Health Survey Research Methods

Elly Kaizar   Elizabeth Stasny

2008 Summer Program in Applied Biostatistics and Epidemiological Methods

July 14-18, 2008

Outline

Part I

Introduction to Sampling

Example: Ohio Family Health Survey

"The OFHS obtained detailed data regarding Ohio residents' access to health insurance coverage, general health status, and their perceptions about, and access to, health care."

Possible Questions:

- What proportion of Ohio residents have trouble accessing needed health care?
- What is the average cost of health care for Ohio families?
- How much money in total to Ohioans pay for health care?
- What proportion of Ohio families with private health insurance have that insurance provided by an employer?
- Are Ohio families headed by a minority less likely to be covered by health insurance?

What?

- What is a sample?
  - A Sample is a subset of a Population
  - Population = All families residing in Ohio in 2003-2004.
  - Sample = Some number of these families.
- What might we do with a sample?

Why?

Why would one want to consider a sample?

- Save time
- Save money
- No choice
- Better information

How?

- Convenience sample
  families that come to the health department to get flu shots
- Systematic sample
  every 10th family in the phone book
- Judgement sample
  families that I choose to be ‘representative’ of Ohio famiiles
- Probability sample
  randomly chosen families

‘Representing’ the Population

- Not necessary (or necessarily desirable) for the sample to be a small version of the population.
- Each sampled unit will represent the characteristics of a known number of units in the population
- In a **Probability Sample**, the probability of inclusion for each unit is known and nonzero.

Example

Part of the Ohio Family Heath Survey selected families for inclusion in the survey by randomly sampling from all residential Ohio-based land line telephone numbers.

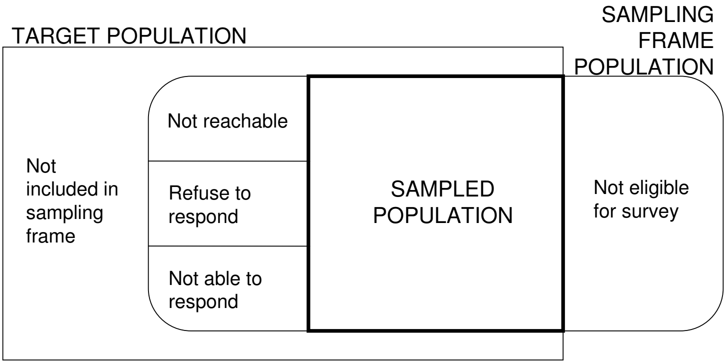
In a **Probability Sample**, the probability of inclusion for each unit is known and nonzero.

Is this a probability sample?

Vocabulary

- **Observation Unit:** An observation on which a measurement may be taken (Family)
- **Target Population:** All observation units we want to study (All families living in Ohio)
- **Sampling Unit:** The unit we actually sample (Phone number)
- **Sampling frame:** List of sampling units (All phone numbers)

Coverage


TARGET POPULATION
SAMPLING FRAME POPULATION
Not included in sampling frame
Not reachable
Refuse to respond
Not able to respond
SAMPLED POPULATION
Not eligible for survey

Nonsampling Errors

Nonsampling Errors are biases and variability that come from causes other than the sampling scheme.

- Coverage error
- Nonresponse bias
- Measurement bias
- Corrupted data (e.g, incorrect data entry)

Sampling Error

Sampling Error is the variability that comes from taking a sample rather than a census.

Desirable Properties

- The appropriate sample quantity is unbiased
  On average, you estimate the right thing
- The appropriate sample quantity is measured with small variability
  If you repeated the survey, your estimate would not change a lot

How are surveys different from experiments?

Experiments

- Theoretical infinite population
- iid observations

Sample Surveys

- Real finite population
- Not necessarily iid

Both want good estimates in terms of bias and variance
Both use ideas like blocking to reduce variance

Part II

Simple Random Sampling: Means, Proportions and Totals

Notation

- $\mathcal{U}$ = the finite population (the Universe)
  (All Ohio families*)
- N = the number of units (families) in the population
  (Number of families in Ohio)
- S = the sample
  (Families sampled for this survey)
- n = the number of units in the sample
  (Number of families in this survey = 39,953 completed surveys*)

Simple Random Sample (Without Replacement)

A **Simple Random Sample (SRS)** is a sample where every possible subset of n sampling units has the same probability of being the sample.

Conditions:

- Sample size (n) is fixed
- No unit can be selected more than once
- Probability of selection is equal for all units
- Joint probability of selection is equal for all pairs, triplets, etc. of units

If we assume all Ohio families have one telephone, we can think of a survey that randomly dials telephone numbers as *approximately* a SRS.

More Notation

Each unit in the sample is associated with some characteristic(s) or attributes(s) that we want to measure:

Unit #	1	2	...	N
Attribute 1	x_1	x_2	...	x_N
Attribute 2	y_1	y_2	...	y_N
Attribute 3	z_1	z_2	...	z_N

Example Attributes

x	Number of adult members of the household
y	Number of child members of the household
z	Yes/No: the household contains children
	$= 1$ if yes, 0 if no

Estimands

Estimands = Population Values = the values we would like to estimate

- Average/Mean
  (Average yearly family expenditure for health care)
  (Average number of adults in a household; $\bar{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i$)
- Proportion
  (Proportion of adults with dental insurance)
  (Proportion of households with children; $p = \frac{1}{N}\sum_{i=1}^{N} z_i$)
- Total
  (Total number of children with health insurance)
  (Total number of adults in Ohio; $t = \sum_{i=1}^{N} y_i$)

Estimators: Simple Random Sample

Population	Estimate
$\bar{y}_U = \frac{1}{N} \sum_{i \in \mathcal{U}} y_i$	$\bar{y} = \frac{1}{n} \sum_{i \in \mathcal{S}} y_i$
$p = \frac{1}{N} \sum_{i \in \mathcal{U}} z_i$	$\hat{p} = \frac{1}{n} \sum_{i \in \mathcal{S}} z_i$
$t = \sum_{i \in \mathcal{U}} y_i$	$\hat{t} = N \frac{1}{n} \sum_{i \in \mathcal{S}} y_i$

Attributes

- Population quantities ($\bar{y}_U$, p, t)
  - Do not depend on the sample that we choose
  - Fixed, not random
- Individual unit quantities (x_i, y_i, z_i)
  - Do not depend on the sample that we choose
  - Fixed, not random

What is random?

Sampling Distribution

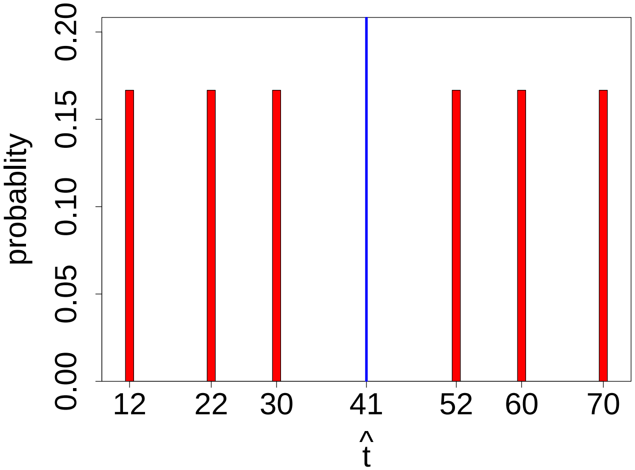
The **Sampling Distribution** describes the values of the sample statistics you would get over all possible samples from the population (using the same sampling scheme).

