



# Accounting for complex survey design in modeling usual intake

Kevin W. Dodd, PhD  
National Cancer Institute

## Slide 1

Hello and welcome to today's webinar, the fourth in the Measurement Error Webinar Series.

I'm Regan Bailey from the Office of Dietary Supplements at the National Institutes of Health. I'll be moderating today's webinar, in which we'll continue with the theme of estimating usual intake distributions with Dr. Kevin Dodd.

A few notes before we get started: The webinar is being recorded so that we can make it available on our Web site. All phone lines have been muted and will remain that way throughout the webinar. There will be a Q&A session following the presentation; you can use the Chat feature to submit a question. And, finally, a reminder: You can find the slides for today's webinar on the Web site that has been set up for series participants. The URL has been sent out via the listserv and appears in the note box at the top left. Other resources available on that page include the glossary of key terms and notation, and the recordings of the first three webinars.

As I mentioned, our presenter for today's webinar is Dr. Kevin Dodd. Kevin is a mathematical statistician in the Biometry Research Group, Division of Cancer Prevention, at the National Cancer Institute. Kevin was involved in the development of the Iowa State University method and the National Cancer Institute method for modeling usual intake. He has worked extensively with population-level survey data, most often the National Health and Nutrition Examination Surveys. Today Dr. Dodd will talk about accounting for complex survey design in estimating usual intake distributions. Dr. Dodd.

# 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



## Slide 2

I want to remind everyone that this series is organized by collaborators from a diverse collection of institutions, as shown here. The nutritionists and statisticians listed on this slide have been working together on the topic of measurement error in self-report dietary intake data for many years now.

# measurement ERROR webinar series



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

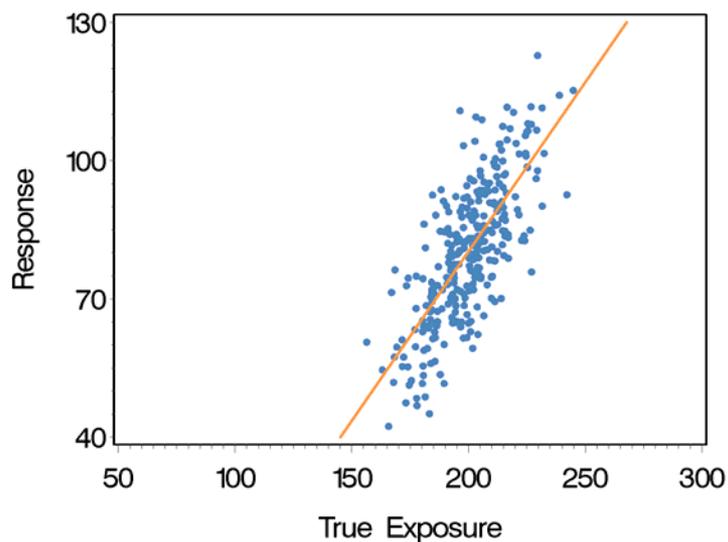
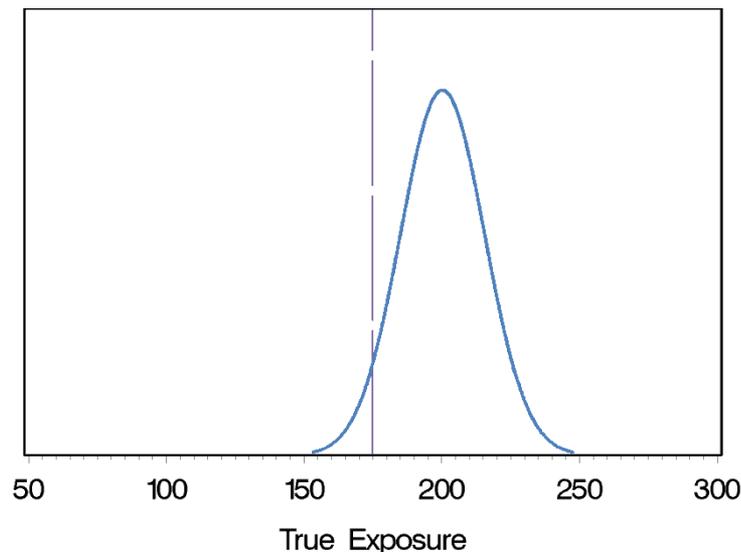
In recognition of his internationally renowned contributions to the field of nutrition epidemiology and his commitment to understanding measurement error associated with dietary assessment.

### Slide 3

The series is dedicated to the memory of our colleague, Arthur Schatzkin, who was instrumental to the progress this area of research has made over the years.

# Two main areas of interest

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



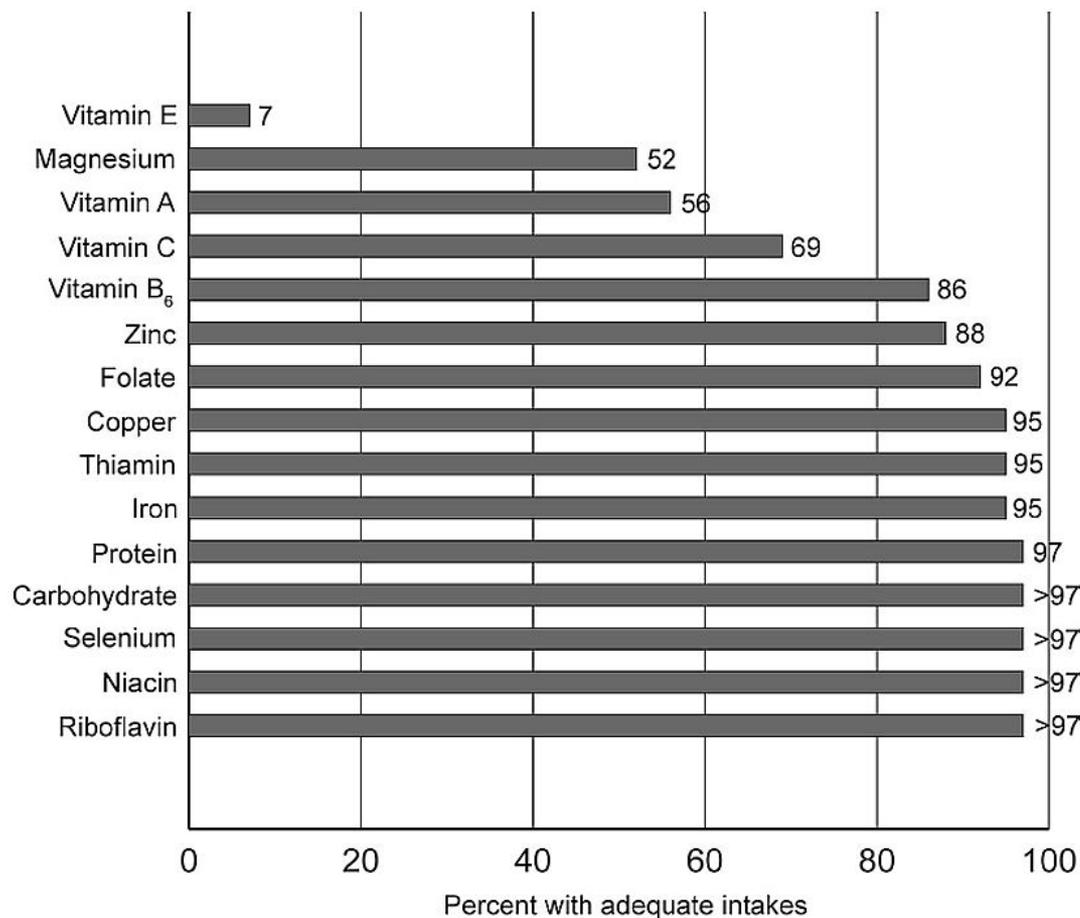
- Estimating diet-health relationships: regression coefficients

## Slide 4

We have previously mentioned that the webinar series as a whole covers two main areas of interest. The first is in describing usual intake distributions, where we want to know the mean usual intake in the population, or the 25<sup>th</sup> percentile, or the proportion of the population with intake below some threshold. The second topic area is concerned with estimating the relationship between diet and some health outcome. In this case, we might want to know the slope of the regression line. Today's webinar introduces concepts that are mostly applicable for the first case, with distributions, but also may impact the way you would address the second case.

# 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

## Slide 5

I'll start off by showing you this figure taken from the latest Dietary Guidelines Advisory Committee report, which shows, for a bunch of nutrients, the estimated fraction of the U.S. population with usual intake that meets or exceeds the Estimated Average Requirement, or EAR, for those nutrients, based on their reported food consumption (i.e., not incorporating the effect of dietary supplements). This table, along with a lot of other information, was used to formulate guidance about how best to eat a healthy diet.

I want to draw your attention to this little line of text in the lower left corner that says where the data that went into this figure actually came from. They came from the dietary portion of the 2001-2002 National Health and Nutrition Examination Survey, which I, at least, consider to be the U.S. flagship diet-and-health survey. The NHANES is a massive undertaking—a large-scale survey that is designed to be nationally representative, where the data are collected from all over the United States according to a complex sampling plan. Now, these sorts of complex national-level surveys happen in other countries as well, and provide the bulk of population monitoring data. In fact, when a lot of these methods we've been talking about for estimating usual intake distributions were being developed, there was always the sense that, eventually, they would be applied to these kinds of survey data.

# 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

## Slide 6

The webinars so far have primarily focused more on the development of the methods and how they would be applied to a data set of, say, 24HRs, but have pretty much skipped over talking about exactly how researchers are supposed to get their hands on these data. In today's webinar, I'm going to focus on details related to data collection; specifically, the details of how collecting data using survey sampling methods affects the analysis you can do with these modeling approaches.

# 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.

## Slide 7

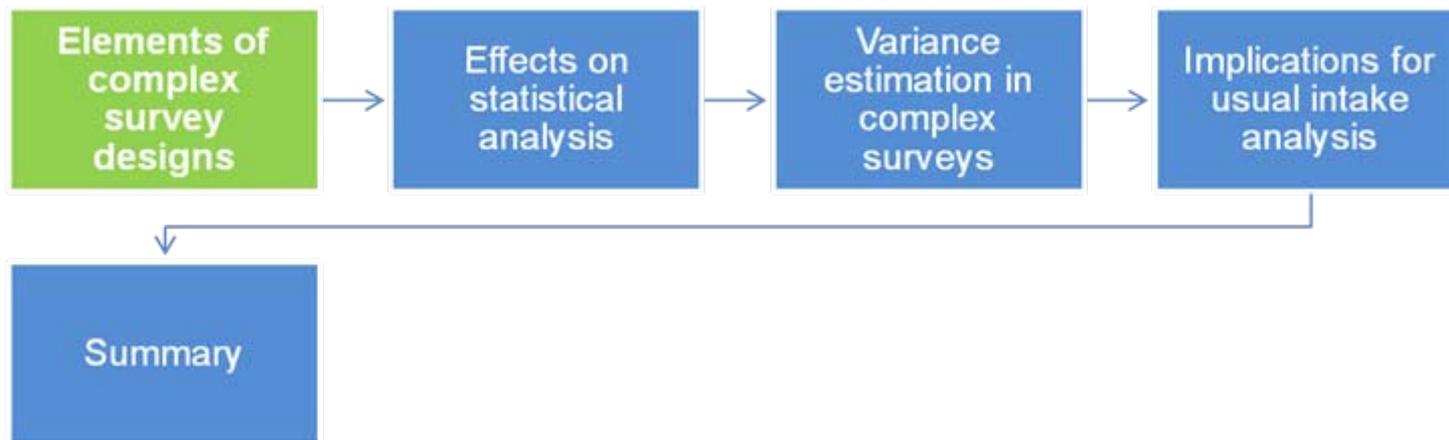
There are two broad objectives for this webinar, listed here. The first objective is to describe three general elements that, separately or in combination, comprise a complex survey sampling design. These three elements are stratification, clustering, and weighting, and each element has to be accounted for in analysis of survey data. The second objective is to identify some statistical techniques for variance estimation that are very important in analysis of survey data in general and in the analysis of usual intake distributions in particular.

# 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

## Slide 8

The outline shown here shows in a little more detail the order in which I'll address the two objectives I just laid out on the previous slide.



# ELEMENTS OF COMPLEX SURVEY DESIGNS

## Slide 9

First, let's talk about what makes a survey sampling design complex.

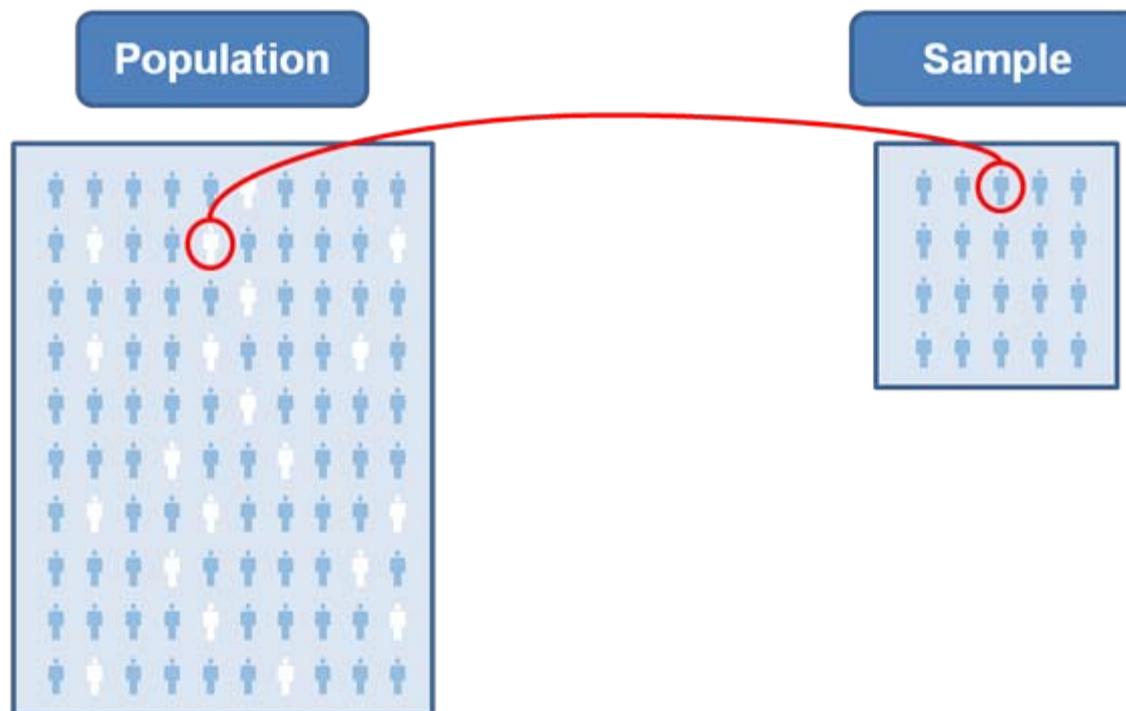
## Simple random sampling

- 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

## Slide 10

Intuitively, a complex design is one that is the opposite of simple. Most of the time, when statisticians are coming up with a new way to analyze data, they start with the simple case, called, appropriately enough, simple random sampling. Under simple random sampling, the data you have at hand were selected from the population of interest completely at random, so that every member of the population has an equal chance of being selected. Now, before you can pull someone from the population, you have to get a list of the population members. In survey sampling language, this list of members is called the “sampling frame.” If your population of interest is a small, homogeneous group, it is pretty easy to obtain a simple random sample and, as I just suggested, it will also be easy to analyze the data you get, because most statistical analysis techniques were developed for that case.

