# Multiple Frame Methods and Designs for Combining Data Sources

IASS Webinar, 26 March 2025

Sharon L. Lohr

Slides & references available at www.sharonlohr.com

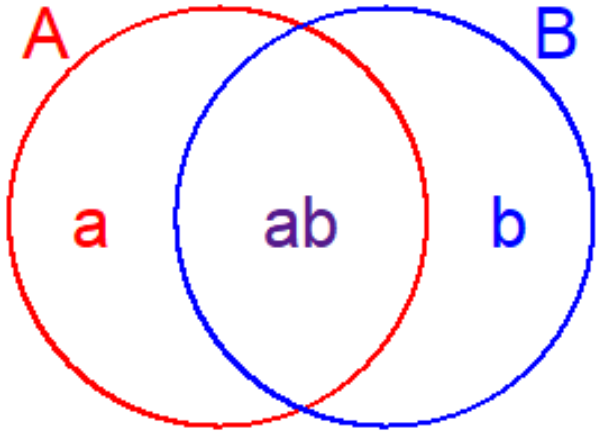
# Motivation

- Probability sample response rates decreasing, costs increasing
- Can we use data from alternative sources to improve:
  - Coverage
  - Precision of national estimates
  - Estimates for subpopulations
  - Cost efficiency
- Combine data sources using Multiple Frame (MF) approach
- Follow-up of ideas in Lohr (2025); Rao & Lohr (2025)

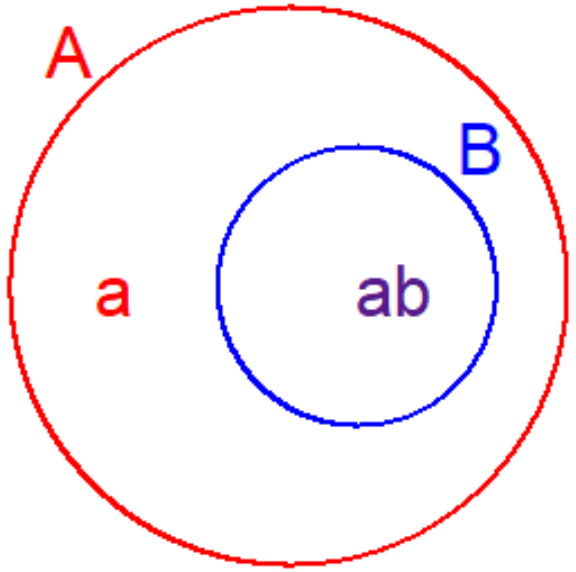
# Outline

- Classical MF surveys: assumptions and estimators
- Advantages
- What if assumptions not met
- MF survey design to
  - Improve coverage and precision
  - Detect violations of assumptions
  - Provide robustness for data sources that may change
- Research problems

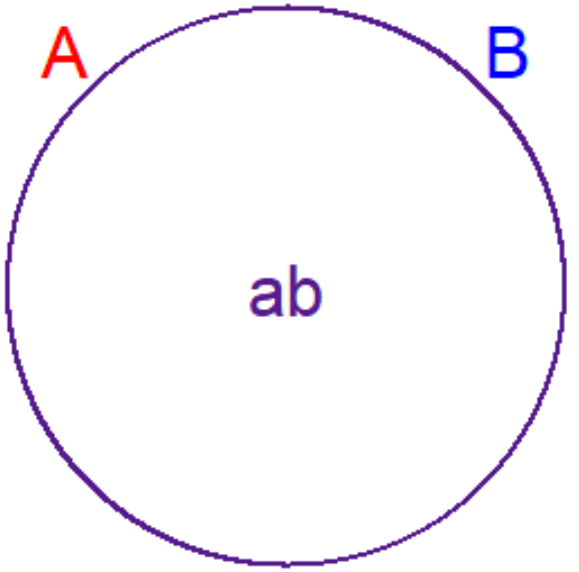
# Three Dual-Frame Structures




Frames A and B are both incomplete but overlap



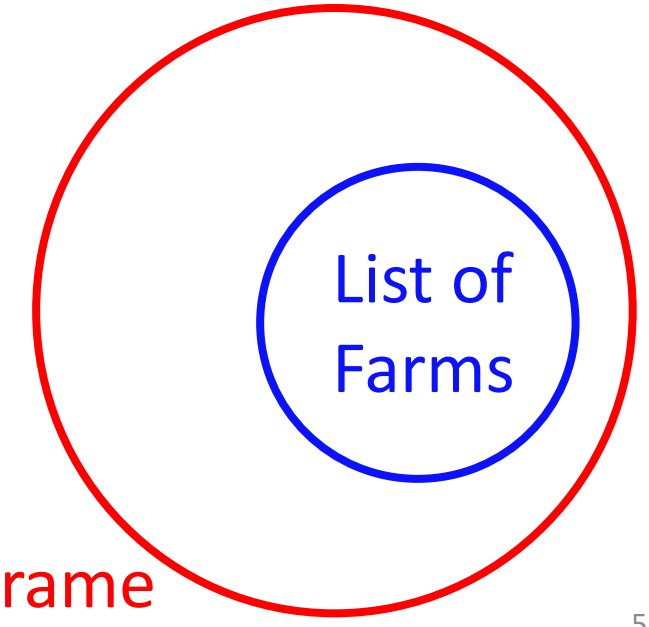

Frame A complete; B is proper subset




Frames A and B cover same population

# MF: Take independent samples $S_A$, $S_B$ from A, B

- Hartley (1962)
  - Defined notation for frames A, B and domains a, b, ab
  - Derived point estimators, variances, optimal allocation when $S_A$, $S_B$ are probability samples
- Motivation: agricultural surveys
- Ferraz et al. (2023): MF in agriculture
- Kott & Vogel (1995): business



# Data Linkage or MF Survey?

- If there is sufficient identifying information
  - Link units from Frames A, B, and identify duplicates
  - Create new frame from unique records in A, B
  - Take single frame (SF) sample from new frame
- SF is simpler, more flexible, & more efficient than MF survey
  - Can use more efficient stratification
  - Avoid complications of merging estimates, separate calibrations
- If you can augment coverage by linkage, preferred path
- But requires detailed info in frames to link

