National Cancer Institute

measurementERRORwebinar series

Estimating usual intake distributions for dietary components consumed episodically

Janet A. Tooze, PhD, MPH
Wake Forest School of Medicine

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

Slide 1

Hello. I'm Sharon Kirkpatrick from the U.S. National Cancer Institute. Welcome to the third webinar in the Measurement Error Webinar Series. In webinar 2, Kevin Dodd provided an overview of methods of estimating usual intake distributions for non-episodically consumed foods. Today we will continue with the theme of estimating usual intake distributions but this time with a focus on episodically consumed foods with Dr. Janet Tooze.

A few quick notes before we get started with the presentation:

1. The webinar is being recorded so that we can make it available on our Web site.
2. All phone lines have been muted and will remain so throughout the webinar.
3. There will be a question and answer period following the presentation. If you would like to submit a question, you can do so using the Chat feature at the left of your screen.
4. A reminder: Various resources, including the slides from this session and the glossary of key terms and notation, are available on the webinar series Web site. The URL is available in the note box at the top left of your screen.

Now, let's move on to today's presentation. We are fortunate to have Dr. Janet Tooze as a member of the Surveillance Measurement Error Working Group at the National Cancer Institute and a presenter in this series. Dr. Tooze is an Associate Professor in the Department of Biostatistical Sciences, Division of Public Health Sciences, at Wake Forest School of Medicine. She has expertise in longitudinal data analysis and nonlinear mixed effect models, with specific applications to diet and physical activity assessment. Her work to develop a statistical model for repeated measures data with excess zeroes provided a foundation for the development of the National Cancer Institute or NCI method. As I mentioned earlier, in today's presentation Dr. Tooze will discuss the estimation of usual intake distributions for episodically consumed dietary components. In contrast to webinar 2 in which Dr. Dodd provided an overview of different methods of estimating usual intake distributions, Dr. Tooze will focus specifically on the application of the NCI method. Dr. Tooze.

Presenters and collaborators

Sharon Kirkpatrick
Series Organizer

Regan Bailey

Dennis Buckman

Raymond Carroll

Kevin Dodd

Laurence Freedman

Patricia Guenther

Victor Kipnis

Susan Krebs-Smith

Douglas Midthune

Amy Subar

Fran Thompson

Janet Tooze

DEPARTMENT OF HEALTH & HUMAN SERVICES · USA

NATIONAL INSTITUTES OF HEALTH

NATIONAL CANCER INSTITUTE

Office of Dietary Supplements
National Institutes of Health

USDA

Wake Forest™ School of Medicine

The Gertner Institute

STATISTICS
TEXAS A&M UNIVERSITY

Slide 2

I’d like to begin by acknowledging the multidisciplinary team of presenters and collaborators who contributed to the development of the NCI method and to this webinar series.

measurement ERROR webinar series

This series is dedicated to the memory of

Dr. Arthur Schatzkin

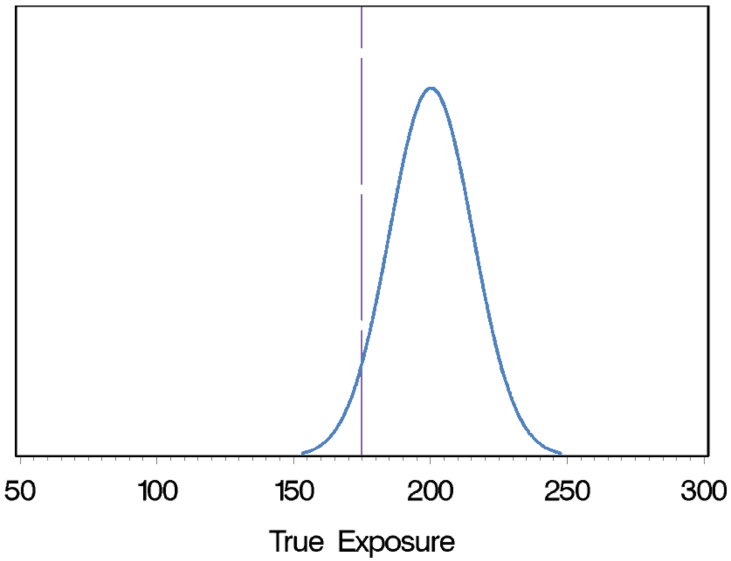
In recognition of his internationally renowned contributions to the field of nutrition epidemiology and his commitment to understanding measurement error associated with dietary assessment.

