National Cancer Institute

measurementERRORwebinar series

Accounting for complex survey design in modeling usual intake

Kevin W. Dodd, PhD
National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Presenters and Collaborators

Sharon Kirkpatrick
Series Organizer

Regan Bailey	Laurence Freedman	Douglas Midthune
Dennis Buckman	Patricia Guenther	Amy Subar
Raymond Carroll	Victor Kipnis	Fran Thompson
Kevin Dodd	Susan Krebs-Smith	Janet Tooze

Wake Forest School of Medicine · The Gertner Institute · Statistics, Texas A&M University

measurementERRORwebinar series

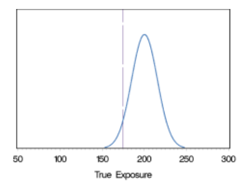
This series is dedicated to the memory of
Dr. Arthur Schatzkin

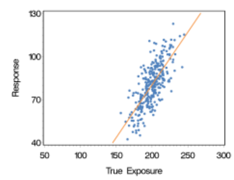
In recognition of his internationally renowned contributions to the field of nutrition epidemiology and his commitment to understanding measurement error associated with dietary assessment.

Introduction

Two main areas of interest

- Describing usual intake distributions: mean, percentiles, proportion above or below a threshold



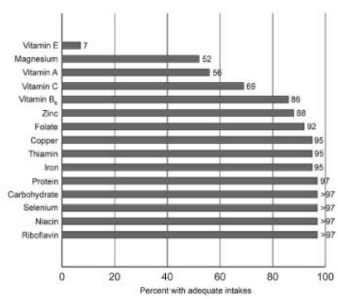


- Estimating diet-health relationships: regression coefficients

Introduction

Data used in population monitoring

Prevalence of adequate usual intake of selected nutrients from food sources only



Source: What We Eat in America, NHANES 2001-02

Introduction

Motivation

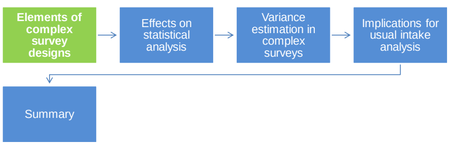
- Previous webinars
  - Focused on methods development/application
  - Skipped over details related to data collection
- This webinar
  - Focuses on details related to data collection
  - Specifically, how collecting data using survey sampling methods affects analysis

Objectives

- Identify considerations in the analysis of dietary data collected as part of a complex survey, including stratification, clustering, and weighting.
- Identify methods of variance estimation for complex survey samples and describe how these are incorporated into the estimation of usual intake distributions.

Outline

- Elements of complex survey designs
- How these elements affect statistical analysis
- Variance estimation in complex surveys
- Implications for usual intake analysis using survey data
- Summary

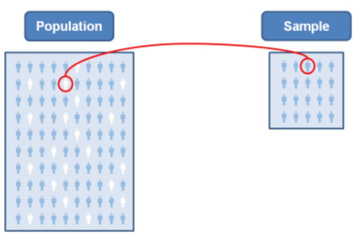


ELEMENTS OF COMPLEX SURVEY DESIGNS

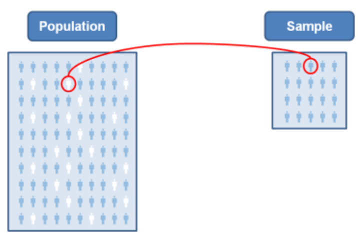
Simple random sampling

- Statistical methods often derived assuming data come from a **simple random sample** (SRS)
- Every member of population (enumerated in the **sampling frame**) equally likely to be sampled
- For small, homogeneous groups simple random samples are practical to obtain and analyze

Selecting a simple random sample



Selecting a simple random sample



- In practice, data are often collected using complex survey methods, not simple random sampling

Why use a complex sampling design?

- Control data collection costs in
  - Drawing the sample
  - Collecting data on sampled individuals
- Improve precision of subpopulation estimates

Elements of complex sampling designs

- Stratification
- Clustering
- Weighting

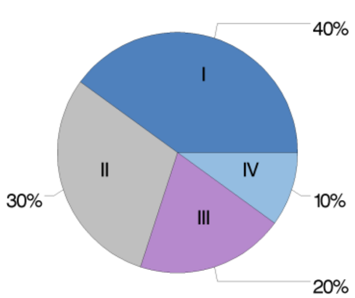
Elements of complex sampling designs

- **Stratification**
- Clustering
- Weighting

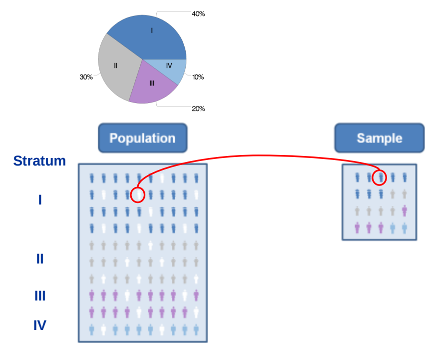
What is stratification?

