompson (1952)

estimator:

$$\hat{T}_{HT} = \begin{cases} y_{(\text{Sambo})} / 0.99 (!!); \\ 4900 y_{(i)}, \text{ if Sambo not chosen (!!!)} \end{cases}$$

HT estimator is unbiased on average but always crazy!

HT model is clearly hopeless here ...

# What went wrong?

- HT estimator optimal under an implicit HT model that  $y_i / \pi_i$  have the same distribution
- That is clearly a bad model given this design ...
- Which is why the estimator is silly

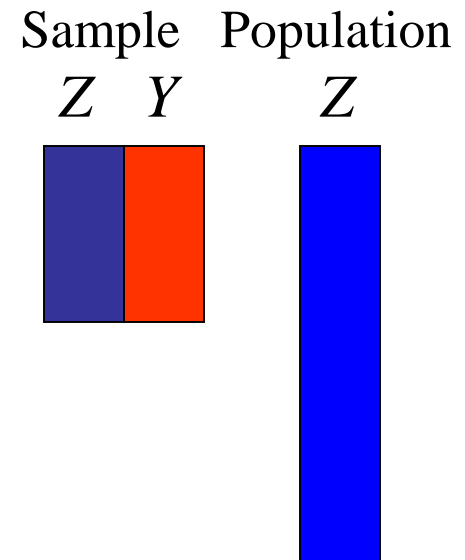
## Ex 2. PPS Sampling, $Z = \text{size}$

$$\bar{y}_{\text{HT}} = \frac{1}{N} \left( \sum_{i=1}^n y_i / \pi_i \right); \pi_i = \text{selection prob (HT)}$$

A modeling alternative to  $\bar{y}_{\text{HT}}$  is  
the prediction estimate

$$\bar{y}_{\text{pred}} = \frac{1}{N} \left( \sum_{i=1}^n y_i + \sum_{i=n+1}^N \hat{y}_i \right)$$

from a more flexible model relating  $Y$  to  $Z$



Zheng and Little (2004, 2005) fit a penalized spline model, and show superior performance to HT in simulations



# Making the HT model more flexible

Mean of HT model:  $E(y_i | z_i) = \beta z_i$  (linear through origin)

A. Polynomial regression:  $E(y_i | z_i) = \beta_0 + \beta_1 z_i + \dots + \beta_k z_i^k$

B. Model mean of  $Y$  as a smooth flexible function of  $Z$

e.g. Penalized Spline (Ruppert and Carroll, 2000) with linear basis:

Set  $m$  knots  $\kappa_1, \dots, \kappa_m$  at known values of  $Z$

(e.g. equally-spaced percentiles of distribution of  $Z$ )

$$E(y_i | z_i) = \beta_0 + \beta_1 z_i + \sum_{j=1}^m \alpha_j (z_i - \kappa_j)_+,$$

$$u_+ = \begin{cases} u, & \text{if } u > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\alpha_j \stackrel{\text{iid}}{\sim} N(0, \tau^2), j = 1, \dots, m.$$

This is a linear mixed model:

$\beta_0, \beta_1$  are *fixed* effects

$\alpha_1, \dots, \alpha_m$  are *random* effects

# Making the HT model more flexible

Variance of HT model:  $\text{Var}(y_i | z_i) = \sigma^2 z_i^2$

Replace by  $\text{Var}(y_i | z_i) = \sigma^2 z_i^k$ ,

$k$  an unknown parameter to be estimated

# Fully Bayes model specification

Set  $m$  knots  $\kappa_1, \dots, \kappa_m$  at known values of  $Z$

$$(y_i \mid z_i, \{\alpha_j\}, \beta_0, \beta_1, \sigma^2, \tau^2, k)$$

$$\sim_{\text{iid}} N\left(\beta_0 + \beta_1 z_i + \sum_{j=1}^m \alpha_j (z_i - \kappa_j)_+, \sigma^2 z_i^k\right)$$