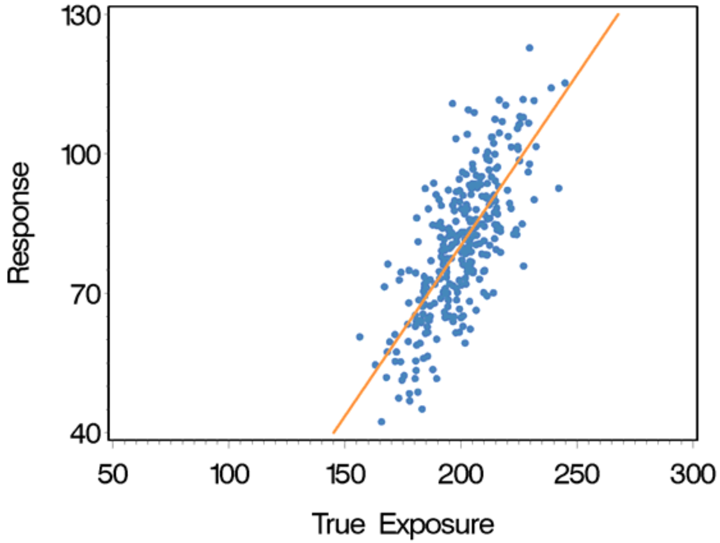
Slide 3

This series is dedicated to the memory of Dr. Arthur Schatzkin. This seems especially appropriate for my webinar today, as it was a talk I heard by Dr. Schatzkin about dietary measurement error ten years ago that inspired me to start working in this field. It is my hope that, in some small way, this webinar series inspires you to begin or to continue research related to dietary measurement error and its implications.

Two main areas of interest

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





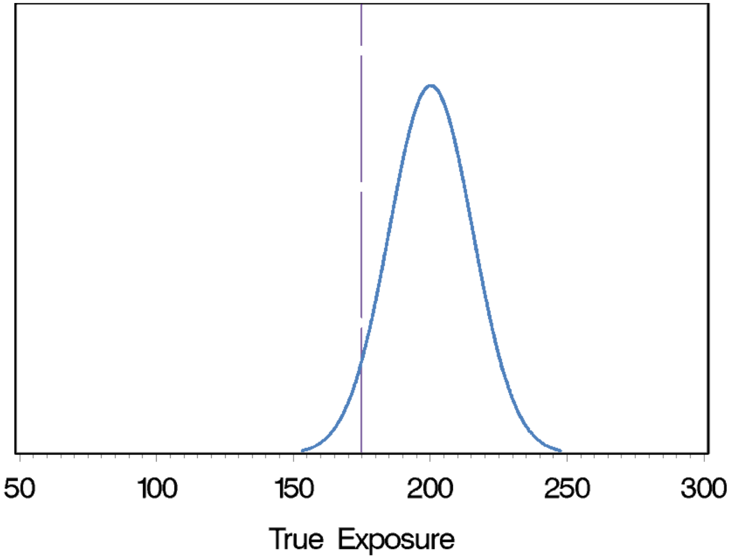
- Estimating diet-health relationships: regression coefficients