- Grouping individuals in the population that share specific (generally demographic) characteristics
- Identifies subpopulations of *a priori* interest
  - E.g., pregnant and lactating women, children, low-income individuals

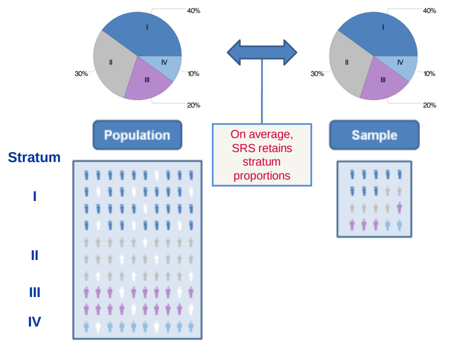
Hypothetical population with four strata



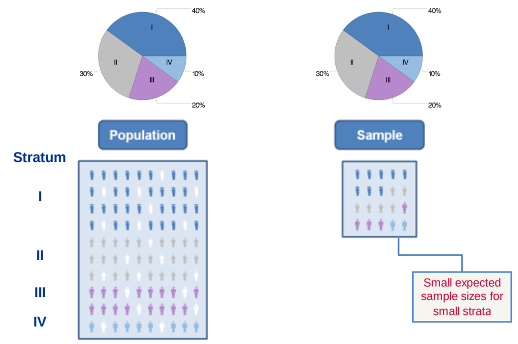
Simple random sample from a stratified population



Simple random sample from a stratified population

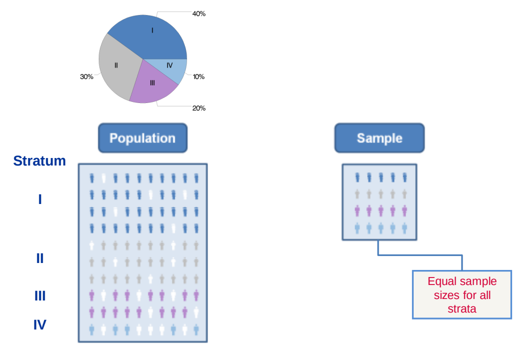


Simple random sample from a stratified population

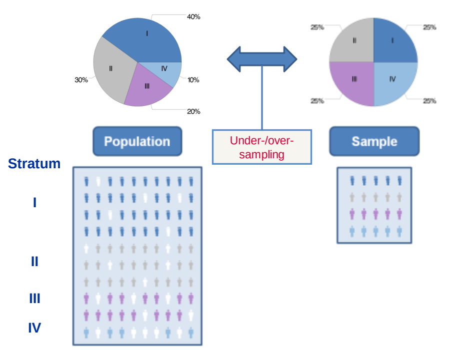


Alternative: balanced stratified random sampling

Alternative: balanced stratified random sampling



Alternative: balanced stratified random sampling



Tradeoffs of stratification

- Balanced sampling across strata yields
  - More precise estimates for small strata,
  - But less precise estimates for large strata
- Stratified sampling need not be balanced
  - If all subpopulations are not of equal interest

Elements of complex sampling designs

- Stratification
- **Clustering**
- Weighting

What is clustering?

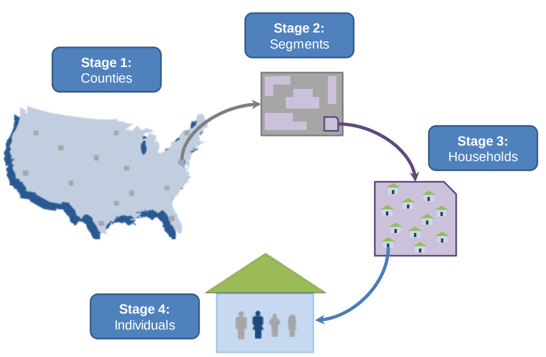
- Sampling of multiple individuals within the same (usually geographic) area
- Helps control data collection costs associated with travel

Stratified cluster sampling

Effects of clustering

- Observations from individuals sampled from the same cluster tend to be correlated
  - Loss of precision
- Multistage designs with several levels of clustering possible
- First-level clusters (Primary Sampling Units; PSUs) tend to induce largest portion of sampling variability

Multistage sampling



Advantages of multistage sampling

- Allows stepwise development of sampling frame:
  - Enumerate counties in the US, then census block groups within selected counties, then households within selected block groups
  - Eliminates the need for master list of households
- Can greatly reduce data collection costs

Elements of complex sampling designs

- Stratification
- Clustering
- **Weighting**

What is weighting?