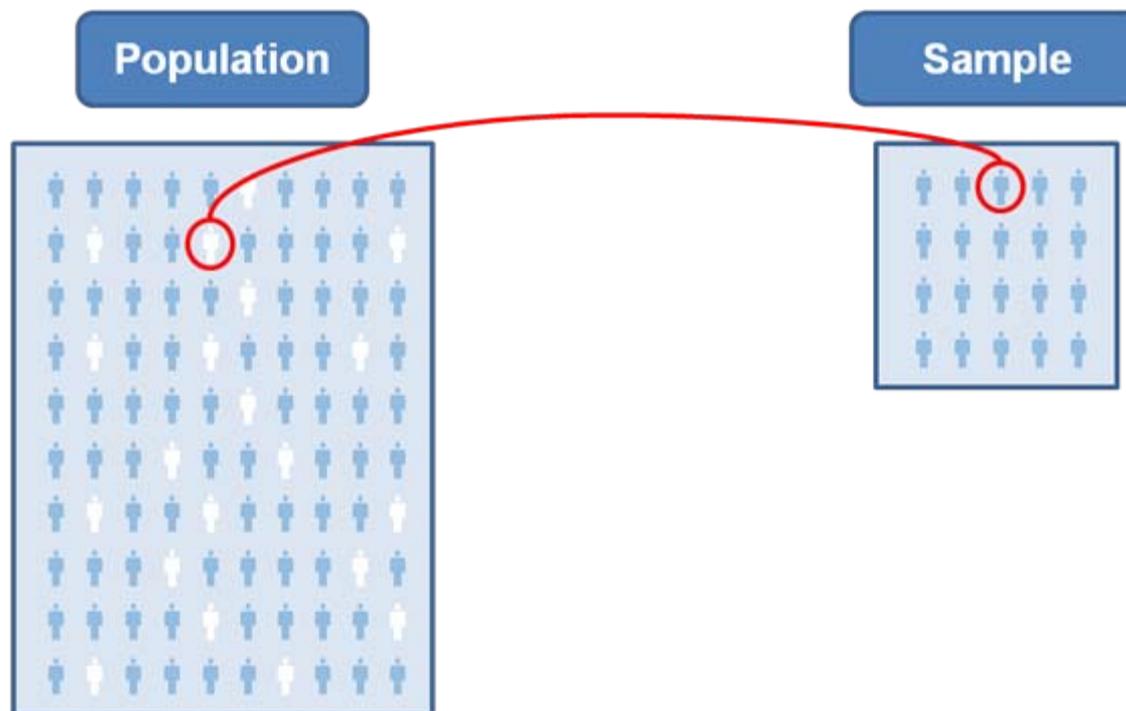
# Selecting a simple random sample



## Slide 11

Here is a diagram showing how you might select a simple random sample of 20 people from the population represented by the 100 symbols in the box on the left. Once you have a list of all 100 members in the population, which in this case is just the numbers from 1 to 100, with 1 corresponding to the top left position and 100, to the bottom right, you just randomly jumble all the numbers and take the first 20 of the jumbled set. The chance that any one person is selected is  $20/100$ , or 20%. In the population box, I've highlighted the individuals in white that comprise the sample shown in the box on the right. You see that the third person in my sample was person number 15. Now, this might be practical in some cases, where your population is easily accessible.

# Selecting a simple random sample



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

## Slide 12

But in practice, especially at a national level, data don't come to you as a simple random sample; instead, data are collected under a complex survey design.

# Why use a complex sampling design?

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

### Slide 13

Well, why would you use a complex survey design? Often, it comes down to cost—either the cost involved with drawing up the desired sampling frame and figuring out how to contact potential survey participants, or the cost of traveling around the country actually collecting the data for participants who agree to be in your study. Another reason for using a complex sampling plan is to improve precision of subpopulation estimates, by making sure you have enough people in certain demographic subgroups to allow separate estimation within those subgroups.

# Elements of complex sampling designs

- Stratification
- Clustering
- Weighting

## Slide 14

A minute ago I told you about these three elements that can make a sampling plan complex—stratification, clustering, and weighting. I'm going to talk about each one in turn, but you'll often see more than one, or even all three, elements used in the development of a complex sampling plan.

# Elements of complex sampling designs

- **Stratification**
- Clustering
- Weighting

## Slide 15

First is stratification.

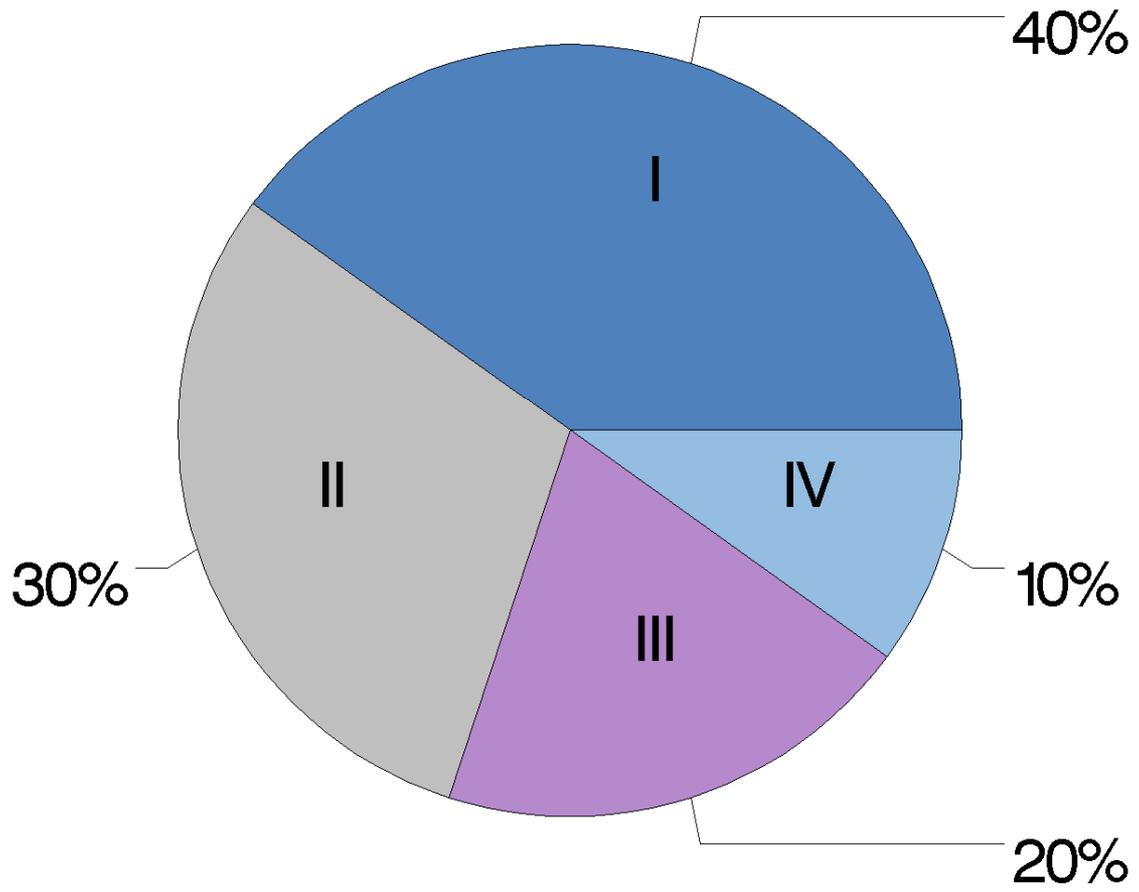
## What is stratification?

- 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

## Slide 16

In the last two webinars, we've talked about performing "stratified" analyses to get subpopulation estimates. Well, that's what we are talking about here. Stratification means grouping individuals in the population into more homogeneous groups that share specific, usually demographic, characteristics. Doing so is usually in the context of identifying subpopulations of *a priori* interest, such as pregnant and lactating women, or children, or low-income individuals.

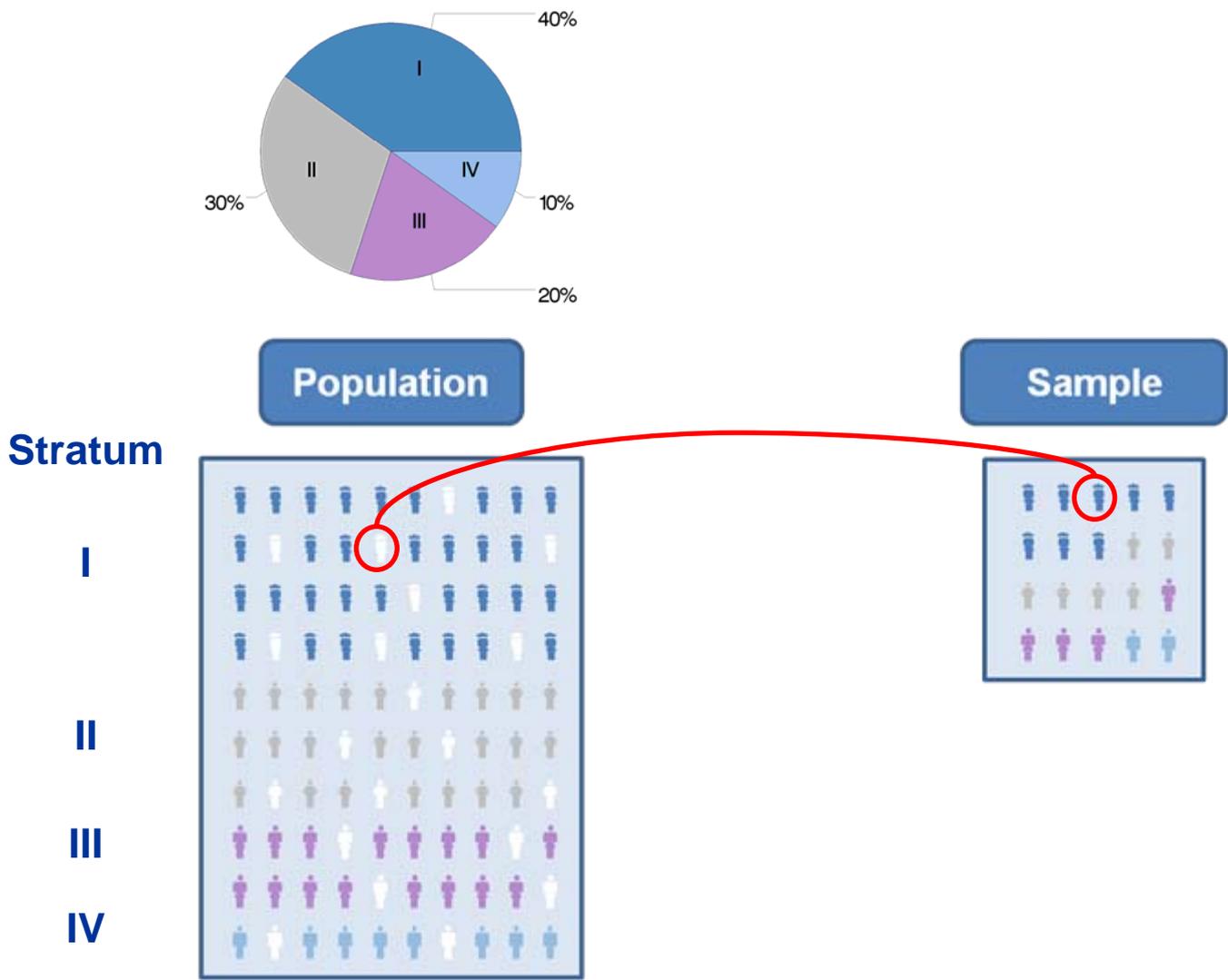
# Hypothetical population with four strata



## Slide 17

For the next few slides, I'm going to be showing some examples of how you might incorporate stratification into your sampling plan. For my example, suppose that the population of interest comprises four important subpopulations, or strata, of different sizes, as shown here. Forty percent of this hypothetical population fall into stratum I; 30 percent, into stratum II; 20 percent, into stratum III; and only 10 percent, into stratum IV.

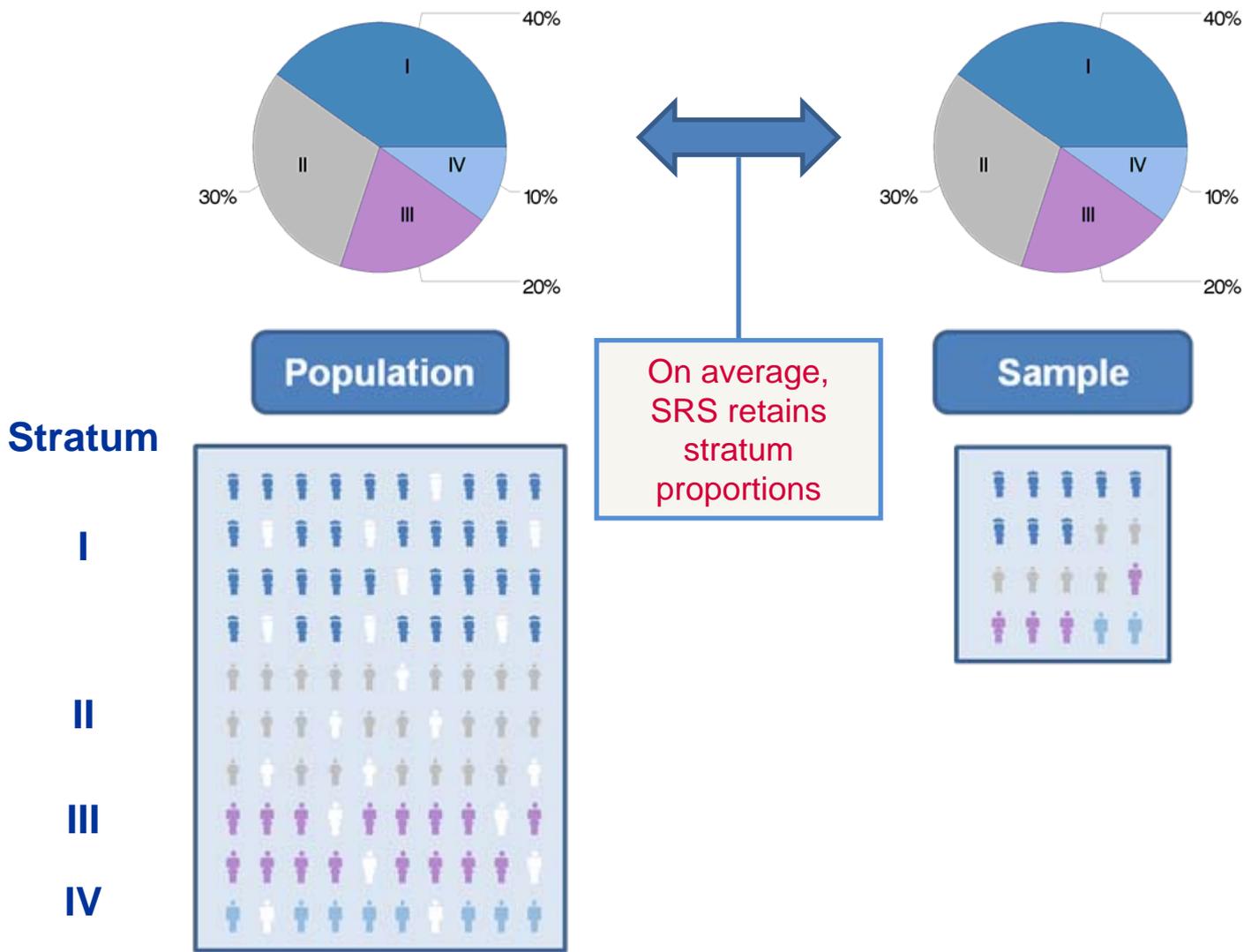
# Simple random sample from a stratified population



## Slide 18

Now my diagram from a few slides back is a little different. Now I have indicated that my population has four different kinds of people in it, by using different symbols and colors to identify people in the different strata. The sampling frame is still the numbers 1 to 100, but now members 1-40 are in Stratum I, members 41-70 are in stratum II, and so on. We can draw the same people from this population that we did before when we pretended the population was homogeneous and all we needed was a simple random sample; here, the third person chosen is still person 15. But now, I keep track of the stratum membership for each sampled person.

