National Cancer Institute

measurement ERROR webinar series

Estimating usual intake distributions for dietary components consumed daily by nearly all persons

Kevin W. Dodd, PhD
National Cancer Institute

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

Slide 1

Hello. I'm Sharon Kirkpatrick from the Risk Factor Monitoring and Methods Branch at the U.S. National Cancer Institute and it's my pleasure to welcome you to the second webinar in the Measurement Error Webinar Series. In our first webinar, we reviewed introductory concepts related to measurement error and provided an overview of the series. Today we will begin the second section of the series, which focuses on estimating usual intake distributions, and have the pleasure of hearing from Dr. Kevin Dodd.

Just a few notes on logistics before we get started with Kevin's presentation: Today's 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 question-and-answer period following the presentation—if you would like to submit a question, please do so using the Chat feature at the left of your screen.

We continue to be very gratified by the response that the webinar series has generated. Participants have been quite enthusiastic about obtaining slide sets prior to each webinar and also about accessing recordings shortly after. We are not able to post these materials on our public web site until the recordings have been transcribed and the documents are formatted to meet accessibility and other standards so we have set up a temporary page to house these materials for participants. The URL for this page was sent out via the listserv and also appears in the note box at the top left. We will aim to post slide sets one day prior to each webinar and recordings 1 or 2 days after each webinar. This web page also includes links to other webinar resources, including the glossary of key terms and notation. With today's webinar, we will begin to get into the specifics of the statistical methods. Given that we've organized the series for a broad audience, Kevin is aiming to ease you into statistical models and equations but we do suggest having the glossary handy as a reference during the webinars. You can find the guide to notation on page 15.

We realize that some participants experience a bit of a lag with slides advancing whereas others do not experience any lag at all. The presenters will aim to go at a pace that allows for a brief delay. You can modify the settings based on your connection using Manage My Settings at the top left.

One final note: I have received some inquiries about continuing education credits for participation in the webinars. We are not able to offer a process for CE credits but if you are a health professional, I'd encourage you to get in touch with your regulatory body to find out whether you can obtain credits by demonstrating your participation in the series in some way.

Now, let's move on to today's presentation. As I mentioned, our presenter for today is Dr. Kevin Dodd. Kevin is a mathematical statistician in the Biometry Research Group,

Division of Cancer Prevention, at the U.S. National Cancer Institute. For the past 20 years, Kevin has been leading the way in the development of statistical methods for analyzing short-term dietary intake data from population surveys. He was instrumental in the development of the Iowa State University method for estimating usual nutrient intake distributions and, more recently, of the National Cancer Institute, or NCI, method for modeling usual intake of episodically consumed foods. Recently, he has been focused on the estimation of total nutrient intakes from diet and supplements, as well as analysis of self-report and objective measures of physical activity. Today, Kevin will discuss estimation of usual intake distributions for dietary components that are consumed daily by most persons. Kevin.

[This page intentionally blank.]

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

Let me start off by reiterating that this series is organized by collaborators from a diverse collection of institutions, 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 the last several years.

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 very committed to moving this area of research forward.

Objectives

- Participants will have an understanding of:
  - Considerations in estimating usual intakes of nutrients and foods consumed nearly daily by nearly all persons
  - Assumptions made in current approaches to estimating usual intake distributions
  - Statistical modeling techniques and data requirements for estimating usual intake distributions

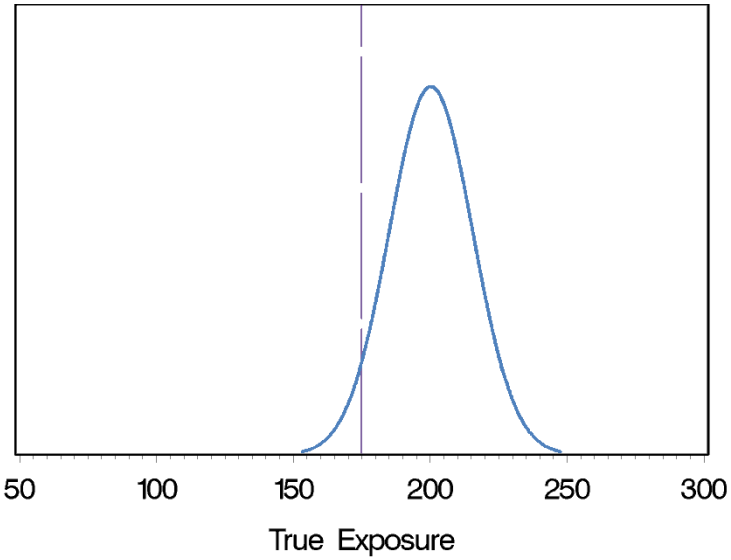
Slide 4

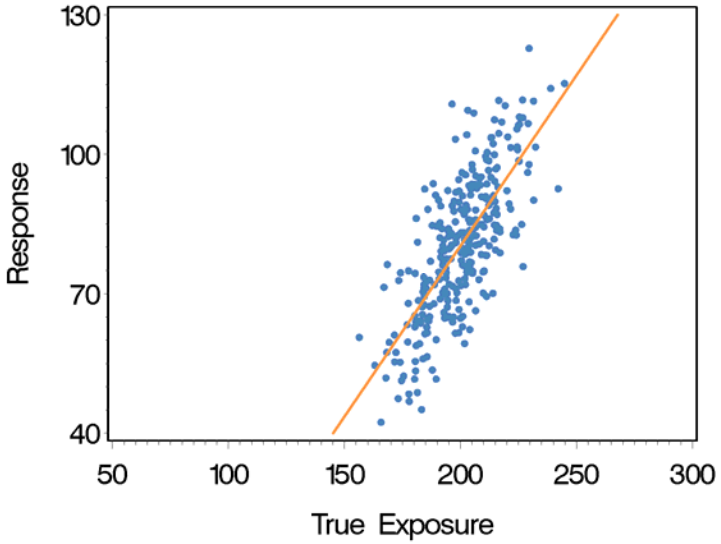
The objective for this presentation is to tell you about considerations in estimating usual intakes of dietary components that are consumed nearly every day by nearly all persons. This is in contrast to next week's webinar, which will focus on estimating usual intake distributions for episodically consumed dietary components.

There have been several approaches developed over the years to attack this problem. I'll be talking about the basic assumptions that are common to all of them, and I'll be talking about the statistical modeling techniques and data requirements in a general manner. In next week's webinar, Dr. Janet Tooze will take a different approach, focusing more specifically on the application of the National Cancer Institute method.

Two main areas of interest

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





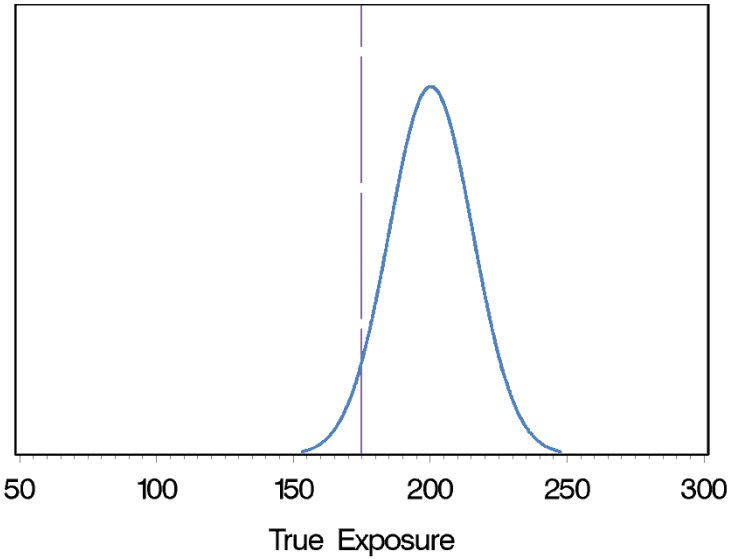
- Estimating diet-health relationships: regression coefficients

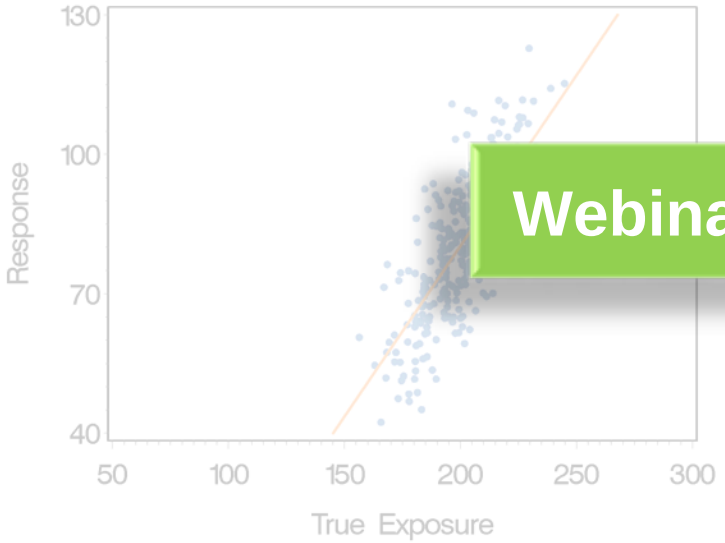
Slide 5

As Sharon mentioned in the first webinar, the full webinar series will cover two main areas of interest. The first is describing usual intake distributions, where interest centers on estimating means, percentiles, and proportions above or below a threshold. The second is estimating relationships between diet and health outcomes, where interest centers on regression coefficients.

Two main areas of interest

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





Webinars 6-8, 12

ting diet-health
nships:
regression coefficients

Slide 6

Today, I'm going to be focused on covering the first area of interest. Webinars 6 through 8 and webinar 12 will explore the second area of interest in depth.

Daily versus episodic consumption

- **Consumed nearly daily by nearly all persons**
  - E.g., vitamin C, total grains, total vegetables, solid fats, added sugars
- **Consumed episodically by most persons**
  - E.g., vitamin A, whole grains, dark green vegetables, fish

Slide 7

Also reviewing from the first webinar in this series, we classify dietary components into two broad categories based upon how often they are consumed.

Daily versus episodic consumption

- **Consumed nearly daily by nearly all persons**
  - E.g., vitamin C, total grains, total vegetables, solid fats, added sugars
- **Consumed episodically by most persons**
  - E.g., vitamin [illegible] dark green vegetables, fish

Webinar 3

Slide 8

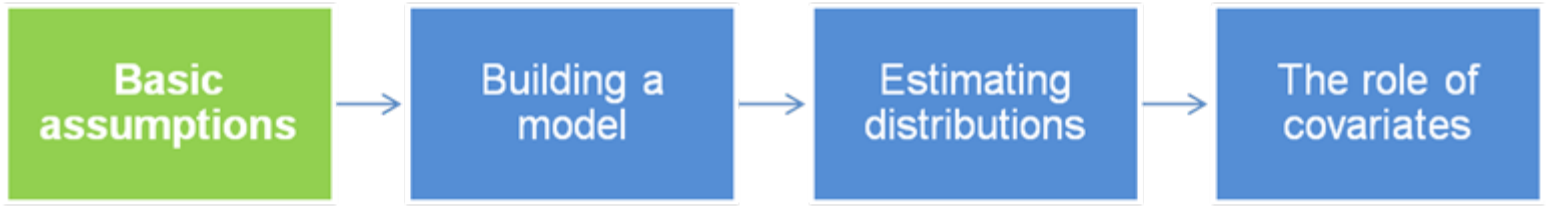
As Sharon mentioned, I'll be focusing on the category of dietary components that are consumed nearly daily by nearly all persons in the population. Most nutrients fall into this category but so do certain foods and food groups such as total grains, total vegetables, solid fats, and added sugars. As noted, Dr. Tooze will address episodically consumed dietary components in the next webinar.

Outline

- Basic assumptions
- Building a statistical model
- Estimating distributions from the model
- The role of covariates

Slide 9

Here is what we are going to cover today. First, I'll spend some time talking about the basic assumptions that underlie the development of statistical models for estimating usual intake distributions. Then I'll talk about specifics on how the statistical models build upon these assumptions, and how we can estimate distributions from these statistical models. I'll finish up by talking about how the basic models that have been developed can be augmented by allowing for the inclusion of covariates and the flexibility that this allows.



BASIC ASSUMPTIONS

Slide 10

Starting with basic assumptions….

Focus is on usual intake

Usual intake = long-term average daily intake

- Reflects idea that nutritional goals should be met over time, but not necessarily every day
- Provides a measure of total (chronic) exposure
  - Not addressing issues of acute exposure here

Slide 11

Our focus throughout the series is on usual intake. What do I mean by usual intake? Usual intake is defined as long-term average daily intake of a dietary component. This concept is relevant because it reflects the idea that nutritional goals should be met over time, but not necessarily every day. This concept also provides a measure of total or chronic exposure. Clearly, there may be instances where acute exposure is of interest—we are not going to cover these cases in the webinar series.

Challenge

Usual intakes are not directly observable

- Self-report dietary assessment instruments measure usual intake with error
- If ignored, this error can bias results
- Statistical modeling methods can be used to correct this bias

Slide 12

Usual intake is generally not directly observable in a free-living population. Instead, we generally ask people about their diet and use that information to make inferences about usual intake. However, self-report dietary assessment instruments are not perfect; they measure usual intake with error. If ignored, this error can bias results (whether it be the estimation of a distribution, or estimation of the relationship between diet and some health outcome). As a statistician, I'm glad to report that statistical modeling methods can be used to correct this bias.

Assessment strategies fall between two extremes

Usual intake = long-term average daily intake

- Focus on **long-term** aspect
  - Food Frequency Questionnaire (FFQ)
- Focus on **daily** aspect
  - 24-hour recall (24HR)

Slide 13

Self-report assessment strategies generally fall between two extremes: either the “long-term” aspect of the usual intake definition like a food frequency questionnaire, or the daily aspect of the definition, like a 24hour recall. However, both types of instruments attempt to get some sort of average—either as an explicit part of the questioning as in the FFQ or by averaging repeat 24HRs. Food records or diaries collected over multiple days fall somewhere in the middle.

Two classes of measurement error in instruments

- **Random:** Average of repeats = true value
  - Instrument is **accurate**, or **unbiased**
  - May not be **precise**

- **Systematic:** Average of repeats ≠ true value
  - Instrument is **inaccurate**, or **biased**
  - Systematic bias can occur in many ways

Slide 14

I just said that self-report instruments measure usual intake with error. Before I go on, I want to review what was discussed in webinar 1 about measurement error in assessment instruments. There are two types of measurement error in instruments. The first type is random measurement error. For an instrument affected by only random measurement error, the average of many repeats approximates the true value. We say such an instrument is accurate or unbiased, but it may not be precise; that is, it may require the average of many repeats to get close to the true value.

On the other hand, systematic measurement error cannot be addressed simply by averaging across many repeats. An instrument affected by systematic measurement error is inaccurate or unbiased, and as we saw in the previous webinar, there are several ways that systematic error can occur.

We also saw that these types of measurement error don't occur in isolation and that a self-report instrument may be affected by both random and systematic errors.

Potential sources of error in instruments: FFQ

− Cognitively challenging

− Limited food list/portion size choices

\+ No need for repeated application (high reproducibility)

Slide 15

For example, potential sources of error in a food frequency questionnaire, or FFQ, are shown here. For one thing, recalling intake over a long period of time is cognitively challenging because of issues like seasonality and avoiding the telescoping effect of more accurate recall of more recent intake. Furthermore, the way an FFQ is laid out with a finite food list and few selections for portion sizes can lead to reporting errors. In particular, the food list may need tailoring for particular populations; for example, the same food list would likely not be appropriate in both South Korea and Scotland. On the other hand, an advantage of the FFQ is that there is relatively high reproducibility, so that repeated applications don't provide much additional information.

Potential sources of error in instruments: 24HR

+ Less cognitively challenging

+ Open-ended format

– Repeats required to deal with day-to-day variation in intake

Slide 16

24-hour recalls, on the other hand, are less cognitively challenging because you only are asked to recall information about your previous day, not the last 30 days or year. The open-ended format of a recall avoids the finite food list problem. However, because data from a recall provide only a snapshot in time, we know we need repeats to deal with day-to-day variation in intake to get a sense of long-term intake. In one sense, the FFQ asks the individual to do their own averaging—for example, over 30 days—whereas for the recall, this averaging is done by the researcher using statistical methods. Typically, the number of repeats per person is strictly limited by budgetary and respondent burden constraints, so simply averaging those repeats may not get you acceptably close to true usual intake, so the modeling becomes necessary.