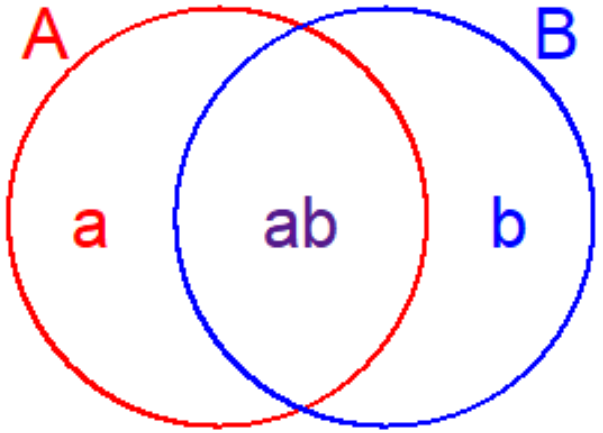
# MF Survey methods can be used when …

- Frames cannot be linked
  - Need to know only if each $S_A$ unit is in B, and if each $S_B$ unit is in A
  - Don't need to know **which** frame unit matches sampled unit
- Frames contain different types of units
  - A has land areas; B lists farms
  - A has households; B has persons
  - A is address frame; B uses adaptive sampling design
- Only summary statistics are available
  - Can purchase statistics about b, ab from vendor

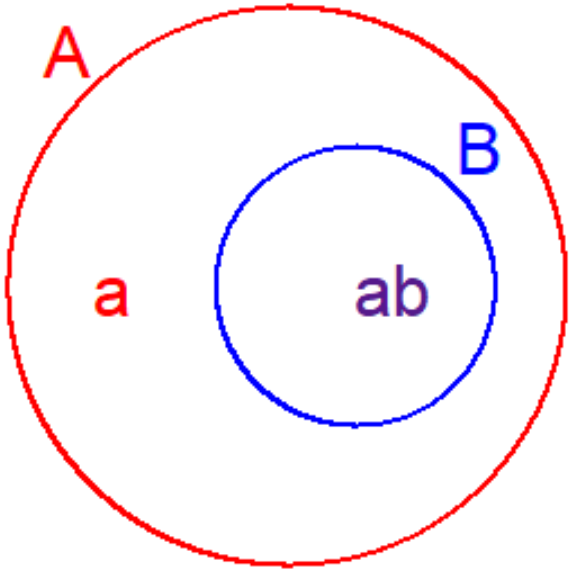
- Research problem: compare SF with imperfect linkage to MF design

# Advantages of MF

- Improve population coverage when one or both frames incomplete

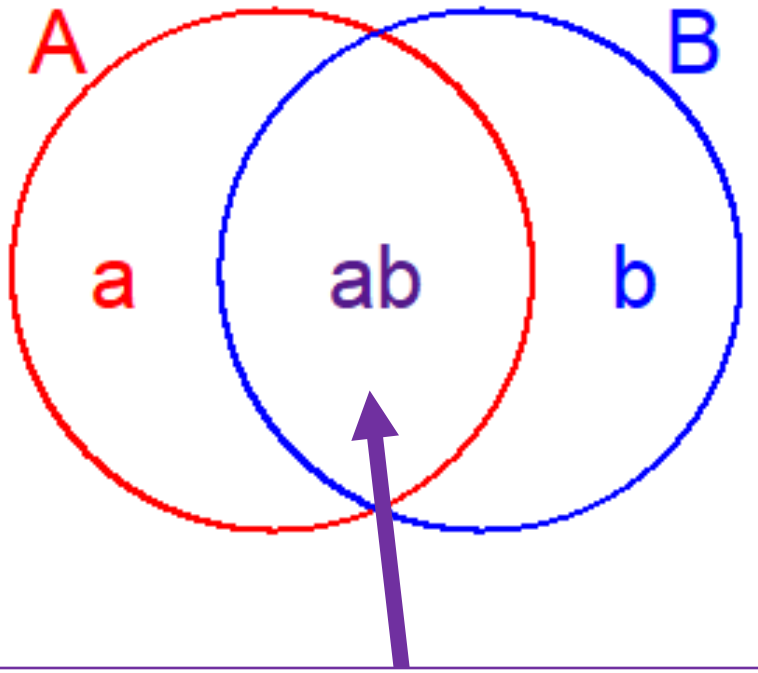



- Cost savings & efficiency gains. If data collection is cheaper for B, can reduce costs by relying on B for observations in ab
- Increase *n* for rare or hard-to-find subpopulations concentrated in B



# Estimation for Classical MF Surveys

- If assumptions met, main problem is to account for overlap
- Domain ab in both samples
- Adjust weights for multiplicity
- $\hat{Y} = \hat{Y}_a + \alpha\hat{Y}_{ab} + (1-\alpha)\hat{Y}_{ab} + \hat{Y}_b$
- Reduces weights in ab
- Or poststratify each term to domain counts $N_a, N_{ab}, N_b$
- Lots of estimators (Lohr 2011, 2021)




Adjust $w_i, w_j$ so sum of weights $\approx N_{ab}$

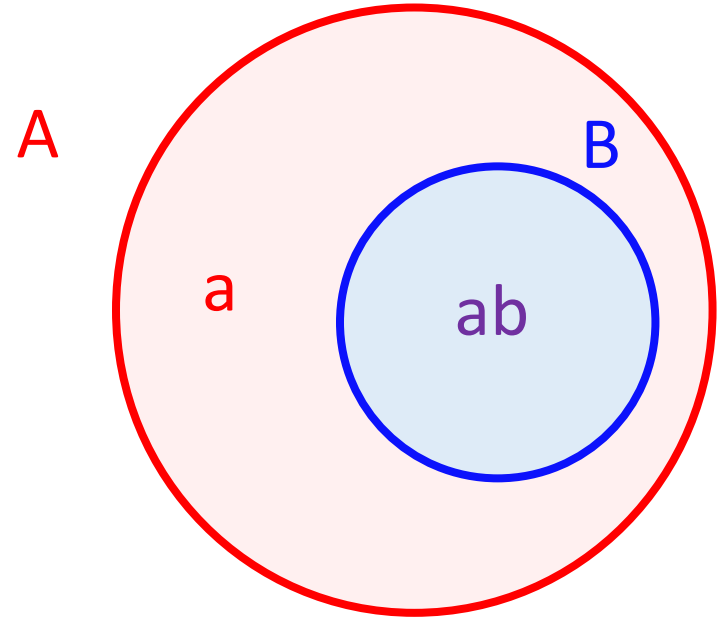
# Assumptions for Classical MF Surveys

A. Union of frames **covers** population
B. Samples from frames are selected **independently**
C. Domain membership **known** for each sampled unit
   - At minimum, need to be able to estimate multiplicity (Mecatti, 2007)
D. Estimators of population totals from each sample in each domain are **unbiased**

   Met if
   - Full-response probability sample taken from each frame
   - No measurement error: $y_i$ (Sample $j$) = $y_i$ (Sample $k$)

# Extension for Big Data

- $S_B$ is census from B (Lohr, 2014; Kim & Tam, 2021)
- Full coverage because of $S_A$ ; $S_B = \mathrm{B}$ can be convenience sample, administrative data, cell phone location data, …
- If MF assumptions met, $S_A$ is needed only to estimate $Y_a$; estimates from $S_B$ have no error