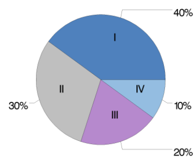
- Indicates how many individuals in the population a sampled individual "represents"

What is weighting?

- Indicates how many individuals in the population a sampled individual "represents"
- Each individual's **sample weight** is equal to the inverse of the final probability of being selected from the population

$$\text{sample weight} = \frac{1}{\text{final probability}}$$

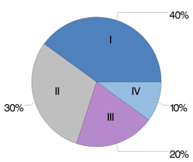
Weighting for a stratified sample of size 100



- Total population size: 1 million

Stratum (Size)	Sample Size	Prob 1000	Weight/10000
I (400K)			
II (300K)			
III (200K)			
IV (100K)			
Total (1M)	100		

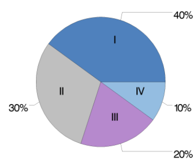
Weighting for a stratified sample of size 100



- Total population size: 1 million
- Want to draw a sample of size 25 from each stratum

Stratum (Size)	Sample Size	Prob 1000	Weight/10000
I (400K)	25		
II (300K)	25		
III (200K)	25		
IV (100K)	25		
Total (1M)	100		

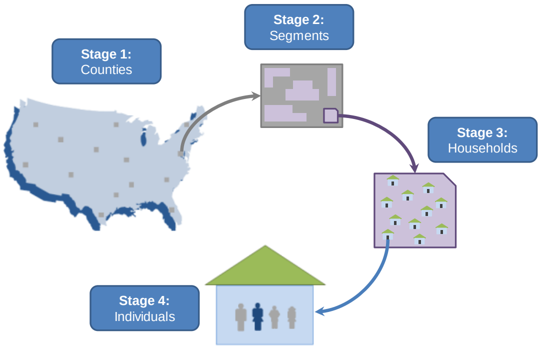
Weighting for a stratified sample of size 100



- Total population size: 1 million
- Want to draw a sample of size 25 from each stratum

Stratum (Size)	Sample Size	Prob 1000	Weight/10000
I (400K)	25	25/400	1.6
II (300K)	25	25/300	1.2
III (200K)	25	25/200	.8
IV (100K)	25	25/100	.4
Total (1M)	100		

Weighting for multistage samples is complicated



Weighting for multistage samples is complicated

- Each individual's **sample weight** is equal to the inverse of the final probability of being selected from the population

$$\text{sample weight} = \frac{1}{\text{final probability}}$$

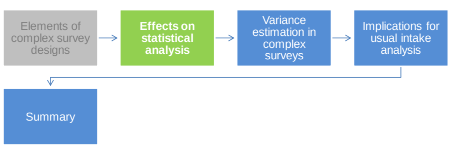
final probability = probability of county being selected
probability of segment being selected from county
probability of household being selected from segment
probability of individual being selected from household

Additional considerations for weighting

- Can incorporate
  - Differential selection probabilities due to stratification and clustering
  - Differential nonresponse probabilities
- Weighted counts of sampled individuals with particular demographic characteristics often set to reproduce "known" population counts – **poststratification**

Summary

- Complex survey methods often used to collect data used for nutrition monitoring
- Stratification, clustering, and weighting are elements of complex sampling schemes
  - Stratification balances precision of subpopulation estimates
  - Clustering decreases sampling costs, but also precision
  - Weighting accounts for stratification/clustering



EFFECTS ON STATISTICAL ANALYSIS

All survey design elements must be accounted for

- Weighting required to minimize bias in survey-based population estimates
- Stratification, clustering, and weighting affect standard errors of estimates