Comparison of measurement error structures

24-hour recall (24HR)

- Larger within-person random error
- Smaller systematic error

Food frequency questionnaire (FFQ)

- Smaller within-person random error
- Larger systematic error

Slide 17

As seen in the previous webinar, there is a body of evidence coming from biomarker-based validation studies that suggest that FFQs and recalls have different measurement error structures. Dr Larry Freedman will discuss these validation studies further in webinar 6, so I will just give the highlights here. In light of what I just said about the reproducibility of FFQs and the day-to-day variation expected in true diet, it is not surprising that the validation studies suggest that 24HRs have more within-person random error. Nor is it surprising that these studies suggest that systematic error is larger in FFQs compared with 24HRs.

The methods I will be discussing today all are based on scenarios where the available data come from replicated short-term instruments, such as a 24HR. Historically, large-scale nutrition surveys in the U.S. and other countries have used such instruments instead of FFQs. These decisions were made, and several of the methods I'll be discussing were developed, long before the current crop of biomarker validation studies came into the picture, so let's examine whether the choice was a good one in light of the biomarker study evidence. I'm not saying that validation studies are not without limitations or that 24HRs are error free, but based on the evidence we have on hand, what implications does the error structure of the 24HR have for estimating usual intake distributions?

Rationale for using short-term instruments

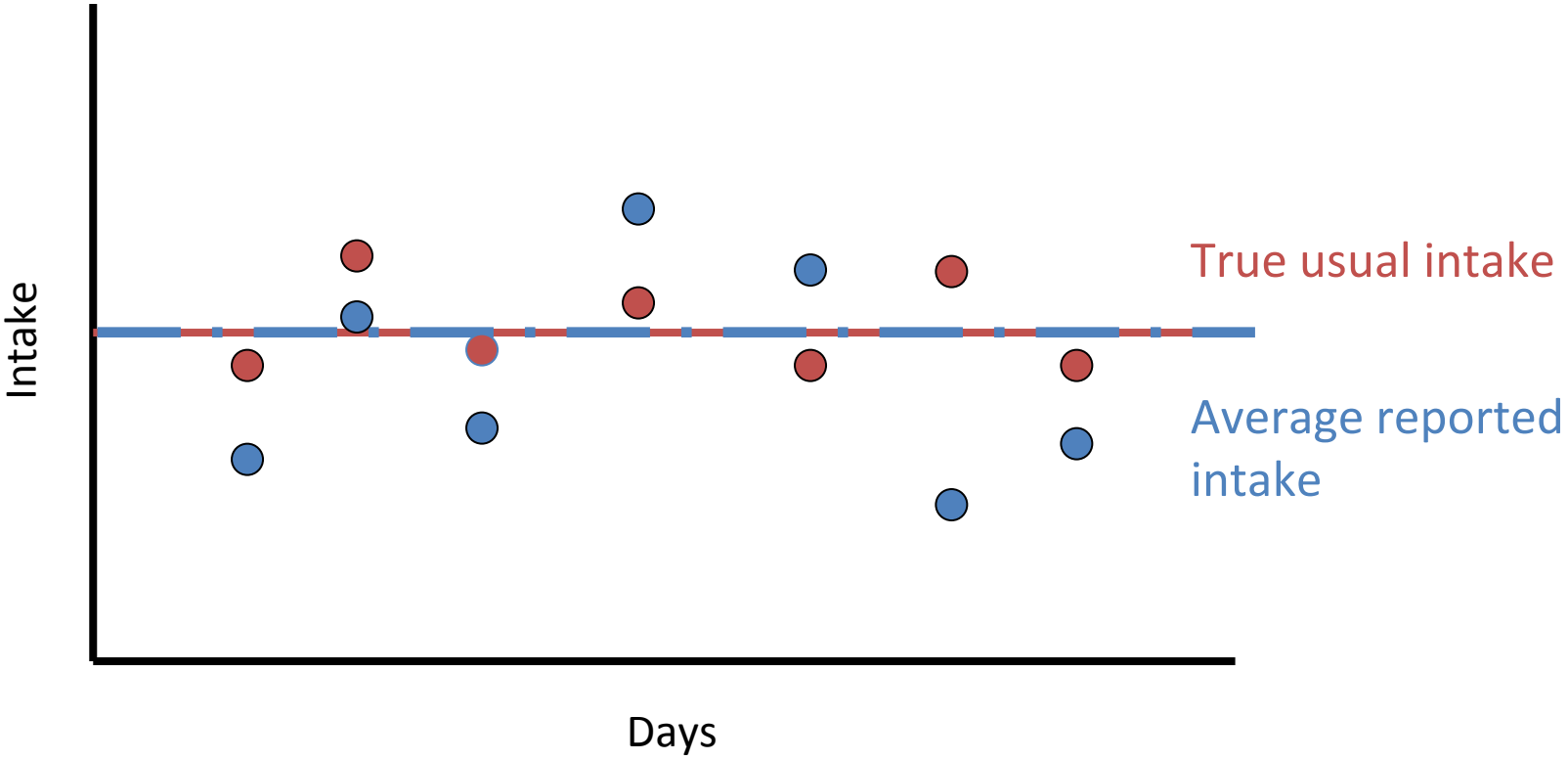
- Effects of random error can be mitigated by averaging repeats
  - Modeling can perform the same task
- Effects of systematic error cannot be mitigated unless we have an additional instrument
- Therefore, usual intake distributions based on 24HRs should be closer to those of truth than those based on FFQ

Slide 18

As we've described, the effects of random error can be mitigated by averaging across repeats. If the number of repeats is limited, statistical modeling can be used to remove the effect of random error analytically. On the other hand, the effects of systematic error cannot be dealt with in the same manner. In fact, as other webinars in the series will show, in order to deal with systematic error, we need to have an additional instrument to act as a reference. On the basis of these propositions, usual intake distributions based on 24-hour recall data should be closer to the distribution of true usual intake compared with those estimated from FFQ data. Note that I say closer rather than identical to because we acknowledge that recalls do have some systematic error.

Main assumption

24HR unbiased for individual-level usual intake



Slide 19

The methods that we will be discussing throughout the series are based on this core assumption that 24 hour recalls have only random error and are thus unbiased for individual usual intake. This assumption is illustrated by this graph, where the red circles denote true intake for a given individual across days and the blue circles represent reported intake for that individual on those same days. Our assumption is that averaging the blue circles across days gives us the same estimate as if we could average the red circles across days.

Working assumption

- Unbiasedness of 24HR is a **working assumption**

- Required to proceed with development of methods

- May be more or less justified depending on dietary component of interest

Slide 20

I reiterate that this is a working assumption. For the biomarkers used in the validation studies that were discussed in webinar 1, there is systematic error in the recall data and therefore the unbiasedness assumption does not completely hold in practice. However, statistical methods for dealing with random error are well established, whereas we have no way of getting at the systematic error except in maybe two or three cases. Until we are able to identify reference instruments for more dietary components, we are forced to make this assumption to proceed with the development of methods. This assumption may be more or less justified depending on the dietary component of interest.

BUILDING A MODEL

Slide 21

Now we're ready to start talking about the methods developed under this assumption.

Typical data scenario

A small number of replicated 24HRs collected on each of many individuals

Notation

- Observations denoted by R_{ij} , usual intake by T_i
- Individuals indexed by subscript $i = 1,2,\ldots,N$
- Repeats indexed by subscript $j = 1,2,\ldots,J$

Slide 22

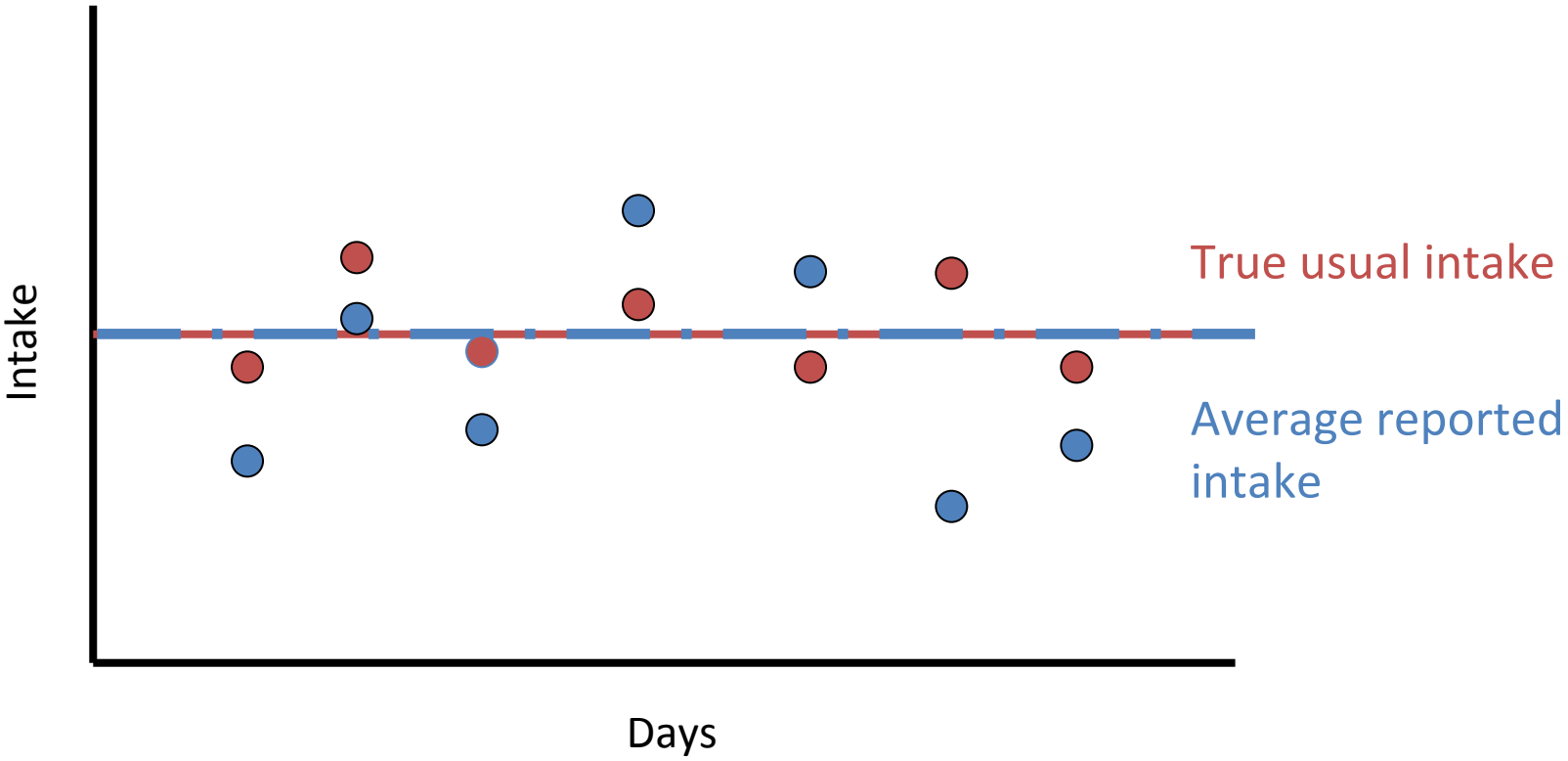
The typical data scenario that led to these models is a small number of replicated recalls collected from a large sample of individuals; usually, the recalls are separated by several days. As mentioned before, I'm going to be using equations, but I will explain them in words. For those interested in the math or thinking about collaborating with a statistician to implement some of these methods at a later date, the equations may be instructive.

The notation to be used throughout the talk and throughout the series is here. We denote observations by Rij (think R for reported intake) and usual intake by Ti (think T for truth). The subscript i refers to the individual and runs from 1 up to capital N for the total sample size, and the subscript j indexes the repeats; for example, 1 is for the first recall, 2 is for the second recall and there are maximum J recalls per person. Usually J is 2; for example, in the U.S. National Health and Nutrition Examination Survey where two recalls are attempted for each respondent.

Implications of unbiasedness assumption

24HR unbiased for individual-level usual intake

$$T_i = \mathrm{E}[R_{ij} \mid i]$$



Slide 23

So you've seen this figure already; let's look at it again using our statistical notation. When we say that 24HRs are unbiased for usual intake, we mean that while any one recall does not directly capture usual intake, the misestimation errors cancel out on average. Statistically, we write this assumption as shown here. Here, we define true intake for individual i by this formula, where E stands for expectation, and should be thought of as "average"; the vertical bar stands for "conditional on" or "given," so that this equation reads "Ti equals the expectation of Rij given I." Throughout the talk I'll be using mean, average, and expectation interchangeably to refer to the same general idea.

I want to reiterate that unbiased does not mean "exact"; the 24Hrs in this diagram exhibit considerable variation around the usual intake line. That is, there is substantial within-person variation in replicated 24HRs, partially due to day-to-day variation in true intake and partially due to other sources of random error. Again, this picture is for a single individual.

Implications of unbiasedness assumption

- The mean usual intake for the population is another kind of average:

$$\mu = \mathrm{E}[T_i] = \mathrm{E}[\mathrm{E}[R_{ij} \mid i]]$$

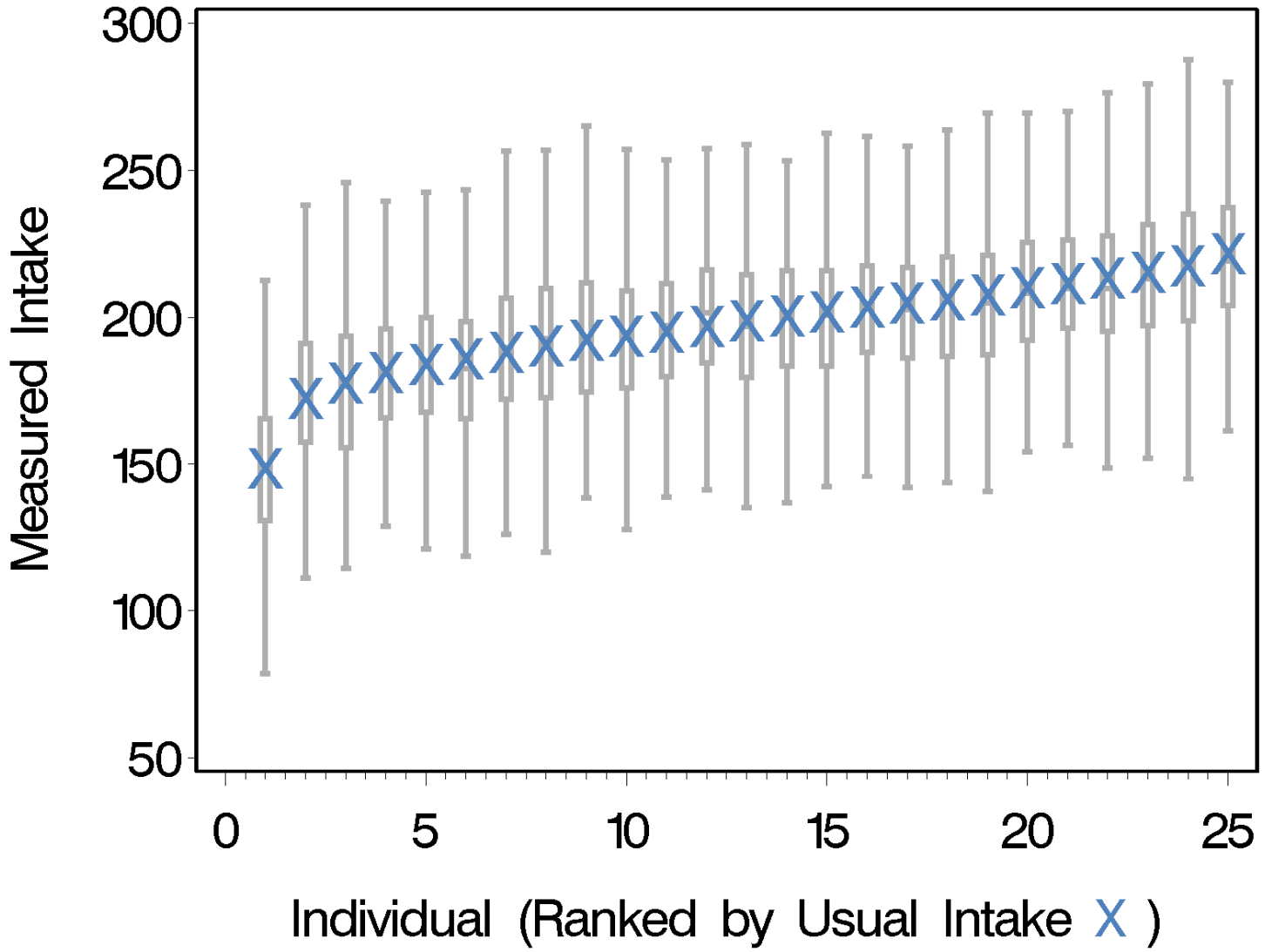
- The population mean usual intake can be estimated as the average of within-person average 24HRs

Slide 24

Now I'm moving to the implications at the group level. The mean usual intake for a group or population is another kind of average; in other words, another kind of expectation. The takeaway message of this slide is written in the second bullet but I'll take a few moments to explain the formula. The mean usual intake for the group, denoted by the Greek letter mu, equals the average of many individual usual intakes, Ti. From the previous slide, we can substitute the definition of Ti into this equation. What we end up with is mu as a double expectation—the expectation of the average of reported intakes per person. The outer expectation is across people and the inner expectation is across days within persons.