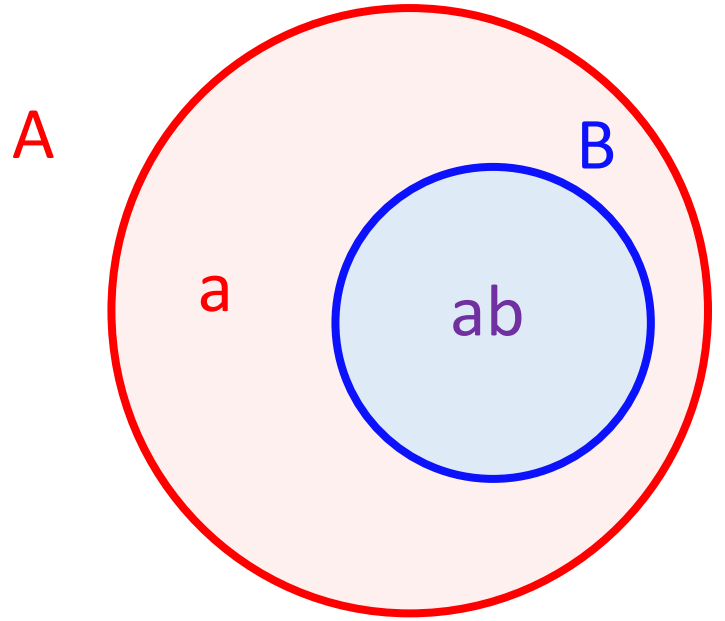
# $S_B = \mathrm{B}$: If assumptions met

- Optimal estimator is

$$\hat{Y} = \hat{Y}_a + \alpha \hat{Y}_{ab} + (1 - \alpha)\hat{Y}_{ab}$$

$$\alpha = -\mathrm{Cov}(\hat{Y}_a, \hat{Y}_{ab})/\mathrm{V}(\hat{Y}_{ab})$$

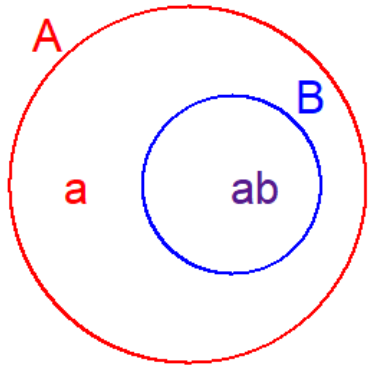
- Leverages covariance (usually small) to increase precision for a
- Weights can be negative or < 1
- Or use $\hat{Y}_a\left(\frac{N_a}{\hat{N}_a}\right) + \hat{Y}_{ab}$



# Does the classical MF theory extend?

- Classical MF surveys are designed to be blended
  - Same questions
  - Protocols as similar as possible
  - Measure domain membership (e.g. ask about cell/landline use)
- When use “found” data for B
  - Different measurement concepts
  - How accurate is domain membership for units in $S_A$ ?
  - No control over data quality, collection protocols in B

# What can go wrong?




- Domain misclassification
  - Sample $S_A$: a units erroneously placed in ab may get reduced weights
- Measurement error
  - Is *y* measured in $S_A$ the same as *y* measured in $S_B$? Different questions, definitions, time periods, modes, measurement error properties ….
  - Is *y* measured the same way for all subpopulations?
  - Is auxiliary information **x** (used for calibration) measured the same way?
- Missing data or duplicate responses
  - Nonresponse in either or both samples
  - Convenience data for $S_B$ may have poor quality or duplicate responses

# Domain Misclassification

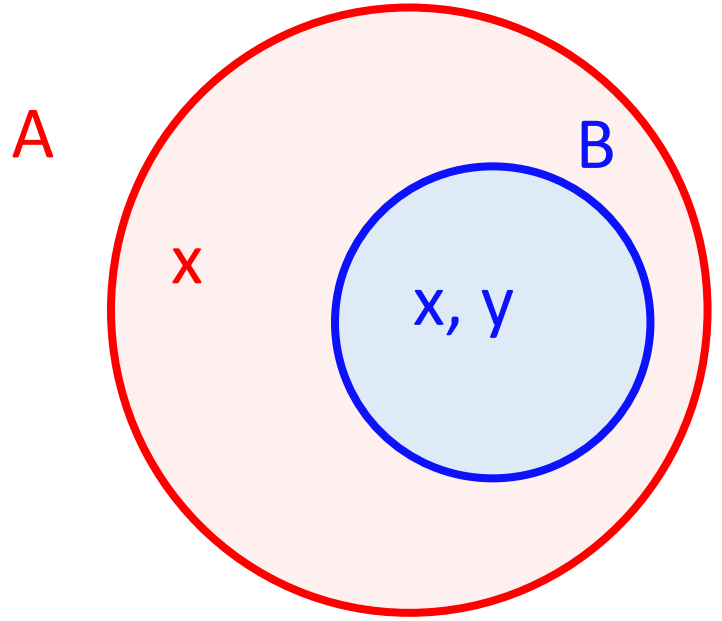
- Even small amount of domain misclassification can lead to bias
- Bias depends on
  - Differences among domain means
  - Misclassification probabilities
- Remedies and diagnostics?
  - Estimate misclassification probabilities from external source (Lohr, 2011)
  - Match samples to evaluate frame overlap (Dever, 2018)
  - Estimate probability unit $i$ belongs to domain $d$ (Kim & Tam, 2021)
  - Stack samples and estimate probability (Savitsky et al., 2023)

# Measurement Error

- Examples
  - Annual survey from A; B has monthly price data
  - Income y is defined differently in the two sources
  - Convenience sample B has bogus responses (Kennedy et al., 2024)
  - Administrative data (Judson & Popoff, 2005):
    - Variables used for program administration are subject to quality control
    - 'Add on' variables superfluous to agency's mission should be used with caution
    - "20 years of household level water consumption data … all entered by people who didn't care."
- Kim & Tam (2021): $y_i^* = \beta_0 + \beta_1 y_i + \epsilon_i$
- Binary variable: misclassification probabilities (Tam et al., 2022)
  - $P(y_i^* = 1 \,|y_i = a, \boldsymbol{x}), P(y_i^* = 0 \,|y_i = a, \boldsymbol{x})$

# Extensions of MF (see references)

- Inverse estimated probability weighting
  - Use $S_A$ to estimate selection probs for B
  - $\hat{\pi}_i = \hat{P}(i \in B \,|\boldsymbol{x})$
  - $\hat{Y}_{sel} = \sum_B y_i/\hat{\pi}_i$
  - Does selection mechanism depend on y?
  - Assumes all $\pi_i > 0$.
- Mass imputation
  - $\tilde{y} = \hat{g}(\boldsymbol{x})$, developed on B
  - $\tilde{Y}_a = \sum_a w_i\tilde{y}_i$, $\tilde{Y}_{ab} = \sum_{ab} w_i\tilde{y}_i$
  - $\hat{Y}_{imp} = \tilde{Y}_a + \alpha\tilde{Y}_{ab} + (1-\alpha)\hat{Y}_{ab}$
  - Does model apply to a? (Lu, 2014)