Toy Example

Population: 4 coins (penny, nickel, dime, quarter)
Sample size: 2
Attribute of interest: value
Estimand: total (truth=41¢)

Sample	Average Estimate ($\bar{y}$)	Total Estimate ($\hat{t} = 4\bar{y}$)
penny, nickel	(1+5)/2 = 3¢	12¢
penny, dime	(1+10)/2 = 5.5¢	22¢
penny, quarter	(1+25)/2 = 13¢	52¢
nickel, dime	(5+10)/2 = 7.5¢	30¢
nickel, quarter	(5+25)/2 = 15¢	60¢
dime, quarter	(10+25)/2 = 17.5¢	70¢

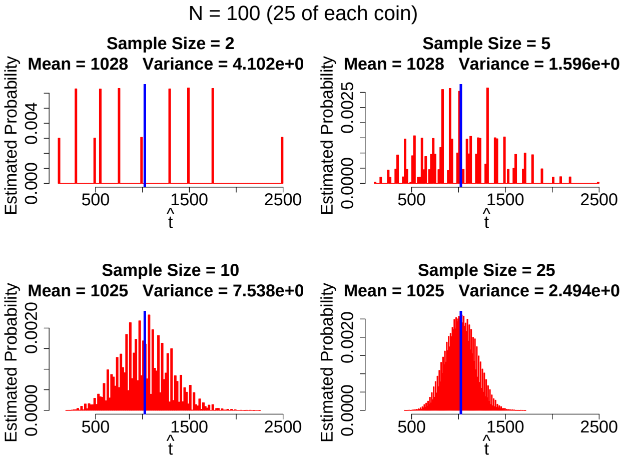
Example, cont.



Mean = 41

Variance = 441

Example, cont.



Bias: Simple Random Sample

Recall Expectation for a mean:

$$\mathrm{E}\left[\bar{y}\right] = \sum_k \bar{y}_k \Pr\left(\mathcal{S} = \mathcal{S}_k\right)$$

Definition of Bias for a mean:

$$\mathrm{Bias}\left[\bar{y}\right] = \mathrm{E}\left[\bar{y}\right] - \bar{y}_U$$

- $\bar{y}$ is an unbiased estimator of $\bar{y}_U$
- $\hat{p}$ is an unbiased estimator of p
- $\hat{t}$ is an unbiased estimator of t

Variance: Simple Random Sample

Definition of variance of a mean:

$$
\begin{aligned}
\mathrm{V}[\bar{y}] &= \mathrm{E}\left[(\bar{y} - \mathrm{E}[\bar{y}])^2\right] \\
&= \sum_k (\bar{y}_k - \mathrm{E}[\bar{y}])^2 \Pr(\mathcal{S} = \mathcal{S}_k)
\end{aligned}
$$

Variance: Simple Random Sample

Estimator	True Variance	Estimated Variance
Mean, $\bar{y}$	$\frac{S^2}{n}\left(1-\frac{n}{N}\right)$	$\frac{s^2}{n}\left(1-\frac{n}{N}\right)$
Proportion, $\hat{p}$	$\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$	$\frac{\hat{p}(1-\hat{p})}{n-1}\left(1-\frac{n}{N}\right)$
Total, $\hat{t}$	$N^2\left(\mathrm{V}\left[\bar{y}\right]\right)$	$N^2\left(\hat{\mathrm{V}}\left[\bar{y}\right]\right)$

$$S^2 = \frac{1}{N-1}\sum_{i\in\mathcal{U}}\left(y_i-\bar{y}_U\right)^2, \qquad s^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(y_i-\bar{y}\right)^2$$

Finite Population Correction

Note that the difference between the variance of a mean for an experiment (infinite population)

$$\mathrm{V}[\bar{y}] = s^2/n$$

And the variance of a mean for a survey sample (finite population)

$$\mathrm{V}[\bar{y}] = s^2/n \left(1 - \frac{n}{N}\right)$$

is the **Finite Population Correction**:

$$\left(1 - \frac{n}{N}\right)$$

Inference: Confidence Intervals

- Central Limit Theorem for large n, N and N-n

$$\frac{\bar{y} - \bar{y}_{\mathcal{U}}}{\sqrt{\widehat{\mathrm{V}[\bar{y}]}}} \sim \mathrm{Normal}(0, 1)$$

- 100(1-α)% Confidence Interval:

$$\bar{y} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{V}[\bar{y}]}} \quad \text{OR} \quad \bar{y} \pm t_{n-1,\alpha/2}\sqrt{\widehat{\mathrm{V}[\bar{y}]}}$$

- Margin of Error $=$ 1/2 the width of a 95% CI

$$z_{\alpha/2}\sqrt{\widehat{\mathrm{V}[\bar{y}]}} \quad \text{OR} \quad t_{n-1,\alpha/2}\sqrt{\widehat{\mathrm{V}[\bar{y}]}}$$

Example

Convert OFHS to a toy SRS example:

- Only use data from Butler County
- Pretend completed interviews = sampled households
- Pretend we know the population size N = 123,082 (from Census)

Example Code

- Mean
  - by hand
  - using built-in tools
- Proportion (two ways)
- Total (whoops - this doesn't work!)

An Alternative Formulation

$$\begin{aligned}
\bar{y} &= \frac{1}{n}\sum_{i\in\mathcal{S}} y_i &&= \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= \frac{\frac{N}{n}\sum_{i=1}^{n} y_i}{N} &&= \frac{\sum_{i=1}^{n}\frac{N}{n} y_i}{\sum_{i=1}^{n}\frac{N}{n}} \\
&= \frac{\sum_{i=1}^{n}\frac{N}{n} y_i}{\sum_{i=1}^{n}\frac{N}{n}} &&= \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}
\end{aligned}$$

$$w_i = \frac{N}{n} \quad \text{for all } i \in \mathcal{S}$$

What's a Weight?

- For SRS, $w_i = N/n$.
- Inverse probability of being selected (n/N)
- Number of units in the population that sampled unit i represents
  - The sum of the weights = population size
  - $\sum_{i \in \mathcal{S}} w_i = \sum_{i=1}^{n} \frac{N}{n} = n\frac{N}{n} = N$

Nice properties for means

For a SRS:
Multiplying by a constant does not change the point estimation.

Let $w_i^* = cw_i$

$$\frac{\sum_{i=1}^{n} w_i^* y_i}{\sum_{i=1}^{n} w_i^*} = \frac{\sum_{i=1}^{n} cw_i y_i}{\sum_{i=1}^{n} cw_i} = \frac{c\sum_{i=1}^{n} w_i y_i}{c\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

Example Code

- Exact weights
- Proportional weights

Part III

Simple Random Sampling: Ratios, Domains, Regression

Unit-Level Ratios

Define a new variable $z_i = y_i / x_i$

$$\bar{z}_U = \frac{1}{N} \sum_{i=1}^{N} z_i$$

- Proportion of income spent on health care in each family
  - y_i = income spent on health care
  - x_i = total family income
- Health care costs per person in each family
  - y_i = health care costs
  - x_i = number of people in family
- Proportion of household members that are under 18
  - y_i = number of children in household
  - x_i = number of people in household

Population-level ratios

$$B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$$

- Proportion of population income spent on health care in Ohio
  - y_i = income spent on health care
  - x_i = total family income
- Total health care costs per Ohioan
  - y_i = health care costs
  - x_i = number of people in family
- Proportion of Ohio population that is under 18
  - y_i = number of children in household
  - x_i = number of people in household