It turns out that under our unbiasedness assumption, you can estimate the mean usual intake for a population just by averaging 24HRs.

Within- and between-person variation



Slide 25

But, often, we don't just care about the mean. We also care about other properties of the distribution of usual intake; for example, how usual intakes differ across people. Let's think about looking at many individuals. This is a graph of simulated data, with the x axis representing 25 people ranked by their true usual intakes, represented on the graph by blue X's. The grey boxplots show the spread of the individual measurements across days, reflecting the within-person variation we have been talking about. So, there is variation in true intake across people as well as variation in measured intake within persons. If you sample only a few days per individual, the overall variation of your observed 24HRs comes from a combination of within-person and between-person variation.

Limitations of unbiasedness assumption

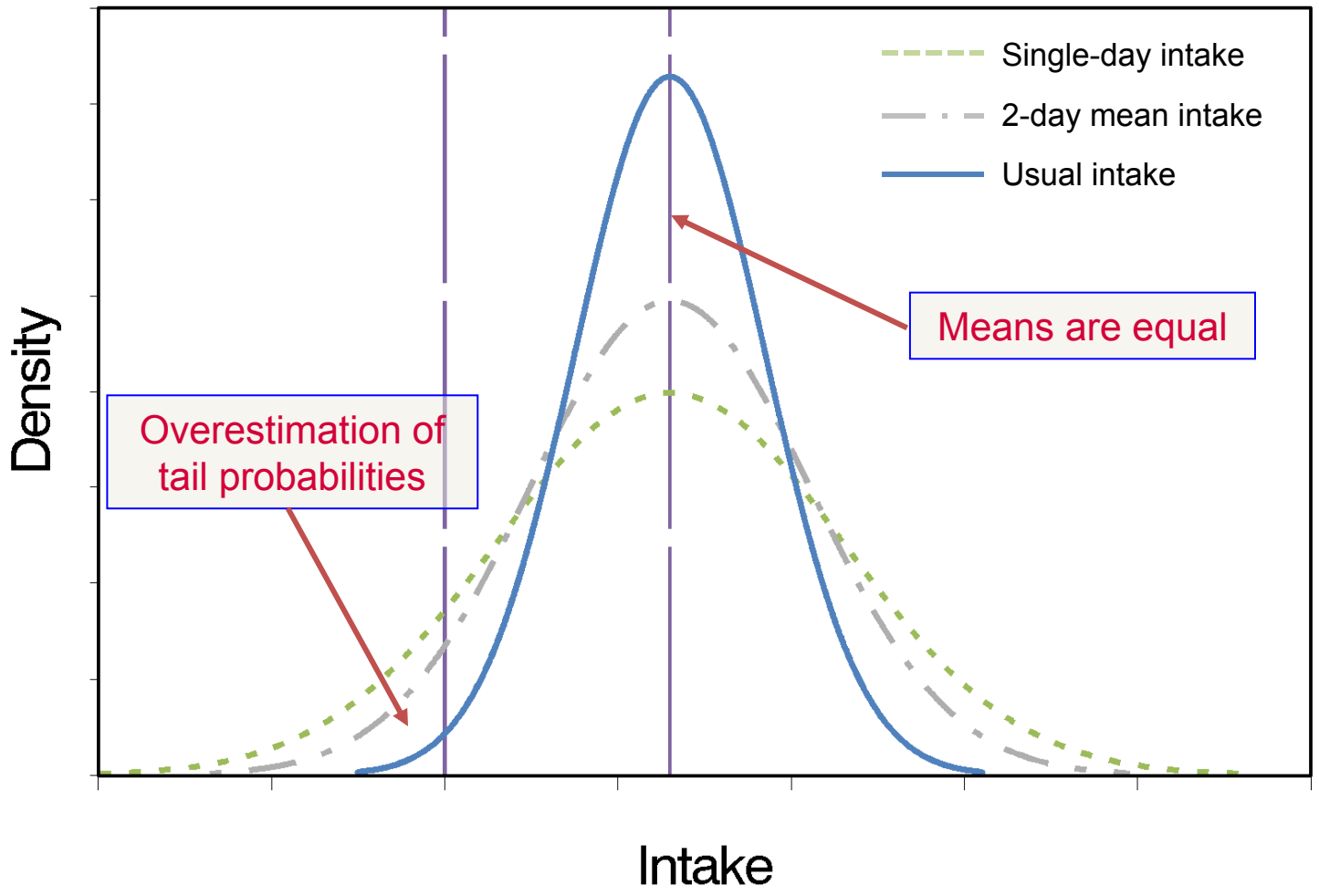
What about characteristics of the usual intake distribution other than the mean?

- With only a few repeats, averaging only removes some of the within-person variation
- Distributions of averages are too wide relative to usual intake distributions

Slide 26

We have already assumed that each 24HR for a person measures his or her usual intake with error, and have alluded to the fact that taking an average of many such 24HRs dilutes this error. However, when only a few repeats are available, averaging only removes a portion of the within-person variation, so that the distribution of averages is too wide relative to the usual intake distribution we want to estimate, even if the means coincide. The more 24HRs we have to compute each per-person average, the closer to the usual intake distribution we will get.

Effect of within-person variation



Slide 27

This graph illustrates what I've just been saying, for three different distributions. The distribution of single-day observations is graphed with the green dashed line; the distribution of 2-day averages is drawn with a gray broken line; and the distribution of usual intake, which has been corrected for excess within-person variation, is drawn with a blue solid line. The means of the three distributions are the same; however, the 1-day and 2-day distributions are too wide. This results in overestimation of the tail probabilities, which is a problem if we are interested in prevalences above or below a threshold, such as the vertical line on the left.

Effect of within-person variation

- Population mean usual intake may be well estimated by simple averaging methods

- Percent of population with usual intake below/above cutoff values may be very biased – **modeling necessary**

Slide 28

The implication is that although mean usual intake for a population may be estimated using a single recall per person or the averaging of a few recalls, this is not true for other properties of the distribution. Estimation of the percent of a population of interest with usual intake above or below some cutoff value may be very biased if we use only the average of a few recalls per person. However, this problem was recognized back in the 1980s, leading to the development of statistical modeling methods to deal with it.

What does “modeling” entail?

- A way of filling in gaps in information using statistical techniques
- In this case, pooling limited information from sampled individuals
- Requires assumptions

Slide 29

So what is statistical modeling and what does it involve? It is a way of filling in gaps in information using statistical techniques. In this case, we don't have enough repeat measures on each individual for averaging to be sufficient, so we use statistical models to pool the data we do have from the entire sample. Now, applying modeling requires making some assumptions in addition to the unbiasedness assumption we've already discussed. I'll talk about these additional assumptions in a few minutes.

Foundation of the model

- Each recall is usual intake plus a deviation

$$R_{ij} = T_i + (R_{ij} - T_i) = T_i + \varepsilon_{ij}$$

Slide 30

It follows from our unbiasedness assumption that we can consider each recall Rij as the sum of an individual's usual intake Ti plus a deviation represented here by epsilon ij.

Foundation of the model

- Each recall is usual intake plus a deviation

$$R_{ij} = T_i + (R_{ij} - T_i) = T_i + \varepsilon_{ij}$$

Within-person deviation

Slide 31

These deviations epsilon ij are within-person deviations and…

Foundation of the model

- Each recall is usual intake plus a deviation

$$R_{ij} = T_i + (R_{ij} - T_i) = T_i + \varepsilon_{ij}$$

Within-person deviation

- Within-person deviations cancel out across days

Slide 32

…our unbiasedness assumption implies that they cancel out over time; in other words, average to zero across days.

Foundation of the model

- Each usual intake is the population mean intake plus a deviation

$$T_i = \mu + (T_i - \mu) = \mu + u_i$$

Slide 33

Similarly, each person's usual intake can be expressed as the population's mean usual intake plus a second kind of deviation, represented here by ui.

Foundation of the model

- Each usual intake is the population mean intake plus a deviation

$$T_i = \mu + (T_i - \mu) = \mu + u_i$$

Between-person deviation

Slide 34

ui’s are between-person deviations…

Foundation of the model

- Each usual intake is the population mean intake plus a deviation

$$T_i = \mu + (T_i - \mu) = \mu + u_i$$

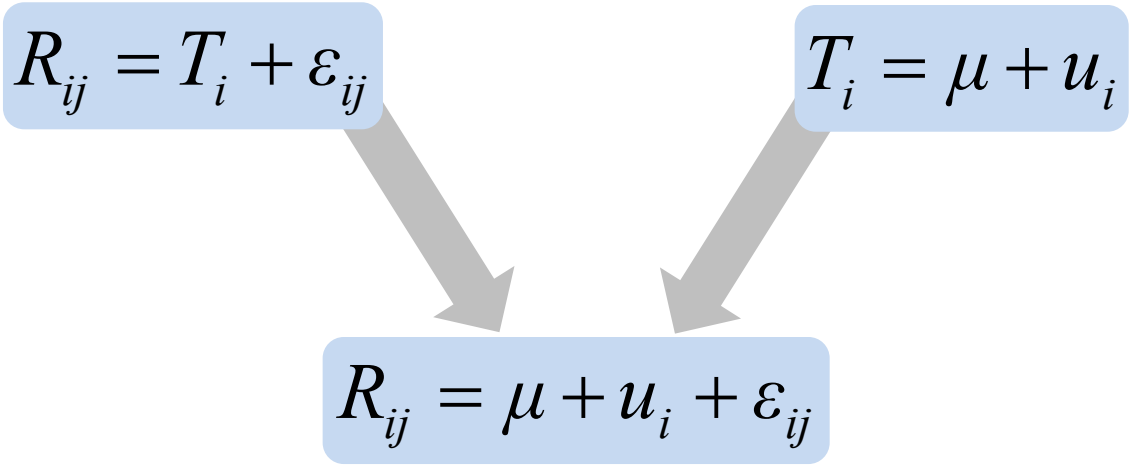
Between-person deviation

- Between-person deviations cancel out across the population

Slide 35

…that cancel out across individuals in the population.

Foundation of the model



- Population mean μ is a **fixed parameter**
- Both types of deviations are **random variables** with
  - Zero expectation
  - Unknown variances, distributions

Slide 36

Now, let's put these two concepts together into a statistical model. Here, we obtain an expression for each 24HR as the sum of a population mean and two deviations.

The population mean is a fixed parameter but both types of deviations are considered random quantities based on an assumption of a probabilistic mechanism for selecting people from the population and recall days from the set of all possible days. From our main assumption of unbiasedness, we know that both types of deviation have mean zero because they cancel out over days and people, but it doesn't explicitly tell us what the variances of the deviations are, nor does it tell us their distributions. Now here's where we start bringing additional assumptions into the picture.

Common variance assumption

- Sample variance among the 24HRs for a person estimates his within-person variance
  - Very few “degrees of freedom”, not very precise
- Assume same magnitude of within-person variation across individuals
  - Pool individual estimates to get more precise estimate

Slide 37

Now, it follows from the unbiasedness assumption that the sample variance among 24-hour recalls for a particular person estimates his or her particular within-person variance. However, in the typical data scenario with few recalls per person, there are very few degrees of freedom for each of these within-person variance estimates and, as a result, they are not very precise. The first assumption we make is a common variance assumption, which posits that there is the same magnitude of within-person variance across persons. This allows us to obtain a pooled variance estimate that is more precise. Some methods stop with this common variance assumption, but many others make further assumptions that we'll talk about next.

Distributional assumptions

- Statistically convenient to assume that both types of deviations follow a parametric probability distribution

- The normal distribution is a common choice
  - Naturally parameterized by mean and variance
  - Dependence between deviations can be completely modeled via correlation

Slide 38

A few moments ago, I mentioned that both the within- and between-person deviations are random. Related to this, a second assumption that we make, based on statistical convenience, is that both types of deviations follow a parametric probability distribution. A common choice is the normal distribution, which we say is naturally parameterized by its mean and variance. In other words, once you know the mean and variance, you know everything about the normal distribution. This is handy since we are so far just focusing on means and variances. Another nice feature of this normal assumption is that any dependence between deviations can be completely modeled using correlation. This can be among within-person deviations for the same person or among between-person deviations. For distributions other than the normal, correlation only models the linear relationship between random variables, not the entirety of the relationship.

Basic statistical model for 24HRs

Within-person deviations are:

- Normally distributed, with a common variance
  - Can be relaxed, if desired
- Independent from those of other people
- Independent from those of the same person
  - Can be relaxed, e.g., if 24HRs are consecutive

Slide 39

Drawing upon what we've discussed, the basic statistical model for 24 hour recalls posits that within-person deviations are normally distributed with a common variance. This can be relaxed if desired—for example, if you think that two different groups have two different within-person variances—but I won't be discussing this scenario in this webinar. We also assume that the within-person deviations for one person are independent of those of other people and, finally, that each one is independent of the others for the same person. Again, this last assumption can be relaxed if, for example, 24 hour recalls are taken on consecutive days.

Basic statistical model for 24HRs

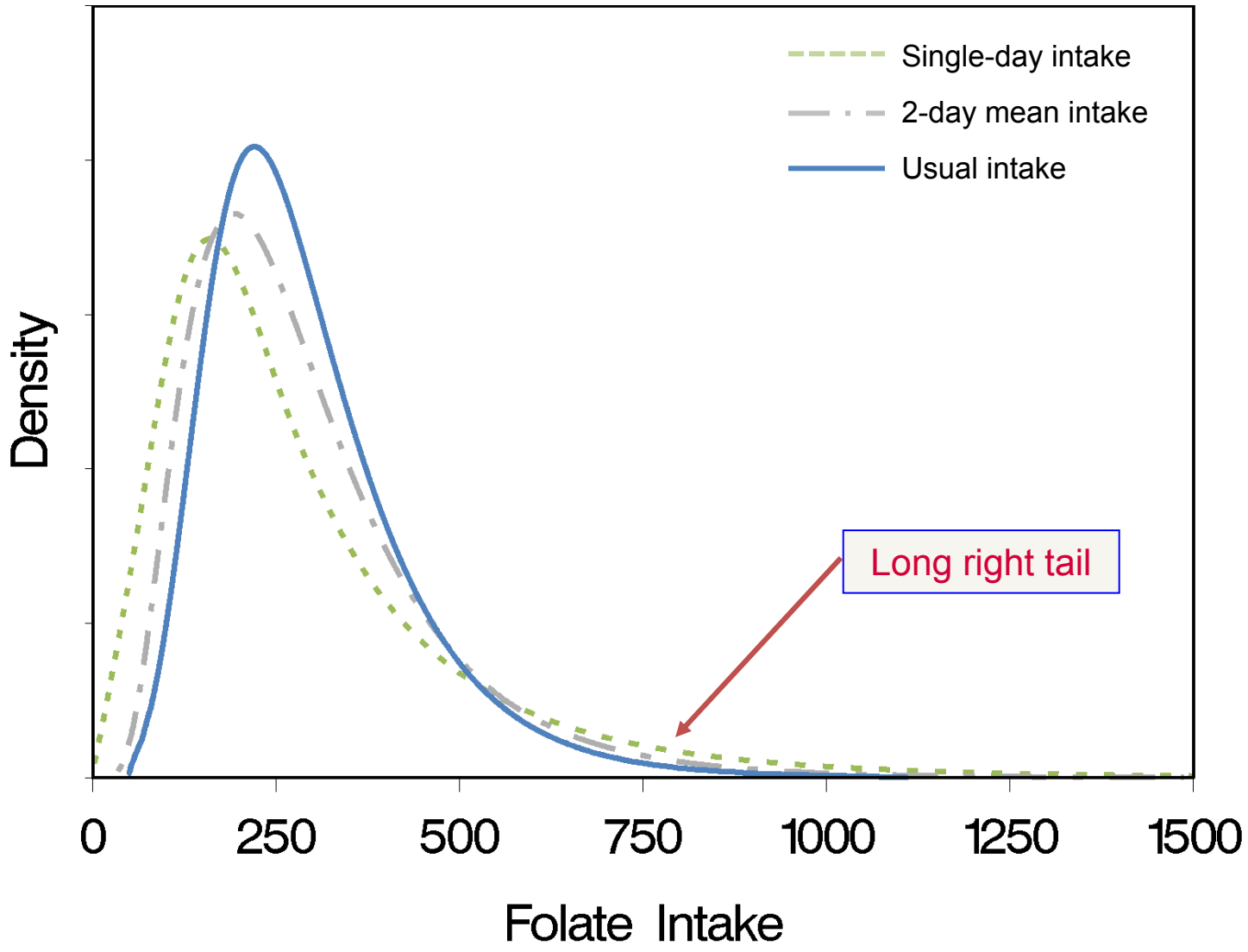
Between-person deviations are:

- Normally distributed, with a common variance
  - Can be relaxed, if desired
- Independent from those of other people

Slide 40

Moving on to the between-person deviations, again, these are assumed to be normally distributed with a common variance, and this assumption can be relaxed if desired. As well, the between-person deviation for one person is assumed to be independent of those of other persons. So far, we've assumed normal distributions for all of these deviations. Under these assumptions, the recalls ought to be normally distributed themselves; however, in practice this is not the case.