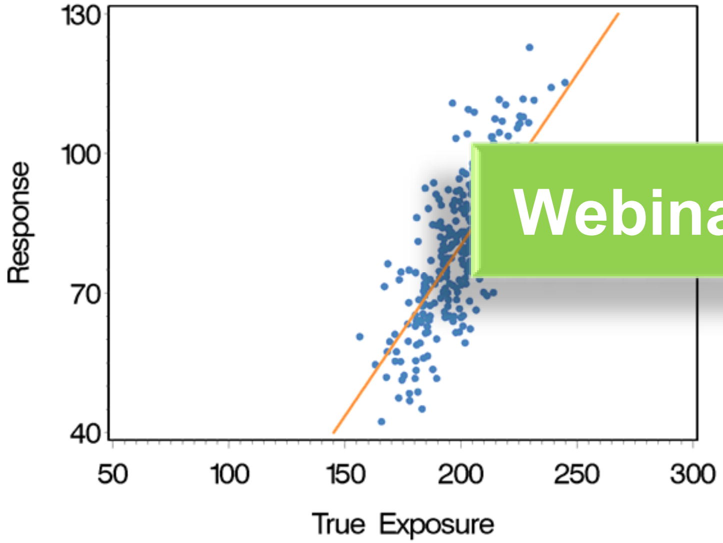
Slide 4

This slide probably looks familiar if you've watched the last two webinars. As has been mentioned, the webinar series will cover two main areas of interest: describing usual intake distributions, and estimating diet-health relationships. In the first case, we are interested in distributions and associated statistics, such as means, percentiles, and proportions above or below a threshold such as a nutrient requirement or food group recommendation. In studying diet-health relationships, we're interested in regression coefficients that describe the relationship between a dietary exposure and an outcome, such as an odds ratio, a relative risk, or a slope.

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 5

In this webinar, similar to the last webinar, I will be focusing on the first area—that of describing the usual intake of distributions—and future webinars will address diet-health relationships.

Estimating distributions of usual intake

- There is interest in monitoring a population’s usual intake of foods and nutrients
  - Informs research
  - Establishes population norms
  - Guides public policy

Slide 6

We are interested in estimating distributions of usual intake of foods and nutrients in many arenas, such as informing research, establishing population norms, and guiding public policy.

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

In the first webinar, Dr. Sharon Kirkpatrick introduced the concept that the regularity with which a dietary component is consumed among a population of interest is a key concept in modeling usual intake. We can think about two different types of dietary components, nutrients and foods that are consumed nearly daily by nearly all persons, such as vitamin C, total grains, and total fruits and vegetables, and those that are consumed episodically by most persons, such as nutrients that are concentrated in a few foods like vitamin A and food groups that are not commonly consumed every day by many individuals such as whole grains and dark green vegetables.

Daily versus episodic consumption

- Consumed nearly daily by nearly all persons
  - E.g., vitamin C, total grains, total vegetables, solid fats, added sugars

Webinar 2

- **Consumed episodically by most persons**
  - E.g., vitamin A, whole grains, dark green vegetables, fish

Slide 8

Unlike the last webinar, we will focus on episodically consumed dietary components today, which pose some unique challenges to statistical methods.

Learning objectives

- Define key concepts of food consumption related to usual intake estimation
- Identify challenges for estimating usual intake for episodically-consumed dietary constituents
- Explain statistical modeling approaches
- Apply NCI macros

Slide 9

The specific learning objectives for the webinar today are to define the key concepts of food consumption related to usual intake estimation, to identify the challenges for estimating usual intake, specifically for episodically consumed dietary constituents, and to explain the statistical modeling approaches that are used to overcome these challenges.

Finally, I will apply NCI macros to an example of estimating the distribution of an episodically consumed dietary constituent to illustrate their use.

Key concepts

Challenges

Statistical modeling

Example

Simulations

Conclusions

KEY CONCEPTS

Slide 10

[No notes.]

Key concepts

- Consumption patterns vary across dietary constituents
- Usual intake is comprised of probability to consume and consumption-day amount
- Dietary intake data are often skewed
- Current dietary assessment measures are prone to error

Slide 11

These are the key concepts that guide the estimation of usual intake for episodically consumed dietary components. First, as I’ve already alluded to, consumption patterns vary across dietary constituents. Second, usual intake is composed of the product of the probability of consumption and the consumption-day amount.

In addition, dietary data are often skewed, and, as was discussed in the first webinar, current dietary assessment measures are prone to error.

I will address each of these key concepts in more detail.

Key concepts

- **Consumption patterns vary across dietary constituents**
- Usual intake is comprised of probability to consume and consumption-day amount
- Dietary intake data are often skewed
- Current dietary assessment measures are prone to error

Slide 12

[No notes.]