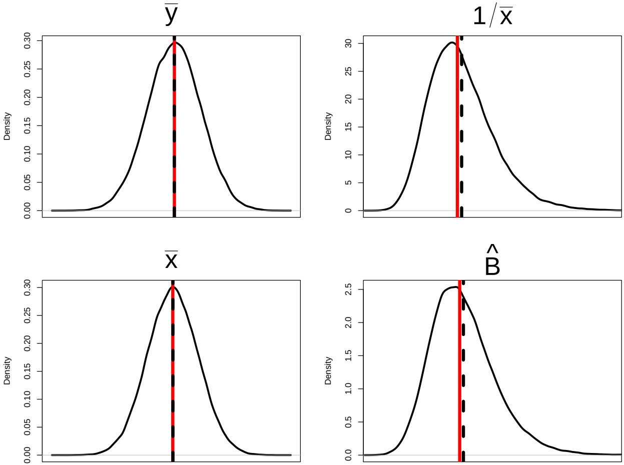
Ratio Estimator

$$\hat{B} = \frac{\bar{y}}{\bar{x}}$$

Attributes:

- Bias
- Variance

Attributes: Bias



Attributes: Bias

Approximated using a Taylor Series Expansion:

$$\text{Bias} = \mathrm{E}\left[\hat{B}\right] - B \approx \left(1 - \frac{n}{N}\right)\left(\frac{1}{n\bar{x}_{\mathcal{U}}^2}\right)\left(BS_x^2 - RS_xS_y\right)$$

$$= \frac{1}{\bar{x}_{\mathcal{U}}^2}\left[B\mathrm{V}\left(\bar{x}\right) - \mathrm{Cov}\left(\bar{x}, \bar{y}\right)\right]$$

$$R = \text{Population Correlation Coefficient}$$

$$= \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}_{\mathcal{U}}\right)\left(y_i - \bar{y}_{\mathcal{U}}\right)}{\left(N-1\right)S_xS_y}$$

Attributes: Bias

Bias is small when:

- The sample size n is large
- The sampling fraction n/N is large
- $\bar{x}_U$ is large
- S_x is small.
- The correlation R between x and y is close to 1

Attributes: MSE

$$\begin{aligned} \text{MSE} &= \mathrm{E}\left[\left(\hat{B} - B\right)^2\right] \\ &\approx \frac{1}{\bar{x}_U^2}\mathrm{E}\left[(\bar{y} - B\bar{x})^2\right] \\ &= \frac{1}{\bar{x}_U^2}\mathrm{V}\left[(\bar{y} - B\bar{x})\right] \end{aligned}$$

Attributes: MSE

Approximate MSE is small when:

- Same criteria for the bias:
  - The sample size n is large
  - The sampling fraction n/N is large
  - $\bar{x}_U$ is large
  - The correlation R between x and y is close to 1
- Deviations about the line $y = Bx$ are small

Attributes: Variance

Using a TS expansion, one can show:

$$V\left[\hat{B}\right] \approx \left(\frac{1}{n\bar{x}_{\mathcal{U}}^2}\right)\left(1-\frac{n}{N}\right)\left[\frac{\sum_{i=1}^{N}(y_i - Bx_i)^2}{(N-1)}\right]$$

$$= \left(\frac{1}{n\bar{x}_{\mathcal{U}}^2}\right)\left(1-\frac{n}{N}\right)\left[S_y^2 - 2BS_xS_yR + B^2S_x^2\right]$$

We estimate the variance:

$$\widehat{V\left[\hat{B}\right]} = \left(\frac{1}{n\bar{x}_{\mathcal{U}}^2}\right)\left(1-\frac{n}{N}\right)\left[\frac{\sum_{i\in\mathcal{S}}\left(y_i - \hat{B}x_i\right)^2}{(n-1)}\right]$$

$$= \left(\frac{1}{n\bar{x}_{\mathcal{U}}^2}\right)\left(1-\frac{n}{N}\right)\left[s_y^2 - 2\hat{B}s_xs_yr + \hat{B}^2s_x^2\right]$$

Example: Ratio

Other Uses for the Ratio Estimator

- Estimate the population total when the population size is unknown
- Increase the precision (decrease the variance) of estimated means and totals
- Adjust estimates to reflect known demographic totals (later)
- Adjust for nonresponse (later)

Estimate the population total when the population size is unknown

You must know the population total of something also measured in the survey: t_x (total Ohio population)

$$t_y = Bt_x = \frac{\bar{y}_U}{\bar{x}_U} t_x = \frac{\frac{1}{N}\sum_{i \in U} y_i}{\frac{1}{N}\sum_{i \in U} x_i} \left(\sum_{i \in U} x_i \right) = \sum_{i \in U} y_i$$

$$\hat{t}_{ry} = \hat{B} t_x$$

Example: Ratio for unknown population size

From Census: Butler County population is 332,807

Increase the precision (decrease the variance) of estimated means and totals

Ratio estimate for a mean:

$$\hat{\bar{y}}_r = \hat{B}\bar{x}_U$$

Estimated variance of ratio estimate:

$$\frac{s_e^2}{n}\left(1 - \frac{n}{N}\right)$$

$$e_i = y_i - \hat{B}x_i$$

Recall the estimated variance for a non-ratio mean:

$$\frac{s_y^2}{n}\left(1 - \frac{n}{N}\right)$$

Example: Ratio for unknown population size

From Census: Butler County has an average of 2.61 persons/household

When do we do better?

$$\text{MSE}\left[\hat{\bar{y}}_r\right] \leq \text{MSE}\left[\bar{y}\right]$$

if and only if

$$R \geq \frac{BS_x}{2S_y} = \frac{\text{CV}(x)}{2\text{CV}(y)}$$

$$\text{CV}(y) = \frac{\sqrt{\text{V}\left[y\right]}}{y}$$

- A straight line through the origin
- Variance of y about the line is proportional to x

Domain Estimation

Domain = subpopulation

Examples:

- Of low income families, what percentage have dental insurance?
- Of households with children, what is the average number of adults?

Domain Estimator

Population quantity:

$$\bar{y}_{U_d} = \frac{1}{N_{U_d}} \sum_{i \in \mathcal{U}_d} y_i$$

Natural estimator:

$$\bar{y}_d = \frac{1}{n_d} \sum_{i \in \mathcal{S}_d} y_i$$

Note that n_d is random!

Domain Estimation as Ratio Estimation

Let:

$$u_i = \begin{cases} y_i \text{ if } i \in \mathcal{U}_d \\ 0 \text{ if } i \notin \mathcal{U}_d \end{cases}$$

$$x_i = \begin{cases} 1 \text{ if } i \in \mathcal{U}_d \\ 0 \text{ if } i \notin \mathcal{U}_d \end{cases}$$

Then:

$$\bar{y}_d = \frac{\sum_{i \in \mathcal{S}_d} u_i}{\sum_{i \in \mathcal{S}_d} x_i}$$

is a ratio estimator.