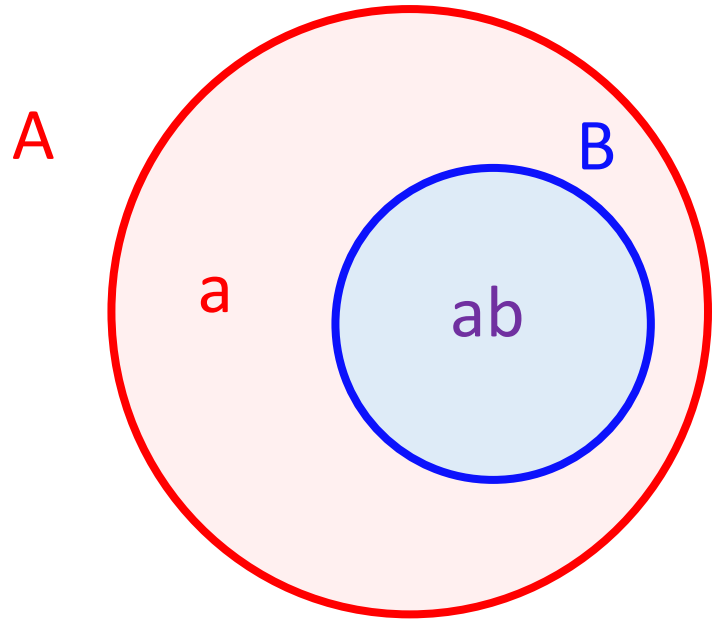



$S_B$ measures **x** and y

$S_A$ measures only **x**

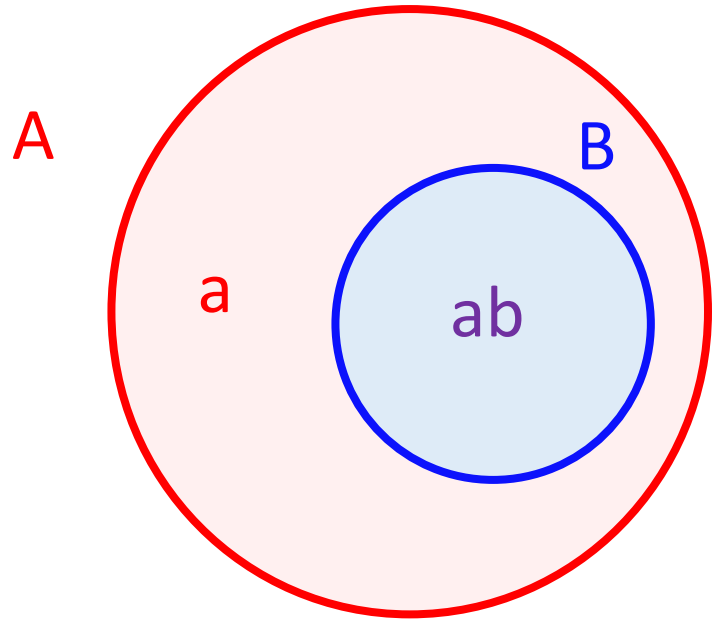
# Ang et al. (2025)

- $S_B$ is census from B
- Compare estimators
  - Single frame (from A), B alone, $\hat{Y}_{sel}$, $\hat{Y}_{imp}$, screening DF designs (only a units from $S_A$),
  - With and without measurement error in B
  - Simulated establishment population
- DF estimators:
  - $\hat{Y}_a + \hat{Y}_{ab}$
  - Calibrated to population totals



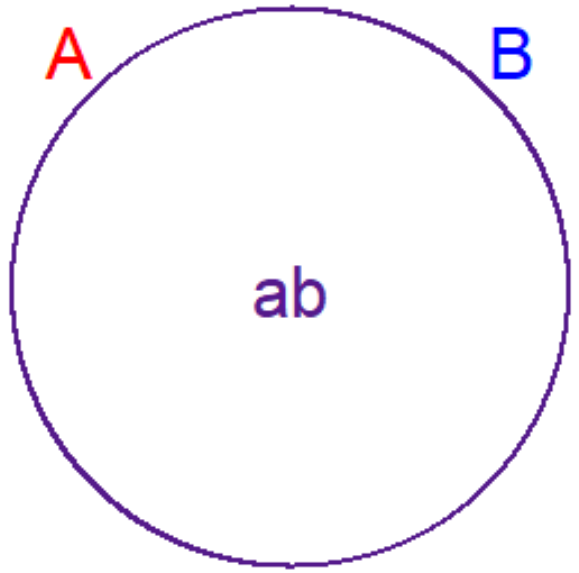
# Ang et al. (2025)

- DF estimators:
  - $\hat{Y}_a + \hat{Y}_{ab}$
  - Calibrated to population totals
- No measurement error in $S_B$
  - DF estimators unbiased, have low MSE
  - DF corrects for undercoverage of B
- Measurement error in $S_B$
  - DF estimators biased
  - Larger MSE than SF survey from A