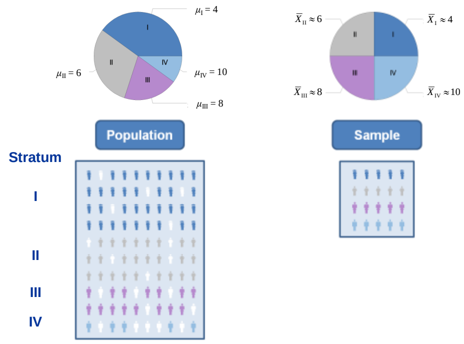
Effects on statistical analysis

All survey design elements must be accounted for

- **Weighting required to minimize bias in survey-based population estimates**
- Stratification, clustering, and weighting affect standard errors of estimates

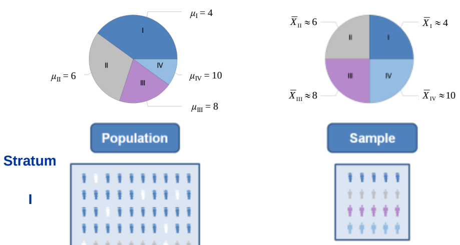
Effects on statistical analysis

Weighting required to account for bias



Effects on statistical analysis

Weighting required to account for bias



- Unweighted sample mean dominated by large values in oversampled strata

Effects on statistical analysis

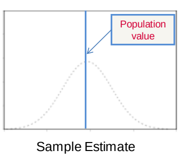
All survey design elements must be accounted for

- Weighting required to minimize bias in survey-based population estimates
- **Stratification, clustering, and weighting affect standard errors of estimates**

Effects on statistical analysis

What is the standard error of an estimate?

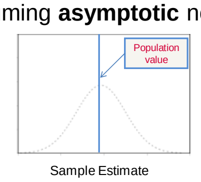
- Reflects variation expected across repeated sampling of the population
  - Most samples yield estimates close to true population value, a few samples yield estimates far away
  - Sampling distributions are often normal (CLT)



Effects on statistical analysis

What is the standard error of an estimate?

- The **standard error** (s.e.) is the standard deviation of the sampling distribution
  - More independent pieces of information ⇒ smaller standard errors
- Used to construct significance tests, confidence intervals assuming **asymptotic** normality



Effects on statistical analysis

Standard errors estimated from sample

- In practice, only one sample is obtained
  - Standard errors must be estimated from the data at hand
- Basic statistical theory provides estimation methods for standard errors of "smooth" statistics
  - Means
  - "Mean-like": regression parameters, ratios

measurementERRORwebinarseries

49 Accounting for complex survey design in modeling usual intake

Effects on statistical analysis

Standard errors estimated from sample

- Estimating standard errors for percentiles is especially challenging
  - Not "mean-like" for purposes of CLT
  - Sampling distributions less well-behaved
- May require alternative methods for tests/CIs
  - Standard error still reflects variation over repeated sampling

measurementERRORwebinarseries

50 Accounting for complex survey design in modeling usual intake

Effects on statistical analysis

Standard errors in complex surveys

- Theoretical derivation based on asymptotic normality of weighted cluster means within strata
- Not all statistical software is fully "survey-aware"
  - "Weighted analysis" might not be sufficient
  - Stratification/clustering may also be important

measurementERRORwebinarseries

51 Accounting for complex survey design in modeling usual intake

Effects on statistical analysis

Stratification/clustering reduces degrees of freedom

- Stratification and clustering result in fewer independent pieces of information

degrees of freedom = (number of clusters) – (number of strata)

- For example, NHANES 2003-6 has
  - 20,470 individuals
  - 60 clusters, 30 strata ⇒ 30 d.f.

measurementERRORwebinarseries

52 Accounting for complex survey design in modeling usual intake

Effects on statistical analysis

Total calcium intake for women in NHANES 2003-6

- Subset of 2601 women ages 31-70 with reliable data on first 24HR
- Parameter of interest: population mean calcium intake from foods and dietary supplements
- Estimates based on combination of data from 24HR and dietary supplement questionnaire

measurementERRORwebinarseries

53 Accounting for complex survey design in modeling usual intake

Effects on statistical analysis

Total calcium intake for women in NHANES 2003-6

- Multiple ways to compute the estimate and its standard error using SAS
  - UNIVARIATE ignoring the weights
  - UNIVARIATE with a WEIGHT statement
  - UNIVARIATE with a FREQ statement
  - SURVEYMEANS with STRATA, CLUSTER, and WEIGHT statements
- Only the last way incorporates all design factors

measurementERRORwebinarseries

54 Accounting for complex survey design in modeling usual intake

Total calcium intake for women in NHANES 2003-6

Procedure Used to Estimate Mean Intake	Est. Mean	Std. Error	Assumed d.f.
UNIVARIATE	1027	13	2600
UNIVARIATE + WEIGHT	1115	14	2600
UNIVARIATE + FREQ	1115	0.08	70667993
SURVEYMEANS	1115	27	30