Example

Regression Mean Estimator

Regression Model:

$$y_i = B_0 + B_1 x_i$$

Predict the population mean:

$$y_U = B_0 + B_1 x_U$$

Estimate the population mean:

$$\hat{\bar{y}}_{reg} = \hat{B}_0 + \hat{B}_1 \bar{x}_U$$

Coefficient Estimates

Estimates of coefficients are the same as in usual linear regression:

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$$

$$\hat{B}_1 = \frac{\sum_{i \in \mathcal{S}} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i \in \mathcal{S}} (x_i - \bar{x})^2}$$

Attributes

$$\begin{aligned} \text{bias} &= -\text{Cov}\left(\hat{B}_1, \bar{x}\right) \\ \text{variance} &= \frac{s_e^2}{n}\left(1 - \frac{n}{N}\right) \\ &\quad e_i = y_i - \left(\hat{B}_0 + \hat{B}_1 x_i\right) \end{aligned}$$

MSE

$$\begin{aligned} \text{MSE}\left(\hat{\bar{y}}_{reg}\right) &\approx \left(1-\frac{n}{N}\right)\frac{1}{n}S_y^2\left(1-R^2\right) \\ &= \frac{S_d^2}{n}\left(1-\frac{n}{N}\right) \\ & d_i = y_i - \left[\bar{y}_U + B_1\left(x_i - \bar{x}_U\right)\right] \end{aligned}$$

Approximate MSE is small when:

- The sample size n is large
- The sampling fraction n/N is large
- The correlation R between x and y is close to 1 **or -1**
- S_y is small

Example

Summary of Mean Estimation Variance

$$\hat{\mathrm{V}}\left[\bar{y}\right] = \left(1 - \frac{n}{N}\right)\frac{s_y^2}{n}$$

$$\hat{\mathrm{V}}\left[\hat{\bar{y}}_r\right] = \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}; \quad e_i = y_i - \hat{B}x_i$$

$$\hat{\mathrm{V}}\left[\hat{\bar{y}}_{reg}\right] = \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}; \quad e_i = y_i - \left(\hat{B}_0 + \hat{B}_1 x_i\right)$$

Summary of Estimation

Mean/Proportion/Total Estimation:

Population Characteristics			
Linear	R	Intercept	Best Estimator
Yes	1	0	Ratio or Regression
Yes	1	$\neq 0$	Regression
Yes	-1	anything	Regression
Yes	0	anything	SRS
No	anything	anything	SRS

Also use ratio estimation to estimate:

- Ratios
- Totals when you don't know the population size

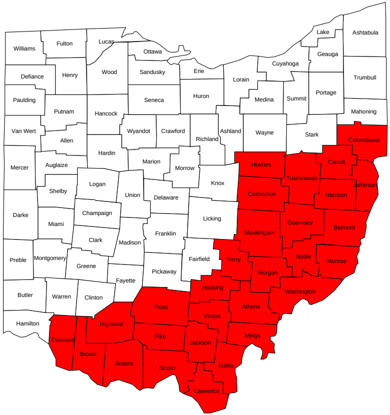
Part IV

Stratified Sampling

Review SRS

- Pros
  - Easy to calculate estimates
  - Easy theory about estimates
- Cons
  - Must know either N or probability of sampling
  - May be inefficient
  - May not represent the population the way you think

"Bad" SRS samples

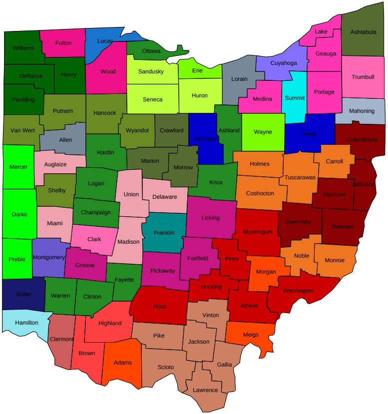


- SRS samples can be "bad"
- Subpopulation estimation
  - Impossible
  - Not useful precision

Strata: Definition

- Strata = Groups
  - Non-overlapping
  - Constitute the whole population
  - We must know the stratification variable for all units in the population before we sample!
- Sampling
  - Independent probability sample within each strata
  - If SRS, called stratified random sampling

Ohio Family Health Survey



Goal: Estimate insurance rates within each county at a certain precision

Problem: Too expensive

Solution: Estimate insurance rates within groups of similar counties

Note that this is different from domain estimation.

Attributes

- This is different from domain estimation because in this case, the subgroup membership is known in advance.
- Stratified sampling is like doing a whole bunch of SRS and then putting them together to analyze.
- This works because the sampling within each of the strata is done independently from each other.
- **Does not produce an SRS overall!**

Estimands and Estimators: Within Strata

Within each stratum, estimators are the same as before:

Strata numbered $h = 1 \ldots H$

y_{hj} = value of jth unit in stratum h

Within-Stratum h Population	Within-Stratum h Estimate
$\bar{y}_{Uh} = \frac{1}{N_h} \sum_{j \in \mathcal{U}_h} y_j = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$	$\bar{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_j = \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$
$p_h = \frac{1}{N_h} \sum_{j \in \mathcal{U}_h} z_j = \frac{1}{N_h} \sum_{j=1}^{N_h} z_{hj}$	$\hat{p}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} z_j = \frac{1}{n_h} \sum_{j=1}^{n_h} z_{hj}$
$t_h = \sum_{j \in \mathcal{U}_h} y_j = \sum_{j=1}^{N_h} y_{hj}$	$\hat{t}_h = N_h \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_j = N_h \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}$

Estimands: Across Strata

	Stratum h	Whole Pop.	Relationship
size	N_h	N	$N = \sum_{h=1}^{H} N_h$
total	$t_h = \sum_{j=1}^{N_h} y_{hj}$	t	$t = \sum_{h=1}^{H} t_h$
mean	$\bar{y}_{h\mathcal{U}} = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$	$\bar{y}_{\mathcal{U}}$	$\bar{y}_{\mathcal{U}} = \frac{t}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{h\mathcal{U}}$
prop	$p_h = \frac{1}{N_h}(\#\text{ with attribute}) = \frac{1}{N_h} \sum_{j=1}^{N_h} z_{hj}$	p	$p = \sum_{h=1}^{H} \frac{N_h}{N} p_h$

Estimators: Across Strata

	Stratum h	Whole Samp.	Relationship
size	n_h	n	$n = \sum_{h=1}^{H} n_h$
total	$\hat{t}_h = N_h \bar{y}_h = \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}$	$\hat{t}_{str}$	$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h$
mean	$\bar{y}_h = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj}$	$\bar{y}_{str}$	$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$
prop	$\hat{p}_h = \frac{1}{n_h}(\#\text{ with attribute}) = \frac{1}{n_h} \sum_{j \in \mathcal{S}_h} x_{hj}$	$\hat{p}_{str}$	$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$

Expected Value/Variance Review

$$
\begin{aligned}
\mathrm{E}[X+Y] &= \mathrm{E}[X] + \mathrm{E}[Y] \\
\mathrm{E}[aX+b] &= a\mathrm{E}[X] + b \\
\\
\mathrm{V}[X+Y] &= \mathrm{V}[X] + \mathrm{V}[Y] \\
&\quad +2\mathrm{Cov}[X,Y] \\
\mathrm{V}[aX+b] &= a^2\mathrm{V}[X]
\end{aligned}
$$

X and Y are random variables
a and b are constants

Bias

Example: Mean

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

$$\begin{aligned} E\left[\bar{y}_{str}\right] &= E\left[\frac{1}{N}\hat{t}_{str}\right] && \text{definition} \\ &= \frac{1}{N} E\left[\hat{t}_{str}\right] && N \text{ is constant} \\ &= \frac{1}{N} t && \hat{t}_{str} \text{ is unbiased} \\ &= \bar{y}_{\mathcal{U}} && \text{definition} \end{aligned}$$

Sample mean, proportion, and total are unbiased for population mean, proportion and total.

Variance

Example: Mean

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

$$\begin{aligned} V\left[\bar{y}_{str}\right] &= V\left[\frac{1}{N}\hat{t}_{str}\right] && \text{definition} \\ &= \frac{1}{N^2} V\left[\hat{t}_{str}\right] && \text{N is constant} \\ &= \frac{1}{N^2}\sum_{i=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} && \text{variance of } \hat{t}_{str}\text{, } \textbf{independence} \\ &= \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} && \text{algebra} \end{aligned}$$

$$\widehat{V\left[\bar{y}_{str}\right]} = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$$

Compare Variance to SRS

Recall the variance of a mean for SRS:

$$\left(1 - \frac{n}{N}\right) \frac{S^2}{n}$$

Variance for of a mean using stratified sampling:

$$\sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

S_h tends to be smaller than S. This usually leads to a reduced variance for the population-wide estimate.