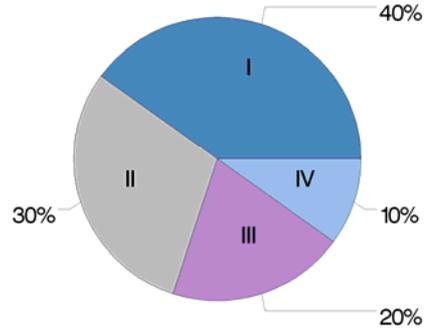
# Simple random sample from a stratified population



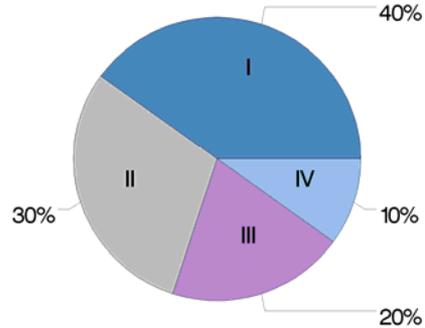
## Slide 19

Now, one nice thing about the simple random sampling idea is that it results in a sample that reflects the population; our sample retains the same proportions in each stratum as are present in the population, at least on average. Depending on the size of the population, the size of the sample, and the number and size of the strata, you can get unlucky with a sample and end up with a sample that isn't exactly a miniature version of the population, but that didn't happen in this example case.

# Simple random sample from a stratified population



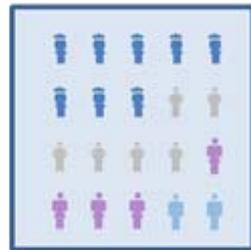
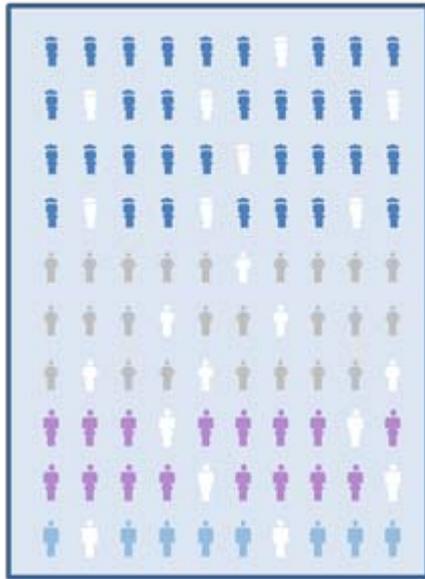
Population



Sample

Stratum

I  
II  
III  
IV



Small expected sample sizes for small strata

## Slide 20

But the downside of the simple random sampling approach is that you can end up with very small sample sizes for small strata, so your resulting subpopulation estimates won't be based on much data and will be imprecise. Here, your estimate for stratum IV would be based on just two people—the absolute minimum number that would allow you to compute a variance estimate. And under a different randomization, you might not have ended up with anybody from stratum IV.

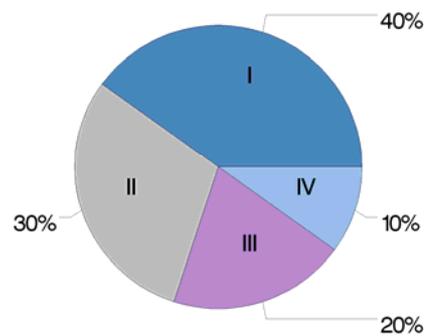
# Alternative: balanced stratified random sampling



## Slide 21

One way to help maintain a given precision for all your subpopulation estimates would be to ensure that you draw the same number of people from each stratum of interest. This slide shows a different sample selection scheme, where my sample of size 20 now draws 5 people per stratum. Here, the second person in the sample is population member 16.

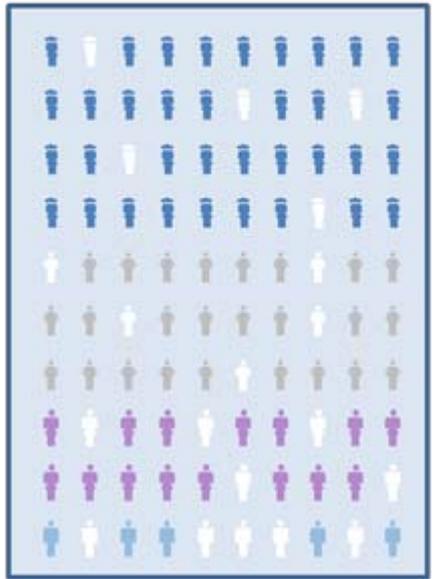
# Alternative: balanced stratified random sampling



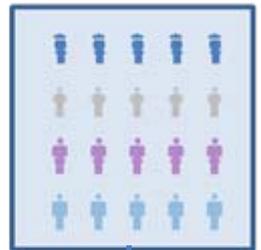
Population

Stratum

I  
II  
III  
IV



Sample

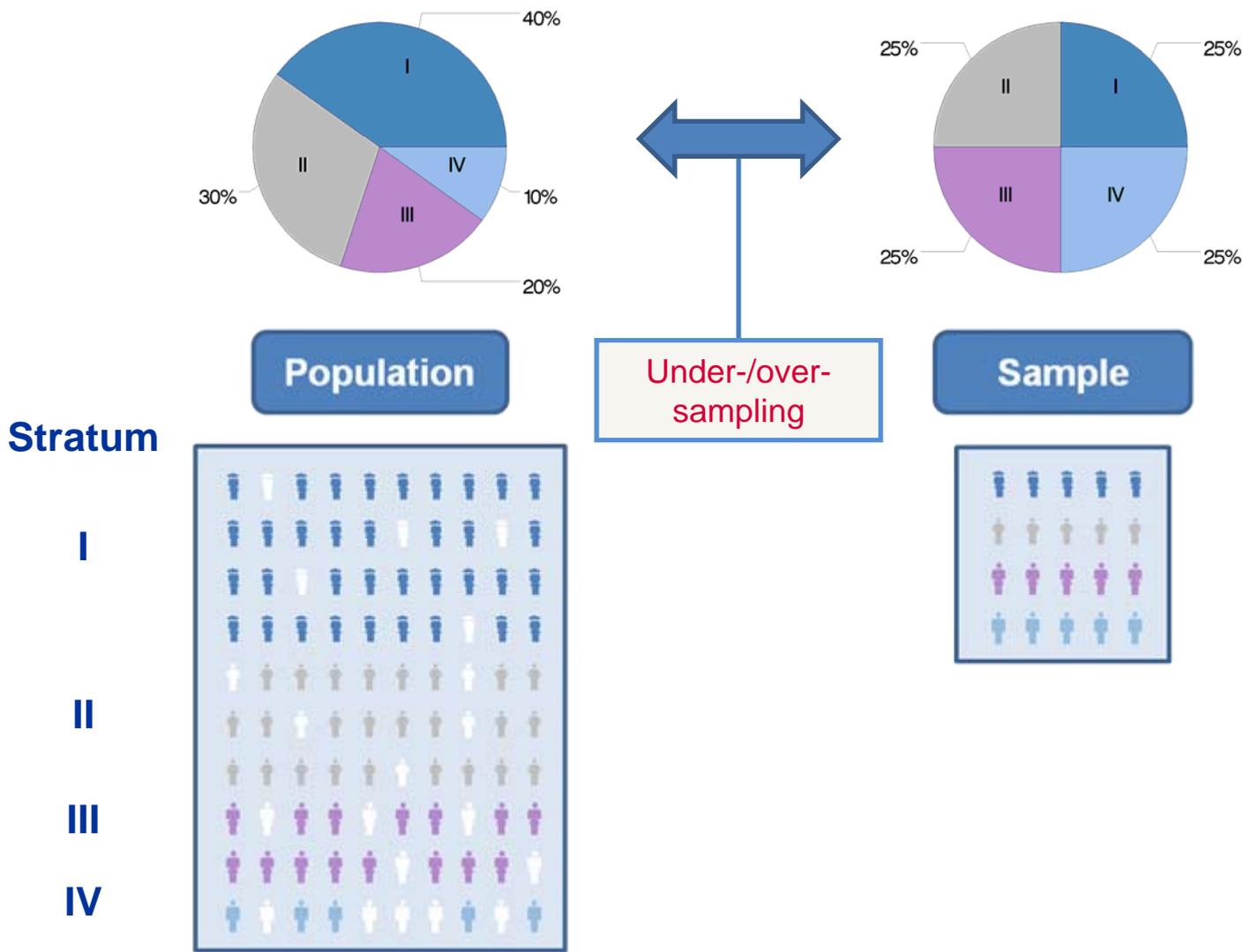


Equal sample sizes for all strata

## Slide 22

Under this sample selection plan, we do have equal sample sizes for all strata, so we have the same amount of information with which to make subpopulation estimates. This doesn't guarantee equal precision of your subpopulation estimates—for example, if some strata have more variability than others, you won't get equal precision—but at least you can avoid the situation where you get really unlucky and some strata are completely missing representation in your sample. But are there any downsides to this approach?

# Alternative: balanced stratified random sampling



## Slide 23

Well, possibly: Under this hypothetical balanced stratified sampling plan, we have lost the property we said was nice about simple random sampling, because now the stratum distribution in our sample doesn't match that of the population. We have actually undersampled Strata I and II, while oversampling Strata III and IV. If we want to get a whole-population estimate as opposed to subpopulation estimates, we will have to account for this under-/oversampling.

## 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

## Slide 24

So there are tradeoffs to be made when you introduce stratification into your sampling design. Compared with simple random sampling, you can end up with more precise estimates for small strata, but at the expense of having less precise estimates for large strata.

Now, this was just an example showing one way stratification could work. In practice, stratified sampling doesn't have to be balanced like I showed in the example. There are more tradeoffs that could be made; for example, you might be in a situation where some subpopulations are of more interest than others and, therefore, need more sample allocated to them.

# Elements of complex sampling designs

- Stratification
- **Clustering**
- Weighting

## Slide 25

Now I'll talk about the second kind of element often incorporated into complex sampling plans—clustering

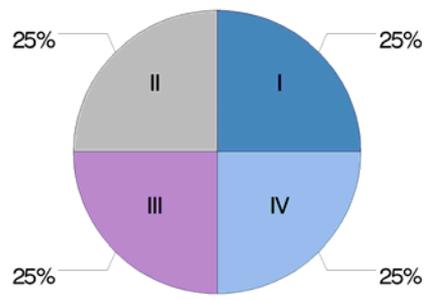
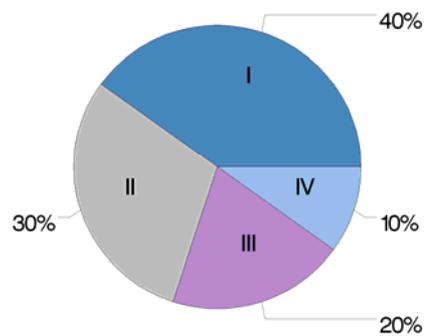
# What is clustering?

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

## Slide 26

Clustering refers to sampling of multiple individuals within the same area. Usually, by area we mean geographic, like a neighborhood or household, but it could also mean sampling people within the same family, regardless of location—for instance, if you are looking at genetic factors. One reason that clustering is often used in complex samples, especially in the U.S. studies like NHANES, is that it cuts down on data collection costs associated with travel. You can set up an exam center in a particular city and get a lot of people to come to the center, as opposed to having to travel all over the country to get one person at a time.

# Stratified cluster sampling

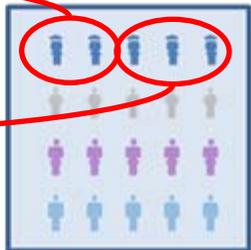
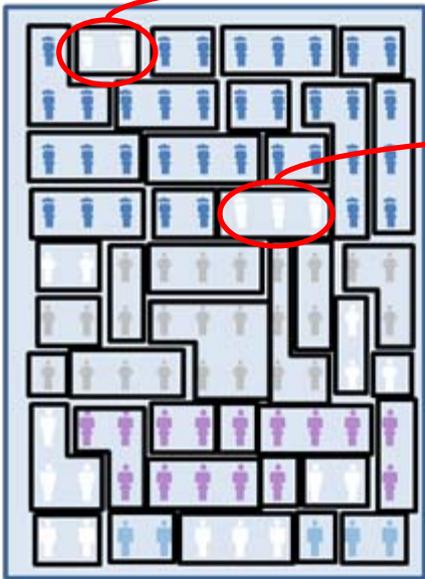


Population

Sample

Stratum

I  
II  
III  
IV



## Slide 27

We'll return again to our stratified population example, but now I've overlaid the black, irregularly shaped boxes that separate people into smaller groups, or clusters. Now with cluster sampling, I don't take individual people; I take all the members of a cluster at a time. So my five people in the first row of the sample box on the right come from taking all the members of the two circled clusters on the left.