Total calcium intake for women in NHANES 2003-6

Procedure Used to Estimate Mean Intake	Est. Mean	Std. Error	Assumed d.f.
UNIVARIATE	1027	13	2600
UNIVARIATE + WEIGHT	1115	14	2600
UNIVARIATE + FREQ	1115	0.08	70667993
SURVEYMEANS	**1115**	**27**	**30**

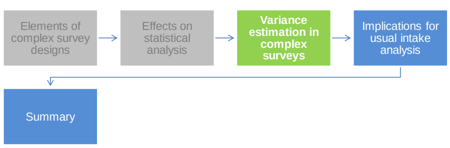
- Mean underestimated by ~8% if weights ignored
- Standard errors underestimated if not all design factors are properly accounted for

Statistical methods for complex surveys limited

- Inference based on t-tests easiest to extend to complex surveys
  - Asymptotic normality, standard error formulae established for many mean-like statistics
- Other statistical methods more difficult to extend
  - E.g., likelihood ratio tests

Summary

- Stratification, clustering, and weighting must be accounted for in analysis of survey data
- Many statistical techniques have no survey analogues
- Inference may need to be simplistic, e.g., t-tests
  - Need proper estimates of standard errors



VARIANCE ESTIMATION IN COMPLEX SURVEYS

Variance estimation techniques

- Taylor linearization
- Resampling methods
  - Bootstrap
  - Jackknife
  - Balanced Repeated Replication (BRR)

Variance estimation techniques

- **Taylor linearization**
- Resampling methods
  - Bootstrap
  - Jackknife
  - Balanced Repeated Replication (BRR)

Taylor linearization

- Used by default in most “survey-aware” software
  - “Textbook” formulae for standard estimators
- Hard to extend to more complex estimators in general survey designs
  - Monte Carlo-based usual intake percentiles (as in NCI method) especially problematic

Variance estimation techniques

- Taylor linearization
- **Resampling methods**
  - **Bootstrap**
  - **Jackknife**
  - **Balanced Repeated Replication (BRR)**

Resampling methods

- Emulate resampling of **population** by resampling from the **sample at hand**
  - Sample is treated as “population in miniature”
  - Reflects definition of sampling distribution
- Will first illustrate for the bootstrap method in the non-survey setting

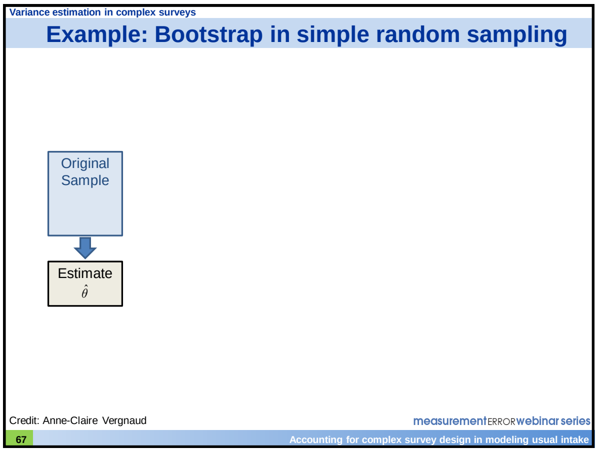
Variance estimation techniques

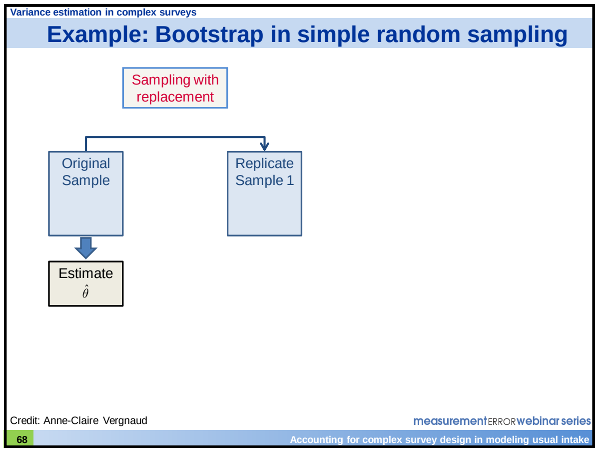
- Taylor linearization
- **Resampling methods**
  - **Bootstrap**
  - **Jackknife**
  - **Balanced Repeated Replication (BRR)**