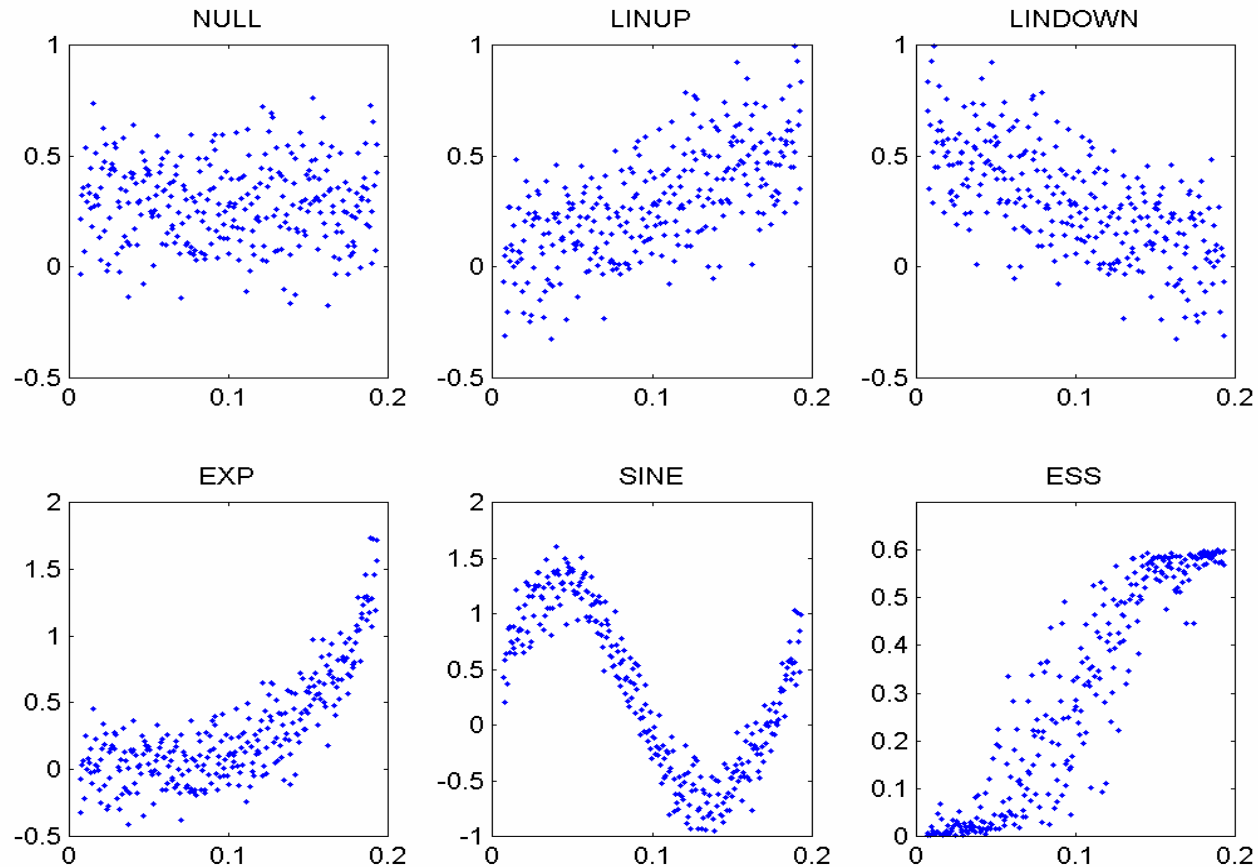
$$\alpha_j \mid \tau^2 \underset{\text{iid}}{\sim} N(0, \tau^2), j = 1, \dots, m.$$

Priors:

$$\pi(\beta_0, \beta_1, \sigma^2, \tau^2) = \text{const.} \sigma^{-2}$$

$$\pi(k \mid \beta_0, \beta_1, \sigma^2, \tau^2) = 1/4 \quad (-2 \leq k \leq 2)$$