## Effects of clustering

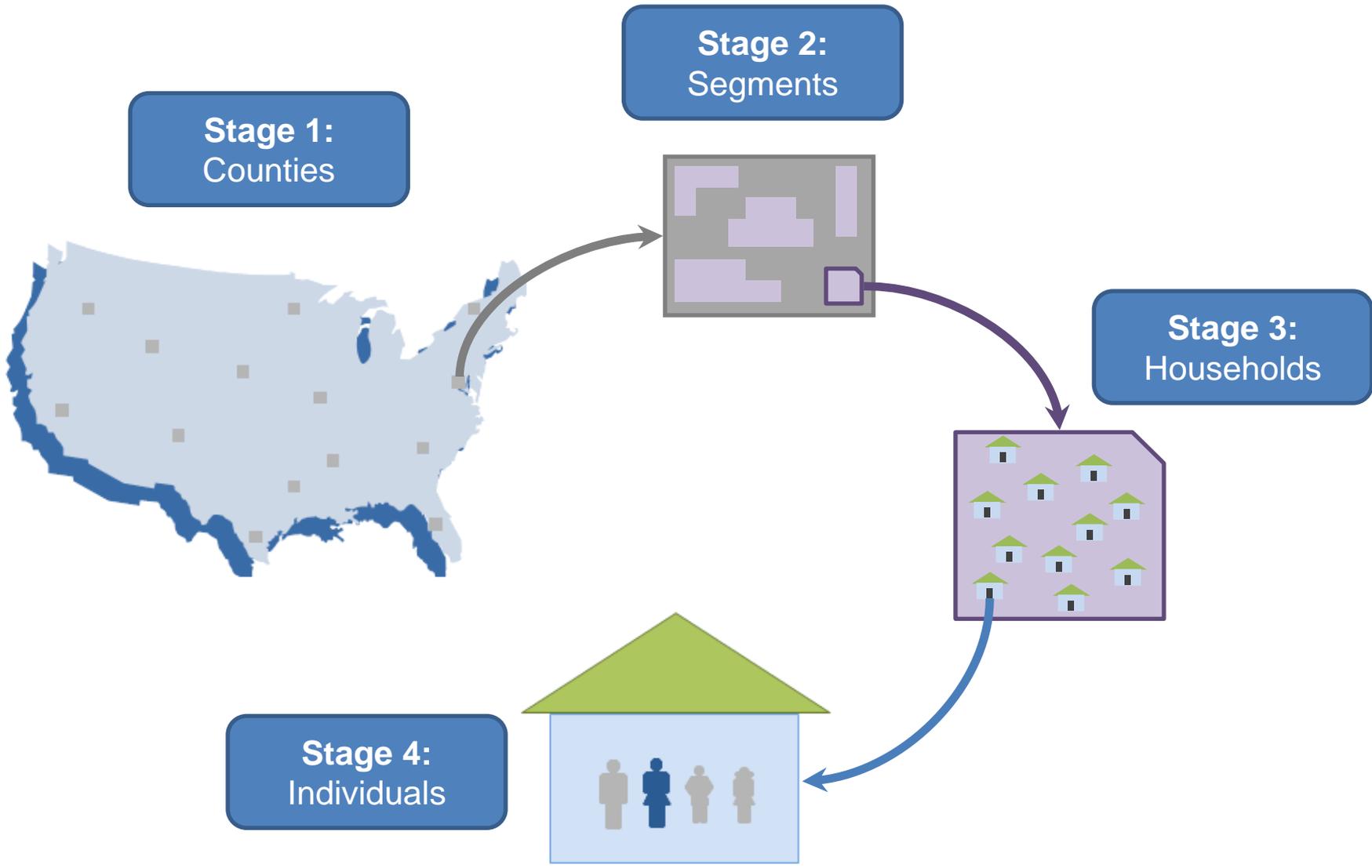
- Observations from individuals sampled from the same cluster tend to be correlated
  - Loss of precision
- Multistage designs with several levels of clustering possible
- First-level clusters (Primary Sampling Units; PSUs) tend to induce largest portion of sampling variability

## Slide 28

One important thing to note is that observations from individuals sampled from the same cluster tend to be correlated. This is in contrast to simple random sampling, where individual observations are treated as independent pieces of information. The correlation in cluster sampling reduces the effective number of these independent pieces of information, leading to a loss of precision in your estimates. You may have the same number of people in your cluster sample, but some (hopefully small) portion of the information is redundant.

It is possible to extend this clustering idea to several levels, leading to what is called “multistage” sampling, where you subsample clusters of smaller units from your first-level clusters. These first-stage clusters are sometimes called Primary Sampling Units, abbreviated as PSUs. When you do your statistical analysis, usually the first level of clustering drives the precision of your estimates, so in practice, analyses of multistage samples sometimes ignore the effects of the other stages of clustering.

# Multistage sampling



## Slide 29

This is a schematic representation of a multistage design, similar to that used in the U.S. NHANES survey. The first stage of sampling selects counties within the U.S., then selects smaller geographic regions called segments, then goes after households within those selected segments, and finally samples individuals in the selected households.

## 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

## Slide 30

I said before that incorporating survey design elements can decrease the costs associated with drawing up your sampling frame and figuring out how to contact prospective participants. With multistage sampling designs like the one in NHANES, you can build up your sampling frame in a stepwise manner. First, you can get a list of counties and draw your first-stage sample units. Then, you can go get more details on just the counties you selected to pick your second-stage units made up of census block groups. Once you have that second-stage list, you can focus on getting details of the households just in those block groups, and then focus on getting contact information for those households. This eliminates the need to go get a complete listing of the contact information for every household in the country.

So you can cut down on the costs of drawing up your sampling frame and getting people to agree to be in your survey. You can also greatly reduce the cost of collecting the data from participants. For example, NHANES participants come to a Mobile Exam Center, or MEC, that is a collection of trailers that get hauled all over the U.S. Keeping the number of stops down to a reasonable number lets NHANES direct their resources into getting data, not into getting fuel for the trailers.

# Elements of complex sampling designs

- Stratification
- Clustering
- **Weighting**

## Slide 31

The last element of complex survey designs that I want to talk about is weighting.

# What is weighting?

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

## Slide 32

Weighting gives an indication of how many people in the population a given sampled individual represents.

## What is weighting?

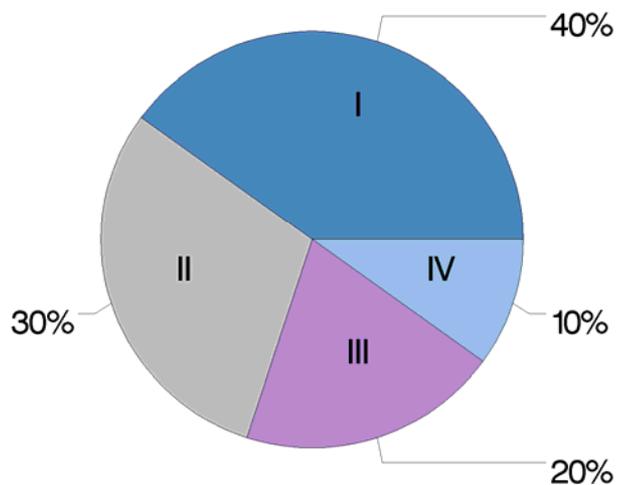
- Indicates how many individuals in the population a sampled individual “represents”
- Each individual’s **sample weight** is equal to the inverse of the final probability of being selected from the population

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

### Slide 33

Each individual in the final sample obtained under a survey design has a sample weight, which is equal to the inverse of the final probability of being selected from the population. For a simple random sample like in our example from before where we took 20 people from a population of size 100, every one of those sampled individuals have the same sample weight of 5, which is the inverse of the uniform selection probability of  $2/10$ , or  $1/5$ . Treating each person as if he/she represented five people in the population is a way of thinking about how you would “expand” your sample to represent the population. For general complex survey designs, sampling weights are more variable.

# Weighting for a stratified sample of size 100



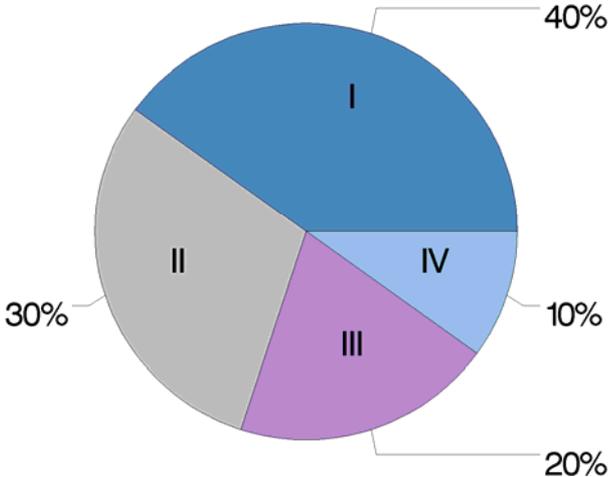
- Total population size: 1 million

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

## Slide 34

For an example of how to compute sampling weights for survey data, let's return to our idea of a stratified population, with four strata appearing in these proportions. To make things a little more realistic, let's pretend that our population size is 1 million, not 100, and that we are going to be taking a sample size of 100. This table shows how the million people are allocated to the different strata.

# Weighting for a stratified sample of size 100



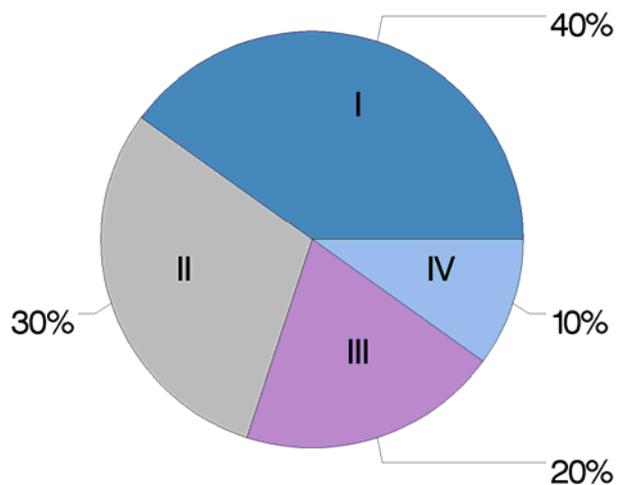
- Total population size: 1 million
- Want to draw a sample of size 25 from each stratum

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

## Slide 35

Now, let's see what happens if we pull a balanced sample of 25 people from each stratum.

# Weighting for a stratified sample of size 100



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

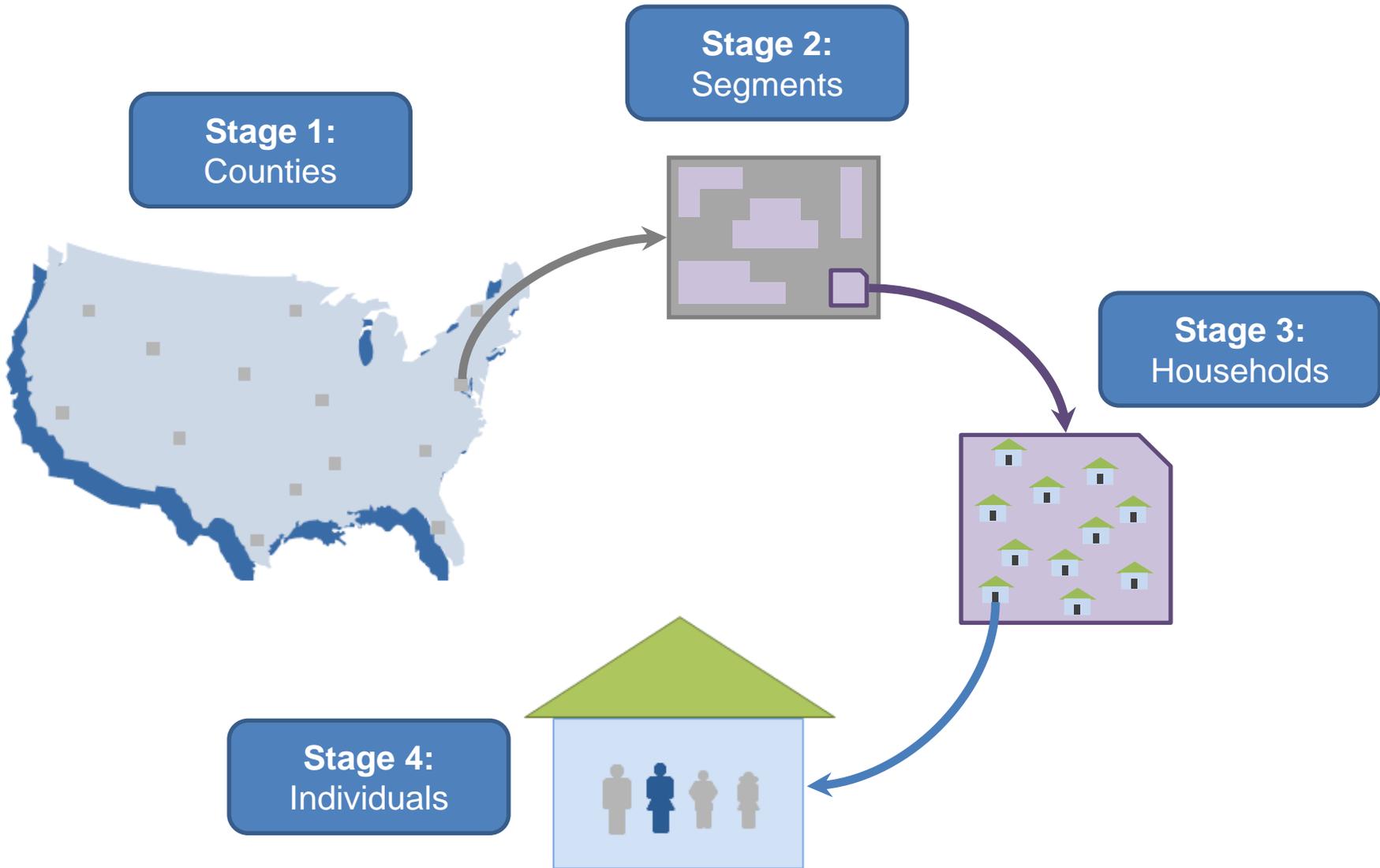
Stratum (Size)	Sample Size	Prob×1000	Weight/10000
I (400K)	25	25/400	1.6
II (300K)	25	25/300	1.2
III (200K)	25	25/200	.8
IV (100K)	25	25/100	.4
Total (1M)	100		

## Slide 36

The final probabilities of selection for each sampled person from each stratum are shown in the third column, scaled by a factor of 1,000, and the fourth column shows the resulting weight, here scaled by 10,000. It turns out that if you are looking to estimate means or distribution functions, you are fine using weights scaled this way, because the scaling factor cancels out in the estimation. In some cases, however, like when you want to estimate the actual number of people in the population with some characteristic, you have to account for the scaling factor.

I want to point out that in this simple example, all the sampled members of a given stratum have the same sampling weight. In practice, for a really complex survey, this will not be the case; instead, each person will be likely to have a unique weight value.

# Weighting for multistage samples is complicated



## Slide 37

This is because 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

## Slide 38

It is still true that the individual sampling weights are the inverse of final selection probabilities but, now, the final selection probability has to take into account all the stage-specific probabilities of selection, as shown in the lower equation.

# 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**

### Slide 39

There are additional reasons why individual sampling weights tend to be unique. As part of the weighting process, we've already shown that stratification and clustering affect the weights, but you could also build in adjustments for different probabilities of nonresponse across the different types of clusters.

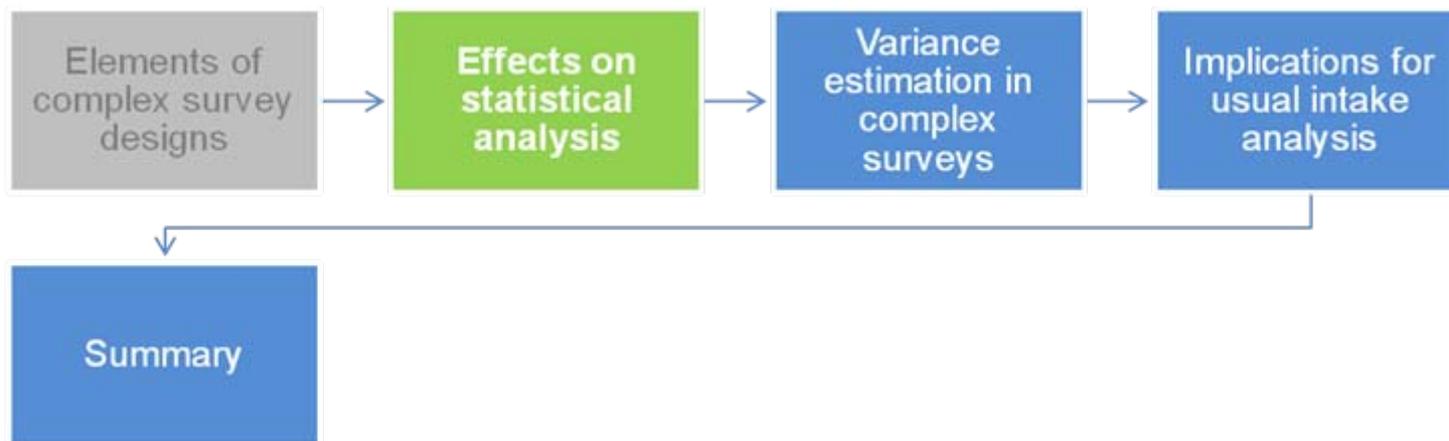
I mentioned earlier that sometimes it is of interest to estimate the total number of people in the population with a particular characteristic, and that doing this has implications for the calculation of weights. Now, you want your sample to be representative of the population, which is often taken to mean that you want your weighted sample counts to reproduce demographic characteristics of your population, say, the numbers falling into sex/age or race-ethnicity groups, which you might "know" from a census. There is a technique called "poststratification" that is used to ensure that your sampling weights satisfy this goal.