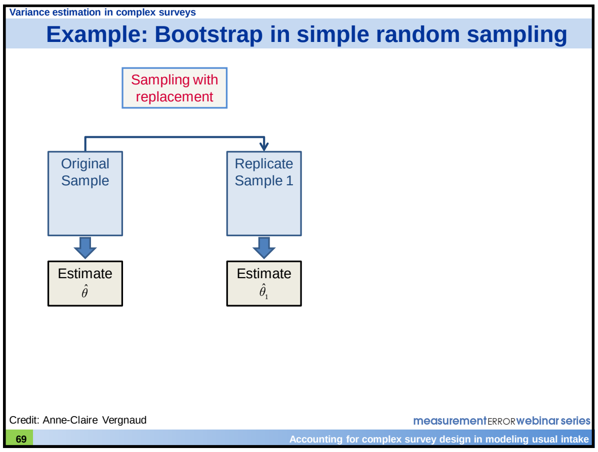
Example: Bootstrap in simple random sampling

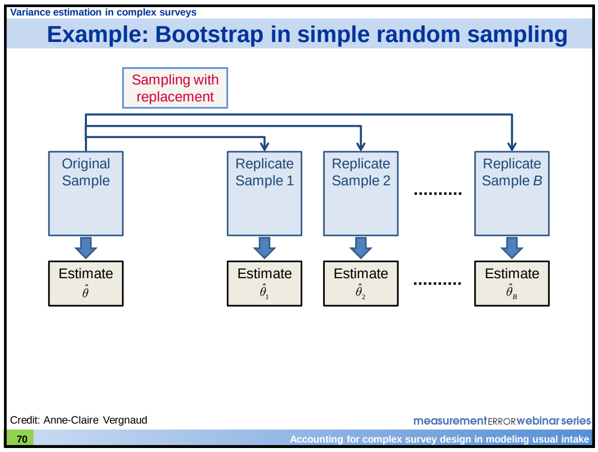
Original Sample

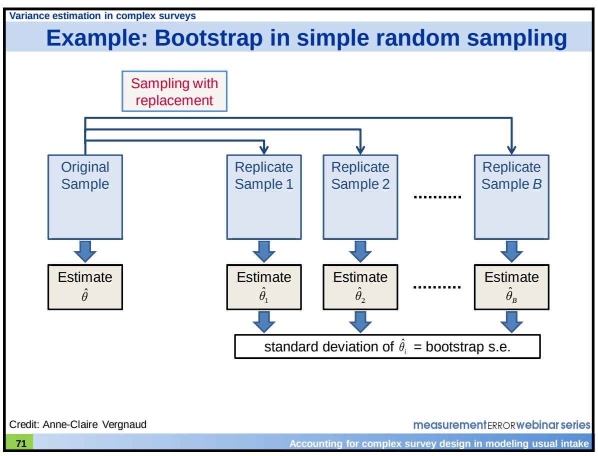
Credit: Anne-Claire Vergnaud

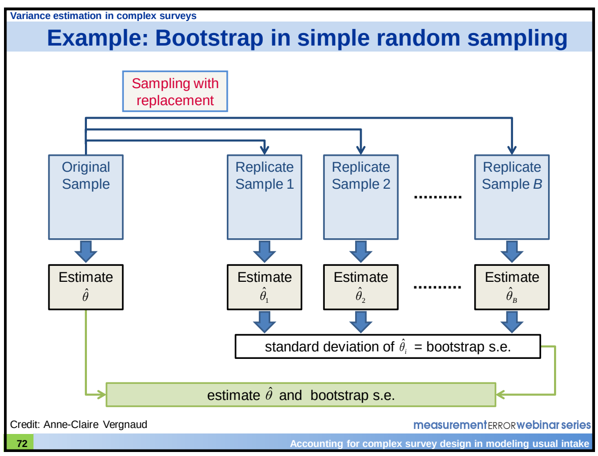

Variance estimation in complex surveys
Example: Bootstrap in simple random sampling
Original Sample
Estimate $\hat{\theta}$
Credit: Anne-Claire Vergnaud
measurementERRORwebinarseries
67
Accounting for complex survey design in modeling usual intake


Variance estimation in complex surveys
Example: Bootstrap in simple random sampling
Sampling with replacement
Original Sample
Replicate Sample 1
Estimate $\hat{\theta}$
Credit: Anne-Claire Vergnaud
measurementERRORwebinarseries
68
Accounting for complex survey design in modeling usual intake