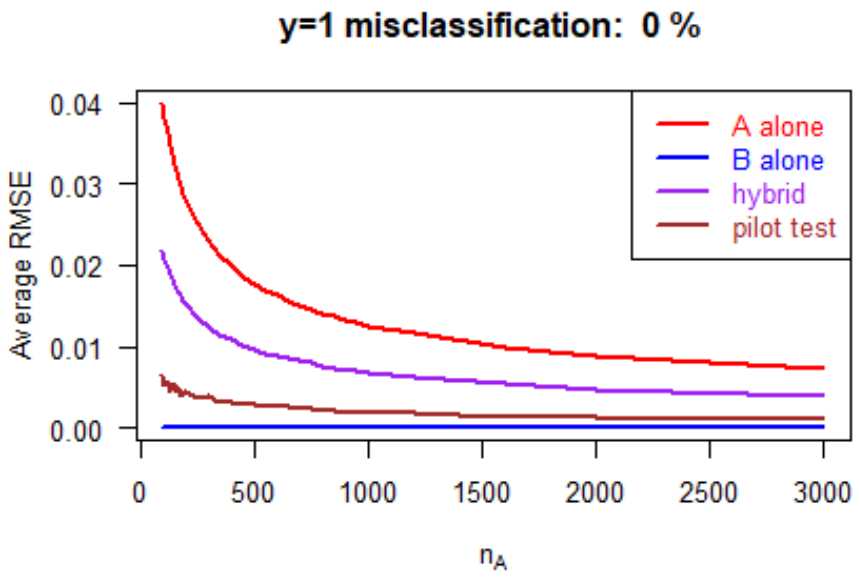
# Effect of Bias

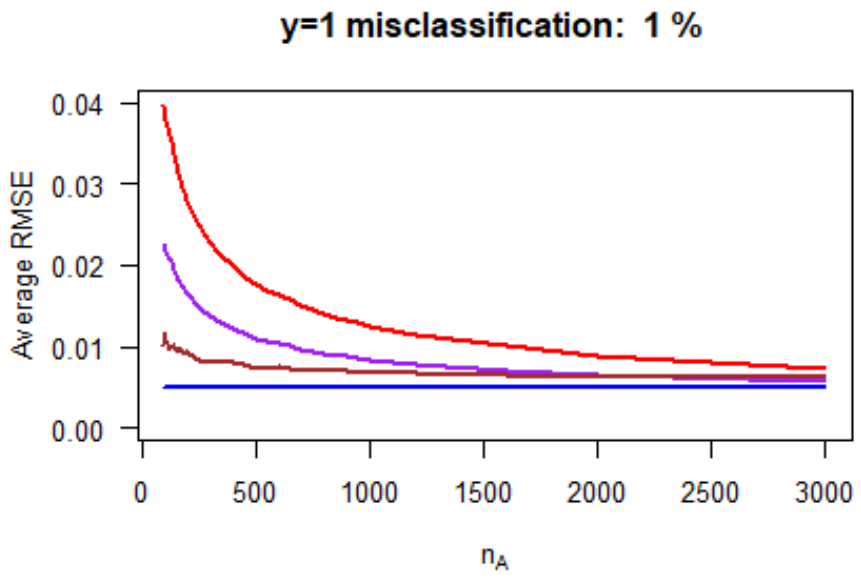
- Binary response (p=.5), y in $S_A$, y* in $S_B$
- In B, misclassify v% of observations with y=1
- Simulation: Average RMSE for
  - $S_A$ alone (simple random sample)
  - $S_B$ alone
  - Pilot sample from A size $n_A/4$
    - If means not significantly different, use $\bar{y}_B$
    - Else use mean of sample of size $3n_A/4$ from A
  - Hybrid: Elliott & Haviland (2007)
    - $\alpha = (\bar{y}_A - \bar{y}_B)^2/[(\bar{y}_A - \bar{y}_B)^2 + \bar{y}_A(1 - \bar{y}_A)/n_A]$
    - More reliance on $S_A$ when estimated bias large