# Simulation: PPS sampling in 6 populations



# Estimated RMSE of four estimators for N=1000, n=100

Population		model	wt	gr
NULL	Normal	<b>20</b>	33	21
	Lognormal	32	44	<b>31</b>
LINUP	Normal	<b>23</b>	24	25
	Lognormal	<b>25</b>	30	30
LINDOWN	Normal	30	66	<b>29</b>
	Lognormal	<b>24</b>	65	28
SINE	Normal	<b>35</b>	134	90
	Lognormal	<b>53</b>	130	84
EXP	Normal	<b>26</b>	32	57
	Lognormal	<b>40</b>	41	58

# 95% CI coverages: HT

Population	V1	V3	V4	V5
NULL	90.2	91.4	90.0	90.4
LINUP	94.0	95.0	95.0	95.0
LINDOWN	89.0	89.8	90.0	90.6
SINE	93.2	93.4	93.0	93.0
EXP	93.6	94.6	95.0	95.0
ESS	95.0	95.6	95.4	95.2

V1 Yates-Grundy, Hartley-Rao for joint inclusion probs.

V3 Treating sample as if it were drawn with replacement

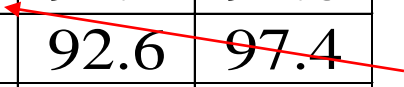
V4 Pairing consecutive strata

V5 Estimation using consecutive differences

# 95% CI coverages: B-spline

Population	V1	V2	V3
NULL	95.4	95.8	95.8
LINUP	94.8	97.0	94.6
LINDOWN	94.2	94.2	94.6
SINE	<b>88.0</b>	92.6	97.4
EXP	94.4	95.2	95.6
ESS	97.4	95.4	95.8

Fixed with  
more knots



V1 Model-based (information matrix)  
V2 Jackknife  
V3 BRR

# Why does spline model do better?

- Assumes smooth relationship – HT weights can “bounce around”
- Predictions use sizes of the non-sampled cases
  - HT estimator does not use these
  - Often not provided to users (although they could be)
- Little & Zheng (2007) also show gains for model when sizes of non-sampled units are not known
  - Predicted using a Bayesian Bootstrap (BB) model
  - BB is a form of stochastic weighting



# Hajek (ratio) estimator:

## A common alternative to HT

Horvitz-Thompson:  $\bar{y}_{HT} = \frac{1}{N} \sum_{i=1}^n (y_i / \pi_i)$

Hajek:  $\bar{y}_{HK} = \frac{1}{\hat{N}} \sum_{i=1}^n (y_i / \pi_i), \hat{N} = \sum_{i=1}^n (1 / \pi_i)$

$(\bar{y}_{HK} = \bar{y}_{HT} \text{ when } \hat{N} = N)$

Question: when  $\bar{y}_{HK} \neq \bar{y}_{HT}$ , which is better?

# A common alternative to HT

$\bar{y}_{\text{HK}}$  is projection estimator for Hajek model:

$$y_i \mid \pi_i \sim_{\text{iid}} N(\mu, \sigma^2 \pi_i)$$

$$(\bar{Y}_{\text{proj}} = \hat{\mu} = \sum_{i=1}^n (y_i / \pi_i) / \sum_{i=1}^n (1 / \pi_i) = \bar{y}_{\text{HK}})$$

So Hajek is better when this model is a better fit to the data

Note that the more general model

$$y_i \mid \pi_i \sim_{\text{iid}} N(\beta_0 + \beta_1 \pi_i, \sigma^2 \pi_i^k)$$

includes HT and HK model as special cases

Could fit this model and let the data decide ...