Complications of skewed distributions



Slide 41

Instead, distributions of intake tend to be skewed, sometimes severely so, with a long right tail indicating that some individuals in the population have very high intakes relative to the rest. This can be partially explained by the fact that there is no upper limit on intake but there is a lower limit of zero. On a given day, it is feasible for someone with a usual intake of 100 units of a dietary component to consume over 200 units of that component. On the other hand, on a given day the person can only consume 100 units less than their usual intake because they can't go below zero. Similarly, at the population level, you can have many individuals with intakes well above the population average but there is a limit to how far below the population average people's usual intake can go.

Now, it turns out that statistical modeling with skewed distributions can be difficult, and that is one reason why most methods in statistics are developed for normally distributed data. One common way to deal with this aspect of 24HR data is to apply a transformation to the raw data to make the normality assumption for the deviations more tenable. The transformation is applied before the modeling step, and then the resulting distributions have to be converted back to the original scale.

Common nonlinear transformations

Name	Functional Form	Inverse Form
Log	$g(R\,;\,\gamma) = \ln(R)$	$g^{-1}(r\,;\,\gamma) = \exp(r)$
Power(γ)	$g(R\,;\,\gamma) = R^{\gamma}$	$g^{-1}(r\,;\,\gamma) = r^{1/\gamma}$
Box-Cox(γ)	$g(R\,;\,\gamma) = (R^{\gamma} - 1)/\gamma$	$g^{-1}(r\,;\,\gamma) = (\gamma r + 1)^{1/\gamma}$
Box-Cox(γ,δ)	$g(R\,;\,\gamma) = [(R + \delta)^{\gamma} - 1]/\gamma$	$g^{-1}(r\,;\,\gamma) = (\gamma r + 1)^{1/\gamma} - \delta$

- Large values affected more than small ones
- Other transformations possible
  - Should be one-to-one (invertible)

Slide 42

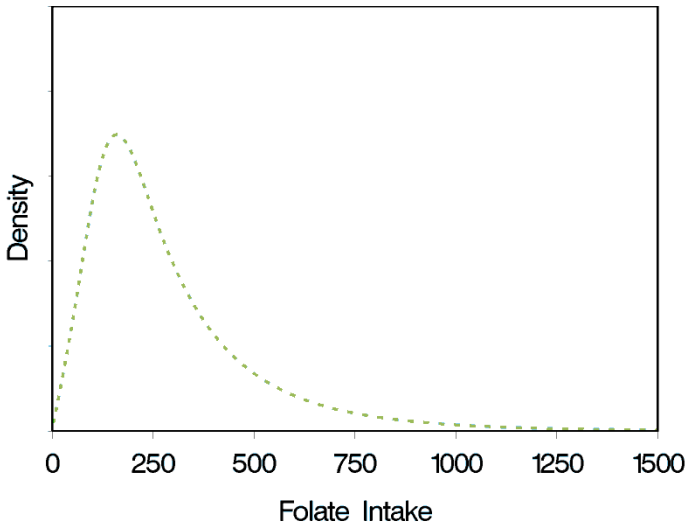
This table shows the functional form of several kinds of transformations commonly used to deal with right-skewed data. In general, we take our original data, R, and we apply one of these transformations. That is, we might take the log or the square root of each intake value. These formulas will be seen repeatedly throughout the webinar series, and this table tries to get across the concept of a family of transformations, such as Log, Power, or Box-Cox, where each member of the family is characterized by a parameter, gamma. Different values of gamma indicate different transformations within the given family. For example, the square root transformation is a power transformation with gamma equal to ½; the cube root has a gamma of ⅓; and the fourth root has gamma of ¼. The Box-Cox family of distributions, here shown in the one-parameter and two-parameter varieties, is a generalization of the power and log transformations. The nice thing about the Box-Cox is that it includes the log as a limiting case where gamma goes to zero.

One important takeaway here is that all of these transformations are nonlinear—they affect large values more than small ones and serve to pull in the right tail so the resulting distribution is symmetric.

Another important takeaway is that while other transformations are possible, any transformation used should be one-to-one, because we have to be able to convert back to the original scale, which involves applying an inverse transformation. In other words, if we use a square root transformation for a dietary component measured in milligrams, we want to eventually be able to talk about usual intakes in milligrams, not in the square root of milligrams. Note that the inverse transformations, the functional forms of which are shown in the third column, are also nonlinear.

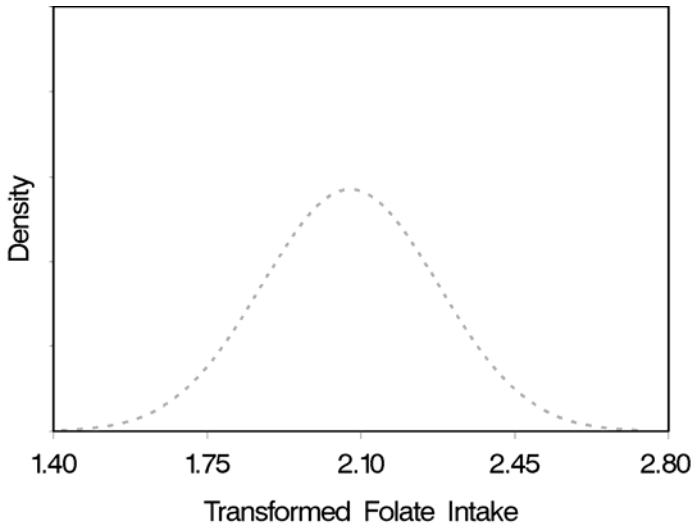
Extended model for 24HRs

Original Scale



Transform

Transformed Scale



$$T_i = E[R_{ij} \mid i]$$

$$g(R_{ij}) = \mu + u_i + \varepsilon_{ij}$$

$$u_i \sim N(0, \sigma_u^2), \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

$$\text{Corr}(u_i, \varepsilon_{ij}) = 0 \text{ for all } i \text{ and } j$$

Slide 43

Schematically, this leads to an extended model for 24 hour recalls, where we apply a transformation to our original data shown in the graph on the left and make our normality assumptions on deviations on the transformed scale shown in the graph on the right. This is written in statistician speak in the equations below. The first equation reiterates the assumption we made earlier that the average of 24hour recalls estimates true usual intake for an individual. In other words, you don't need to apply transformations for this to hold. This convention, that 24 hour recalls are unbiased on the original scale, has been used in almost all of the methods developed to date for estimation of usual intake distributions from short-term instruments. The next three lines in the blue block show that we are conducting our modeling and making our distributional assumptions on the transformed scale. We use the tilde notation and the italic N to signify the normally distributed assumption, and we put the mean and variance of the normal distribution in the parentheses. A consequence of the normality assumption is that zero correlation implies the independence of deviations as we discussed earlier.

A warning about means of transformed data

- Averaging transformed data is **not the same** as transforming averages of raw data if the transformation is nonlinear

$$\mathrm{E}[g(R)] \neq g(\mathrm{E}[R])$$

- Taylor series argument:

$$\mathrm{E}[g(R)] = g(\mathrm{E}[R]) + \tfrac{1}{2} g''(\mathrm{E}[R])\,\mathrm{Var}(R) + \text{extra terms}$$

- Extra terms involve “higher-order moments”

Slide 44

Now, before we go on, I want to give you one warning. Remember that these modeling methods are an alternative way to doing averaging when we don't have a lot of data per individual. We are doing statistically what we can't do with limited data. Remember that a usual intake is a mean and what we're trying to get is a distribution of these means.

When transformations are involved, you have to be careful working with these means. It turns out that taking the mean of transformed data is not the same as transforming the original mean when the transformation is nonlinear. In other words, if we transform our raw intake data and take an average, this is not the same as applying the same transformation to the average of the raw data. The same goes for inverse transformations which are nonlinear. Suppose you take all of your 24hour recalls and average them—this will give you the usual intake for the population. However, if you take the square root of the recalls, then average them, then take the inverse of the square root, you will not get the same value. Just as this inequality doesn't hold for an overall mean, it also applies to theoretical means per person, which are what make up the usual intake distribution.

The first equation here makes this point in statistician speak, by saying that the expectation of g(R) is not equal to g of the expectation of R. If you really want the expectation of g(R), the second equation uses a Taylor series argument to show that you need to include additional terms made up of derivatives of the transformation and higher-order moments. We don't necessarily need to appreciate all of the details here, but we will see this equation later when we discuss backtransformations to return our data to the original scale. I introduce the concept of moments here, and I'll talk more about them later. Moments are characteristics of the distribution involving expectations of squares, cubes, etc. The mean is the first moment; the variance, the second; skewness, the third; kurtosis, the fourth; and then we stop giving them names. One important statistical fact is that a distribution is uniquely characterized by the collection of its moments; if all of the moments for two distributions match, then the two distributions are identical in all respects—in particular, in terms of percentiles.

Summary

- Unbiased ≠ error-free
- Within-person variation → overdispersion
- Model built using additional assumptions
  - Common variance components
  - Distributional assumptions (optional)
- Skewed distributions of intake may be handled with transformations

Slide 45

To summarize this section of the webinar, we started with the assumption that 24 hour recalls are unbiased for usual intake but we recognize that this does not mean that 24 hour recalls are error free. Within-person random error in 24HR data leads to overdispersion, or wider distributions compared with true usual intake. We also saw that we can build a statistical model using additional assumptions involving variance components and distributional assumptions, and stated that the skewed distributions of intake may need to be transformed as part of the modeling process

ESTIMATING DISTRIBUTIONS

Slide 46

We’re now going to look at how we use the model that we have built to help us estimate distributions of usual intake.

Data requirements

- Two or more 24HRs on at least a subsample
- Replicate 24HRs should be far apart in time to maximize information
- Distribution of 24HRs should be "**normalizable**"
  - Unimodal, no spikes at extreme values

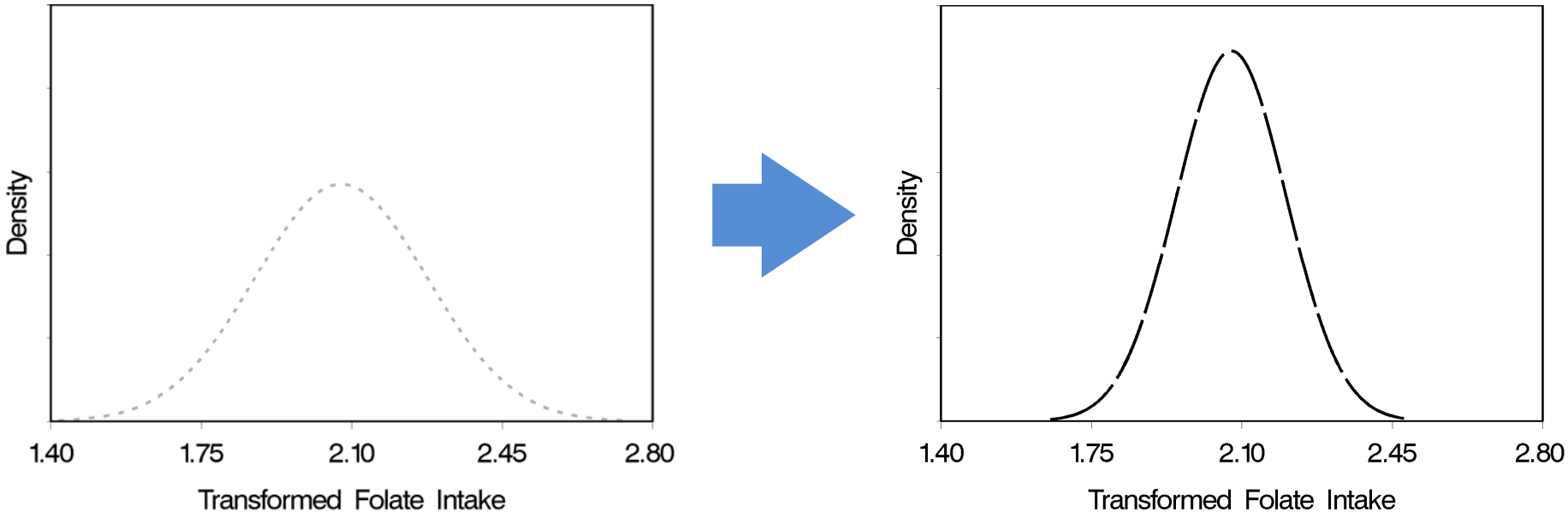
Slide 47

Starting out with data requirements, using these methods assumes that we have one recall for all individuals in a sample but two or more for at least a subsample. The replicate recalls should be far apart in time, usually interpreted as several days, to maximize information; that is, to avoid the so-called “leftover effect” where what you eat one day is correlated with what you had the days before or after. Furthermore, in light of what we’ve discussed, the distribution of 24 hour recalls should be normalizable, meaning that a transformation can be applied to approximate normality. In practice, this means that the distribution of recall data should ideally be unimodal (i.e., with only one hump like the normal distribution) and should not have any spikes or clumps of identical values near the tails of the distribution. In webinar 3, Dr. Tooze will discuss the scenario when, for example, a large fraction of the recalls have a value of zero, which is by definition at the extreme lower limit of possible intake.

General approach – no transformations (yet!)

- Separate within- from between-person variation

- Estimate usual intake distribution that exhibits only between-person variation



Slide 48

The general approach, which I'll talk about for the next few minutes in the simple case where normality transformations are not involved, is to use our model to separate within- from between-person variation and then to estimate a usual intake distribution that exhibits only the between-person variation. This is illustrated here by showing that we want the wide distribution on the left that reflects both within- and between-person variation to be adjusted for within-person variation, resulting in the narrower distribution on the right.

Two general approaches

- Model-Assisted (M-A) – rescales observed individual mean distribution

- Model-Based (M-B) – estimates distributions from theoretically-derived quantities

Slide 49

There are two general approaches to going about this shrinking of the distribution. The first method, that I will call model-assisted, rescales the individual mean distribution and I'll show you the details in a moment. This is called model-assisted because it attempts to let the data do the talking and makes as few assumptions as possible. Later, I'll explain the model-based approach, where distributions are estimated from theoretically derived quantities. These quantities depend very much upon the exact distribution hypothesized for the within- and between-person deviations. This will become clearer shortly.