Dietary constituents

- What makes up your diet
  - Foods
  - Food groups
  - Components of foods
    - Macronutrients
    - Micronutrients

Many are “episodically consumed”

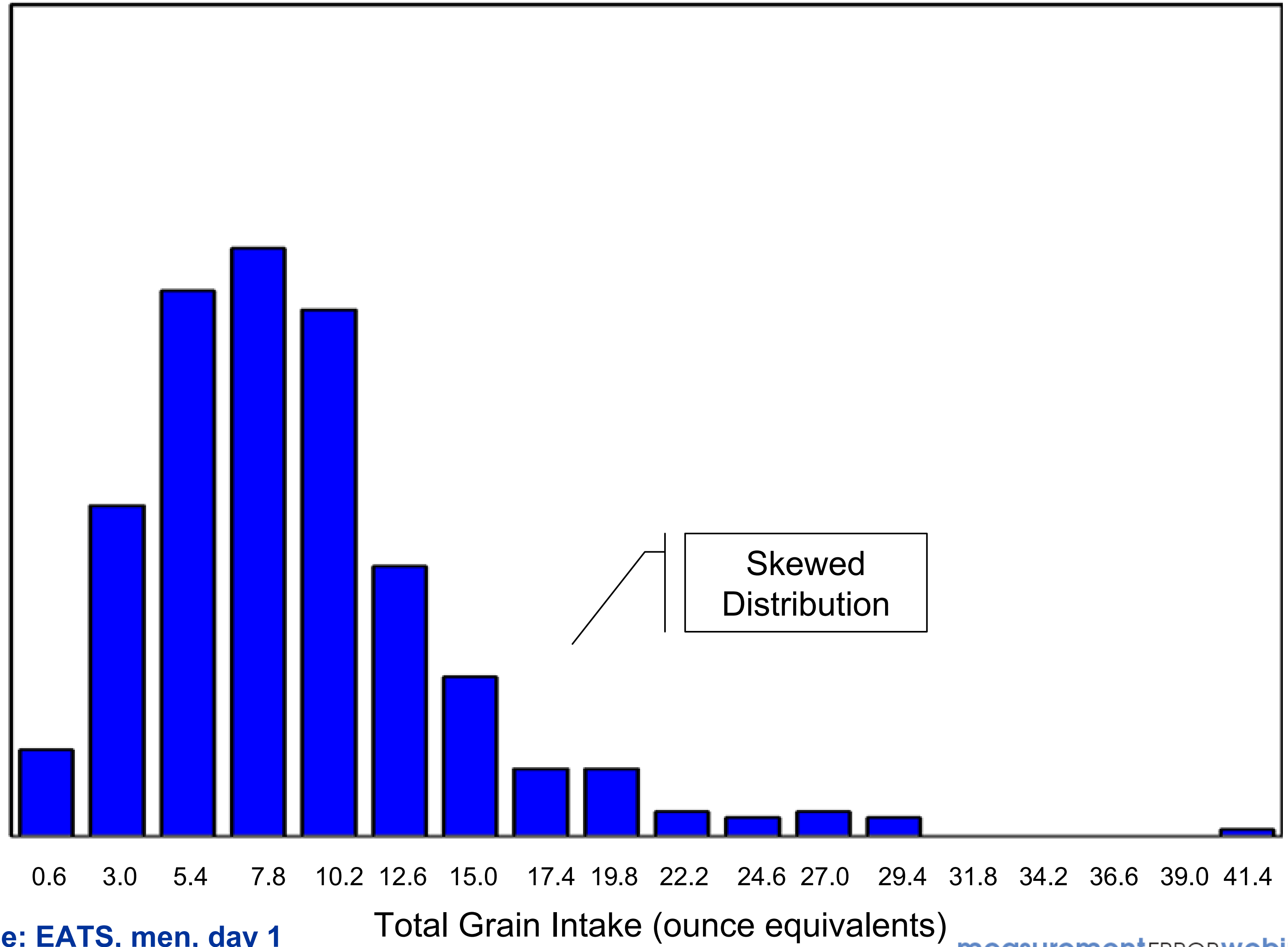
Most are consumed daily by most persons

Slide 13

Dietary constituents are simply the components of a diet, which may be described as individual foods, food groups, or components of foods such as macronutrients and micronutrients.

In general, foods and many food groups are episodically consumed; that is, they are consumed by many individuals on a less-than-daily basis. In contrast, many nutrients are consumed daily by most persons in the population. Of course, there are exceptions to these categorizations; for example, some food groups such as total grains and fruits and vegetables are consumed on a daily basis by most individuals, and some nutrients, such as vitamin A, are episodically consumed.