Toy Example

Population:

1	2	3	4
11	12	13	14
21	22	23	24
31	32	33	34

Stratify by row or column?

Toy Example, cont.

Recall:

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} (y_{hi} - \bar{y}_h)^2$$

Population:

					$S^2_{h,row}$
	1	2	3	4	1.7
	11	12	13	14	1.7
	21	22	23	24	1.7
	31	32	33	34	1.7
$S^2_{h,col}$	167	167	167	167	

Toy Example, cont.

Recall:

$$V[\bar{y}] = \sum_{h=1}^{H} \frac{N_h^2}{N^2}\left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h}$$

Suppose I take a sample of size 2 from each stratum:

Row: $V(\bar{y}_{str}) = \sum_{i=1}^{4} \frac{4^2}{16^2}\left(1 - \frac{2}{4}\right)\frac{1.7}{2} = 0.106$

Column: $V(\bar{y}_{str}) = \sum_{i=1}^{4} \frac{4^2}{16^2}\left(1 - \frac{2}{4}\right)\frac{167}{2} = 10.4$

Ideal Strata

Ideal strata have small within-strata population variance

- Implies small within-strata sample variance
  - Means units within strata are similar to each other with respect to the survey questions
  - homogeneous within
- Implies large between-strata population variance
  - Mean units within different strata are different from each other with respect to the survey questions
  - heterogeneous between

Example: Ohio Family Health Survey

Weights Revisited

Recall weights from SRS:

- For SRS, $w_i = N/n$.
- Inverse probability of being selected (n/N)
- Number of units in the population that sampled unit i represents
  - The sum of the weights = population size
  - $\sum_{i \in \mathcal{S}} w_i = \sum_{i=1}^{n} \frac{N}{n} = n\frac{N}{n} = N$

Using weights to estimate a mean:

$$\bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

Weights for Stratified Sample

Within each stratum, we have a SRS, and so the weight is the same as for the SRS:

- $w_{hj} = \dfrac{N_h}{n_h}$.
- Inverse probability of being selected (n_h/N_h)
- Number of units in the population that sampled unit represents
  - The sum of the weights = population size
  - $\displaystyle\sum_{h=i}^{H}\sum_{j=1}^{n_j} w_{hj} = \sum_{h=i}^{H}\sum_{j=1}^{n_j} \frac{N_h}{n_h} = \sum_{h=i}^{H} n_h \frac{N_h}{n_h} = \sum_{h=i}^{H} N_h = N$

Using weights to estimate a mean

$$
\begin{aligned}
\bar{y}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h &= \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{N} \\
&= \frac{\sum_{h=1}^{H} N_h \frac{1}{n_h} \sum_{j=1}^{n_h} y_{hj}}{N} \\
&= \frac{\sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{N_h}{n_h} y_{hj}}{N} \\
&= \frac{\sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{N_h}{n_h} y_{hj}}{\sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{N_h}{n_h}} \\
&= \frac{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj}}
\end{aligned}
$$

Using weights without strata notation

$$\bar{y}_{str} = \frac{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j=1}^{n_h} w_{hj}} = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i}$$

where

$$w_i = \frac{N_{h_i}}{n_{h_i}}$$

and h_i is the stratum to which the unit i belongs.

County Stratification in the OFHS

"ODJFS and ODH set a new statistical constraint for the sampling methodology: that counties, or clusters of similar counties, have sufficient sample size to produce reliable estimates of the health insurance status of children under the age of eighteen, with a sampling error of no more than $\pm$ 5% at the 95% level of confidence. ORC Macro calculated that with approximately 35% of households across the State of Ohio containing at least one child, and taking into account estimates of child health insurance status from the 1998 FHS, a sample size of 800 completed interviews would be necessary in counties, or county clusters."

County-Level Estimation Constraints:

- A minimum of 800 completed interviews in each stratum (county or county cluster)
- A minimum of 50 completed interviews in each county

Example

Summary

Stratification pros:

- Protect yourself from a really bad sample
- Convenient to administer
- Obtain data of specified precision for subgroups
- Smaller variance of estimates

Pseudo-Stratification in the OFHS

The OFHS also sampled extra Hispanic and Asian households:

- Created two additional lists of telephone numbers associated with traditionally Hispanic and Asian surnames.
- Independently sampled telephone numbers from these lists.

Are these true strata capturing the Hispanic and Asian population?

Do I need to use domain estimation to estimate the percentage of Hispanic households that have dental insurance?

Part V

Cluster Sampling

The Elephant in the OFHS Room

So far we have been making inferences only about households. Remember why:

- Assume each household in Ohio has exactly one telephone
- Then, the sampling unit (telephone number) is identical to the household
- Household is the effective sampling unit
- We have only learned how to make inference about the sampling unit

What proportion of Ohio adults have health insurance?

Appropriate question:

> *A1: Are you covered by health insurance or some other type of health care plan?"*

Estimates we have learned all focus on estimation for the household. This question is about an individual within a household.

Definition: Cluster Sampling

- Divide population into groups
  - non-overlapping
  - constitute whole population
- Select n of these groups
- Sample every unit within the selected groups

Example: National Immunization Survey

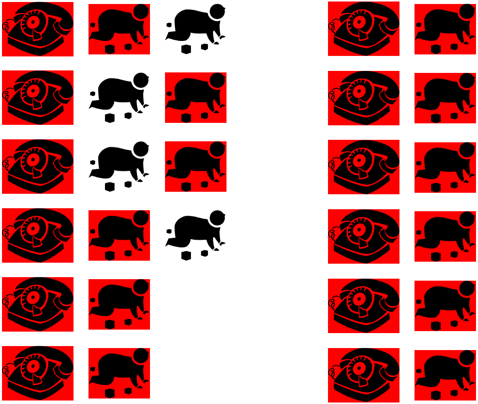
NIS is interested in the immunizations of all children aged 19-35 months.

NIS is a random digit dial telephone survey, stratified by state and major city.

In each selected household, information is collected about all the resident children in the given age range.

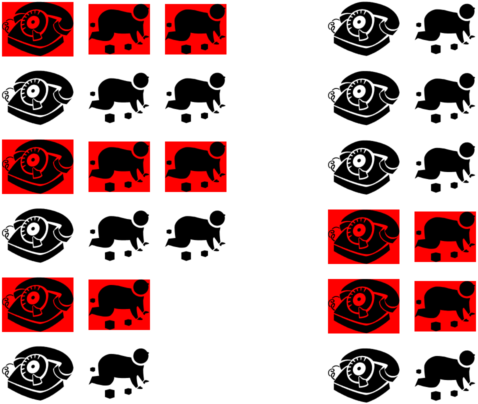
Toy Example: Stratified Sampling

Strata = Household



Toy Example: Cluster Sampling

Cluster = Household



Pros and Cons

- Pros
  - Convenient (time/money)
  - Eliminates the need to have a sampling frame that actually includes all observation units
    - Only need a list of clusters (sampling units)
      Sample household instead of person
      Sample city block instead of person
      Sample class instead of students
- Cons
  - Increases the variance of estimates because the observation units included in the sample are not independent

Notation

N	# clusters in population
n	# sampled clusters
M_i	# units in cluster i
$K = \sum_{i=1}^{N} M_i$	# units in the population
$\sum_{i \in \mathcal{S}} M_i$	total # units in the sample

Estimands

		y_{ij}	value for the j^{th} unit of the i^{th} cluster
t_i	$=$	$\sum_{j=1}^{M_i} y_{ij}$	total of units in the i^{th} cluster
$\bar{y}_{i\mathcal{U}}$	$=$	$\frac{t_i}{M_i}$	mean of units in the i^{th} cluster
t	$=$	$\sum_{i=1}^{N} t_i$	population total
	$=$	$\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$	
$\bar{t}_{\mathcal{U}}$	$=$	$\frac{t}{N}$	average cluster total
$\bar{y}_{\mathcal{U}}$	$=$	$\frac{t}{K}$	population mean per unit

Estimators

Population Value	Sample Estimator
t	$\hat{t} = N\bar{t} = N\left(\frac{1}{n}\sum_{i \in \mathcal{S}} t_i\right) = N\left(\frac{1}{n}\sum_{i \in \mathcal{S}}\sum_{j=1}^{M} y_{ij}\right)$
$\bar{y}_{\mathcal{U}}$	$\hat{\bar{y}} = \frac{\text{estimated total}}{\text{number of units in population}} = \frac{\hat{t}}{K} = \frac{\hat{t}}{\sum_{i=1}^{N} M_i}$

$\hat{t}$ and $\hat{\bar{y}}$ are unbiased estimators for t and $\bar{y}_{\mathcal{U}}$.