Zheng and Little spline model is even more flexible...

at the expense of more parameters to estimate

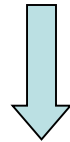
## Ex 3: Bayes for binary $Y$

- Inference for finite population proportion – slides from Qixuan Chen's thesis presentation (Chen, Elliott and Little, 2010)
- She also worked on estimating percentiles of a distribution (Chen, Elliott and Little, 2012)

# Design-based vs. model-based (cont.)

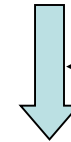
## Design-based estimators

- design unbiased
- potentially very inefficient
- variance estimation is cumbersome, and CI may deviate from nominal level at small sample size



## Parametric model-based estimators

- subject to bias when the underlying model is misspecified
- efficient if model is correct
- variance estimation is more straightforward



Zheng and Little  
(2003, 2005)

## Robust Bayesian predictive estimators

- robust to model misspecification
- efficient
- variance or CI is estimated from posterior distribution, and the confidence coverage is close to the nominal level

# Probit p-spline regression model

- Probit truncated polynomial p-spline model (Ruppert, Wand, and Carroll 2003):

$$\Phi^{-1}\left(P\left(y_i = 1\right)\right) = \beta_0 + \sum_{k=1}^p \beta_k \pi_i^k + \sum_{l=1}^m b_l \left(\pi_i - k_l\right)_+^p$$
$$b_l \sim N\left(0, \tau^2\right), l = 1, \dots, m; i = 1, \dots, N$$

- the constants  $k_1 < \dots < k_m$  are  $m$  selected fixed knots.

# BPSP model for binary $Y$

- Gibbs sampling to obtain posterior distributions

- Model  $y$  via a normal latent variable

$$y_i^* \sim N\left((X\beta + Zb)_i, 1\right), \quad y_i = I(y_i^* > 0)$$

- prior distributions  $\beta_i \sim N(0, 10^6)$ , or  $\{\beta_i \propto 1\}$

$$\tau^2 \sim IG(A, B), \text{ or } \{\tau \propto 1\}$$

- posterior distributions:

$$(\beta, b) | \tau^2, y^* \sim MVN_{m+p+1} \left( \left( C^T C + D / \tau^2 \right)^{-1} C^T y^*, \left( C^T C + D / \tau^2 \right)^{-1} \right)$$

$$\tau^2 | \beta, b \sim IG \left( A + m / 2, B + \|b\|^2 / 2 \right), C = [X, Z], D \text{ a diagonal matrix}$$

with  $p+1$  values of  $10^{-6}$  followed by  $m$  1's on the diagonal

- can also be implemented using WinBUGS. (Crainiceanu, Ruppert, and Wand 2005)

# BPSP estimator (cont.)

- The posterior distribution of the population proportion can be simulated by generating a large number  $D$  of draws of the form

$$p^{(d)} = N^{-1} \left( \sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j^{(d)} \right)$$

- Bayesian p-spline predictive (BPSP) estimator: average of these draws.
- The  $100(1 - \alpha)\%$  credible interval: split the tail area  $\alpha$  equally between the upper and lower endpoints.

# Other estimators

- The Hájek estimator (discussion of Basu (1971))

$$\hat{p}_{HK} = \left( \sum_{i \in s} y_i / \pi_i \right) / \left( \sum_{i \in s} 1 / \pi_i \right)$$

- The parametric model-based estimators

$$\hat{p}_M = N^{-1} \left( \sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j \right)$$

$\hat{y}_j$  = prediction from linear logistic or probit model

- The generalized regression (GR) estimators (Lehtonen and Veijanen 1998)

$$\hat{p}_{GR} = N^{-1} \sum_{j=1}^N \hat{y}_j + \left( \sum_{i \in s} (y_i - \hat{y}_i) / \pi_i \right) / \left( \sum_{i \in s} 1 / \pi_i \right)$$