Daily consumed: Total grains

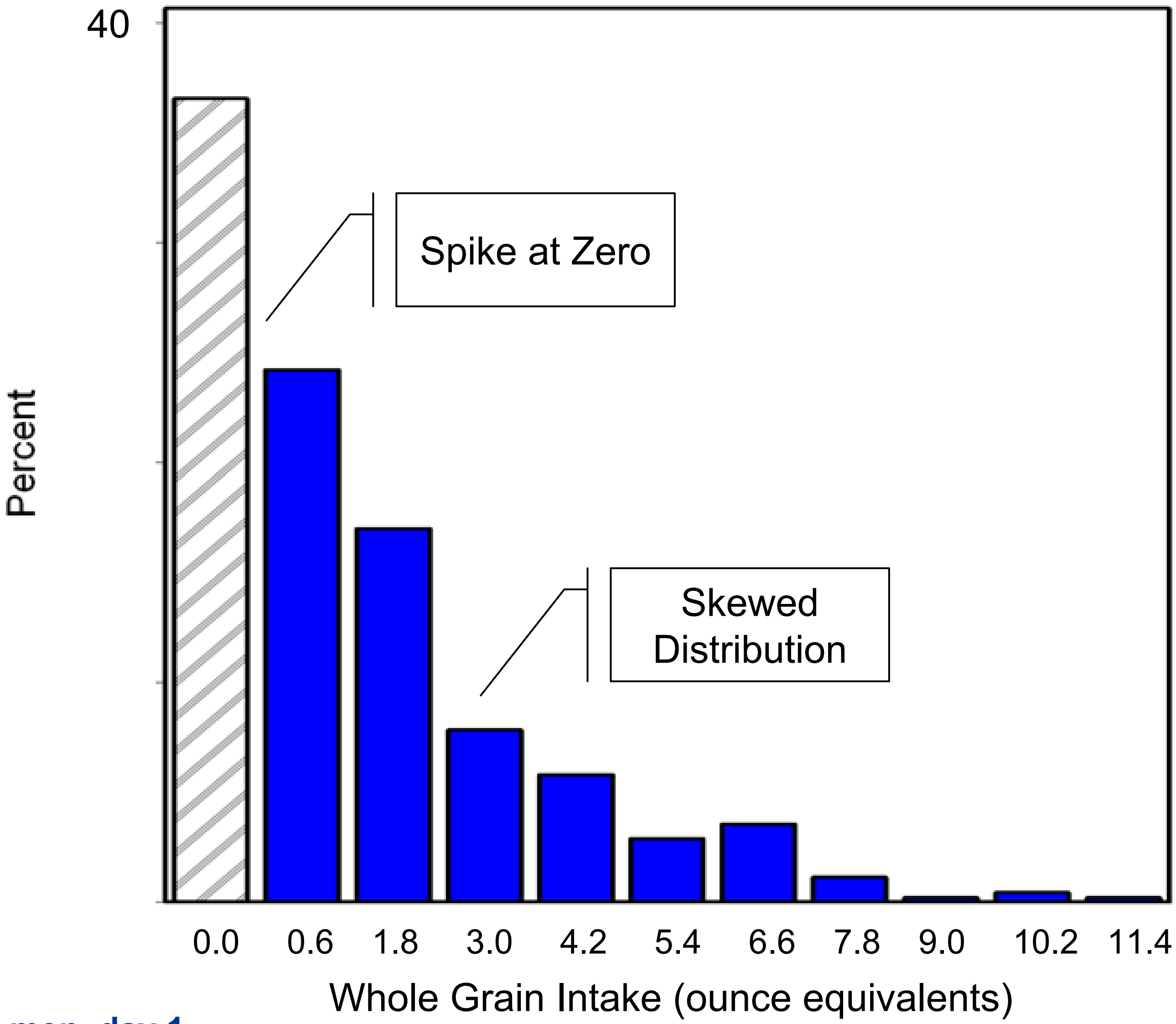


Source: EATS, men, day 1

Slide 14

Here is an example of a food group that is consumed daily by most persons, total grains. This plot is a histogram of total grain intake in ounce equivalents reported by men on one 24 hour recall. You can see here that all of the men reported some grain consumption, even if it was a small amount. We also see a common characteristic of dietary intake data, a distribution that is skewed to the right, with one man consuming over 41 ounce equivalents of total grains per day.