# 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

## Slide 40

So to summarize this first part of the talk, I've told you that complex survey methods are often used to collect data used for nutrition monitoring of large populations, and that these complex survey methods may use some or all of the techniques of stratification, clustering, and weighting. Stratification provides a way of balancing precision of subpopulation estimates, while clustering can decrease your sampling costs at the expense of precision. Weighting is usually necessary to account for the effects of stratification and clustering, but may also be used for additional adjustments to make sure your sample is representative of your population.



# EFFECTS ON STATISTICAL ANALYSIS

## Slide 41

Now, I'm going to talk about some ways you have to modify your approach to statistical analysis when your data are from a complex sample instead of a simple random sample that most of your statistical techniques were originally developed for.

## 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

## Slide 42

I've mentioned before that incorporating some or all of the complex survey design elements can affect the precision of your estimates, and I'll talk more about it later in the context of standard errors, but I also want to talk about the companion to precision, which is bias. In practice, you need to account for the effects of all of the survey design elements that went into collecting the data you have at hand.

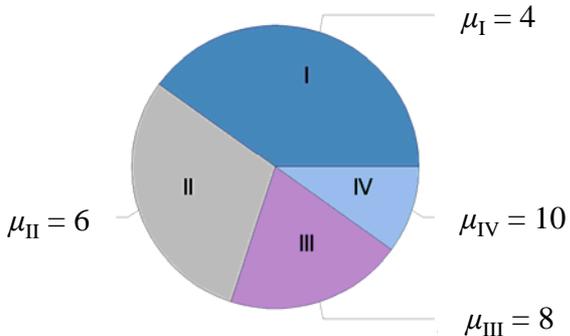
## All survey design elements must be accounted for

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

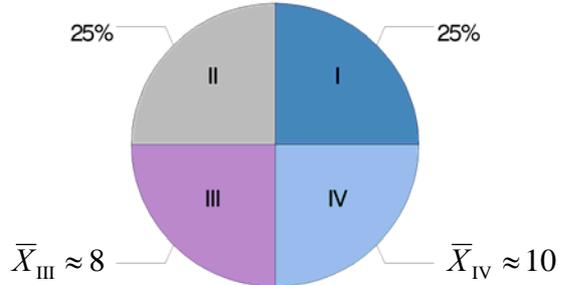
### Slide 43

I said before that weighting helps ensure that your sample is representative of your population; well, this means that you need to take the weighting into account to make sure that your sample-based estimates are unbiased for the population parameters they are supposed to be estimating.

# Weighting required to account for bias



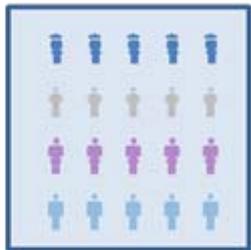
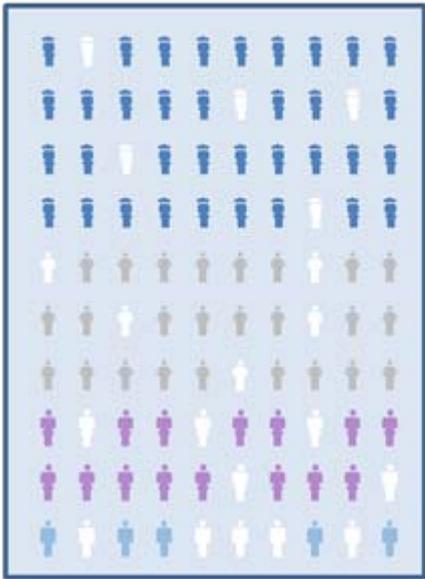
Population



Sample

Stratum

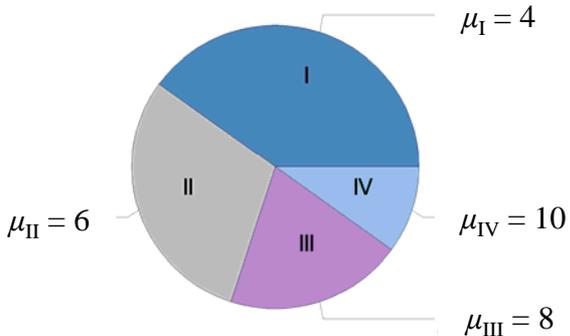
I  
II  
III  
IV



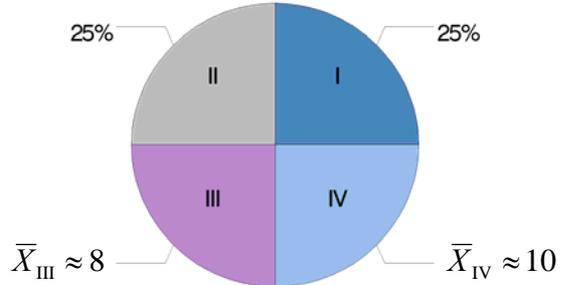
## Slide 44

To show you what I mean about how bias can come about, let's return to our hypothetical stratified population of size 100, from which we selected a sample of 5 people from each stratum. Now, on the left-hand pie chart, I've assumed that the true mean of some characteristic of interest differs by stratum, with Strata I and II having small values ( $\mu = 4$  and  $6$ ) while Strata III and IV have larger values,  $8$  and  $10$ . Now, under our balanced stratified sampling scheme, we expect to get sample means within strata that are close to the corresponding values in the population, shown on the right-hand pie chart by having  $\bar{X}$ -bars approximately equal to the same  $4$ ,  $6$ ,  $8$ , and  $10$  values.

# Weighting required to account for bias



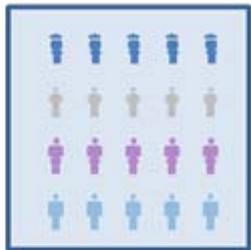
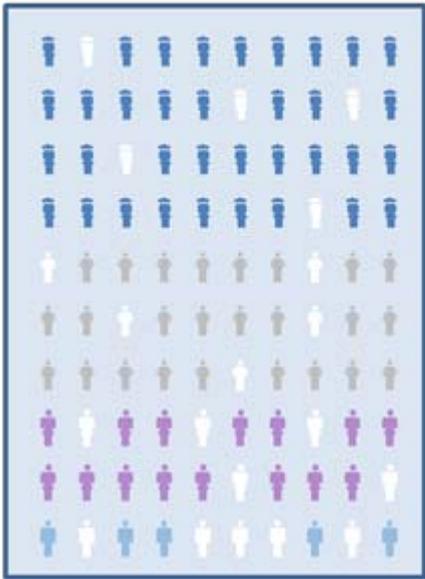
Population



Sample

Stratum

I  
II  
III  
IV



- Unweighted sample mean dominated by large values in oversampled strata

## Slide 45

However, if we just take the unweighted sample mean, we will get a biased estimate of the population mean, because our estimate will be dominated by the large values in the oversampled strata. Though it's not shown here, the population mean is 6, but the unweighted sample mean would tend to average 7, a 16 percent overestimate. I'll show you a real-life example from NHANES later on in this section of the webinar, where we'll see this bias issue come up again.

## All survey design elements must be accounted for

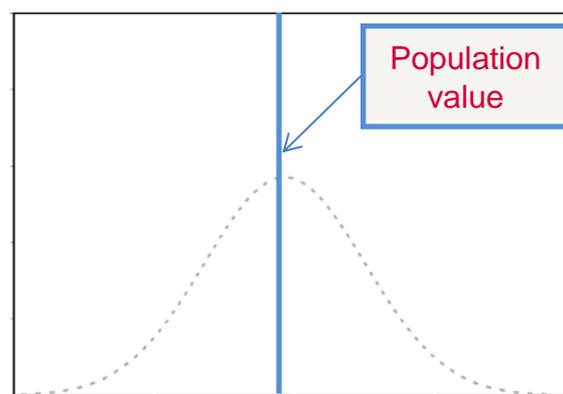
- Weighting required to minimize bias in survey-based population estimates
- **Stratification, clustering, and weighting affect standard errors of estimates**

## Slide 46

But, first, I'm going to sharpen up my previous statements about how survey sampling affects the precision of estimates, by addressing specifically how the design elements affect the standard error of estimates

# 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)



Sample Estimate

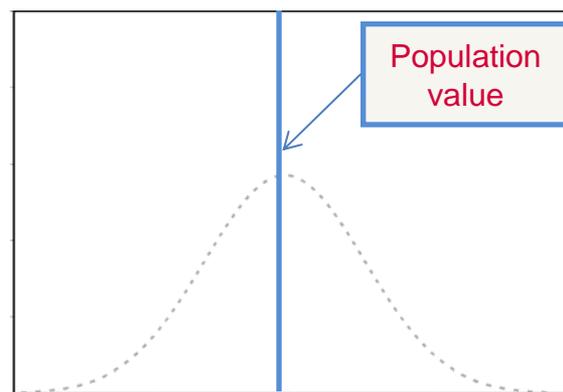
## Slide 47

Well, what exactly is the standard error of an estimate? In spirit, it's a measure that reflects the variation expected across repeated sampling of a population. That is, if you pick a sample from the population and compute some estimate of a population parameter, like the mean, then go and select a different sample from the same population and repeat the process, you will get a different value for the estimate, because you picked different people the second time. If your sampling scheme is doing what it is supposed to, most of the samples will yield estimates close to the true population value, while a few samples will yield estimates far away. The picture at the bottom of the slide shows graphically what I just said—that the sampling distribution of the statistic has most of its mass near the population value, and less mass far away.

For many estimators that have been proposed as the result of statistical modeling techniques, it has been shown that, under certain conditions, the sampling distributions are approximately normal. Showing that this property holds is sometimes referred to as establishing a Central Limit Theorem for the estimator. This is the first time I've used this word "estimator" instead of the word estimate, and I want to briefly point out the difference between the two. By estimate, I mean the number that you compute from your sample data, that you hope is close to the population value of interest. However, by "estimator," I mean the general statistical formula used to compute the estimate. For example, you can get an estimate of the median of the usual intake distribution using several statistical methods, such as the ISU method or the NCI method that have been talked about in earlier webinars. The numbers you get, the estimates, might be very close for the different methods, but the procedure that you used to obtain them could be very different. So we talk about the ISU estimator for the median usual intake, or the NCI estimator for the median usual intake.

# 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  $\Rightarrow$  smaller standard errors
- Used to construct significance tests, confidence intervals assuming **asymptotic** normality



Sample Estimate

## Slide 48

By definition, the standard error, often abbreviated s.e., is the standard deviation of the sampling distribution of an estimator. As a consequence of what I said before, the more independent pieces of information that are used to construct each sample estimate, the smaller the standard error. So, if you are sampling 20 percent of your population each time, you will have smaller standard errors than if you are sampling 2 percent of the population.

Also, if a Central Limit Theorem has been established, you can appeal to the asymptotic normality to construct significance tests or confidence intervals. For example, you have probably all at some time or another constructed a confidence interval by adding plus or minus 1.96 times the standard error to a sample estimate—that 1.96 comes from the normal distribution.

# 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

## Slide 49

Now, I have described the concept of a standard error as having to do with repeated sampling of the population but, in practice, you only get one sample. To do significance tests or make confidence intervals, you need to estimate the standard error from the data at hand.

It turns out that basic statistical theory provides ways to estimate standard errors from sample data for a wide range of estimators. When statisticians try to develop these methods, they often use the words “smooth” versus “non-smooth” to differentiate between estimators that have well-behaved asymptotic properties and estimators that do not. Luckily, many estimators fall into the “smooth,” well-behaved category; for example, means and other mean-like statistics such as regression parameters or ratios of means.

## 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

## Slide 50

On the other hand, empirical percentiles such as medians, or functions of them such as interquartile ranges, are classic examples of the sort of “non-smooth” estimators for which standard error estimation is especially challenging. Such estimators do not act sufficiently “mean-like” to permit easy derivation of Central Limit Theorems, and their sampling distributions are less well-behaved. I’m not saying it can’t be done; I’m just saying it is not as easy. So, testing significance or constructing confidence intervals for such estimators may have to be done using alternative methods, where you can’t necessarily rely on asymptotic normality. However, the standard error still retains its interpretation as a measure of variation over repeated sampling.

## 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

## Slide 51

In the general complex survey scenario, when you have strata, clusters, and weights in the picture, behavior of the sampling distribution is often derived by trying to formulate the estimator as a function of weighted cluster means within strata, and then working with the asymptotic distributions of those cluster means. This gets back to the overall theme I started off this section with: It is important to account for all the survey design factors.

I have to give you a word of warning about the software typically used to do statistical analysis—not all of it is fully “survey-aware.” There are a lot of statistical procedures that allow the use of some sort of weight, like weighted least squares analysis, that won’t necessarily do what you need them to. These types of procedures can treat the sample weight as a population expansion factor and just pretend that an observation with a weight of 100 means that 100 observations with identical data were obtained. Or, the weights might be treated as inversely proportional to the variance of the residuals in a regression. These types of “weighted” analyses might not account properly for the stratification or clustering.

# 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  $\Rightarrow$  30 d.f.

## Slide 52

I said earlier that selecting clusters instead of individual units gives you fewer independent pieces of information. In a complex survey, it is the number of clusters that is important, since they are independent units, not the people within the clusters. This concept of counting such independent pieces is formalized in the notion of “degrees of freedom.” Remember from your basic stats course when you were testing the significance of a sample mean, and your sample is small, you had to look up critical values in a t-table, rather than a z-table? And the value you looked up was on a line in the table indexed by one less than the sample size? That number,  $n-1$ , is the degrees of freedom. If your mean comes from a general complex survey, the  $n$  is the total number of clusters, and you have to subtract one degree of freedom for each stratum, as you see in the equation. For the four-year cycle of the NHANES survey from 2003-6, the sample is comprised of over 20,000 individuals. However, when you analyze the NHANES data, you treat them as a stratified cluster sample with 60 clusters, 2 in each of 30 strata. I say you treat it this way—the actual sampling design of NHANES is kept secret to maintain confidentiality—but this two-cluster-per-stratum pseudo-design works well as an approximation. This can make a big difference in the way statistical procedures report their significance tests, if the procedures are not fully survey-aware.

# 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

## Slide 53

Now, I get to the real-life example I promised you earlier. I'm going to look at total calcium intake for a subsample of women from the 2003-6 NHANES. The subsample comprises 2,601 women ages 31-70 who had a reliable first 24HR. My parameter of interest is the population mean intake of calcium, where I want to include calcium that comes from food and calcium that comes from supplements. As Regan will explain in the next webinar, the data I'll use for estimation are based on a combination of 24HR data and data from the NHANES dietary supplement questionnaire.

# 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

## Slide 54

Now, if you recall from webinar 2, you generally don't have to do a lot of complex modeling if all you want is to estimate the mean usual intake for a group. Under the usual assumptions, you can just take the mean of your 24HR data. But in SAS, the statistical analysis software I usually use, I can think of at least four ways to get an estimated mean and standard error for this set of NHANES data. I can use the procedure UNIVARIATE alone or with two options to incorporate the weights, either the WEIGHT option, or the FREQ option. The FREQ option uses the population expansion idea and pretends that each real observation represents multiple identical observations, where the multiple is given by the weight value. Or, I can use one of SAS's fully survey-aware procedures, say SURVEYMEANS, with syntax that specifies all three elements of the NHANES design—strata, cluster, and weight. Only the last way incorporates all the design factors.