$\hat{y}_j$  = prediction from linear logistic or probit model



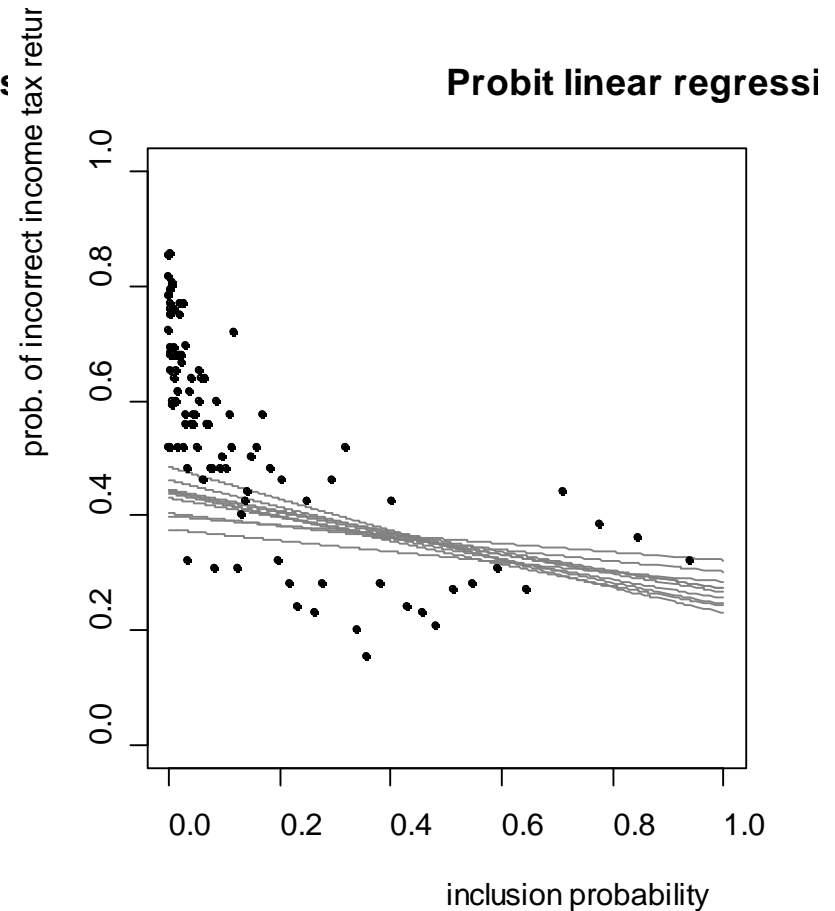
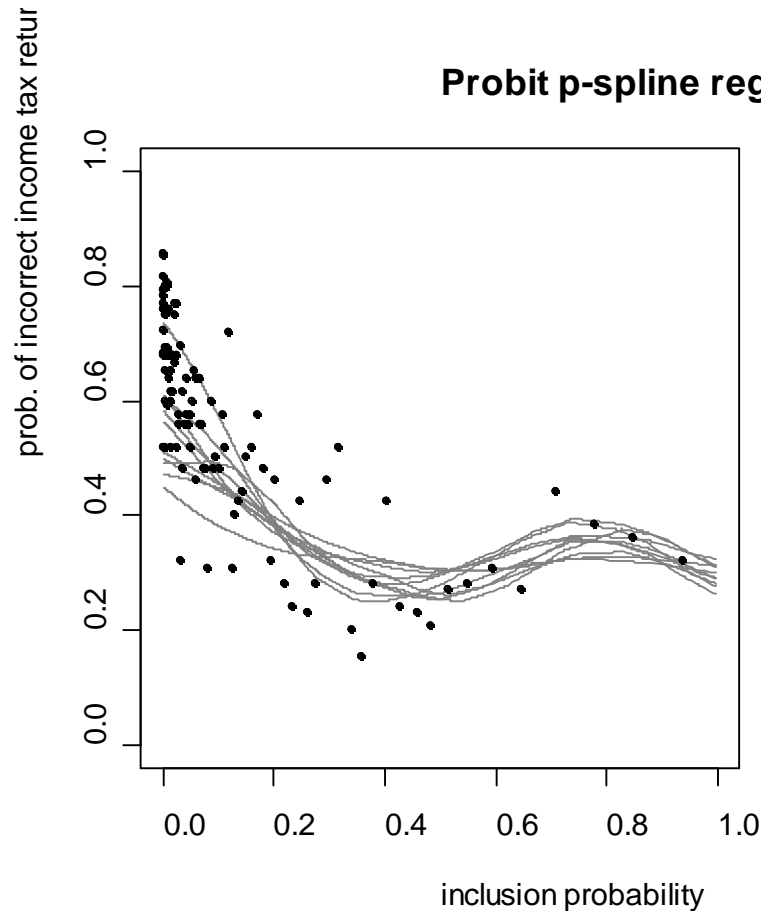
# Simulation studies