Two general approaches

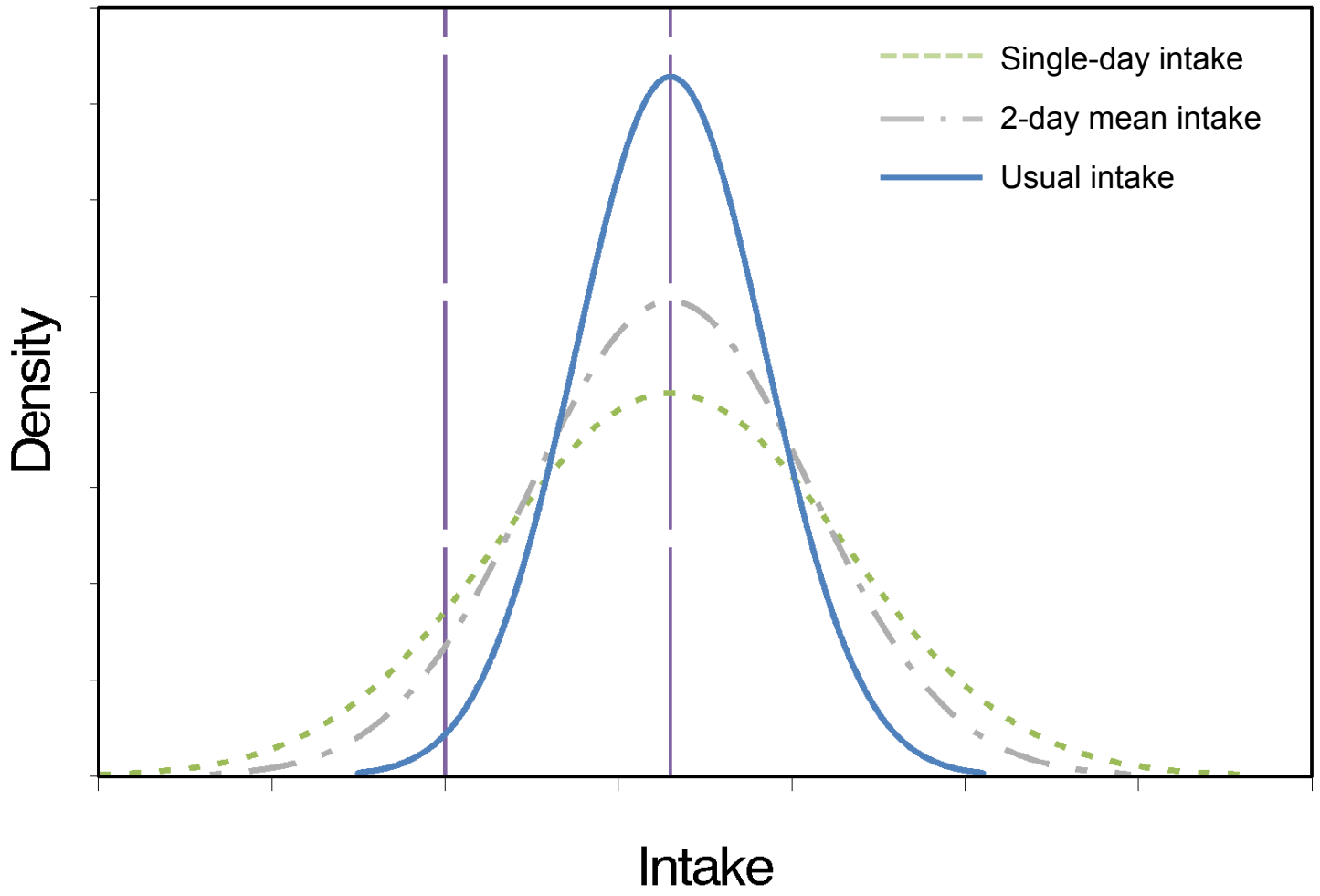
- Model-Assisted (M-A) – rescales observed individual mean distribution

- Model-Based (M-B) – estimates distributions from theoretically-derived quantities

Slide 50

First, model-assisted….

Rationale for the Model-Assisted approach



Slide 51

Recall from earlier in the webinar that the distribution of individual means is centered in the correct spot but has a variance that is too large. We will use this fact to guide our development of an estimated usual intake distribution based upon the observed distribution of individual means.

Rationale for the Model-Assisted approach

$$R_{ij} = \mu + u_i + \varepsilon_{ij}, \quad \mathrm{Var}(u_i) = \sigma_u^2, \quad \mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$$

- For a sample of single 24HRs:

$$\mathrm{E}[R_{i1}] = \mu$$
$$\mathrm{Var}(R_{i1}) = \sigma_u^2 + \sigma_\varepsilon^2$$

- For a sample of *J*-day means:

$$\mathrm{E}[\bar{R}_{i\bullet}] = \mu$$
$$\mathrm{Var}(\bar{R}_{i\bullet}) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{J}$$

Slide 52

Before we get to the equations highlighted in blue, let's review our statistical model for recalls shown at the top of the slide. We are assuming that the between-person deviations have variance sigma squared u and the within-person deviations have variance sigma squared epsilon. Right now, we don't have the tildes and the N to signify normality; we just assume that the deviations have these specified means and variances. I said before that the distribution of recalls has variance that is a combination of these two components. If we just took a sample of single 24 hour recalls, the mean across people is assumed to be mu, the population mean, and the variance of the recalls is just the sum of the variances, as shown here. For a sample of j day means where we take j recalls per person and average them within person to get an average recall, r bar i dot, we still have that the average across people is mu but now the variance of the j day mean distribution is shown here, where the contribution of the within-person variance is reduced by dividing by j in this formula. This shows how averaging a few days of recalls reduces the impact of within-person variation. But we want to completely remove the impact of within-person variation, so what do we do?

Implementing the Model-Assisted approach

$$R_{ij} = \mu + u_i + \varepsilon_{ij}, \quad \mathrm{Var}(u_i) = \sigma_u^2, \quad \mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$$

- Fit model to obtain parameter estimates
- Scale individual means to have desired variance

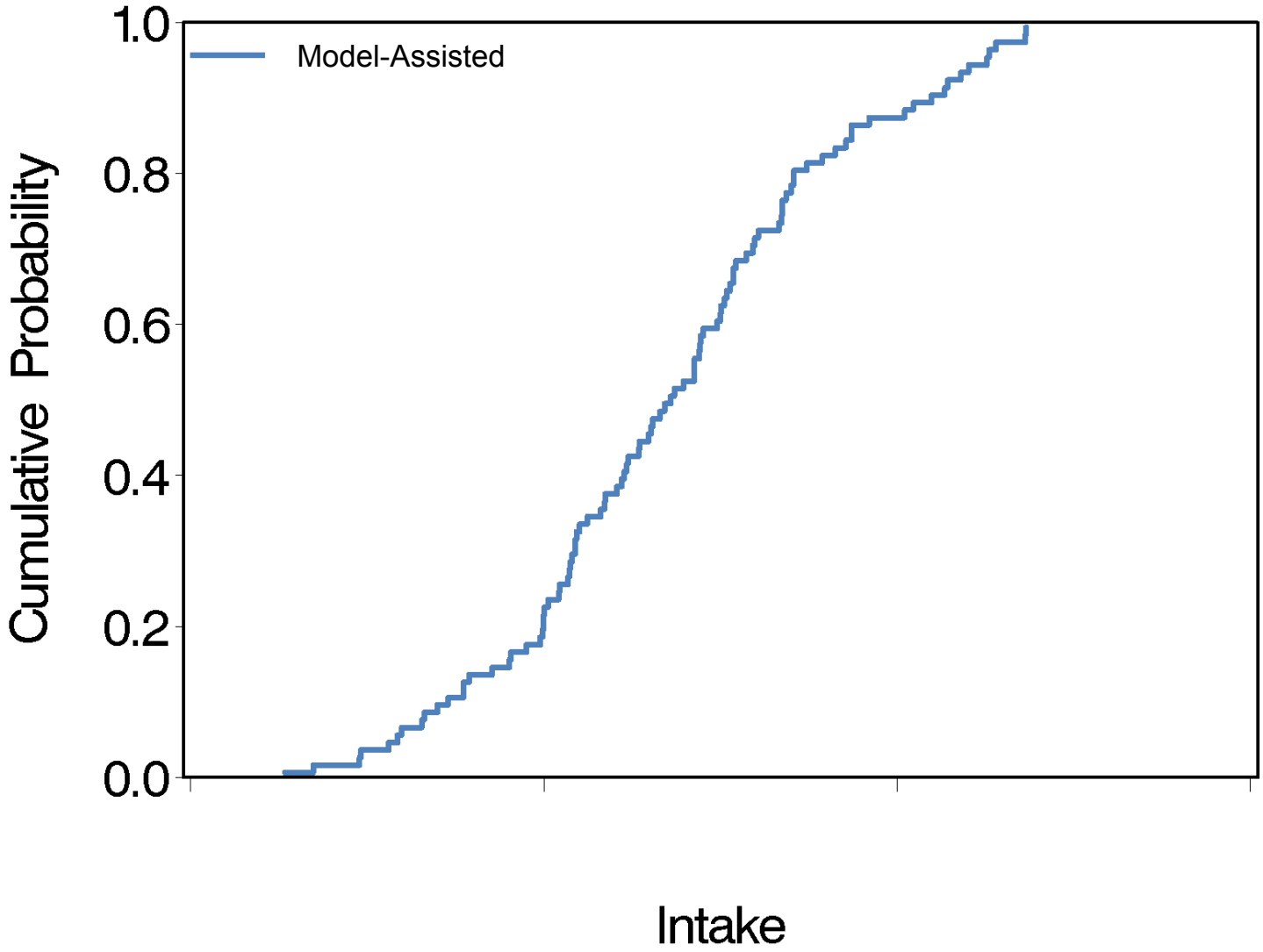
$$r_i = (\bar{R}_{i\bullet} - \hat{\mu}) \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_\varepsilon^2}{J}}} + \hat{\mu}$$

- Use empirical distribution of r_i as estimate of usual intake distribution

Slide 53

First, we fit our model to obtain estimates of the within- and between-individual variance components; we'll add little hats to indicate that these are estimates. Then we take our individual means, r bar i dot, center them to the estimated population mean, and then scale this deviation by this term under the square root sign, which is the ratio of the between-person variance to the variance of the within-person mean distribution we derived on the previous slide. Like most of the deviations we have discussed so far, the rescaled deviations have mean zero across people. The last part of this equation says we then add back the estimated population mean. What we end up with are rescaled means, one per person, denoted by little ri. These rescaled means have a smaller variance than they used to, so I use the small letter r to indicate this. Now, the model-assisted approach says to use the empirical distribution of little ri as an estimate of the usual intake distribution. What do we mean by that?

Features of the Model-Assisted approach



Slide 54

This graph shows an empirical distribution constructed from 100 rescaled means. The Y axis of the graph is in units of cumulative probability, while the X axis is in units corresponding to usual intake (although I have suppressed the actual numbers). Though it may not be obvious yet, we can use this graph to estimate quantiles of usual intake, and also estimate the proportion of the population with usual intake below any particular cutoff. Essentially, the empirical distribution function is a step function with steps of height 1 over N; in this example N is 100, located at each rescaled mean intake, where the rescaled means are ordered from smallest to largest. The percent of people with usual intake below a particular value is estimated by the relative frequency or fraction of shrunken means below that value. For example, to get the 20^{th} percentile of the distribution of usual intake, you can draw a line from the .2 value on the Y axis over to where it intersects the blue line, then draw a line down to the X axis. The point where it touches is the desired estimate at the 20^{th} percentile. This is because there are 20 rescaled means less than or equal to that value in this graph, where N is 100. Similarly, if you pick any point on the X axis, draw a vertical line up to the blue line, then draw a line over to the Y axis, the point where the horizontal line touches is an estimate of the proportion of the population with intake below the corresponding point on the X axis.

Interpretation of scaled means

- The scaled means r_i are not intended to be estimates of individual usual intake

- The distribution of scaled means has the same mean and variance as the distribution of usual intakes in the population
  - Distributions coincide for normal distributions
  - Agreement only approximate otherwise

Slide 55

Note that the rescaled means of the model-assisted approach are not intended to be estimates of individual usual intake. Instead, they are intended to be used to represent a distribution that has the same mean and variance as the distribution of usual intake in the population. When we said that the normal distribution is completely specified by its mean and variance, we meant that if two normal distributions have the same first two moments, then they are identical, and in fact all of their moments coincide. Here, if the distribution of usual intake in the population really is normal, then matching the true mean and variance with that of the rescaled mean distributions means that features of the usual intake distribution, such as percentiles, can be estimated with percentiles from the rescaled mean distribution, like I talked about on slide 44. If the distribution of usual intake in the population is not normal, the agreement is only approximate; you don't recover every feature of the usual intake distribution. The usefulness of the approximation really depends upon how close the usual intake distribution is to normal. That's why we try really hard to apply transformations that approximate normality so that matching the mean and variance is all we need to do.

Features of the Model-Assisted approach

- Data-driven, uses few assumptions
- Only requires separation of variance components
- Precision of empirical percentiles limited
  - There are only *N* jumps in estimated distribution function

Slide 56

To summarize the model-assisted approach, remember that we're trying to use as few assumptions as possible and let the data do the talking. The unbiasedness assumption implies we can easily get the mean and, therefore, all that is left to do is to separate within- from between-person variance components. One drawback of this approach is that it uses an empirical distribution function derived from individual means, which has a limited number of jumps; each person provides only one rescaled mean and therefore the distribution function looks kind of jagged.

Two general approaches

- Model-Assisted (M-A) – rescales observed individual mean distribution

- Model-Based (M-B) – estimates distributions from theoretically-derived quantities

Slide 57

Now we move on to the model-based approach.

Rationale for the Model-Based approach

$$R_{ij} = \mu + u_i + \varepsilon_{ij}, \quad u_i \sim N(0, \sigma_u^2), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

- Distribution of usual intake is specified by estimated model parameters:

$$T \sim N(\hat{\mu}, \hat{\sigma}_u^2)$$

- Probabilities/quantiles can be computed from tabulations of the standard normal distribution

$$\Pr(T \le c) = \Phi\left(\frac{c - \hat{\mu}}{\hat{\sigma}_u}\right)$$

$$q_{p(T)} = \hat{\mu} + \hat{\sigma}_u \, \Phi^{-1}(p) = \hat{\mu} + \hat{q}_{p(\varphi)}$$

Slide 58

For the model-based approach, we assume distributions for the deviations. The equation at the top of the slide now demonstrates this; the tilde and N indicate that ui is normally distributed with mean zero and variance sigma squared u as before. The epsilons are also assumed to be normally distributed. If you look back to slides 52 and 53 in your notes later, you'll see that the model-assisted approach did not explicitly make the normality assumption.

In the model-based approach, we note that the distribution of usual intake is specified by estimated model parameters (i.e., that the distribution of true usual intake is itself normal with mean mu hat and variance sigma squared u hat, where the hats denote estimates as before). Under this assumption, the probabilities or quantiles can be computed directly from tabulations of the standard normal distribution. These equations might be familiar to you from a basic statistics course in which you used the normal probability table. The important thing to focus on here is the second equation that says that the pth quantile of true usual intake, denoted by q sub p (t) can be computed using the estimated model parameters and the quantile from the standard normal distribution, denoted in the middle by phi inverse of p. This means that if you want to know the 25^{th} percentile of usual intake, you take the 25^{th} percentile of the standard normal distribution, multiply by the standard deviation of the usual intake distribution, and add the mean of the usual intake distribution.

Rationale for the Model-Based approach

$$R_{ij} = \mu + u_i + \varepsilon_{ij}, \quad u_i \sim N(0, \sigma_u^2), \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

- Distribution of usual intake is specified by estimated model parameters:

$$T \sim N(\hat{\mu}, \hat{\sigma}_u^2)$$

- Probabilities/quantiles can be computed from tabulations of the standard normal distribution

$$\Pr(T \le c) = \Phi\left(\frac{c - \hat{\mu}}{\hat{\sigma}_u}\right)$$

$$q_{p(T)} = \hat{\mu} + \hat{\sigma}_u \, \Phi^{-1}(p) = \hat{\mu} + \hat{q}_{p(\varphi)}$$

Quantile from the distribution of u_i

Slide 59

I just want to point out here that the last term on the right, q p of phi, is a quantile from the assumed distribution of the deviations, ui.

Implementation using Monte Carlo simulation

- Randomly draw many (say *K*) values from the assumed normal distribution

$$u_k \sim N(0, \hat{\sigma}_u^2)$$

- Create simulated usual intake (pseudo-value)

$$r_k = \hat{\mu} + u_k$$

- Use empirical distribution of r_k as estimate of usual intake distribution

Slide 60

This approach can be implemented by looking stuff up in the normal table, but we'll find it more convenient to use what we call a Monte Carlo simulation approach. Basically, the idea is to randomly draw a whole bunch of values, say capital K, from the assumed normal distribution of the between-person deviations, u. Each one of these draws can be interpreted as a quantile from the between-person deviation distribution, which we highlighted on the last slide. Then we create simulated usual intake, also called a pseudo-value, by adding on the estimated population mean. And now, as we did in the model-assisted approach, we use the empirical distribution of the little r k's to estimate the usual intake distribution. This is in contrast to using r i's as in the model-based approach, which were based upon observations for a particular person.