Variance estimation in complex surveys
Example: Bootstrap in simple random sampling
Sampling with replacement
Original Sample
Replicate Sample 1
Estimate $\hat{\theta}$
Estimate $\hat{\theta}_1$
Credit: Anne-Claire Vergnaud
measurementERRORwebinarseries
69
Accounting for complex survey design in modeling usual intake


Variance estimation in complex surveys
Example: Bootstrap in simple random sampling
Sampling with replacement
Original Sample
Replicate Sample 1
Replicate Sample 2
Replicate Sample *B*
Estimate $\hat{\theta}$
Estimate $\hat{\theta}_1$
Estimate $\hat{\theta}_2$
Estimate $\hat{\theta}_B$
Credit: Anne-Claire Vergnaud
measurementERRORwebinarseries
70
Accounting for complex survey design in modeling usual intake


Variance estimation in complex surveys
Example: Bootstrap in simple random sampling
Sampling with replacement
Original Sample
Replicate Sample 1
Replicate Sample 2
Replicate Sample *B*
Estimate $\hat{\theta}$
Estimate $\hat{\theta}_1$
Estimate $\hat{\theta}_2$
Estimate $\hat{\theta}_B$
standard deviation of $\hat{\theta}_i$ = bootstrap s.e.
Credit: Anne-Claire Vergnaud
measurementERRORwebinarseries
71
Accounting for complex survey design in modeling usual intake


Variance estimation in complex surveys
Example: Bootstrap in simple random sampling
Sampling with replacement
Original Sample
Replicate Sample 1
Replicate Sample 2
Replicate Sample *B*
Estimate $\hat{\theta}$
Estimate $\hat{\theta}_1$
Estimate $\hat{\theta}_2$
Estimate $\hat{\theta}_B$
standard deviation of $\hat{\theta}_i$ = bootstrap s.e.
estimate $\hat{\theta}$ and bootstrap s.e.
Credit: Anne-Claire Vergnaud
measurementERRORwebinarseries
72
Accounting for complex survey design in modeling usual intake

Example: Bootstrap in simple random sampling

- Key to bootstrap is **with-replacement sampling**
- In a given bootstrap sample,
  - Some individuals will appear multiple times
  - Some individuals will not appear at all
- Number of times an individual appears is analogous to a sampling weight

Resampling via weight perturbation

- Resampling operationalized using a set of weights for each sample (replicate and original)
  - In SRS, all weights for original sample are 1
- Eliminates need to store multiple copies of data set with many analysis variables per person

Resampling via weight perturbation

- Resampling in complex surveys operationalized using sets of “perturbed” weights
- Bootstrap, jackknife, BRR methods differ in the
  - Numbers of weight sets needed
  - Ways weight sets are constructed
  - Formulae for computing variability among replicate estimates

Bootstrap in complex surveys

- Bootstrap samples must be drawn according to sampling plan used to draw the original sample
  - Computationally intensive (*B* very large)
- Offers robust method for constructing CIs
  - Bounds based on 95% of empirical distribution of bootstrap estimates
  - May work better for poorly-behaved sampling distributions of “non-smooth” statistics

Use of bootstrap in usual intake estimation

- Recommended for estimating standard errors of complex statistics for Canadian Community Health Survey, Nutrition Cycle 2.2
- Used for estimating standard errors of model parameters and usual intake percentiles calculated using the NCI method
  - Simulation study for SRS (Tooze et al., 2010)
  - Dutch National Food Consumption Survey (Verkaik-Kloosterman et al., in press)

Variance estimation techniques

- Taylor linearization
- **Resampling methods**
  - Bootstrap
  - **Jackknife**
  - Balanced Repeated Replication (BRR)

Jackknife in complex surveys

- Creation of perturbed weight sets
  - One set of weights per cluster
  - Weight set *k* deletes (zero-weights) all the observations in cluster *k*
  - Redistributes missing weight among other observations in same stratum as cluster *k*
  - Leaves weights unchanged for observations in all the other strata

Jackknife in complex surveys

- For surveys with many clusters, many weight sets must be generated
  - Less computationally intensive than bootstrap
- Each set of jackknife weights may need to be poststratified to recover subpopulation sizes

Use of jackknife in usual intake estimation

- Alternative to Taylor linearization for
  - Usual intake model parameters
  - ISU method percentiles
- Not applicable to Monte Carlo-based usual intake percentiles