- **Comparison study**
  - **HK**, design-based Hájek estimator
  - **LR**, design-consistent predictive estimator with the ML predictions from the model  $\text{logit}(p_i) = \beta_0 + \beta_1 \pi_i^{-1}$  (Firth and Bennett 1998)
  - **PR**, predictive estimator with predictions from the Bayesian probit model  $\Phi^{-1}(p_i) = \beta_0 + \beta_1 \pi_i$
  - **PR\_GR**, the generalized regression (GR) estimator with the weighted ML predictions from the model  $\Phi^{-1}(p_i) = \beta_0 + \beta_1 \pi_i$
  - **BPSP**, the BPSP estimator ( $p = 1$  and 15 knots)
  - **BPSP\_GR**, the GR estimator with the posterior means of  $\Pr(Y_i = 1 \mid \pi_i)$  from the BPSP model as predictions

## Simulation study (2)

- Tax auditing data (Compumine 2007)
  - 3,119 income tax returns
  - $Y$ : whether the income tax return is incorrect ( $p=0.517$ )
  - $X$ : the amount of the realized profit
  - PPS sampling using  $X$  as the size variable
  - $n = 300$  or  $600$
  - 1,000 replicates of simulation

# Simulation study (2): Tax auditing data



# Simulation study (2): results

Table 3 Comparison of various estimators for empirical bias, root mean squared error, and average width and noncoverage rate of 95% CI, in the tax return example

Methods	bias*100		RMSE*100		average width*100		noncoverage*100	
	300	600	300	600	300	600	300	600
HK	-2.4	-1.8	12.4	10.2	36	29	14.1	10.2
LR	6.7	5.5	11.9	9.2	27	21	43.5	45.6
PR	-11.6	-10.1	12.4	10.6	18	14	69.8	83.4
PR_GR	-1.2	-0.3	11.5	8.8	33	26	16.1	11.4
BPSP	-6.8	-2.7	9.3	5.2	27	19	14.2	5.0
BPSP_GR	-0.7	0.2	12.0	10.1	34	26	15.9	12.8

\* The variance of GR estimator is estimated using linearization

★ BPSP estimator performs well; PR estimator is biased and has poor confidence coverage because of model misspecification

# Discussion

- The BPSP estimator yields smaller RMSE than the Hájek and GR estimators, despite slightly higher empirical bias.
- The BPSP estimator achieves robustness to model misspecification compared to parametric model-based estimators.
- The BPSP estimator has closer to nominal level confidence coverage and shorter average length of 95% CI than the Hájek and GR estimators.
  - especially when  $p$  is closer to zero or one and few data are selected into the sample in the tails.
  - This suggests the importance of the current research in estimating finite population prevalence of rare events.

# Ex 4. One stratifier $Z_1$ , one post-stratifier $Z_2$

## Design-based approaches

(A) Standard weighting is  $w_i = w_{is} \times w_{ip} (w_{is})$

Notes: (1)  $Z_1$  proportions are not matched!