Episodically consumed: Whole grains



Spike at Zero

Skewed Distribution

Source: EATS, men, day 1

Slide 15

In contrast, here is the consumption of an episodically consumed food group for the same group of men on the same day of recall, whole grains. We see a large spike at zero, which I have indicated with a white striped column, with 36 percent of the men reporting no intake of whole grains for the day of 24 hour recall. Among those men who did report whole grain consumption, we also see a skewed distribution, with one man consuming over 11 ounce equivalents on the recall day.

Key concepts

- Consumption patterns vary across dietary constituents
- **Usual intake is comprised of probability to consume and consumption-day amount**
- Dietary intake data are often skewed
- Current dietary assessment measures are prone to error

Slide 16

[No notes.]

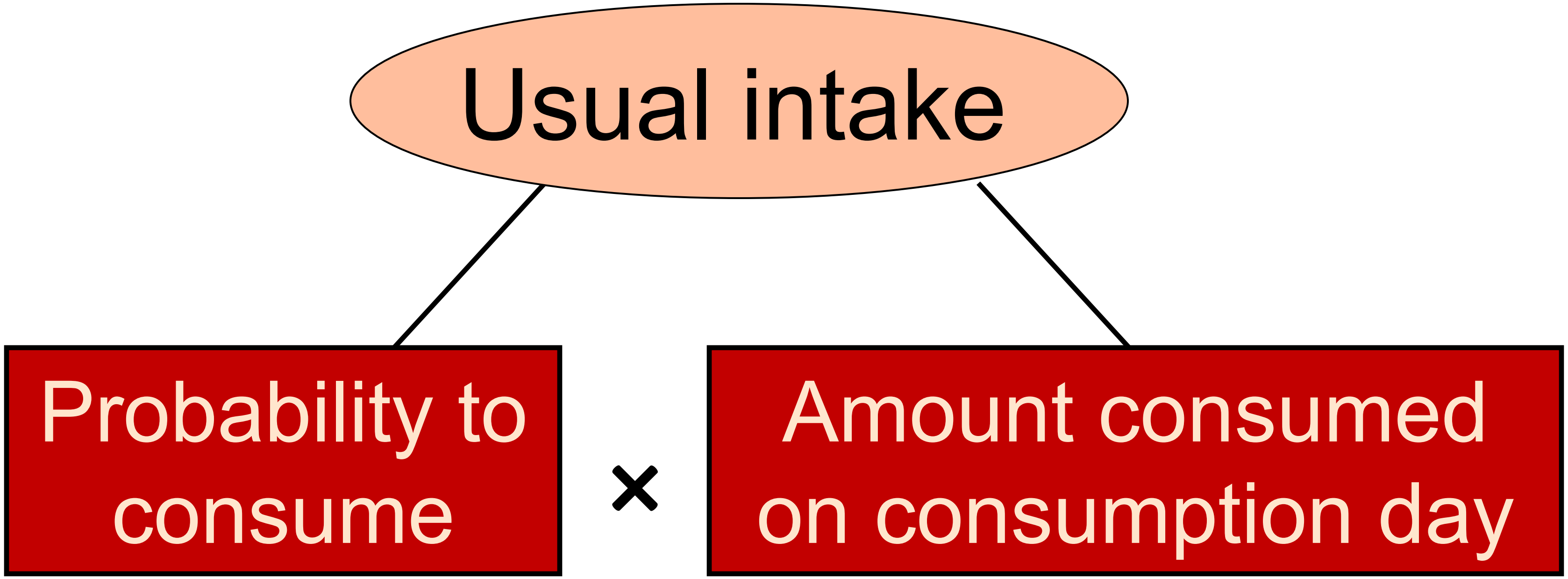
Usual intake

- Usual intake = long-term or habitual intake
- Nutrients are stored in the body
- Nutrient and food intake recommendations should be met over time, not necessarily every day

Slide 17

With dietary surveillance, as I mentioned previously, interest is on usual intake, which is a measure of long-run or habitual intake for an individual. Interest is generally on long-term intake, because nutrients can be stored in the body, and intakes vary from day to day. Therefore, it is usually unnecessary and impractical to achieve nutrient and food intake recommendations every day. There are some notable exceptions to this statement; for example, for alcohol consumption there may be interest in intake for a given day to define activities such as binge drinking, but we make this general assumption for almost all dietary recommendations.