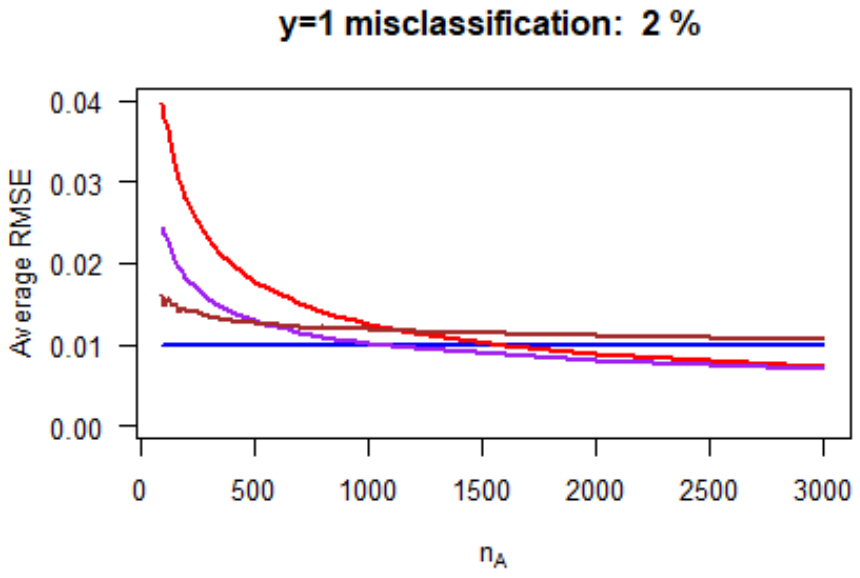


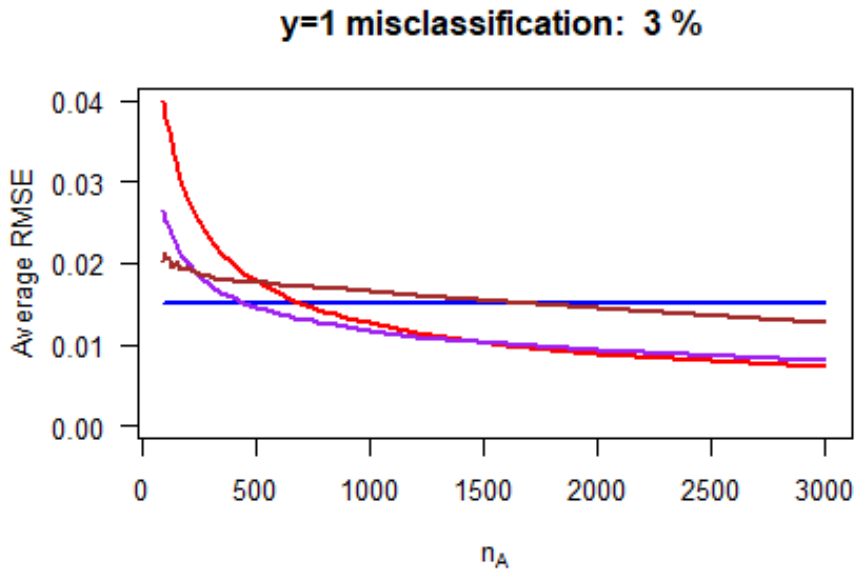
Bias always wins in end

As $n_A \uparrow$, hybrid $\alpha \to 1$

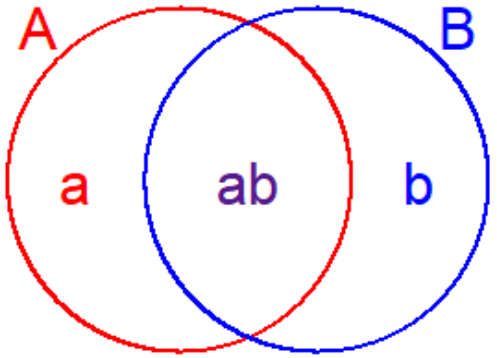


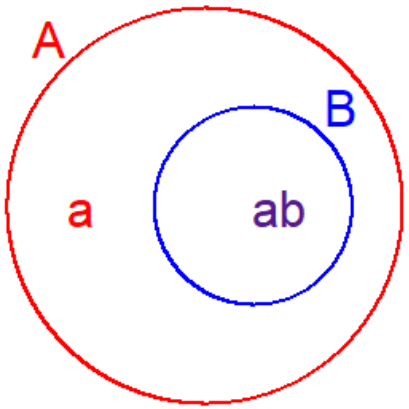



# Design Considerations if Assumptions Met

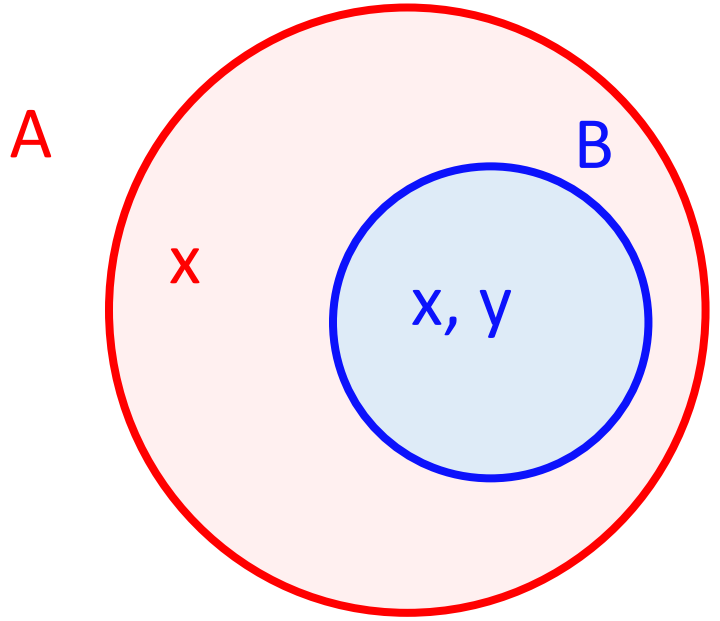



- Is domain membership known before sampling?
  - Use stratified sampling allocation
  - Optimal design: all ab observations from cheaper frame
- Can we screen cheaply for domain membership?
  - Two-phase design with stratification
- Domain membership known only after sampling
  - Optimal design in Hartley (1962)
- Design if nonresponse
  - Brick et al. (2011), Lu et al. (2013), Lohr & Brick (2014)



# Holmberg et al. (2024)

- Optimal design for $S_A$
- Inverse estimated probability weighting
  - $\hat{\pi}_i = \hat{P}(i \in B \mid \boldsymbol{x})$, developed on $S_A$
  - $\hat{Y}_{sel} = \sum_B y_i/\hat{\pi}_i$
- Design $S_A$ to minimize anticipated variance of $\hat{Y}_{sel}$