Simplification

Consider the situation where all the clusters are of equal size.

Estimator	Variance	Estimated Variance
$\hat{t} = N\left(\frac{1}{n}\sum_{i\in\mathcal{S}}\sum_{j=1}^{M} y_{ij}\right)$	$N^2\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}$	$N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}$
$\hat{\bar{y}} = \frac{\hat{t}}{K} = \frac{\hat{t}}{NM}$	$\frac{1}{M^2}\left(1-\frac{n}{N}\right)\frac{S_t^2}{n}$	$\frac{1}{M^2}\left(1-\frac{n}{N}\right)\frac{s_t^2}{n}$

where

$$S_t^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(t_i - \frac{t}{N}\right)^2$$

and

$$s_t^2 = \frac{1}{n-1}\sum_{i\in\mathcal{S}}\left(t_i - \frac{\hat{t}}{N}\right)^2$$

Toy Example Revisited

Population:

1	2	3	4
11	12	13	14
21	22	23	24
31	32	33	34

Cluster by row or column?

Toy Example, cont.

Recall:

$$t_i = \sum_{j=1}^{M_i} y_{ij}$$

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(t_i - \frac{t}{N} \right)^2$$

Population:

					$t_{i,row}$
	1	2	3	4	10
	11	12	13	14	50
	21	22	23	24	90
	31	32	33	34	130
$t_{i,col}$	64	68	72	76	t=280

Toy Example, cont.

Recall:

$$\mathrm{V}\left[\hat{\bar{y}}\right] = \frac{1}{M^2}\left(1 - \frac{n}{N}\right)\frac{S_t^2}{n}$$

Suppose I take a sample of $n = 2$ clusters (each of size $M = 4$):

$$\begin{aligned} \textit{Row}: \quad S_t^2 &= \tfrac{1}{4-1}\textstyle\sum_{i=1}^4 \left(t_i - \tfrac{280}{4}\right)^2 = 8000/3 \\ \mathrm{V}\left[\hat{\bar{y}}\right] &= \tfrac{1}{4^2}\left(1 - \tfrac{2}{4}\right)\tfrac{8000/3}{2} = 41.7 \end{aligned}$$

$$\begin{aligned} \textit{Column}: \quad S_t^2 &= \tfrac{1}{4-1}\textstyle\sum_{i=1}^4 \left(t_i - \tfrac{280}{4}\right)^2 = 80/3 \\ \mathrm{V}\left[\hat{\bar{y}}\right] &= \tfrac{1}{4^2}\left(1 - \tfrac{2}{4}\right)\tfrac{80/3}{2} = 0.417 \end{aligned}$$

Ideal Clusters

Ideal strata have large within-cluster population variance

- Implies large within-cluster sample variance
  - Means units within clusters are different from each other with respect to the survey questions
  - Heterogeneous within
- Implies small between-cluster population variance
  - Means cluster-wide values are similar to each other with respect to the survey questions
  - Homogeneous between

Cluster Variance Attributes

- Cluster sample variances are usually larger than variances associated with a SRS of the same size
  - Members of a cluster are usually similar (homogeneous within)
- Relative precision: (Note: $n_{SRS} = n_{clust}M$ and $N_{SRS} = N_{clust}M$)

$$\frac{V\left[\hat{\bar{y}}_{cluster}\right]}{V\left[\bar{y}\right]} = \frac{\frac{1}{M^2}\left(1 - \frac{n_{clust}}{N_{clust}}\right)\frac{S_t^2}{n_{clust}}}{\left(1 - \frac{n_{SRS}}{N_{SRS}}\right)\frac{S_y^2}{n_{SRS}}} = \frac{S_t^2/M}{S_y^2}$$

- However, one can often afford to collect data on more units, so the increased sample size may offset the cluster effect

Relaxing the assumption: Unequal Size Clusters

- Same estimators for t and $\bar{y}_U$ as before
- The variance of these is likely to be large:

$$\hat{V}\left[\hat{t}\right] = N^2 \left(1 - \frac{n_{clust}}{N_{clust}}\right) \frac{S_t^2}{n_{clust}}$$

$$S_t^2 = \frac{1}{N-1} \sum_{i=1}^{N} (t_i - \bar{t})^2$$

  - Totals are likely to vary a lot over unequal size clusters
- Alternative: Ratio Estimation

Ratio Estimation Revisited

$$B = \frac{t_y}{t_x} = \frac{\bar{y}_U}{\bar{x}_U} = \frac{\sum_{i=1}^N y_i}{\sum_{i=1}^N x_i}$$

x_i are now M_i

y_i are now t_i

$$\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} t_i}{\sum_{i \in \mathcal{S}} M_i} = \frac{\bar{t}}{\bar{M}}$$

$$\hat{t}_r = K \hat{\bar{y}}_r$$

Variance of Ratio Estimates

$$\mathrm{V}\left[\hat{B}\right] \approx \left(\frac{1}{n\bar{M}_{\mathcal{U}}^2}\right)\left(1-\frac{n}{N}\right)\left[\frac{\sum_{i=1}^{N}\left(t_i-\frac{\bar{t}}{\bar{M}}M_i\right)^2}{(N-1)}\right]$$

$$= \left(\frac{1}{n\bar{M}_{\mathcal{U}}^2}\right)\left(1-\frac{n}{N}\right)\left[\frac{\sum_{i=1}^{N}M_i^2\left(\frac{t_i}{M_i}-\frac{\bar{t}}{\bar{M}}\right)^2}{(N-1)}\right]$$

The variance now involves differences of averages rather than differences of totals, and so wont be affected by different sizes of clusters.

Bias/Variance Trade-Off

- Recall that ratio estimates are biased
- We may prefer to give up a little bias to get a lot less variance (Overall MSE smaller)
- Relative precision:

$$\frac{V\left[\hat{\bar{y}}_r\right]}{V\left[\hat{\bar{y}}_{clust}\right]} = \frac{\sum_{i=1}^{N}\left(t_i - \left(\bar{t}_{\mathcal{U}}/\bar{M}\right)M_i\right)^2}{\sum_{i=1}^{N}\left(t_i - \bar{t}_{\mathcal{U}}\right)^2}$$

Example: National Immunization Survey

Two-Stage Cluster Sampling

In the OFHS, the sampling within each stratum is as follows:

- Select a household (via a telephone number)
- Select an adult within that household (via most recent birthday method)

In other words:

- Select a cluster (called a Primary Sampling Unit [PSU])
- Select units within that cluster (called Secondary Sampling Units [SSUs])

This is two-stage sampling.