# 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

## Slide 55

Ok, here are the results. You see going down the second column that the first way, which ignored the weights, yielded a different estimate than the other three that took weights into account. You also see differences in the third column for the estimated standard error, and in the fourth column, where I listed the degrees of freedom that the procedures used to test whether the mean was different from zero.

# 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
<b>SURVEYMEANS</b>	<b>1115</b>	<b>27</b>	<b>30</b>

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

## Slide 56

But, remember, only the last way correctly accounts for all the survey design factors. Putting in the weights, regardless of how you do it, keeps you from making the approximately 8% underestimate that the most naïve approach would produce. Remember the bias issue? But the standard errors can be very wrong. In general, ignoring the sample design produces standard errors that are too small; compare the 13 to the 27 that is the better estimate. But depending on how you treat the weights, you can get wildly different answers, as shown in the middle two rows, which show that the population expansion idea on the third row really falls apart. This is because the NHANES weights suggest that there are about 70 million women ages 31-70 in the population. This is due to that poststratification I mentioned early on.

# 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

## Slide 57

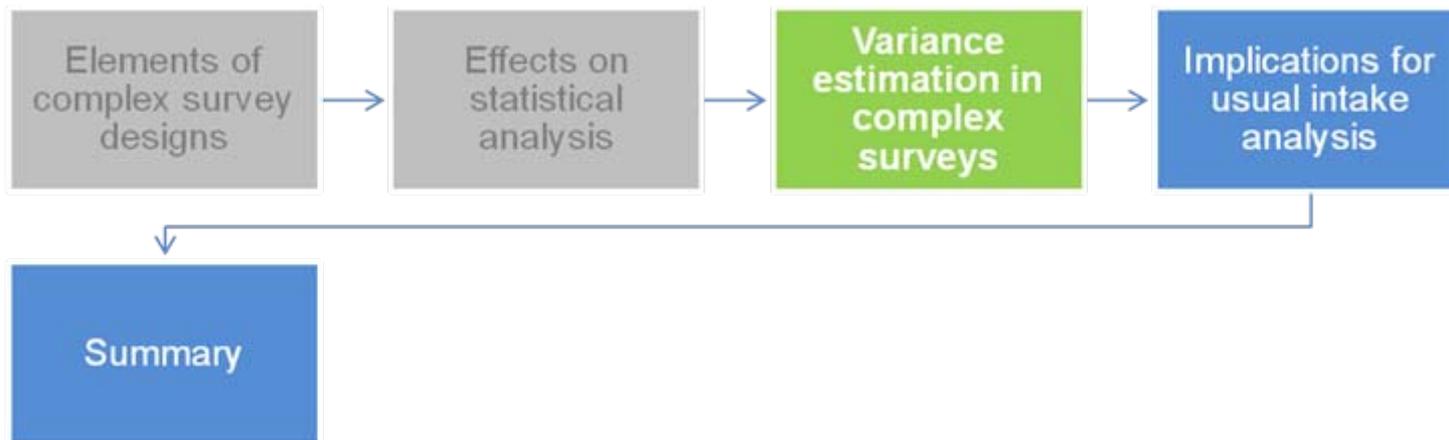
This problem, where not all software is fully survey-aware, is in part due to the fact that it takes a good bit of work to extend statistical techniques developed for simple random samples to the complex survey case. As you might have guessed from the examples I've discussed so far, it turns out that simple estimators like means are the easiest to work with, and inference based on the classic t-test is relatively easy to extend to the survey setting. Survey statisticians have established Central Limit Theorems and standard error formulae for many mean-like statistics, and these have been incorporated into standard survey analysis software packages such as SAS, SUDAAN, Stata, and R. However, some statistical methods are more difficult to extend—for example, I'm not aware of a survey analogue to the likelihood ratio test that is often used to choose between competing models.

# 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

## Slide 58

So the takeaway messages for this section of the webinar are these: It is important to account for any and all of the survey design factors when you analyze survey data, although your analysis tools may be limited because no one has previously extended them to handle complex survey situations. You are probably better off trying to see if simple t-tests, or similar classic methods for which CLTs have been established, can be used to answer your research question because, then, all you need for inference is an estimated parameter and a proper estimate of its standard error.



# VARIANCE ESTIMATION IN COMPLEX SURVEYS

## Slide 59

In keeping with what I just said about needing proper estimates for standard errors, I'm going to turn to the next topic in my outline—variance estimation in complex surveys. This section could just have easily been called “standard error estimation in complex surveys,” but if you ever want to get more information about the topic, you'll have better results searching for “variance estimation” as a keyword.



## Slide 60

If you look at the standard textbooks on this subject, you'll find that there are two basic approaches to variance estimation in complex surveys. The first is to use the method of Taylor linearization, sometimes abbreviated as simply linearization, and the second is to use a so-called resampling method, some examples of which are listed here.



## Slide 61

I'll first say a few words about Taylor linearization.

## 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

## Slide 62

This is the approach implemented by default in most survey-aware software, because it allows the use of textbook formulae to compute standard errors for estimators where some sort of Central Limit Theorem has been previously established in the survey context. Now, for some of the new, intricate estimators we have been describing in previous webinars, which use measurement error models to estimate usual intake distributions, these textbook formulae don't exist—yet, anyway. Remember when I told you that percentiles like medians were the classic example of a problematic, non-smooth estimator? Well it turns out that in the particular case of the Monte Carlo-based estimators of usual intake percentiles as used in the NCI method they are especially problematic to handle with the Taylor linearization approach, so we turned to the alternative...



## Slide 63

...resampling methods. In contrast to linearization, resampling methods don't need a textbook formula for every possible estimator. Rather, they provide an estimation framework that has been shown to work for general classes of estimators, like "smooth" or "non-smooth."

## 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

## Slide 64

The basic idea that underlies the theory of resampling methods is that you can emulate taking repeated samples from your population by repeatedly sampling from the data you have on hand. You basically treat your sample as your population in miniature. This idea reflects the spirit of the standard error being related to the variance of the sampling distribution.

Now, I'm first going to illustrate this idea for the bootstrap method in the non-survey, simple random sampling case, to keep things a little clearer.

# Variance estimation techniques

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

## Slide 65

Here we go.

# Example: Bootstrap in simple random sampling

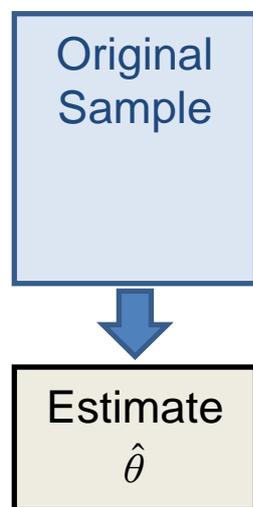


## Slide 66

I want to mention that this illustration of the bootstrap is taken from a presentation by Anne-Claire Vergnaud, and I think it is a really nice illustration. You may see other versions of these slides in later webinars.

OK, to start out with, you have your original sample of data...

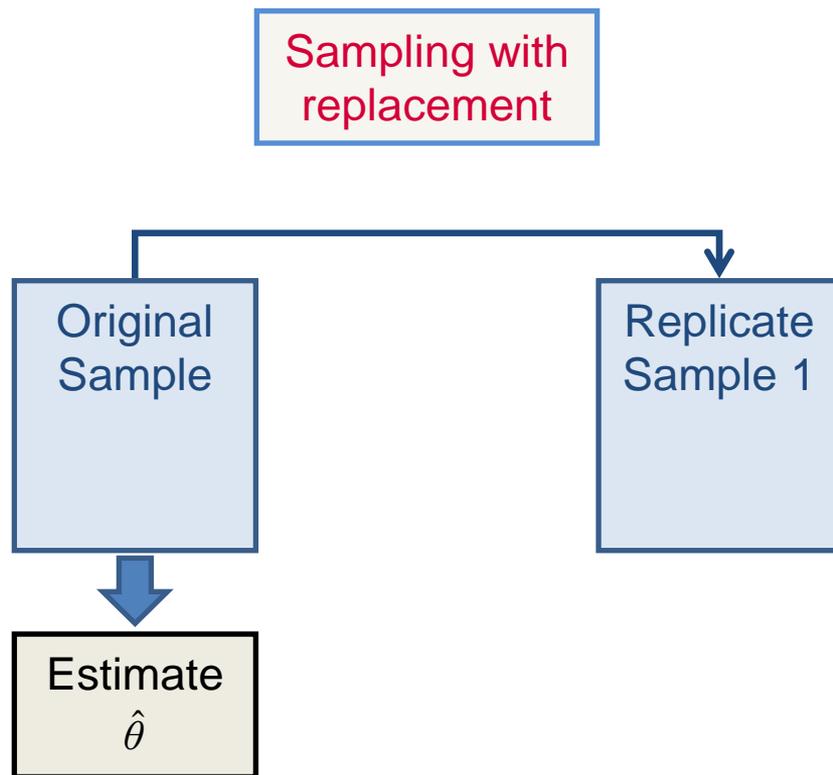
# Example: Bootstrap in simple random sampling



## Slide 67

...and you compute your estimate, call it  $\hat{\theta}$ , of the population parameter,  $\theta$ , that is of interest.

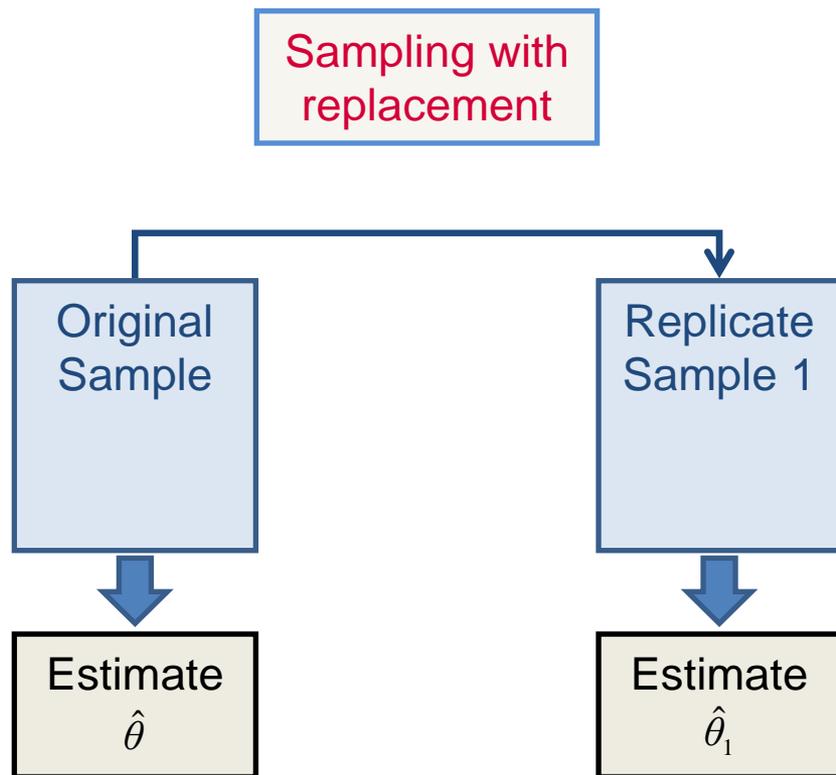
# Example: Bootstrap in simple random sampling



## Slide 68

The bootstrap method says to draw a replicate sample of the same size as your original sample, using with-replacement sampling. If your random sampling scheme involves pulling names out of a hat, every time you draw a name for the replicate sample, you put the name back in the hat for the next draw. This means that some people may appear more than once in the replicate sample and some people might not appear at all.

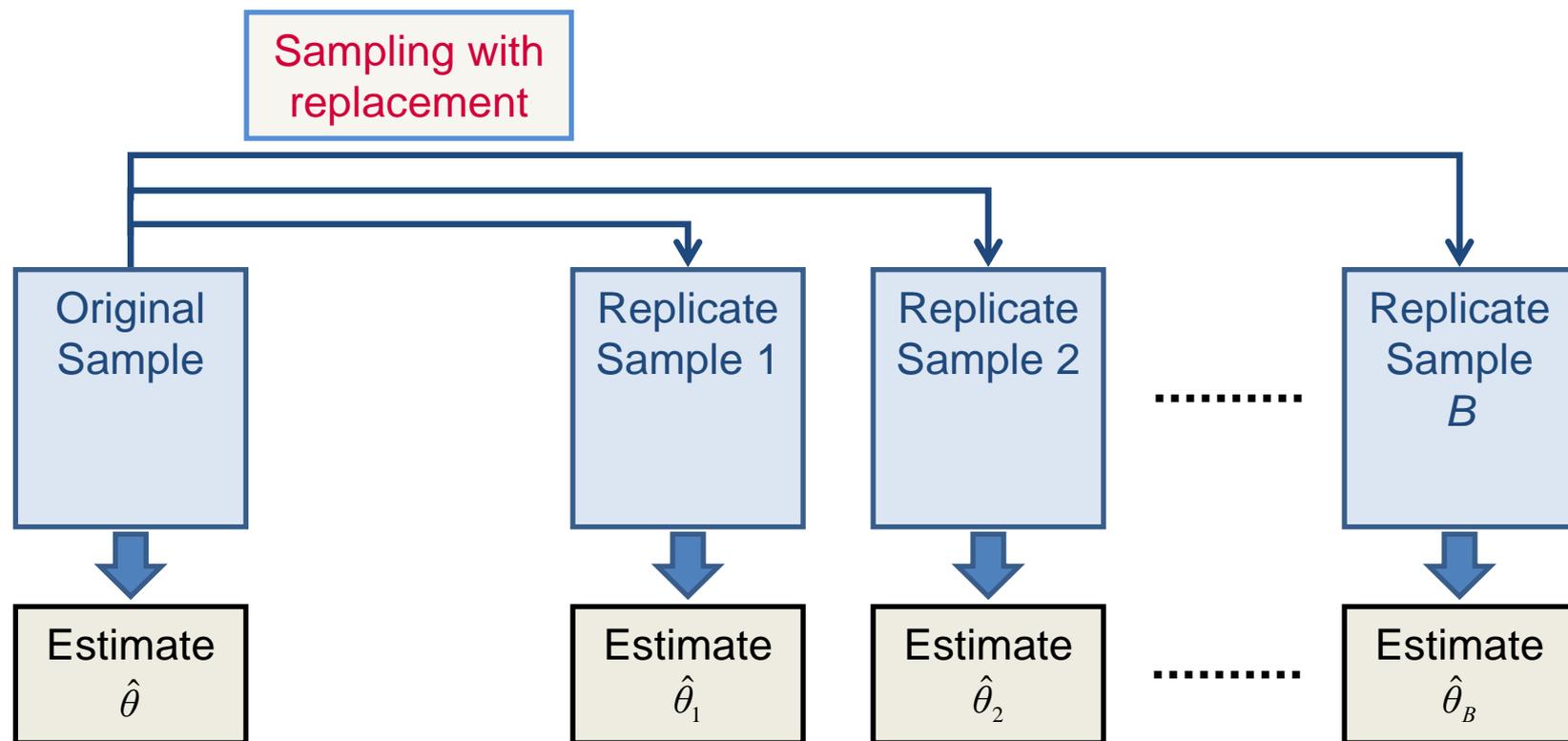
# Example: Bootstrap in simple random sampling



## Slide 69

Then, you use the same estimation procedure to compute an estimate based on the replicate sample; call it  $\hat{\theta}_1$ .