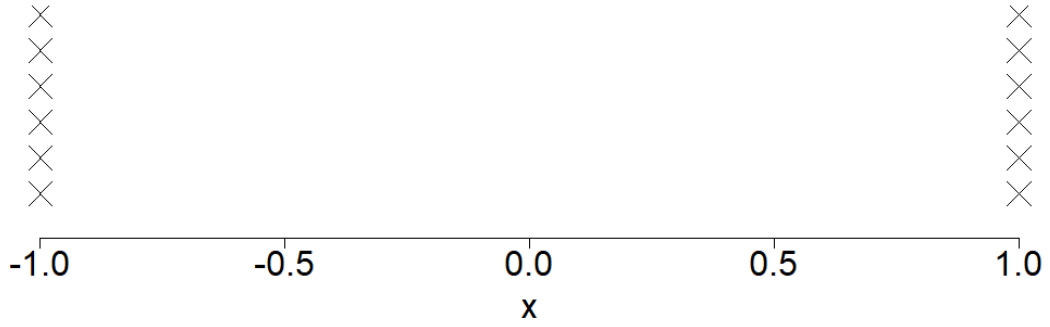


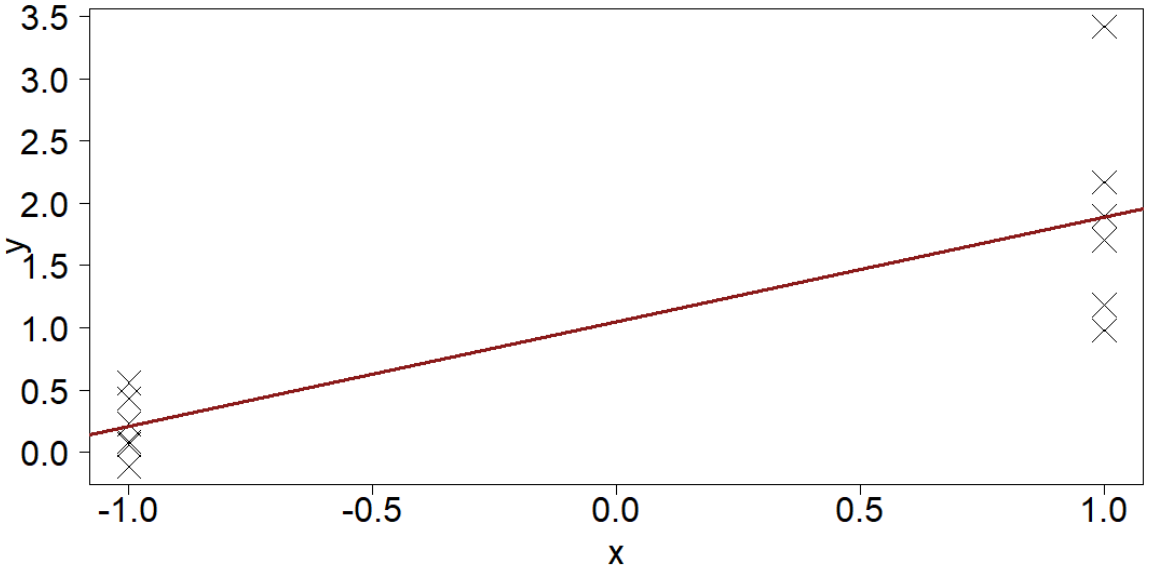

$S_B$ measures **x** and y

$S_A$ measures only **x**

# Optimal Design for Regression (Smith, 1918)

Optimal Design for Straight Line Regression

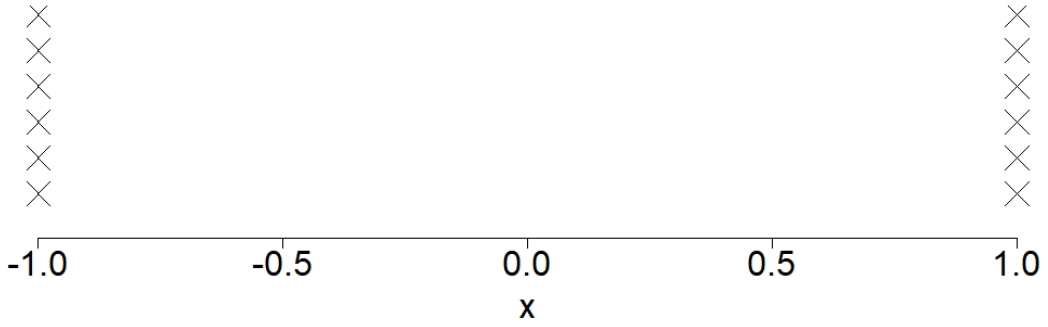


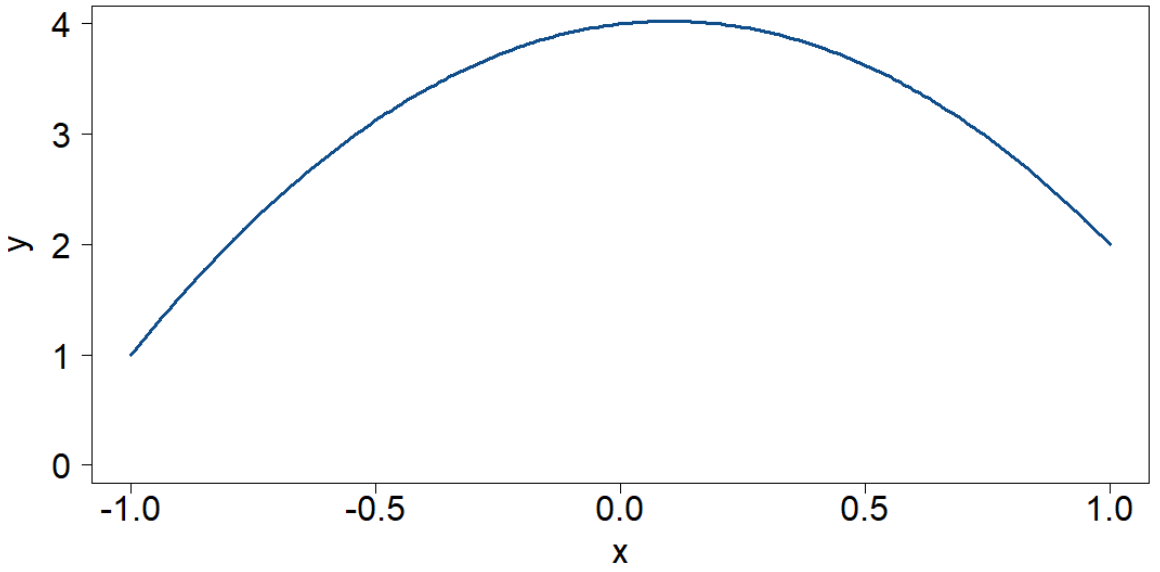



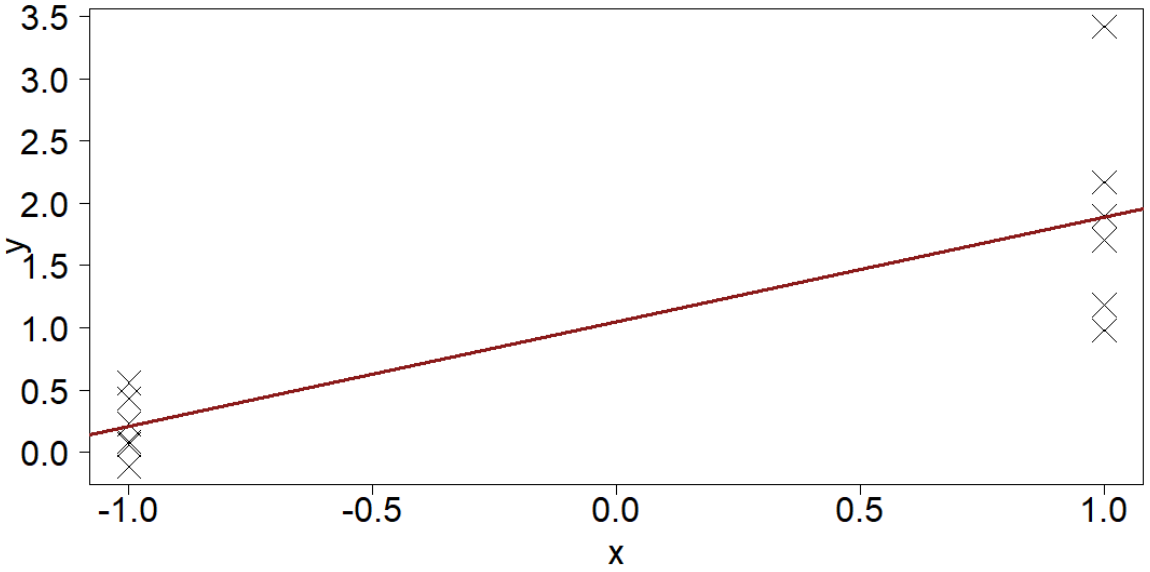

Kirstine Smith

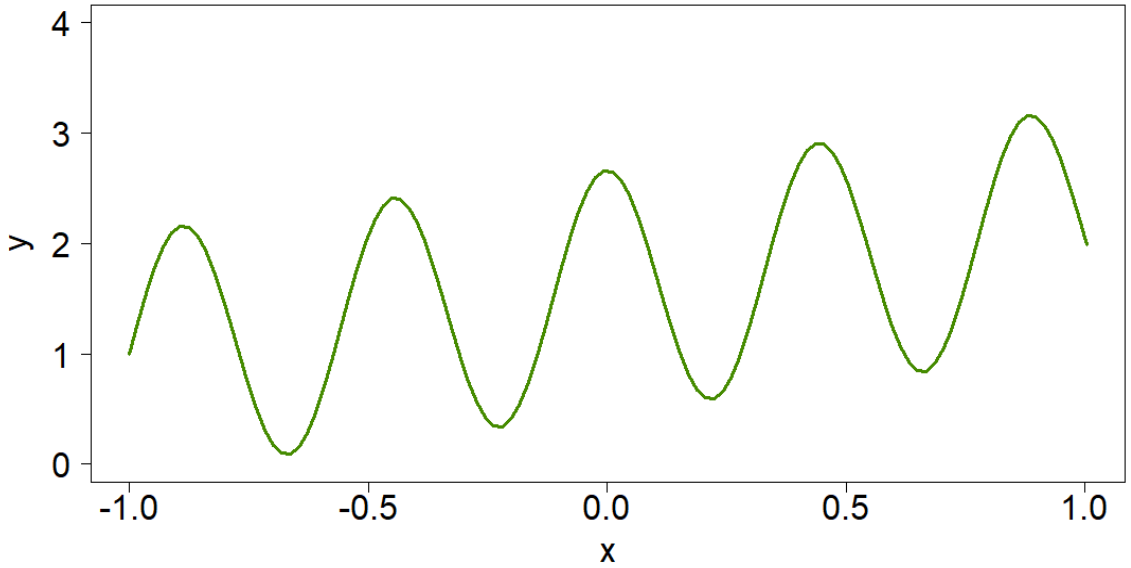
# What if function is

## Optimal Design for Straight Line Regression



# Robust Design for Regression

- Smith (1918): “we shall try to find a compromise between the two desiderata of a low maximum of standard deviation and of a uniform distribution”
- Box & Draper (1959): “Necessity for considering bias as well as variance”
- Box (1982): “assumptions and specifications which motivate alphabetic optimality are … unduly limiting”
- Herzberg (1982): “construction of designs which guard against particular shortcomings”