Variance estimation techniques

- Taylor linearization
- **Resampling methods**
  - **Bootstrap**
  - **Jackknife**
  - **Balanced Repeated Replication (BRR)**

Balanced repeated replication in complex surveys

- Limited to stratified cluster designs with two clusters/stratum
- Most aggressive perturbation of weights
  - Weight set *k* deletes(zero-weights) the observations in half of the clusters, and
  - Doubles the weights for observations in the remaining clusters
  - **Perturbation factors** 0 and 2

Balanced repeated replication in complex surveys

- Fewer weight sets than for jackknife
  - Smallest multiple of 4 greater than number of strata
- Choice of which cluster to zero/double determined by a **Hadamard matrix**
  - Orthogonality property minimizes number of weight sets required
  - “Balances” the influence of each cluster

Balanced repeated replication in complex surveys

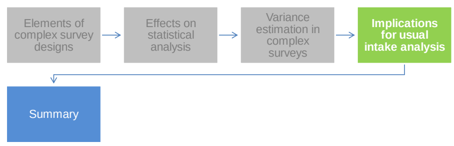
- Standard BRR can be unstable due to extreme perturbations
- Fay's modified BRR uses perturbation factors less extreme than 0 and 2
- Each set of BRR weights may need to be poststratified to recover subpopulation sizes

Use of BRR in usual intake estimation

- Alternative to Taylor linearization for the What We Eat In America (WWEIA) portion of the US National Health and Nutrition Examination Survey (NHANES)
- BRR works for Monte Carlo-based percentiles as well as usual intake model parameters
  - Application of NCI method, including multiple simulation studies and analyses of NHANES

Summary

- "Survey-aware" software typically uses Taylor linearization to estimate standard errors
  - Limited to basic, "mean-like" estimators
  - Low computational burden
- Resampling methods offer an alternative to Taylor linearization for complex estimators



IMPLICATIONS FOR USUAL INTAKE ANALYSIS

Typical research question

What is the usual intake of component X among subgroup Y in my population?

To answer, must consider:

- Estimator of interest
- Method of analysis and its data requirements
- Technique for variance estimation and how to use software to properly implement

Example 1

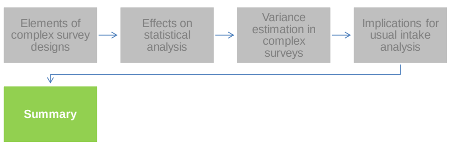
- **Estimator:** mean of usual intake distribution
- **Method/data:** mean, all valid first-day 24HRs from NHANES survey
- **Variance estimation:** Taylor linearization
  - Procedures available in common software
    - SAS
    - SUDAAN
    - Stata

Example 2

- **Estimator:** distribution of usual intake
- **Method/data:** NCI method, all valid 24HRs from NHANES survey
- **Variance estimation:** BRR
  - Need to obtain/construct BRR weights
  - NCI SAS macros

Example 3

- **Estimator:** distribution of usual intake
- **Method/data:** ISU method, all valid 24HRs from Canadian Community Health Survey 2.2
- **Variance estimation:** bootstrap
  - Official bootstrap weight sets from Statistics Canada
  - ISU software
    - SIDE



SUMMARY

Key messages

- Data used for monitoring of usual intakes among populations typically collected using complex survey methods
- Computation of point estimates and standard errors must account for stratification, clustering, and weighting

Key messages

- Standard error estimation can be complicated
  - Means and "mean-like" statistics:
    - Can use Taylor linearization implemented in some software packages
  - Percentiles and other "non-smooth" statistics:
    - May need resampling techniques like bootstrap or BRR implemented in various ways

Key messages

- No "one size fits all" approach to modeling usual intake using data from a complex survey
- Particulars of analyses depend on:
  - Research question
  - Available data
  - Desired modeling method (e.g., NCI method, ISU method)
  - "Survey-aware" features of modeling software
  - Statistical expertise/support

QUESTIONS & ANSWERS

Moderator: Regan Bailey

Please submit questions using the *Chat* function

measurementERRORwebinar series

97 Accounting for complex survey design in modeling usual intake

National Cancer Institute

measurementERRORwebinar series

Next Session

Tuesday, October 18, 2011
10:00-11:30 EDT

Estimating usual total nutrient intake distributions from diet and supplements

Regan Bailey, PhD
Office of Dietary Supplements
National Institutes of Health

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health