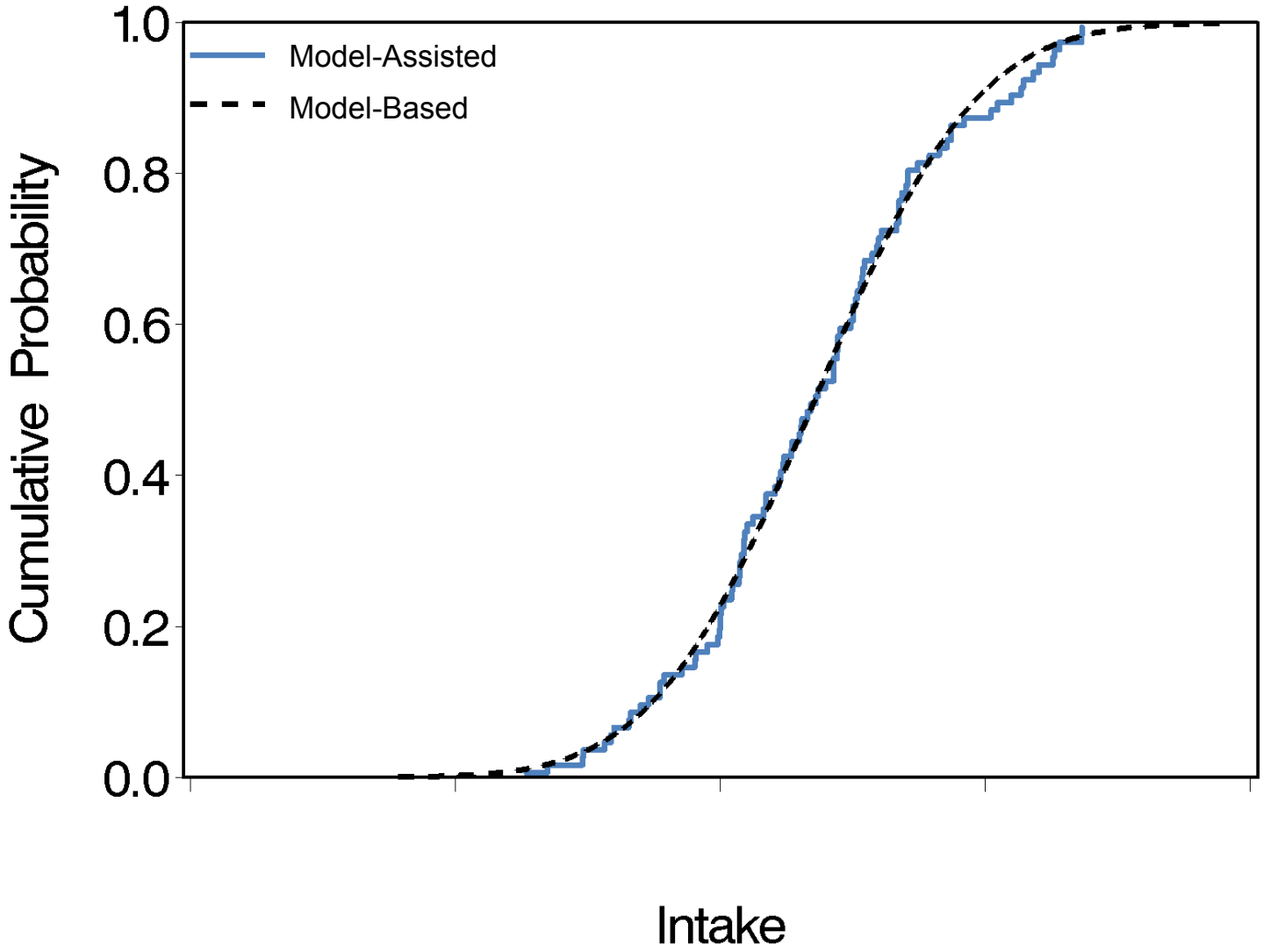
Features of the Model-Based approach

- Less robust, uses more assumptions than M-A
- Assumes distribution of *u* is known
- More precise percentile estimates
  - No limit to smoothness of estimated distribution function

Slide 61

There are some disadvantages to the model-based approach. For one thing, because it uses more assumptions than the model-assisted approach, it is less robust. Primarily, the weakness is that we assume that the distribution of true usual intake is known; in other words, normal. However, this approach can give us much more precise percentile estimates because we can make as many r k's as we want to and if the distribution really is normal, we'll get a very smooth curve…

Features of the Model-Based approach

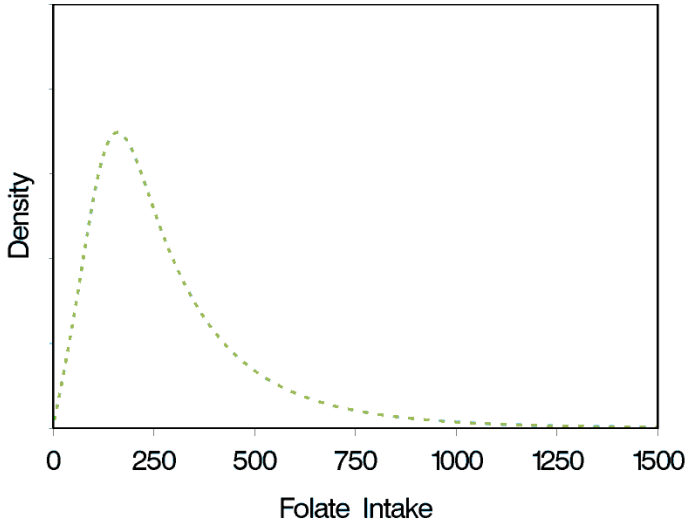


Slide 62

…as shown on this slide.

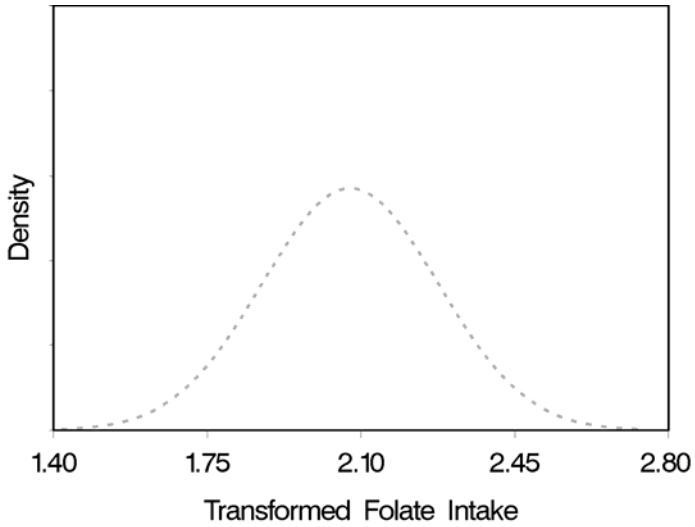
Accounting for nonlinear transformations

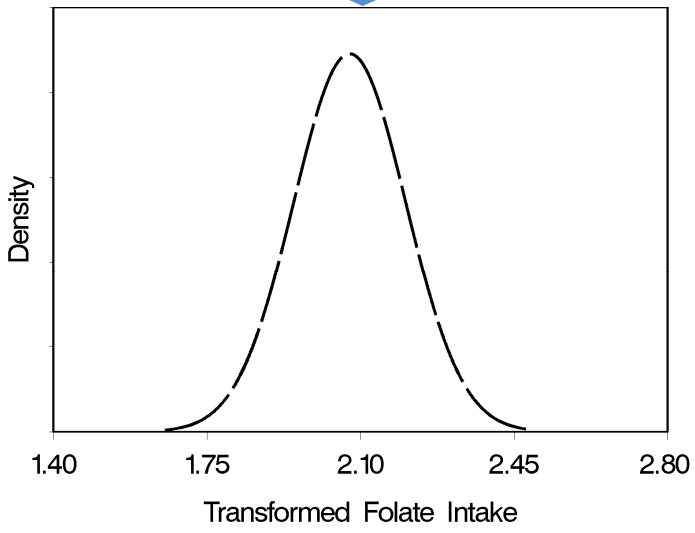
Original Scale



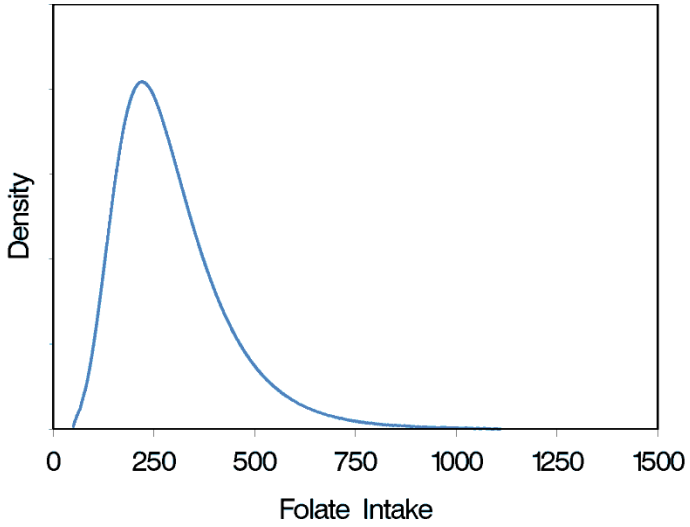
Transform

Transformed Scale





Backtransform



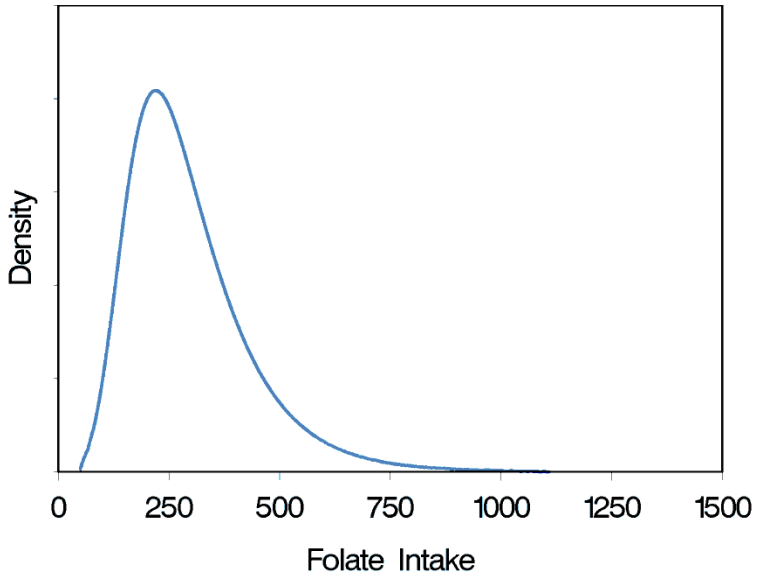
Slide 63

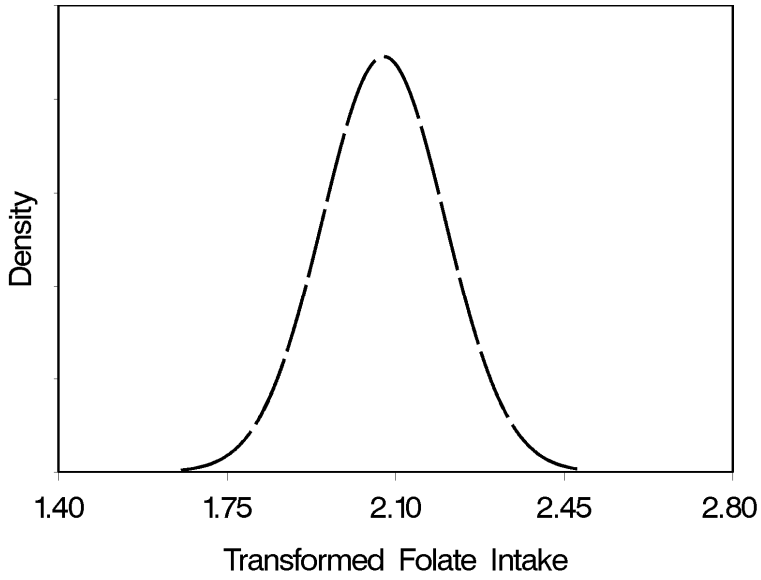
We've been talking a lot about normality here, so let's go back and see how we're going to deal with the fact that dietary data are skewed and we'll require some sort of transformation before we can apply the distribution estimation methods I've been discussing.

This slide gives a schematic overview of the approach we're going to take. We'll take our skewed data on the original scale, apply a transformation such as a log or Box-Cox so that the data look normal, we'll estimate our variance components and shrink the top right distribution by removing the effect of within-individual variation, and then we'll apply what we call a backtransformation to the distribution on the bottom right to return our usual intake distribution to the original scale of interest.

Estimating quantiles when transformations are used

- Goal is to estimate a quantile of usual intake that corresponds to one in the normal distribution that exhibits only between-person variance



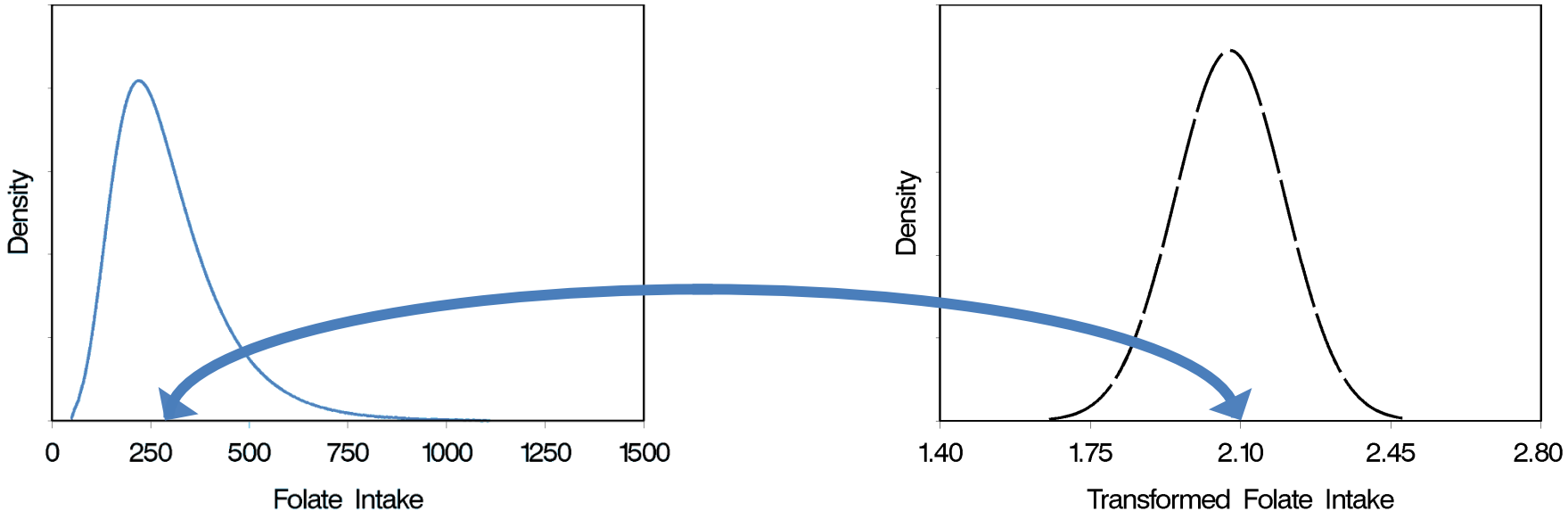


Slide 64

So what do the transformation and backtransformation mean for estimating quantiles? Our goal is to estimate the quantile of usual intake that corresponds to one in the normal distribution that exhibits only between-person variance.

Estimating quantiles when transformations are used

- Goal is to estimate a quantile of usual intake that corresponds to one in the normal distribution that exhibits only between-person variance



Slide 65

So we've already transformed the original data to normality and removed the effect of within-person variation; in a sense, we have averaged out the day-to-day variation. Now we want to map quantiles in the shrunken normal scale distribution to quantiles of the backtransformed usual intake distribution.

Estimating quantiles when transformations are used

- With no transformation used:

$$
\begin{aligned}
q_{p(T)} &= \mathrm{E}[\mu + u + \varepsilon \,|\, u = q_{p(\varphi)}] \\
&= \mu + \mathrm{E}[u \,|\, u = q_{p(\varphi)}] + \mathrm{E}[\varepsilon \,|\, u = q_{p(\varphi)}] \\
&= \mu + q_{p(\varphi)}
\end{aligned}
$$

- Estimated quantile is a linear function

Slide 66

These next couple of slides are a little complicated and are included for the sake of completeness. I am not going to go through the equations in detail; I'm just going to hit the highlights. First, when there is no transformation involved, any estimated quantile of usual intake is simply a linear function of the model parameter, mu, and quantiles from a normal distribution.

Estimating quantiles when transformations are used

- With nonlinear transformation *g* used:

$$q_{p(T)} = \mathrm{E}[g^{-1}(\mu + u + \varepsilon) \,|\, u = q_{p(\varphi)}]$$
$$= \mathrm{E}[g^{-1}(\mu + q_{p(\varphi)} + \varepsilon)]$$

- Estimated quantile is an integral

- Can be calculated/approximated several ways

Slide 67

When a nonlinear transformation g is used, the estimated quantile is not a linear function; instead, it is an integral. This integration can be carried out in several ways.