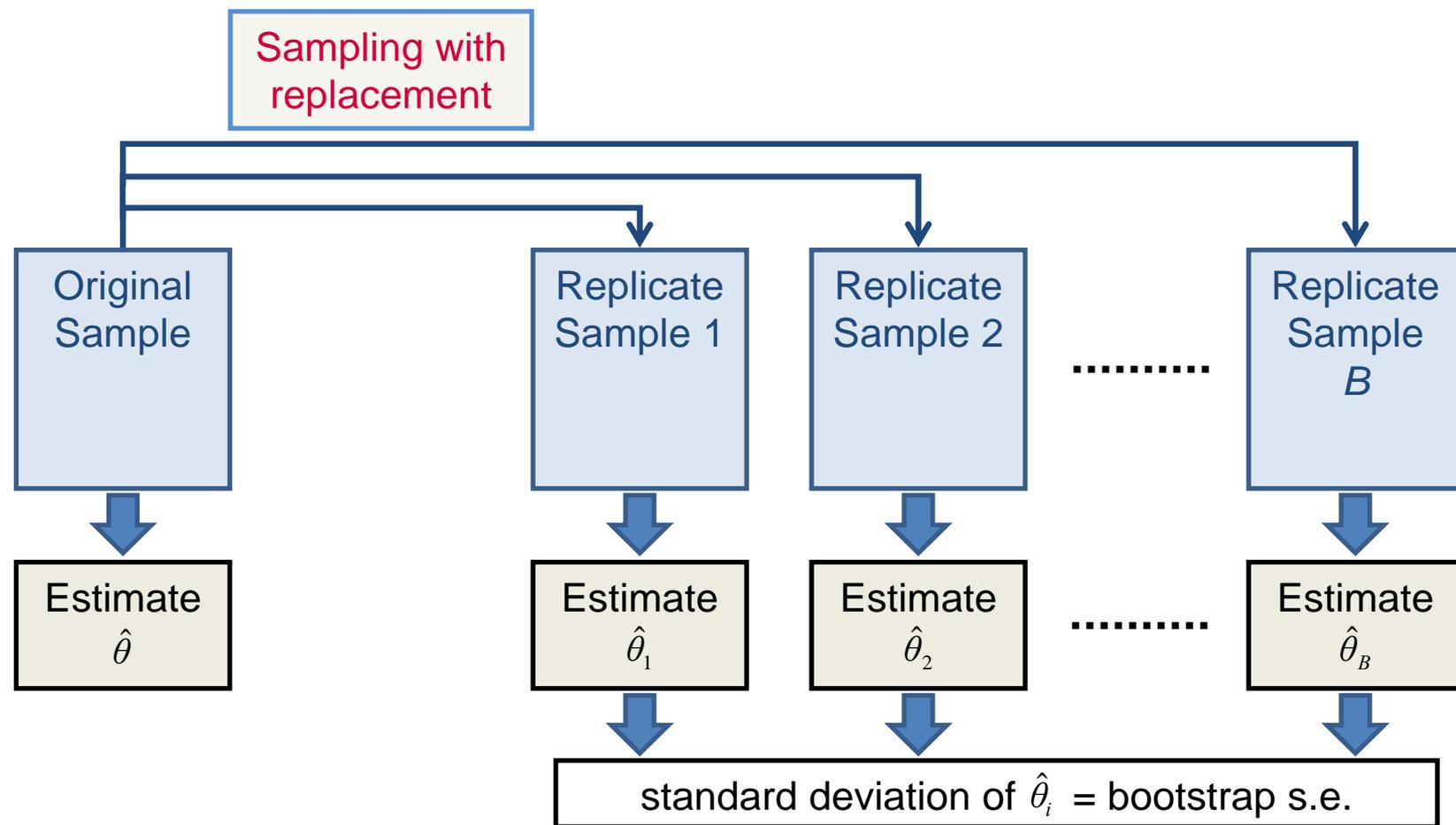
# Example: Bootstrap in simple random sampling



## Slide 70

You repeat this process many, many times, and end up with  $B$  different replicate samples and  $B$  different theta-hats computed from the replicate samples.  $B$  is usually a big number—200, 300, 1000, or more.

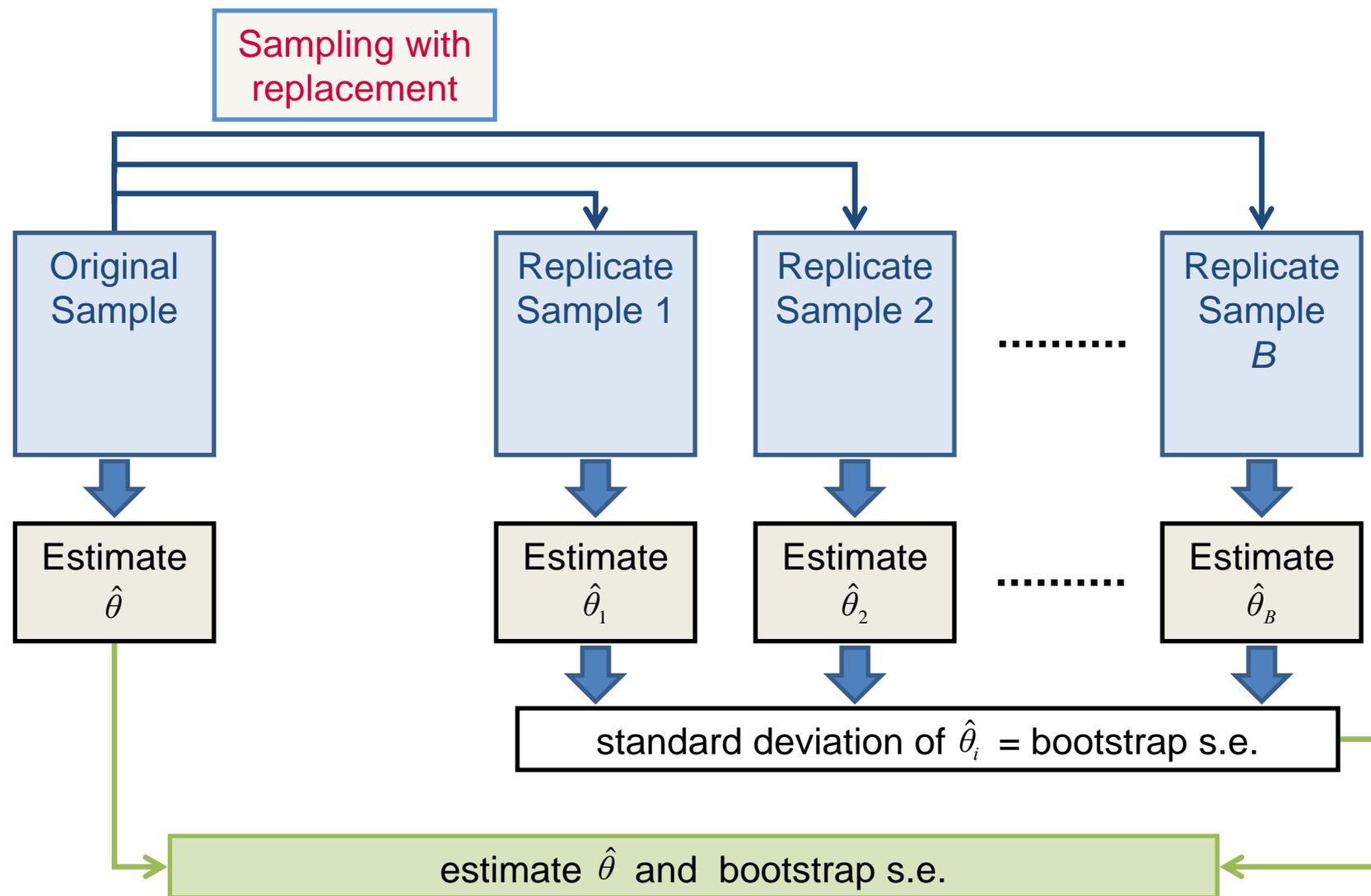
# Example: Bootstrap in simple random sampling



## Slide 71

The sample standard deviation of the theta-hats from the replicates is the bootstrap estimate of the standard error...

# Example: Bootstrap in simple random sampling



## Slide 72

...of the estimate you computed from the original sample.

## 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

## Slide 73

The key to the bootstrap method is with-replacement sampling. If you do your sampling without replacement, you will always get back the original sample. So it is important to allow the situation I mentioned, where in a given bootstrap sample, some individuals appear multiple times and some appear not at all.

Now, the reason I talked about the bootstrap for simple random sampling first is here in the last bullet. I claim that the number of times an individual appears in a bootstrap sample is analogous to a sampling weight, at least in the population expansion sense, because identical observations really do appear multiple times in the replicate samples. It is this property that really helps to implement resampling methods for complex surveys...

## 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

## Slide 74

...because we can operationalize resampling from the original sample by creating sets of weights for the original sample and each of the replicate samples we want to draw. In the bootstrap example for simple random sampling, all the weights for the original sample are one, but the individual weights created for a bootstrap replicate vary according to the number of times the individual showed up, and may be zero.

This feature eliminates the need to store multiple copies of data sets with many analysis variables. You simply need to add one additional weight variable per bootstrap sample, and a set of constant weights for the original sample, and you can use a “weight-aware” procedure to compute the bootstrap estimates you need for estimating the standard error. Now you can think of each set of bootstrap weights as a “perturbed” version of the weight set for the original sample, which were all one.

## 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

## Slide 75

All the resampling methods applied to complex surveys can take advantage of this idea of implementing resampling by “perturbing” the base weights of the original sample. The bootstrap, jackknife, and BRR methods applied to complex surveys differ in the numbers of weight sets needed, the way the weight sets are constructed, and the formulae for expressing the standard error estimate as a function of the variability among the 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

## Slide 76

Now, I illustrated the bootstrap for the simple random sampling case, and I said you could do the with-replacement sampling by pulling names at random out of a hat. If you try to extend the bootstrap to complex surveys, you have to do your bootstrap sampling according to the sampling plan used to draw your original sample. In the general case, that means mimicking the drawing of clusters from strata, and possibly poststratifying all your replicate weight sets. If the number of weight sets,  $B$ , is very large, this can be very computationally intensive to do the sampling, not to mention having to run all the analyses. This computational burden is one reason that the bootstrap is not used as often in complex surveys as some of the other resampling methods.

It is kind of a shame, really, because the bootstrap method has some nice features. One is that it offers a robust method for constructing confidence intervals. You can base the bounds of your CI so that they bracket 95% of the bootstrap estimates, instead of taking plus or minus 1.96 times the standard error. This may work better for poorly behaved sampling distributions of non-smooth estimators like percentiles, or for estimators like proportions where the asymptotic normality approximation is adequate only for very large sample sizes.

## 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)

## Slide 77

I said the bootstrap is not used as often as some other methods, but it does get used. In fact, it is recommended to use the bootstrap for standard error estimation in Nutrition Cycle 2.2 of the Canadian Community Health Survey, and I'm aware of its use in two applications of the NCI method for usual intake modeling. The first was a simulation study that Dr. Janet Tooze did, showing that the bootstrap does work for estimating model parameters and percentile estimates. Based on that result, I recommended using the bootstrap in more recent joint work with Janneke Verkaik-Kloosterman and her colleagues in the Netherlands, where we modeled usual nutrient intakes for young children based on data from the Dutch National Food Consumption Survey.



## Slide 78

I'm going to talk about the remainder of the resampling methods, assuming I'm in the general complex survey case where I have strata, clusters, and weights.

Now, we move to talking about the jackknife method.

## 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

## Slide 79

For the jackknife procedure, you need to create one set of perturbed weights for each cluster in your sample. The main idea of the jackknife is to operate on only a few clusters at a time. The  $k$ -th weight set deletes (or assigns zero weight to) all the observations in cluster  $k$ . When you do this, you lose a chunk of weight because cluster  $k$  is no longer in your sample. The jackknife corrects this by redistributing the chunk of lost weight among other observations in the same stratum as cluster  $k$ . However, it leaves the weights unchanged for observations in clusters from 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

## Slide 80

For surveys with many clusters, you still have to generate a lot of weight sets, but not as many as you would with the bootstrap. And, as I said before, you may want to poststratify each weight set to your known 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

## Slide 81

Now, the theory of the jackknife says that it works for the same class of estimators that Taylor linearization works for—generally, the smooth kind of estimators. For purposes of usual intake estimation, this means that it will work for the model parameters, since they are regression slopes, one of the mean-like kinds of statistics. Back in my graduate school days, I did some simulation studies and theoretical work that showed that the jackknife can also be used to obtain approximate standard errors for usual intake percentiles computed using the ISU method. But similar simulation studies suggested that the jackknife is not suitable for Monte Carlo-based usual intake percentiles such as those used in the NCI method...



## Slide 82

...which is why the NCI method applied to complex surveys like NHANES usually uses the Balanced Repeated Replication, or BRR, method that I'll talk about next.

# 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**

## Slide 83

The BRR method is one that is limited to the case of stratified cluster designs that have two PSUs per stratum. I've already said that the NHANES is analyzed as if this is the case.

Remember, I said the jackknife operates on a few clusters at a time. The BRR is much more aggressive in its perturbation of the weights. In constructing the  $k$ -th weight set, BRR deletes the observations in half the clusters, and has to make up for this by doubling the weights for observations in the remaining clusters. We say that BRR uses perturbation factors of 0 and 2, as compared with the jackknife, which uses perturbation factors of 1 (i.e., no perturbation) for most of the observations. Now, it is true that you want to have at least two clusters per stratum to estimate standard errors, so jackknife perturbation factors can be as big as 2, but may be much smaller if many clusters are selected in some strata. Regardless, because it perturbs a lot of weights at a time, I call BRR more aggressive.

# 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

## Slide 84

It turns out that the BRR method often requires fewer weight sets than the jackknife method. The exact number of weight sets you need for BRR is the smallest multiple of 4 that is greater than the number of strata. That rule is a result of the way the weight sets are constructed. The choice of which cluster in a stratum to zero or double is determined by the elements of what is known as a Hadamard matrix, which is a special matrix with elements plus or minus 1 that is constructed to have orthogonal rows and columns. Hadamard matrices come in orders that are multiples of four, so that is where the rule on the number of weight sets comes from. For the NHANES example where there are 30 strata, you need 32 BRR weights, but you would need 60 jackknife weights.

The orthogonality property of the Hadamard matrix is used to ensure that no single cluster is disproportionately represented in the standard error calculation. The influence of each cluster is balanced; that's where the name balanced repeated replication comes from.

# 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

## Slide 85

The classic BRR method is sometimes too aggressive—you basically throw out half your sample for each weight set, so BRR standard errors can sometimes be unstable. A modified BRR method was developed by Robert Fay at the U.S. Census Bureau that uses less extreme perturbation factors. Instead of zero and 2, Fay's method might use 0.3 and 1.7.

As in the jackknife and bootstrap methods, you may want to perform the same poststratification adjustments to each BRR weight set as you did for the original sample's weight set.