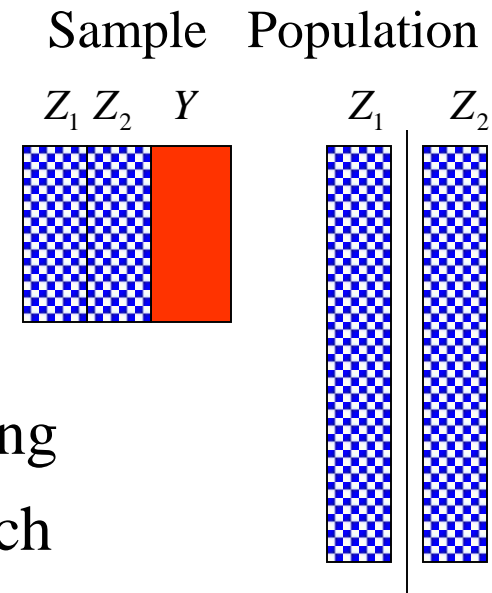
(2) why not  $w_i^* = w_{ip} \times w_{is} (w_{ip})$ ?

(B) Deville and Sarndal (1992) modifies sampling weights  $\{w_{is}\}$  to adjusted weights  $\{w_i\}$  that match poststratum margin, but are close to  $\{w_{is}\}$  with respect to a distance measure  $d(w_{is}, w_i)$ .

Questions:

What is the principle for choosing the distance measure?

Should the  $\{w_i\}$  necessarily be close to  $\{w_{is}\}$ ?



# Ex 3. One stratifier $Z_1$ , one post-stratifier $Z_2$

## Model-based approach

Saturated model:  $\{n_{jk}\} \sim \text{MNOM}(n, \pi_{jk});$

$$y_{jki} \sim \text{Nor}(\mu_{jk}, \sigma_{jk}^2)$$

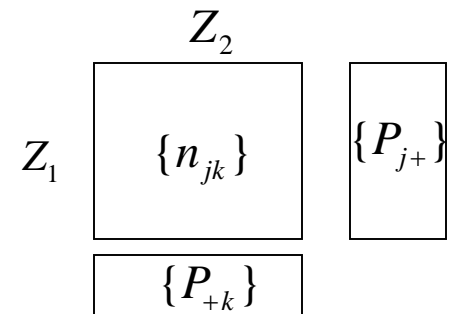
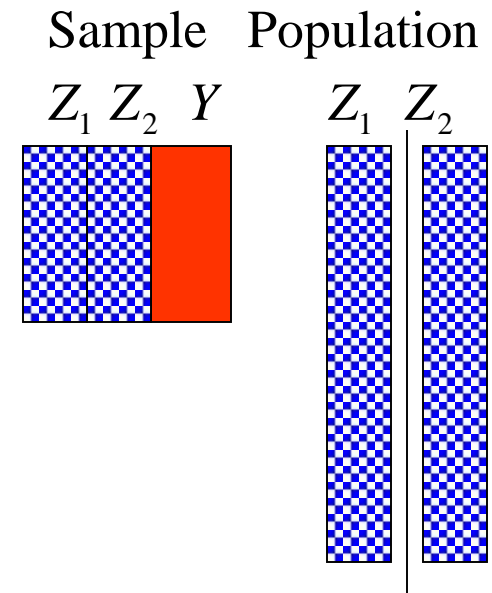
$$\bar{y}_{\text{mod}} = \sum_{j=1}^J \sum_{k=1}^K \hat{P}_{jk} \bar{y}_{jk} = \sum_{j=1}^J \sum_{k=1}^K w_{jk} n_{jk} \bar{y}_{jk} / \sum_{j=1}^J \sum_{k=1}^K w_{jk} n_{jk}$$