Integration provides the "backtransformation"

- Taylor series approximation (Dodd, 2006):

$$q_{p(T)} \approx g^{-1}(\mu + q_{p(\varphi)}) + \tfrac{1}{2}(g^{-1})''(\mu + q_{p(\varphi)})\sigma_{\varepsilon}^{2}$$

- Exact calculation for normal ε (Hoffmann, 2002)

- Numerical integration for known ε distribution
  - Quadrature formulas, e.g., Gauss-Hermite
  - Monte Carlo integration

Slide 68

One approximation is based on the same sort of Taylor series expansion we saw on slide 44 that uses just the first two moments; that is, the mean and the variance of the distributions of deviations. Another method uses an exact calculation for normal epsilons or for the log transformation. In the general case, numerical integration can be used for any specified epsilon distribution. This includes quadrature formulas such as Gauss-Hermite or Monte Carlo integration. The important thing to keep in mind is that one characteristic that differentiates between the methods developed so far for estimating usual intake distributions is how they do this backtransformation.

Estimation approaches when transformations used

- Both Model-Assisted and Model-Based approaches can be extended

- If transformation *g* achieves the desired distribution of ε terms, Taylor series approximation may be poor
  - Alternatives use all moments, not just two

Slide 69

Both the model-assisted and model-based approaches for estimating usual intake distributions can be extended to handle these backtransformations. Some methods extend the model-assisted approach and some extend the model-based approach; this is another differentiating characteristic for the methods. Lastly, if the transformations really do a good job of achieving normality, the Taylor series approximation may not perform very well in some cases because it only uses the first two moments where its competitor methods use all of the moments. This is one reason why the NCI method has recently been updated to use a quadrature method by default rather than the Taylor series.

Evolution of estimation methods

Method	Transformation	Distributions via
NRC (1986)	None*	M-A
Slob (1993)	Log	M-B
BP (1996)	Power	M-A
ISU (1996)	Two-stage	M-B/M-A
NCI (2006)	Box-Cox	M-B/M-A
MSM (2011)	Box-Cox	M-A
SPADE (2012?)	Box-Cox	M-B

* NRC method incorporates transformations under alternative assumptions

Slide 70

Several methods have been developed over the past 25 years. This table lists some of them in order of their appearance and shows which transformations the methods will permit and also the general approach they take to estimating distributions of usual intake. You'll notice that the BP and ISU methods both made their appearance at the same time; this is because both were presented in a single publication by researchers at Iowa State University.

Software availability for estimation methods

Method	Software?	Platform
NRC (1986)	Yes	SAS/C/Windows
Slob (1993)	N/A	N/A
BP (1996)	Yes	SAS/C/Windows
ISU (1996)	Yes	SAS/C/Windows
NCI (2006)	Yes	SAS
MSM (2011)	Yes	R (via Website)
SPADE (2012?)	Yes (beta)	R

Slide 71

For a number of these methods, software has been developed and released to the public so that researchers can apply these methods to their data. This table shows for the methods in the previous slide whether software is available and, if so, which computing platforms the software runs on. The BP and ISU methods are implemented in a single program called SIDE. The additional resources associated with this webinar include references that can be used to get more information on these methods.

Summary

- Within-individual variation is adjusted out, leaving only between-individual variation
- Two approaches to estimate distributions
  - Model-assisted vs. Model-based
- Use of normalizing transformations requires special care in estimating distributions
  - Backtransformations of varying complexity
- Wide range of software implementations

Slide 72

Okay, so to summarize, methods of estimating distributions of usual intakes separate within- from between-person variation and remove the effects of the former. There are two approaches to estimating distributions, model-assisted and model-based, each of which has pros and cons. As we discussed, intake data are often very skewed, leading to the routine use of normalizing transformations; this raises some complex technical issues related to backtransformations. Finally, several methods have been developed over the years to estimate usual intake distributions and there is a wide range of software implementations available for researchers.

THE ROLE OF COVARIATES

Slide 73

In the last section of the talk, I'll focus on the role covariates can play in modeling usual intake distributions.

The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics

Slide 74

First off, why do we consider using covariates? This comes from the interest we often have in estimating usual intake distributions for subpopulations. Nutritional status often depends upon personal characteristics.

Dietary Reference Intakes: Estimated Average Requirements

Food and Nutrition Board, Institute of Medicine, National Academies

Life Stage Group	Calcium (mg.d)	CHO (g/d)	Protein (g/kg/d)	Vit A (µg/d)	Vit C (mg/d)	Vit D (µg/d)	Vit E (mg/d)	Thiamin (mg/d)	Riboflavin (mg/d)
Infants									
0 - 6 mo									
6 - 12 mo			1.0						
Children									
1-3 y	500	100	0.87	210	13	10	5	0.4	0.4
4-8 y	800	100	0.76	275	22	10	6	0.5	0.5
Males									
9-13 y	1,100	100	0.76	445	39	10	9	0.7	0.8
14-18 y	1,100	100	0.73	630	63	10	12	1.0	1.1
19-30 y	800	100	0.66	625	75	10	12	1.0	1.1
31-50 y	800	100	0.66	625	75	10	12	1.0	1.1
51-70 y	800	100	0.66	625	75	10	12	1.0	1.1
> 70 y	1,000	100	0.66	625	75	10	12	1.0	1.1
Females									
9-13 y	1,100	100	0.76	420	39	10	9	0.7	0.8
14-18 y	1,100	100	0.71	485	56	10	12	0.9	0.9
19-30 y	800	100	0.66	500	60	10	12	0.9	0.9
31-50 y	800	100	0.66	500	60	10	12	0.9	0.9
51-70 y	1,000	100	0.66	500	60	10	12	0.9	0.9
> 70 y	1,000	100	0.66	500	60	10	12	0.9	0.9

Slide 75

For example, nutrient requirements such as Estimated Average Requirements may be defined by life-stage groups, as shown in this table.

The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics
- Population monitoring:
  - Characterizing *a priori* "at-risk" subpopulations
  - Proportion not meeting sex/age-specific targets vs. not meeting "average" target

Slide 76

In population monitoring, often, we would like to characterize some *a priori*-defined, at-risk subpopulations. For example, we may wish to estimate the proportion of various subgroups not meeting a target requirement estimate like an EAR that is specific to age and sex. If we couldn't examine subpopulations, the best we could do would be to estimate the proportion not meeting some average target.

One answer is to stratify sampled data

- Run separate analyses on subsamples defined by personal characteristics
  - Population proportion not meeting sex/age targets is weighted average of subpopulation proportions
- Small subsamples lead to less precise estimates

Slide 77

One answer to this problem is to run separate analyses on subsamples defined by personal characteristics. With this approach, the proportion of the overall population of interest not meeting their specific targets is the weighted average of the subpopulation proportions, where the weights are determined by the relative sizes of the subpopulations. However, the drawback to this stratification approach is that taking small subsamples leads to less-precise estimates.

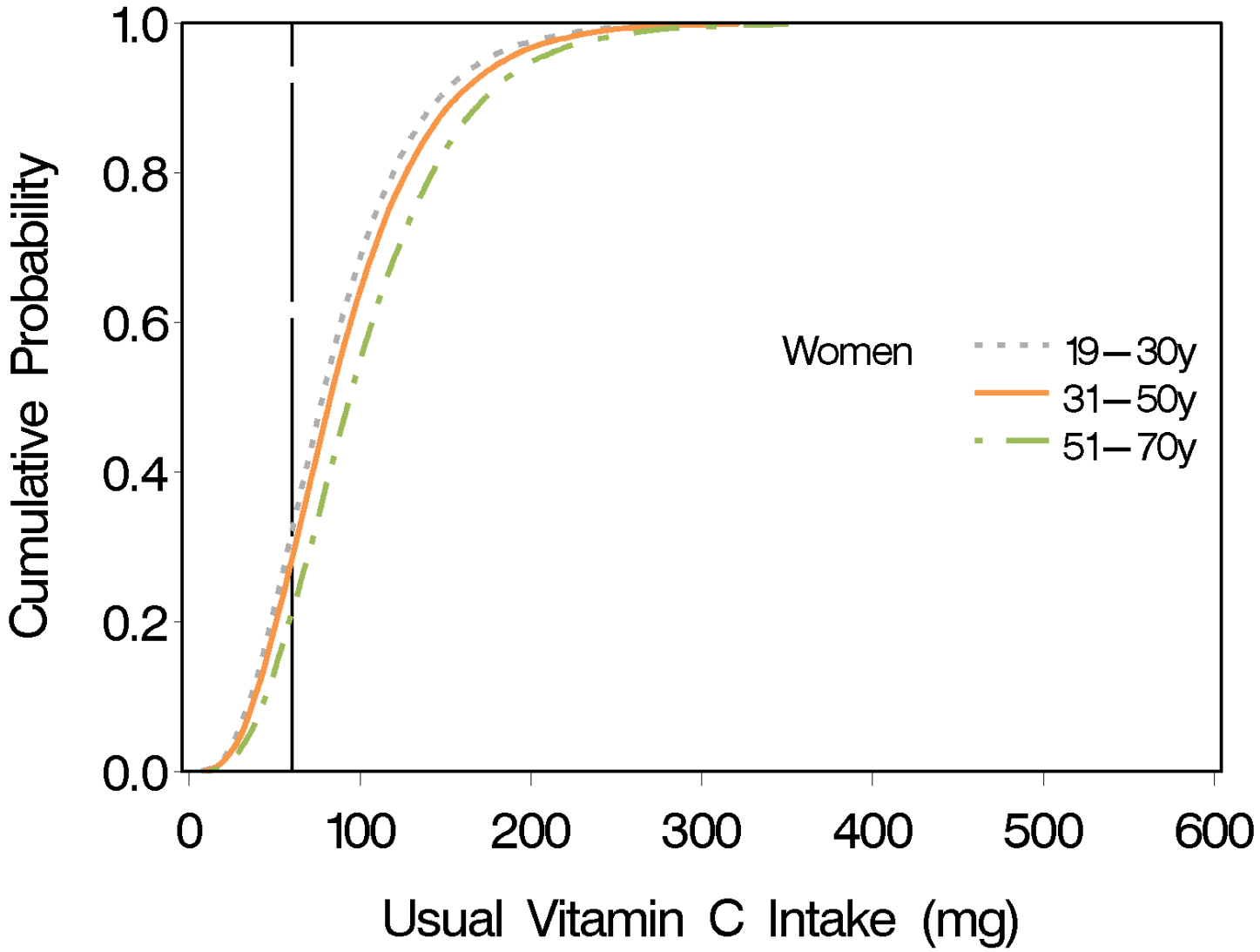
The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics
- Population monitoring:
  - Characterizing *a priori* "at-risk" subpopulations
  - Proportion not meeting sex/age-specific targets vs. not meeting "average" target
- Understanding determinants of diet
  - Identify characteristics associated with higher/lower average intake, e.g., smoking

Slide 78

We might also be interested in determining the factors that influence diet; for example, we may think that smokers have a higher or lower average intake than nonsmokers. We could generate separate estimates for smokers and nonsmokers and test if the distributions of usual intake for the two groups are similar or different. This is another application of the stratification idea.

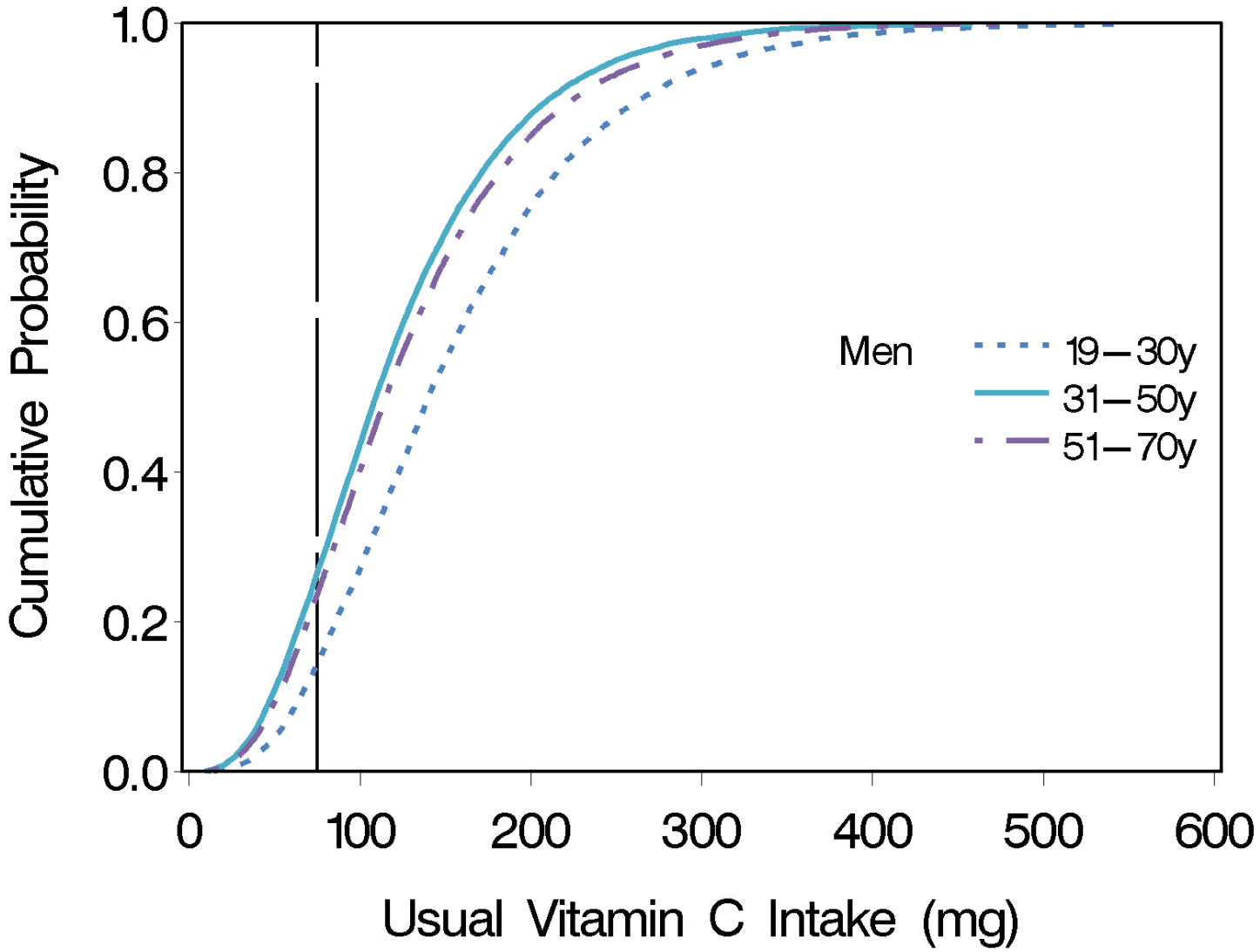
Example: Eating at America's Table Study (EATS)



Slide 79

For example, here, I'm showing separate estimates of usual intake of vitamin C by age group, using data from women in the Eating at America's Table Study. You can see that there are some differences in the proportions meeting the estimated average requirement of 60 mg per day.

Example: Eating at America's Table Study (EATS)



Slide 80

This shows a similar figure for men, showing more pronounced differences across subpopulations. Here, the EAR is 75 mg.

Limitations of stratification approach

- When multiple factors thought to influence diet are considered,
  - Subsample sizes decrease dramatically
  - Analysis burden increases

- Allowing covariates in the statistical models can overcome this limitation

Slide 81

However, there are limitations to the stratification approach. When multiple factors thought to influence diet are considered—in other words, you don't have *a priori* subpopulations identified—the subsample sizes decrease dramatically and the analysis burden increases. For example, one might end up running analyses for 14 or 16 age and sex groups also stratified by 2 or 3 categories of another variable such as smoking status or race/ethnicity. Allowing covariates is a way to overcome this limitation of the stratification approach.

A mixed model formulation

$$R_{ij} = \mu + u_i + \varepsilon_{ij}$$

- Population mean is a **fixed effect**
  - Only one model parameter to estimate
- Deviations are **random effects**
  - Reflect variation from individual persons/days
  - Focus on higher-order moments, e.g., variance
- **Mixed models** include fixed and random effects

Slide 82

So how do we go about including covariates? We start by recognizing that the models we've been talking about have been studied extensively in the statistical literature and are examples of what are called mixed models. Here is the model we've been considering for recalls. In this model, one of the parameters we estimate is the population mean; we call it a fixed effect. The within- and between-person deviations are examples of random effects that reflect variation from individual persons or days. The two parameters needed to account for these effects are variances, and we make distributional assumptions about them, so focus is on the higher-order moments. The fact that we have both fixed and random effects is what makes this a mixed model.