Estimands and Estimators

Unlike one-stage sampling, we must now **estimate** the within-cluster totals t_i with

$$\hat{t}_i = \sum_{j \in \mathcal{S}_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i$$

where m_i is the number of units sampled from cluster i.

$$t = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}$$

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in \mathcal{S}} \hat{t}_i$$

$$\bar{y}_{\mathcal{U}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij}}{\sum_{i=1}^{N} M_i}$$

$$\hat{\bar{y}}_{unb} = \frac{\sum_{i \in \mathcal{S}} M_i \bar{y}_i}{n \bar{M}} = \frac{\sum_{i \in \mathcal{S}} \hat{t}_i}{n \bar{M}}$$

$$\hat{\bar{y}}_r = \frac{\sum_{i \in \mathcal{S}} \hat{t}_i}{\sum_{i \in \mathcal{S}} M_i}$$

Sources of Variability

Now there are two sources of variability:

1. Between PSUs
2. Within each PSU (estimation of $\hat{t}_i$)

Estimated Variance

$$\hat{V}\left(\hat{t}_{unb}\right) = N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i=1}^{N}\left(1-\frac{m_i}{M_i}\right)M_i^2\frac{s_i^2}{m_i}$$

$$\hat{V}\left(\hat{\bar{y}}_{unb}\right) = \frac{\widehat{V\left(\hat{t}_{2stage}\right)}}{K^2}$$

$$\hat{V}\left(\hat{\bar{y}}_r\right) = \frac{1}{\bar{M}^2}\left[\left(1-\frac{n}{N}\right)\frac{s_r^2}{n} + \frac{1}{nN}\sum_{i\in\mathcal{S}}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{s_i^2}{m_i}\right]$$

$$s_t^2 = \frac{\sum_{i\in\mathcal{S}}\left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2}{n-1} \qquad s_i^2 = \frac{\sum_{j\in\mathcal{S}_i}\left(y_{ij}-\bar{y}_i\right)^2}{m_i-1} \qquad s_r^2 = \frac{\sum_{i\in\mathcal{S}}\left(M_i\bar{y}_i - M_i\hat{\bar{y}}_r\right)^2}{n-1}$$

Using Weights for Cluster Estimation

We can use weights for clusters just as for strata.

Recall that the weight is the inverse probability of selection.

$$w_{ij} = 1/\pi_{ij}$$

Weights

- Single-Stage Cluster Sample:
  - $\pi_{ij} = \Pr(\text{Cluster } i \text{ chosen}) = n/N$
  - same as SRS
  - $w_{ij} = 1/\pi_{ij} = N/n$
- Two-Stage Cluster Sample:
  - $\pi_{ij} = \Pr(\text{unit } ij \text{ chosen}|\text{Cluster } i \text{ chosen}) \Pr(\text{Cluster } i \text{ chosen}) = (m_i/M_i)(n/N) = (nm_i)/(NM_i)$
  - $w_{ij} = 1/\pi_{ij} = (NM_i)/(nm_i i)$

Using Weights for Cluster Estimation of a Mean

Using weights is identical to the ratio estimator:

$$\begin{aligned}\hat{\bar{y}}_{weight} &= \frac{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}y_{ij}}{\sum_{i\in\mathcal{S}}\sum_{j\in\mathcal{S}_i} w_{ij}} \\ = \hat{\bar{y}}_r &= \frac{\sum_{i\in\mathcal{S}}\hat{t}_i}{\sum_{i\in\mathcal{S}} M_i}\end{aligned}$$

Example

Intra Cluster Correlation

ICC quantifies the homogeneity within clusters.

Assuming equal sized clusters:

$$
\begin{aligned}
ICC &= 1 - \frac{M}{M-1} \frac{\text{sum of squares within clusters}}{\text{total sum of squares}} \\
&= 1 - \frac{M}{M-1} \frac{\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_{i\mathcal{U}})^2}{\sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_{\mathcal{U}})^2}
\end{aligned}
$$

Relative Precision:

$$
\frac{\mathrm{V}\left[\hat{t}_{clust}\right]}{\mathrm{V}\left[\hat{t}_{SRS}\right]} = \frac{NM-1}{M(N-1)} \left[1 + (M-1)\mathrm{ICC}\right]
$$

Summary

- Pros
  - Convenient (time/money)
  - Eliminates the need to have a sampling frame that actually includes all observation units
- Cons
  - Increases the variance of estimates because the observation units included in the sample are not independently sampled
  - However, this increase in variance may be offset by a cheaply increased sample size

Review of Grouping

	Strata	Clusters
Ideal within	Homogeneous	Heterogeneous
Ideal between	Heterogeneous	Homogeneous
Design for ideal	Usually try to	Usually not
Variance	Decreased	Increased over SRS of same size
Design for Subpopulation Estimation	Yes	No

Part VI

Model Oriented Estimation

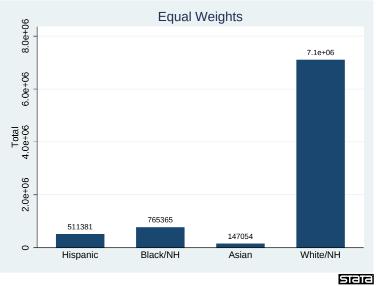
Exploratory Data Analysis

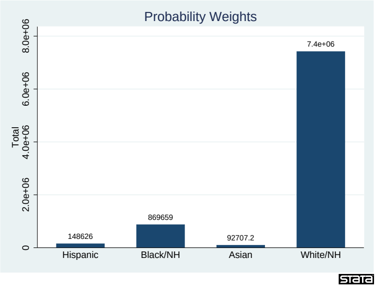
Before analyzing data (especially using a model-oriented approach), it is a good idea to look at the data. Standard plotting methods do not consider the sampling scheme.

Consider:

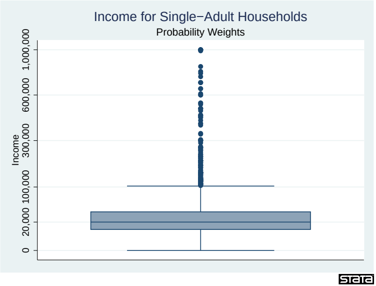
- Bar Plots
- Box Plots
- Scatter Plots

Bar Plots
Effect of Weights

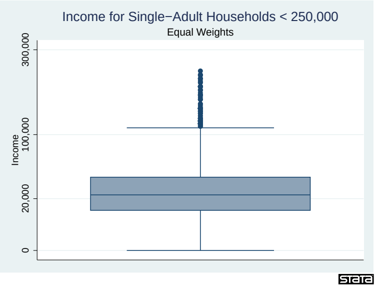


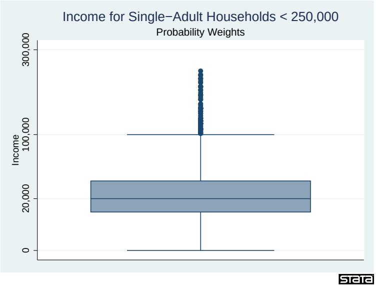


Box Plots
Effect of Weights

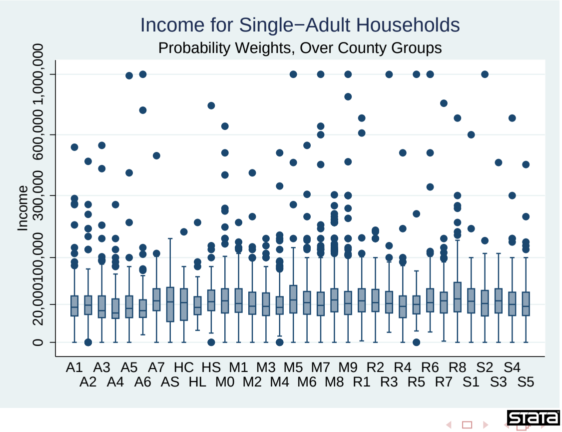


Box Plots
Effect of Weights

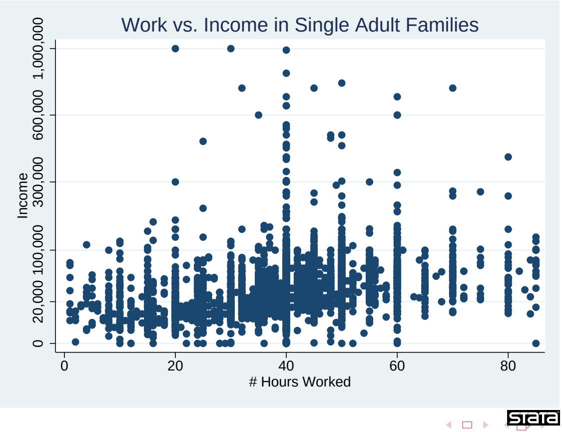




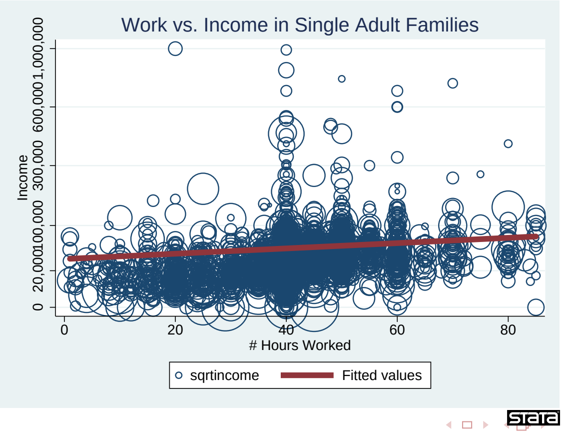
Box Plots
Values over Grouping



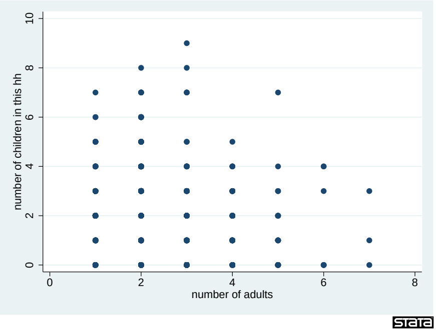
Scatter Plots
Unweighted



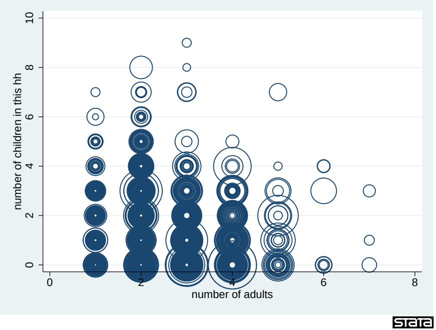
Scatter Plots
Unweighted



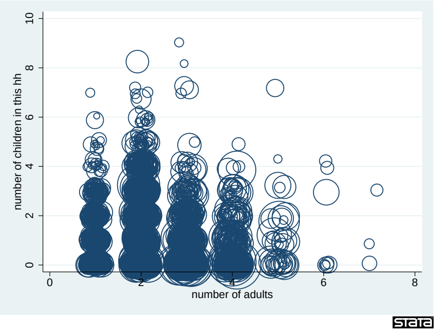
Scatter Plots
Ordinal Variables



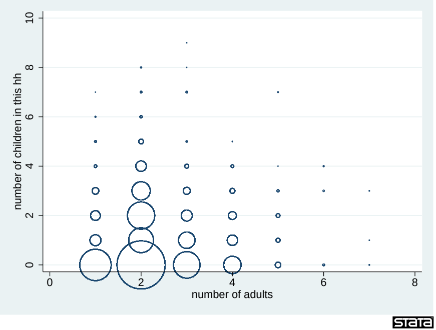
Scatter Plots
Ordinal Variables



Scatter Plots
Ordinal Variables



Scatter Plots
Ordinal Variables



Linear Regression: Finite Population

1. Estimands:
   - $$B_1 = \frac{\sum_{i \in \mathcal{U}} (x_i - \bar{x}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}})}{\sum_{i \in \mathcal{U}} (x_i - \bar{x}_{\mathcal{U}})^2}$$
   - $$B_0 = \bar{y}_{\mathcal{U}} - B_1 \bar{x}_{\mathcal{U}}$$
2. These are the coefficients for the best (least squares) fit straight line through the population values.
3. Coefficient and variance estimates are as before, using weights and survey structure.

Linear Regression

Probability weights are not the same as the weights used in weighted least squares!

Linear Regression: Super-Population

Instead of considering a finite population, you might instead consider a population-generation process that could generate an infinite number of populations. This puts us back into the realm of infinite populations.
Now, we need the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The estimands are β_0 and β_1, and the variability is no longer primarily from the sampling process. The role of the design in estimating these superpopulation values is still hotly debated.

Logistic Regression

Logistic regression is also possible under finite population sampling.

The estimand is again the best logistic fit to the population data. (e.g., the maximum likelihood estimate given all the population data)

Small Area Estimation

Why:

- Data are available from a large-scale sample survey designed to produce good estimates at a high level (such as State-wide)
- We want estimates for "small areas" or small domains, such as small counties or particular sub-groups (e.g., African-Americans in rural Ohio counties)

Problem:

- How can we produce a reliable estimate for the small area from the available data?

Some Small Area Estimators

Let d identify the small area (domain) of interest.

Suppose we wish to estimate the total number of uninsured African American children in Champaign County.

Methods:

- Direct Estimator
- Synthetic Estimator
- Composite Estimator
- Model-Based Estimator

Direct Estimator

Use the same domain methods we've already covered.

$$\hat{t}_{d,(\text{dir})} = N_i \bar{y}_i = \sum_{i \in \mathcal{S}} w_i y_{i(d)}$$

where $y_{i(d)} = y_i$ if the unit is in the small area and $=0$ otherwise.

- Generally very large variance
- Requires at least two observations in the small area

Synthetic Estimator

Assume that the ratio of two variables is constant across the entire sample (or some large subset of the entire sample). Then, use ratio estimation ideas.

$$\hat{t}_{d,(\mathrm{syn})} = \frac{\sum_{i \in \mathcal{S}} w_i y_i}{\sum_{i \in \mathcal{S}} w_i x_i} t_x = \hat{B} t_x$$

The simplest denominator is the sample size: $x_{i(d)} = 1$ if the unit is in the small area and $= 0$ otherwise.

We could use another value, such as unemployment rates.

- Generally small variance, since it is based on a large sample

Composite Estimators

Combine the direct and synthetic estimators to try to get the benefits of both.

$$\hat{t}_{d,(\text{comp})} = \alpha_d \hat{t}_{d,(\text{dir})} + (1 - \alpha_d)\hat{t}_{d,(\text{syn})}$$

for $0 < \alpha_d < 1$

Challenge: Determining an optimal α_d

- If n_d relatively large, α_d closer to 1
- If n_d relatively small, α_d closer to 0

Model Based Estimators

Use a superpopulation model to "borrow strength" from similar areas. To apply this, you usually divide the population into a number of small samples, not consider just one.

The model can be any structure. One possibility:

$$y_{idk} = \beta_d + \delta_i + \epsilon_{idk}$$

where i represents a 'large area' (e.g., county), d represents the different 'small areas' (e.g., minority groups), and k represents single observations (e.g, adults)

δ_i is a 'large area' random effect
ϵ_{idk} is random variability across observations