$n_{jk}$  = sample count,  $\bar{y}_{jk}$  = sample mean of  $Y$

$\hat{P}_{jk}$  = proportion from raking (IPF) of  $\{n_{jk}\}$

to known margins  $\{P_{j+}\}, \{P_{+k}\}$

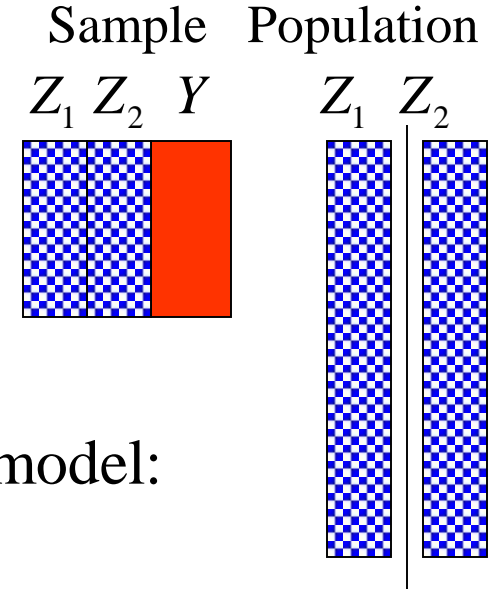
$w_{jk} = n\hat{P}_{jk} / n_{jk}$  = model weight



# Ex 3. One stratifier $Z_1$ , one post-stratifier $Z_2$

## Model-based approach

$$\bar{y}_{st} = \sum_{j=1}^J \sum_{k=1}^K \hat{P}_{jk} \bar{y}_{jk} = \sum_{j=1}^J \sum_{k=1}^K w_{jk} n_{jk} \bar{y}_{jk} / \sum_{j=1}^J \sum_{k=1}^K w_{jk} n_{jk}$$



What to do when  $n_{jk}$  is small?

Model: replace  $\bar{y}_{jk}$  by prediction from modified model:

e.g.  $y_{jki} \sim \text{Nor}(\mu + \alpha_j + \beta_k + \gamma_{jk}, \sigma_{jk}^2)$ ,

$$\sum_{j=1}^J \alpha_j = \sum_{k=1}^K \beta_k = 0, \gamma_{jk} \sim \text{Nor}(0, \tau^2) \text{ (Gelman 2007)}$$

Setting  $\tau^2 = 0$  yields additive model,

otherwise shrinks towards additive model

Design: arbitrary collapsing, ad-hoc modification of weight



# Summary

- HT estimate is design-unbiased, but does not have good (design-based properties) when the “implied” underlying HT model is not a good fit to the data
- Bayes inference under a more flexible model relating  $Y$  to  $Z$  yields better design-based inferences
  - More efficient estimates
  - Better confidence coverage in moderate samples
- Unlike design-based inference, Bayes inference is not asymptotic, and can deliver good frequentist properties in small samples

# References

- Chen, Q., Elliott, M.R. & Little, R.J. (2010). Bayesian Penalized Spline Model-Based Estimation of the Finite Population Proportion for Probability-Proportional-to-Size Samples. *Survey Methodology*, 36, 23-34.
- Chen, Q., Elliott, M.R. & Little, R.J. (2012). Bayesian Inference for Finite Population Quantiles from Unequal Probability Samples. *Survey Methodology*, 38, 2, 203-214
- Ruppert, D. and Carroll, R.J. (2000). Spatially Adaptive Penalties for Spline Fitting. *Australia and New Zealand Journal of Statistics*, 42, 205–223.
- Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge, UK: Cambridge University Press.
- Zheng, H. & Little, R.J. (2004). Penalized Spline Nonparametric Mixed Models for Inference about a Finite Population Mean from Two-Stage Samples. *Survey Methodology*, 30, 2, 209-218.
- Zheng, H. & Little, R.J. (2005). Inference for the Population Total from Probability-Proportional-to-Size Samples Based on Predictions from a Penalized Spline Nonparametric Model. *Journal of Official Statistics*, 21, 1-20.