## 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

## Slide 86

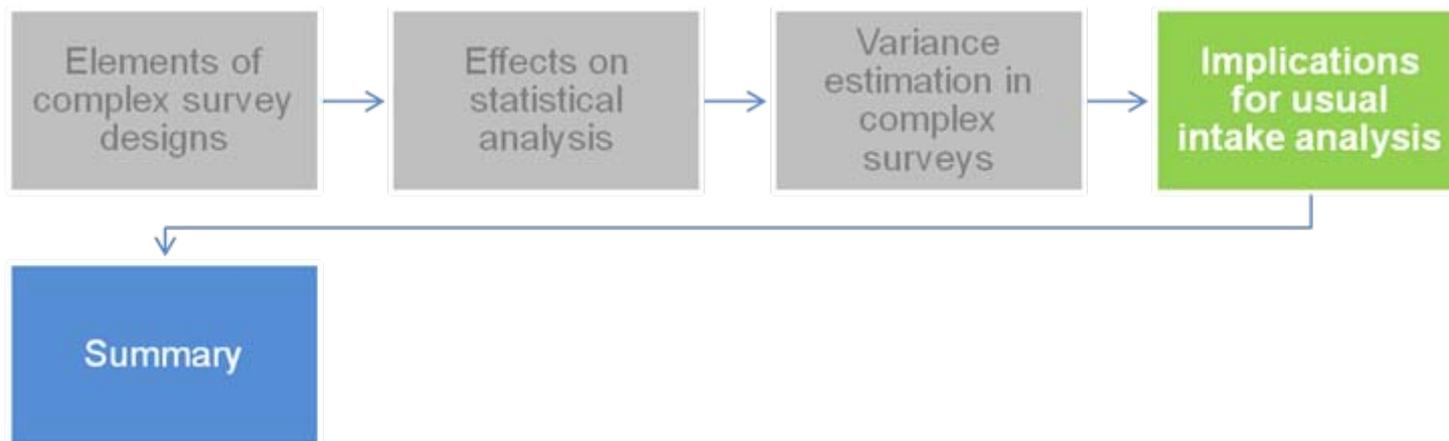
Now, for simple statistics like population means, you can use Taylor linearization (or, equivalently, the jackknife) to do analysis of usual intake data based on dietary data that come from the What We Eat in America portion of the NHANES survey. In fact, I did exactly that in my real-life example earlier in this webinar. Well, the BRR works for those cases, too, but more importantly, BRR also works for those problematic Monte Carlo-based percentile estimates that the other methods did not work for. So, this has been the standard error estimation method of choice when we have applied the NCI method to NHANES data.

# 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

## Slide 87

Here is a summary of what I've said about standard error estimation in complex surveys. The standard "survey-aware" statistical analysis software typically uses Taylor linearization to compute standard errors. While it is limited to basic, "mean-like" estimators, it is really fast because it just uses textbook formulae. However, for complex estimators that are computed using techniques for which textbook formulae aren't available, resampling methods offer an alternative.



# IMPLICATIONS FOR USUAL INTAKE ANALYSIS

## Slide 88

I'm now going to give some examples of what sorts of implications these survey design factors have when you want to actually do an analysis using these complex models for usual intake.

## 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

## Slide 89

Here's a typical research question you might want to answer:

Based on my data, obtained under a complex survey sampling plan, what is the usual intake of component X among subgroup Y in my population?

To answer the question you have to consider:

The estimator of interest: Are you looking to obtain an estimated mean, an estimated distribution, or an estimated regression slope?

Method of analysis and its data requirements: Has the analytic method ever been applied in the context of complex surveys? Do you have the right kind and amount of data for the method to be applied?

Technique for variance estimation and how to use software to properly implement: Can you use standard survey analysis procedures in SAS, SUDAAN, Stata, or R? Or do you have to use a special software package? If the special package is not survey-aware, how do you trick it into giving you results that you can use as part of a resampling approach to estimate standard errors?

# 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

## Slide 90

Here's how those considerations came into play for my earlier example about the mean usual intake of calcium. My desired estimator was the mean of the usual intake distribution, and the analytic method boiled down to computing the (properly weighted) mean of all the valid first-day 24HRs from the NHANES data. Because the analytic method was so basic, I could use the default Taylor linearization approach to estimate standard errors that is part of the common survey-aware software packages like SAS, SUDAAN, or Stata. I think I probably could have done it in R as well.

## 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

## Slide 91

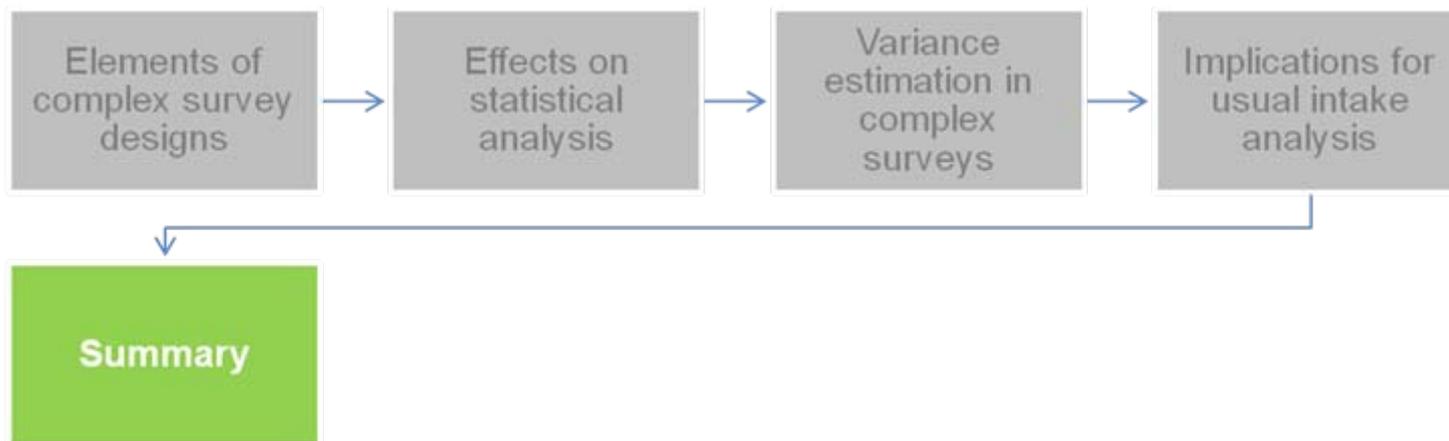
Suppose I wanted to estimate the distribution of usual intake from my NHANES survey data, and I wanted to use the NCI method. The data I need are all the valid 24HRs, not just the valid first-day recalls, and as I have said, for NHANES analysis, I would use the BRR method for variance estimation. The SAS macros to implement the NCI method are partially survey-aware, in that they perform weighted analysis, but you have to do some extra work to account for the stratification and clustering. Part of this extra work goes into constructing the BRR weights; the rest is in running the replicate analyses and computing the standard errors. In the past, researchers in the Agricultural Research Service of the U.S. Department of Agriculture have constructed sets of BRR replicate weights suitable for analysis of the dietary data portion of NHANES, the What We Eat in America portion, for several different combinations of NHANES survey cycles. These weight sets have been provided to various researchers doing these kinds of analysis, and we've been told that there are plans in the works to make these weight sets more widely available. Also, the advanced portion of the NHANES dietary Web tutorial provides a detailed walkthrough of the analysis portion.

## 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

## Slide 92

Alternatively, you might want to estimate the distribution of usual intake from the Canadian Community Health Survey, applying the ISU method to all valid 24HRs from that survey. As I mentioned before, Statistics Canada recommends using the bootstrap procedure and provides official bootstrap weight sets for this purpose. Since you will be using the ISU method, you need to use the ISU software, either the original version called SIDE, or one of the later versions, C-SIDE, or PC-SIDE.



# SUMMARY

## Slide 93

OK, we are coming to end of this webinar.

## 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

## Slide 94

And I want to leave you with some key messages. First, as I mentioned way back at the beginning, the data used for monitoring usual intakes among populations are usually collected using complex survey methods, and I gave you some reasons why. I also showed you how the elements that give such methods the “complex” name affect the way you have to compute point estimates and standard errors. These elements—stratification, clustering, and weighting—all have to be accounted for.

## 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

## Slide 95

I showed you standard error estimation is an important part of dealing with complex survey data, and showed you how complicated it can be. While means and “mean-like” statistics are handled easily by the default Taylor linearization method implemented in survey-aware software packages, there are some statistics like percentiles that are less well-behaved. For these kinds of non-smooth statistics, I described alternative methods based on resampling (the BRR and bootstrap) that could be used. Remember that jackknife is supposed to work only where linearization does.

## 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

## Slide 96

Finally, I want to stress that there is no one-size-fits-all approach to modeling usual intake from complex survey data. You have to tailor your approach to your particular research question, the data you have available, and the method you want to use to do the analysis. You have to know just how survey-aware your software is. If it can't handle the standard error estimation properly for a general complex survey, maybe you can trick it into giving you answers suitable for use in one of the resampling methods. And, finally, your approach is going to depend upon how much statistical expertise or support you can call on. The NHANES dietary Web tutorial, one of the suggested resources for this webinar, may be a good place to start looking.

# QUESTIONS & ANSWERS

Moderator: Regan Bailey

Please submit questions  
using the *Chat* function

## Slide 97

Thank you Kevin. We'll now move on to the question and answer period of the webinar.

## Measurement Error Webinar 4 Q&A

**Question:** For the examples that you provided on slide #56, is the standard error for the survey means procedure similar to the standard error derived from the resampling methods?

Yes, they would be very similar in almost all cases. The only time they're not going to be exactly the same or very close to the same is when—if you do the resampling methods, say, using Jackknife or BRR and you poststratify your weights, which is what's usually done, you'll get a standard [error] estimate that reflects resampling of a survey where you want to poststratify. On the other hand, the SAS survey procedures generally don't account for this poststratification idea, so it's possible that if the poststratification makes a difference in your standard error that your resampling method will give you a smaller standard error than the Taylor series would. But most of the time, they're going to be equivalent, yes.

Mean—that's one of those smooth statistics. All of your sampling methods are going to give you something that's pretty close to the Taylor series method because the estimator is so basic. (*K. Dodd*)

**Someone asked if you could use the NHANES data without the weights, and would they still be considered nationally representative?**

I think that the idea of nationally representative kind of boils down into saying—it kind of requires you to use the weights to really talk about nationally representative data in terms of counts and things like that. Now, some people have considered the idea of saying, "Well, what if I only care about some sort of biological mechanism between diet and health?" And they may justify ignoring the sample design that way and saying, "These are people and this is biology and this is what it is." But most of the time, I think NCHS, the people that run the survey, and I think a lot of other people, would say that you probably should think about using the weights if you really want to be safe in calling your results representative of the U.S. population. (*K. Dodd*)

**Does either the Tooze article or the Dutch publication demonstrate variance estimation for percentiles for episodically consumed foods?**

No, it actually didn't in those two things I looked at because we were looking at nutrients, but I have done some similar studies once before that showed that for episodically consumed foods they seem to—BRR does seem to work well for these percentiles. This is something I did for a

presentation at one of the international conferences for diet and activity methods quite a while ago. So I think that BRR is sort of your go-to variance estimation technique for when you have complex surveys where you want to look at something like percentiles. *(K. Dodd)*

**Are the NCI SAS macros to construct BRR weights fairly straightforward to use? Or would you need some experience in this area to understand and apply them properly?**

Well, the NCI macros themselves don't actually—we don't actually have macros posted yet to do the actual BRR weight construction. As I said, the people at ARS, the Agricultural Research Service, at USDA have in most cases provided the weights for us. I have constructed BRR weights before for parts of the survey and I do have some internal macros that I think as a group we're going to discuss whether or not we need to make those available. But, hopefully, the actual use of the NCI macros for doing the variance estimation is explained quite well in the dietary Web tutorial, the advanced section. So I think that's where you have a lot more information available. *(K. Dodd)*

**I heard you use the phrase, "Taylor linearization" in your presentation, and other people use the phrase "Taylor series." Are those interchangeable and if not, how do they differ?**

They refer to the same general idea. I think in webinar 2 and maybe webinar 3 we talked about expressing the mean or the variance in terms of a function of some data and we used the Taylor series argument to say that a function of this random variable can be written this way. Taylor linearization uses the same sort of idea, so they are very, very similar. *(K. Dodd)*

**While we still have you in the hot seat, here's a few more questions. Can we use the resampling methods to estimate the means? And what advantages and disadvantages are related to that?**

You can use the resampling methods to estimate the mean. They're usually going to give you about the same answer, and so if you have software, you can do it much faster. I guess the disadvantage of using the resampling method is it's much faster to use the Taylor series way than do the resampling in general. But to go back to what I just mentioned a minute ago about what happens when you poststratify your replicate weights, you can get a [standard error estimate] that's a little smaller if you're using a resampling method with poststratified weights versus if you

were using just a Taylor series linearization from a software [package] that didn't take that poststratification into account. *(K. Dodd)*

**In bootstrapping, why do you use the point estimates in the original sample and not the replicate?**

Well, because the bootstrap and the other things are technically variance estimation techniques. They are not necessarily point estimation techniques. The idea is that your estimate based on your weighted data that you get from the original sample is set up to be the best estimate it can be. And the only reason you're using bootstrap or BRR or Jackknife is to provide a variance estimate. Now, if you go back and look in the survey sampling literature, you will find that the original derivation of the Jackknife for complex surveys was actually designed to do some sort of—it was actually designed to do a bias reduction most of the time. And we don't treat it that way for purposes of variance estimation because when we're using it as a general variance estimation technique, all we want to do is use it to estimate the variance. We want to take our point estimate from the full sample and treat that as our estimate of the actual parameter of interest. *(K. Dodd)*

**What makes the Canadian Nutrition Survey fit for the ISU bootstrapping method and not the BRR method?**

Well, if I had my druthers, I would hand that off to Sharon Kirkpatrick, who's actually done a lot of work with the Canadian Survey, but apparently, if I understand it, and Sharon is sitting right here so she'll tell me if I'm right, in order to analyze the Canadian data with bootstrap weights, you actually have to go to the research center and do your results there, and they have things set up in a certain way that it's better to use the bootstrap that they provide as opposed to trying to write up your own, develop your own BRR weights and use those and put them all together in the right way. It's just not—I mean, technically, if you had access, unfettered access, to the data, you probably could do BRR instead of bootstrap, but you would have to construct those BRR weights yourself, whereas [Statistics] Canada provides official bootstrap weights for use with its data. And the reason you don't have unfettered access to the data is the confidentiality reason again coming in. So, in general, people aren't going to have the data they need to actually construct BRR replicate weights because they don't have access to the confidential sampling design. So you end up again coming back to the use of using what's officially provided, which is the bootstrap weights. *(K. Dodd)*

**What sample size do you need to be able to generalize to the U.S. population?**

That is actually a question that is best answered by the people who run the NHANES survey. And I just pointed out that the sample sizes for [that are] about 20,000 people in a two-year cycle. I think it tends to get bigger over time, but the number of people that you need depends upon a lot of competing factors. You need to think about how you're going to collect people but also how you're going to do the multistage sampling, how you're going to allocate your resources to driving all over the country to pick people up to get them to come to the MEC. All these different things have to be incorporated into your final decision, and NHANES or the NCHS people that actually draw the sample—those are the ones that really could better answer this question. All I can say is that when we take the data we have in hand from NHANES and we use it and we estimate our standard errors for it and when the standard errors say that this estimate is not very precise, we don't report it because there weren't enough people to actually answer that particular question for that particular segment of the U.S. population. *(K. Dodd)*

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

## Slide 98

Thank you Kevin and thank you to our audience for joining today's webinar. Please join us next week for webinar 5 when we will be discussing the estimation of distributions of total usual nutrient intakes from diet and supplements.