A mixed model formulation including covariates

$$R_{ij} = \mu(\mathbf{X}) + u_i + \varepsilon_{ij}$$

- Fixed effect part of the model expressed as a **function** of measured covariates **X**
  - Multiple parameters to estimate
  - Allows “structured” variability in group means
- Random effects reflect variation from all other unmeasured characteristics
  - “Unstructured” variability

Slide 83

When we extend this mixed-model formulation to include covariates, we operate on the fixed effect, or mean-focused, part of the model. Here, we write the mean of the group as a function of measured covariates denoted here by the bold X. Now instead of one parameter, we have several parameters to estimate. This extension allows what I call structured variability in our group means—in other words, people with different covariate values might have different intakes and we have an idea of why they are different; that's why it's structured. We have the same random effects in the variance-focused part of the model to reflect variation from all of the other unmeasured characteristics. This part of the model soaks up all of the leftover unstructured variation we can't explain.

Types of covariates

- **Individual-level**: affects true intake on all days, e.g., gender, age, smoker/nonsmoker status

- **Time-dependent**: affects true intake on specific days, e.g., season, weekday

- **Nuisance**: affects reporting error, e.g., interview sequence, mode (telephone vs. in-person)

Slide 84

There are three types of covariates that we consider. Individual-level or time-independent covariates affect true intake on all days; for example, sex, age, and smoking status.

A second kind of covariate is a time-dependent covariate that affects true intake on specific days, such as season or weekday effects.

Lastly, there are nuisance effects that affect reporting error, not true intake. Examples of these are interview sequence or interview mode.

Potential benefits of incorporating covariates

- Allows different means for subpopulations, while pooling information about variance components
  - Point estimates for overall population may be unaffected by covariates,
  - But should be more precise if model holds

Slide 85

Incorporating individual or time-independent covariates into our models allows us to get separate means for subpopulations while continuing to pool information about variance components. This is in contrast to stratification, where the reduced sample size makes the variance component estimates less stable. The point estimates for the overall population may not be affected by covariates because they reflect all of the variation, both structured and unstructured, but these estimates should be more precise if the model really holds (i.e., the subpopulations do have common variance components) because of the pooling of information.

Potential benefits of incorporating covariates

- Can investigate multiple determinants of diet
  - Test significance of main effects/interactions
  - Joint modeling leads to lower analysis burden

Slide 86

Another benefit of incorporating covariates is that multiple determinants of diet can be investigated by testing the significance of main effects or interactions, and you can do this in a single model, thereby reducing the analysis burden. These significance tests could apply to time-independent, time-dependent, or even nuisance effects.

Potential benefits of incorporating covariates

- Overall bias due to nuisance effects can be corrected

- In epidemiologic applications, less **unstructured** variation is better

Webinar 10

Slide 87

As well, including covariates in our models allows us to correct overall bias due to nuisance effects. That is, we can correct for the overall tendency for weekends and weekdays to have different intakes but we can't correct at the individual level if different persons have different magnitudes of these biases. Most importantly for this webinar series, we will discover in webinar 10 that minimizing leftover unstructured variation by allowing covariates makes these models very useful in studies of diet-and-health relationships.

Estimating distributions with covariates in the model

- Model-Assisted: use observed covariate pattern $\mathbf{X}_i$ for *i*-th individual:

$$r_i = (\overline{R}_{i\bullet} - \hat{\mu}(\mathbf{X}_i)) \sqrt{\frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_\varepsilon^2}{J}}} + \hat{\mu}(\mathbf{X}_i)$$

- Model-Based: use a specified covariate pattern $\mathbf{X}_0$ for *k*-th pseudo-value:

$$r_k = \hat{\mu}(\mathbf{X}_0) + u_k$$

Slide 88

To briefly sketch out how the procedures for estimating distributions are affected by allowing covariates, I'll show equations similar to those in earlier slides for the model-assisted and model-based approaches. For the model-assisted approach, the observed covariate pattern X for the ith individual is used to generate the rescaled mean, as shown here. For the model-based approach, a specified covariate pattern, symbolized here by X zero, can be selected for any one of the pseudo-values generated. What are the implications of this?

Estimating distributions with covariates in the model

- Model-Assisted approach retains observed joint distribution of individual-level covariates
  - Some covariate combinations may be rare
  - M-B: draw $\mathbf{X}_0$ at random from observed joint distribution to mimic this behavior

- Model-Based approach also offers a choice to perform **direct standardization**
  - Draw $\mathbf{X}_0$ from a **standard population**

Slide 89

The model-assisted approach, in keeping with the principle of sticking close to the data, will always retain the observed joint distribution of individual-level covariates, even if some of the covariate combinations are very rare. If you want to mimic this behavior in the model-based approach, the specified covariate pattern X zero can be drawn from the observed joint distribution at random. However, the model-based approach also offers a choice to perform what we call direct standardization, where you might be interested in estimating usual intake distributions assuming that the true distribution of covariates in the population is different from the distribution observed in your data. This idea is used, for example, when age-adjusting to some previous decade's demographic distribution. Here, the idealized distribution is called the standard population. Doing direct standardization with the model-based approach means that you draw your X zero at random from the standard population rather than from the observed joint distribution.

Estimating distributions with covariates in the model

- Model-Assisted and Model-Based similar unless
  - Important covariate(s) are omitted, and/or
  - Exact normality does not hold

- Discrepancy between Model-Assisted and Model-Based distributions useful as a diagnostic

Slide 90

Both the model-assisted and model-based approaches should give you similar distributions unless the additional assumptions made for the model-based approach don't hold; for example, if you've left out some important covariates or exact normality does not hold. Examining the discrepancies between the two approaches can be useful as a diagnostic tool.

Direct standardization for time-dependent covariates

- Overall usual intake is weighted average of time-dependent usual intake

- Weights come from the standard population, e.g., for weekend/weekday effects:

Standard Population for Weekdays/Weekend Days

Weekend	Days of Week	Weight
No	MTWT	4/7
Yes	FSS	3/7

Slide 91

This direct standardization approach is also used to adjust for time-dependent covariates; for example, where we say that overall usual intake is a weighted average of time-dependent usual intake where the weights come from a standard population, exemplified here by this table which shows that if you define weekends as Friday, Saturday, and Sunday and weekdays as the rest of the days of the week, the weights are 4/7 for weekday days and 3/7 for weekend days. That is, overall usual intake is 3/7 of the weekend usual intake plus 4/7 of the weekday usual intake.

Explicit adjustments for nuisance effects

- Can be done before fitting the mixed model, or

- In a two-stage process:
  - Include nuisance effects in the mixed model
  - Estimate distributions using group means calculated with nuisance covariates set to fixed reference values, e.g., the first interview, or the in-person interview

Slide 92

Adjusting for nuisance effects can be done before fitting the mixed model by adjusting the raw data to remove the influence of nuisance covariates; this is done in the ISU method, for example. Alternatively, you can do the adjustment in a two-stage process by including the nuisance effects in the mixed models and then estimating distributions using group means calculated with the nuisance covariates set to specific values; for example, always computing the mean as if the recall was the first interview or the in-person interview where multiple modes of administration are used.

Types of covariates allowed in available methods

Method	Covariates Allowed
NRC (1986)	None
Slob (1993)	None
BP (1996)	Nuisance
ISU (1996)	Nuisance
NCI (2006)	Individual, Time-dependent, Nuisance
MSM (2011)	Individual
SPADE (2012?)	Individual, Time-dependent*, Nuisance

* fractional polynomial option for age

Slide 93

Returning to the list we've seen before, this table shows what kinds of covariates are allowed in each method. The earliest methods do not account for covariates at all but methods developed later have the capability of incorporating various types of covariates. In the question-and-answer session in the first webinar, a question about modeling usual intakes for children was asked. I wanted to point out that the program SPADE does allow a special model for the age covariate.

Summary

- Covariates provide an alternative to stratification
- Mixed model allows a combination of structured and unstructured variation
- Both approaches to distribution estimation (M-A and M-B) can be extended to handle covariates of three types: *individual*, *time-dependent*, and *nuisance*
- Not all available methods incorporate covariates; if they do, implementations vary

Slide 94

Okay, so to summarize this section of the webinar, allowing covariates into our models provides an alternative to stratification, which we saw has some drawbacks. The mixed-model formulation allows us to model both structured and unstructured variation. Both the approaches to distribution estimation that we discussed—model-assisted and model-based—can be extended to handle three different types of covariates, individual or time-independent, time-dependent, and nuisance. Not all available methods incorporate covariates and the details of the implementations vary across methods.

QUESTIONS & ANSWERS

Moderator: Sharon Kirkpatrick

Please submit questions using the *Chat* function

Measurement Error Webinar 2 Q&A

Question: **The first question relates to the assumption that the within-person variance is the same across people, which is not likely to be true across ethnic groups. How could violating this assumption affect the model?**

Well, it's obviously going to violate the assumption that we talked about, and depending on how different those variance components are across ethnic subgroups, it's going to influence how much that affects your analysis. But it is nice to be able to—in that case it's quite possible that you do want to try some sort of stratification approach, at least at the race/ethnicity level so that you can allow those variance components to be different, but then allow other covariates—I mean you might want to say let age come in as a covariate so that you don't have to keep stratifying by smaller and smaller cross-cuts based on your covariates. *(K. Dodd)*

Also related to race and ethnic groups: If the 24-hour recalls are adjusted using an FFQ, could this result in over- or underadjusting for specific nutrients or could this adjustment introduce a new type of error into the data? And this person specifically wanted to know about gene-diet interactions and small sample size per subgroup.

That's a very good question. I think this concept of adjusting using FFQs and 24-hour recalls together and adjusting one for the other, is going to be discussed more in a later webinar. And what was the question again. I want to take a look at your notes here.

So I think that's where you'll really see how this stuff is going to come into play. But operationally, I think a lot of times you have to assume that if you have 24-hour recalls and an FFQ, you're going to have to assume that one of them is unbiased and the other provides an additional source of information that may be biased, so you'll usually get—let's see…. *(K. Dodd)*

That goes back to having an FFQ that's appropriate to your population. *(S. Kirkpatrick)*

I think it does. That's the sort of thing that comes back into that. I think there are lots of ways that people try to make FFQs most applicable to their population at hand. *(K. Dodd)*

And if it's not, you might not want to use that for your model. *(S. Kirkpatrick)*

Right, you might not want to, or you might; of course, the same thing also applies for 24-hour recalls. You may need to really do a lot of database

work behind the scenes to make sure that you have good entries about ethnically diverse food patterns. *(K. Dodd)*

Another participant heard that NHANES is moving from two recalls on all participants to a single recall on all participants and a second recall only on a subset. Can you comment on the statistical implications for estimating nutrient inadequacies?

This is where this idea of modeling and pooling information from different individuals really starts making a big impact, because if you have for a given iteration of NHANES that you have very limited information from the second recall for a given cycle, you may have to go back to a previous cycle that had a larger faction of people with two recalls and try to pool information from previous survey years into your estimates for the new data. This is another form of modeling where you have to take special care to do it, but that's probably the approach people are going to want to take, is that they're going to want to try to combine information from multiple waves of the NHANES into some sort of model where you let the wave of NHANES be treated as a covariate or maybe some other ways that you incorporate that pooling idea of information. *(K. Dodd)*

Next, can these methods be translated to food record data; for example, by treating each day of a record as a recall?

I think that, in general, the idea of a recall or a diet—I'm sorry, a food record or a multiday diary is that the average over the, say, a three-day diet—the average over those three days is supposed to be treated as one application of the instrument so that if you want to apply these methods, then you ought to be thinking about replicating the entire instrument over a period of time, separated by a period of time, to make this independence hold. So I think that these methods are directly applicable when you talk about repeating multiple applications of the entire instrument. So you can't just take, or you probably shouldn't just take, each day of a three-day diary and say, "These are three independent days or three separate days." There has been some modeling work done where you really do try to incorporate this idea of consecutive days. The ISU method, for example, does a lot of that, but the direct analog when using diaries as your short-term measures is that you should think about doing replicates of the entire instrument. So I hope that explains and answers your question. *(K. Dodd)*

In C-Side, we have the option of "controlling" for race and age groups, for example, in addition to what you call nuisance variables (e.g., day of

week). You only listed C-Side as handling nuisance variables. Are we not understanding this function of the program?

Well, that's true that C-Side will adjust for race/ethnicity, but it does so by treating the race/ethnicity effect—what it does is it does a sort of direct standardization where it says, "I'm going to adjust my raw data to reflect what happens if everyone had the same race/ethnicity. I'm going to remove the effect of the race/ethnicity differences and talk about the usual intake distribution of sort of the average person in the population. So that's the way that C-Side kind of operationalizes that. What it doesn't do is it doesn't maintain internally to its estimation separate effects for race/ethnicity in the modeling part. So it doesn't directly bring the covariates in. It doesn't directly do a more complicated mixed model like what we saw. But it does try to do the sort of adjustment, so that is one of the feature of Side that is often used but doesn't quite have the same interpretation as the kinds of adjusted distributions we get, or some type of distributions we get, using a mixed-model approach where you have covariates involved. *(K. Dodd)*

For estimating usual intake with covariates, is it best to use a mixed model as opposed to generalized estimating equations?

I think that the generalized estimating equations and the mixed models are basically similar ways of getting at the same idea. I'm not sure there is necessarily a reason to prefer one method over the other, as long as the two things are trying to get at the same quantities. *(K. Dodd)*

This next question relates to the number of recalls and people; specifically, the participant asked whether this modeling applies to greater than ten recalls collected on over 100 individuals. Basically, how many people and how many recalls do you need?

That's also a very difficult question. There is the idea of need vs want, that to do the methods that we've discussed here, you only need two, okay. Every second recall you get adds one degree of freedom to your variance component estimates. There is sort of a hard and fast rule of thumb that says that, you know—it's my personal rule of thumb, anyway, is that I want to have 50 or 100 people that have—I want to make my variance component estimates have 50 or 100 degrees of freedom. So I need to have 50 or 100 second recalls to do that with. Now, if you have a lot more recalls, you get—if you have three recalls, each set of three provides two degrees of freedom for the variance component estimates. So you can start building up a lot of degrees of freedom for the within-person

variance component, but you start losing—but after a while, getting those extra degrees of freedom for the within-person variance component, while having to reduce further the number of people that you can get with that many days, you start getting to the point that the between-individual variance component becomes limited by your sample size. So you really want to have a reasonable number of people and a reasonable number of days per person so that you can—so the degrees of freedom are large enough or there is enough of them to get a good estimate of the things you're trying to estimate. So I really can't say that for any particular purpose, you need to have exactly this many days, but we will see in the next webinar that—how the number of days—well, not in the next webinar but in a later webinar—how the number of days can influence the precision of some of your estimates. *(K. Dodd)*

Next, how is the statistical method validated? Was the estimate of population mean of 24-hour recalls validated against multiple dietary records?

We usually talk about validating the method merely by saying, "If …." Usually, it's through some sort of simulation study where you say, "If your data really act like you think they do and if you can generate many, many, many days of recall and average it, do you get the same distribution of average if you apply the statistical methods to using only a couple of recalls?" So the methods themselves have been validated in that respect for many—over the course of many, many years. But now you're talking about validation in the sense of: Are these estimates that we get with these methods—are they the real usual intake distributions? And I mentioned right off the bat that I am making this unbiased assumption about 24-hour recalls. And I know [it] does [not strictly hold] in practice. *(K. Dodd)*

Let me just say a clarification that they are speaking about validation using real data. (S. Kirkpatrick)

Well, I just think that the issue of whether 24-hour recalls and multiday food records or diaries—the fact that when you do these things and you take real data on both, meaning you take observations on both, and you look at averages or something like that, and you see they don't match up exactly right, that's something that goes along with validating the instrument. It does not have much to do with validating the methods themselves, and that's what I was trying to focus on here today. *(K. Dodd)*

[This page intentionally blank.]

National Cancer Institute

measurement ERROR webinar series

Next Session

Tuesday, October 4, 2011
10:00-11:30 EDT

Estimating usual intake distributions for foods and nutrients consumed episodically

Janet Tooze
Wake Forest University

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

Slide 96

In this webinar, we've covered a lot of ground, explaining how we build statistical models from the ground up and how these models can be used to help us to estimate distributions of usual intake even when we have limited information available per person. As I mentioned earlier, I gave an overview of different methods, and in webinar 3, Dr. Tooze will focus more specifically on the use of a specific method, walking through an example for episodically consumed dietary components using the National Cancer Institute Method.

That brings today's webinar to a close. Please join us next week for webinar 3, when Dr. Janet Tooze will discuss estimation of usual intake distributions for episodically consumed dietary components.