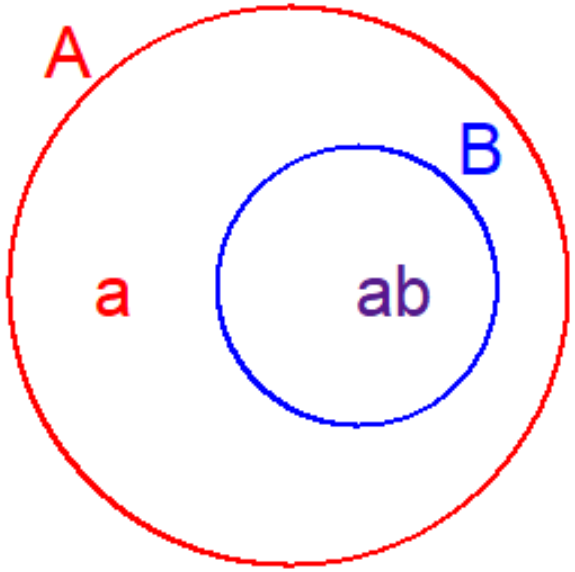
# Robust Design for Regression

- Nonparametric model
  - $y = f(x) + \epsilon,\ f$ is twice continuously differentiable
- Müller (1984): optimal design density is uniform
- Cheng, Hall, & Titterington (1988): sequential design for estimating $f$

- Chaloner & Verdinelli (1995):
  - "The experimental design process should incorporate model uncertainty"
  - Bayesian design puts priors on model uncertainty

# Robust Design for MF survey: Fixed Source B

- Want to use info in B
- But be able to estimate bias
- Need to take some observations from all sources in overlap domains
- Bayesian design: incorporate uncertainty about bias from B
- Sequential approach
  - Pilot sample from A in ab
  - Estimate measurement error
  - Use estimated measurement error to inform main design

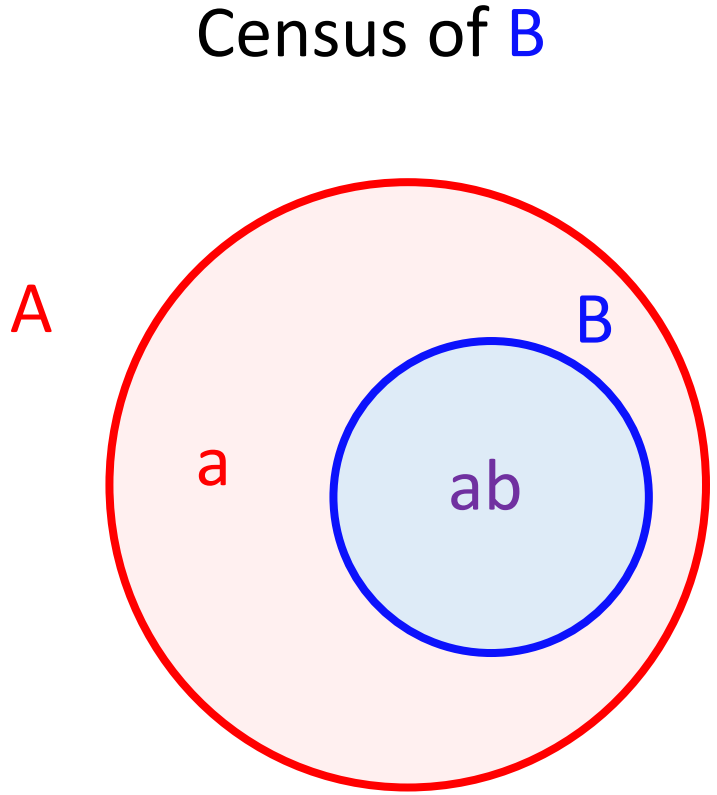


# Design Considerations

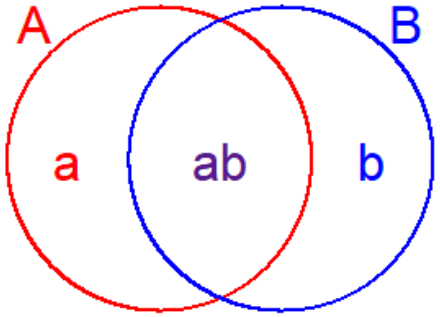
- Assumption: $S_A$ is "Gold Standard" Source
  - These are becoming rarer ….
  - $S_A$ can have nonresponse, measurement error
  - Implications for design, estimates if $S_A$ has errors
- Will each data source continue to be available?
- Or contain the same quality of information?
- Use multiple data sources to monitor quality, assess measurement errors
  - deBroe et al. (2021), National Academies (2023), Coffey et al. (2024)

# Design Considerations

- Relative amounts of information for different domains
  - Greatly unequal weights
  - $w_i = 1, w_i = 60000$
  - Less info for subpopulations concentrated in a



# Estimation for Subpopulations




- Improve coverage
  - Persons experiencing homelessness
  - Persons without telephones
  - B is list or respondent-driven sample from rare population
- Assess data quality for $S_A$, $S_B$
  - Compare subpopulation estimates in ab
- Subpopulations in a, b have less information
- What if screening design used and B information flawed?

# Design Surveys to Facilitate Integration

- Rich frames: Information to predict membership in other sources
  - Can include domain membership in stratification design
  - Error in domain membership from frame affects efficiency
- Collect variables in each sample to allow for data integration
  - Similar to collecting information to adjust for nonresponse
  - Membership in other data sources
  - Rich auxiliary variables **x**
  - Subpopulation membership
- Consider quality of information for subpopulations

# Multiple Goals for Multiple Frame Surveys

- Estimate key population quantities with sufficient accuracy
- Assess nonsampling errors from different data sources
- Provide information to improve future data collections
- Be adaptable for future needs
  - Take advantage of new data sources
  - Continuity of time series
  - Will today's data sources be available tomorrow?
  - Or will they be equally reliable tomorrow?
- Cultivate alternative data sources

# Thank you!

Slides and References
www.sharonlohr.com