Two parts of usual intake



Usual intake

Probability to consume

×

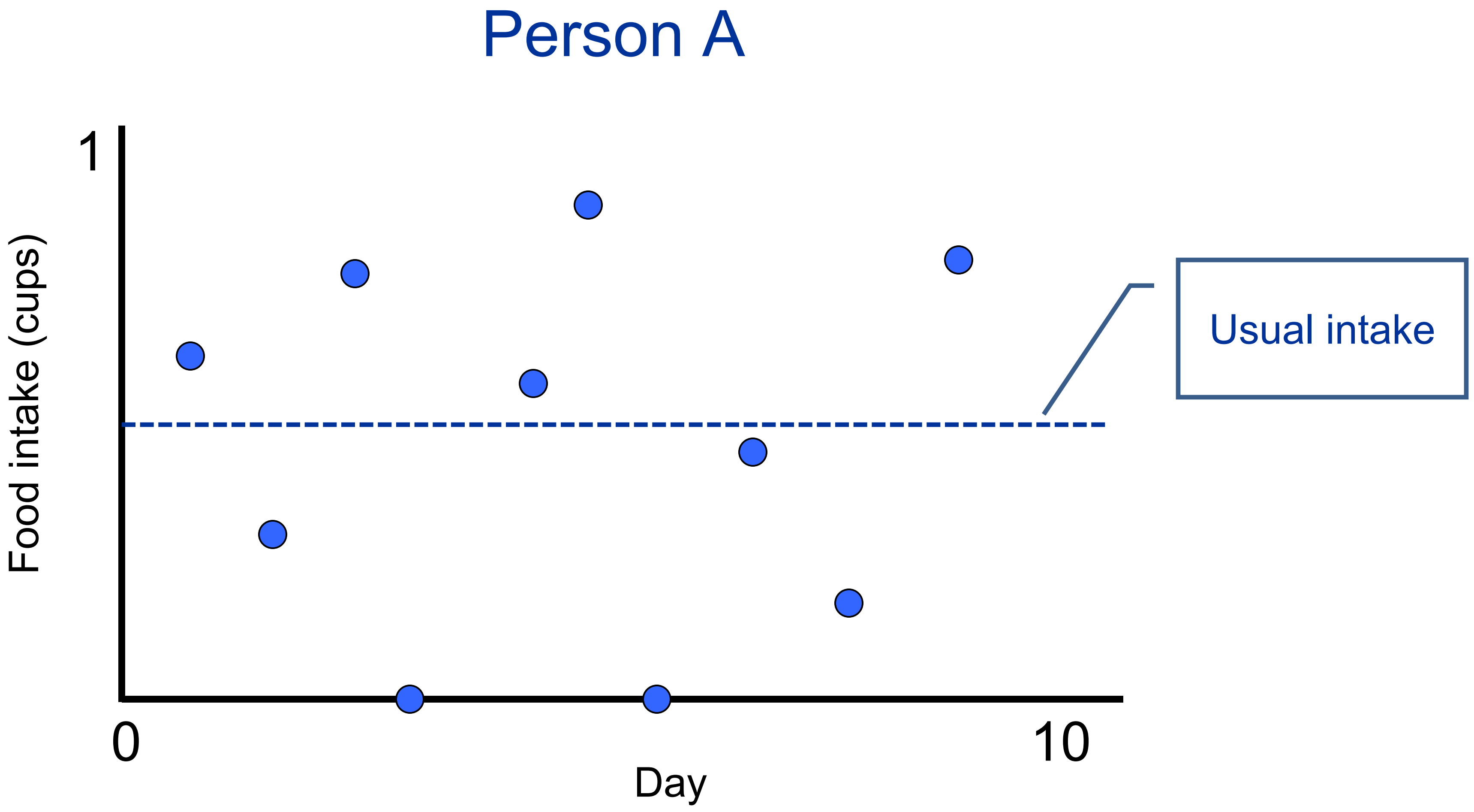
Amount consumed on consumption day

Slide 18

For describing the usual intake of episodically consumed dietary components, it is helpful to break down usual intake into two components. It can be seen that usual intake is the product of the probability of consumption and the consumption day amount.

When a dietary component is consumed every day, like total grains, the probability to consume is 1 or nearly 1 and, therefore, it is only necessary to model the consumption-day amount. However, when foods are not consumed daily, it is useful to separate probability and consumption-day amount.

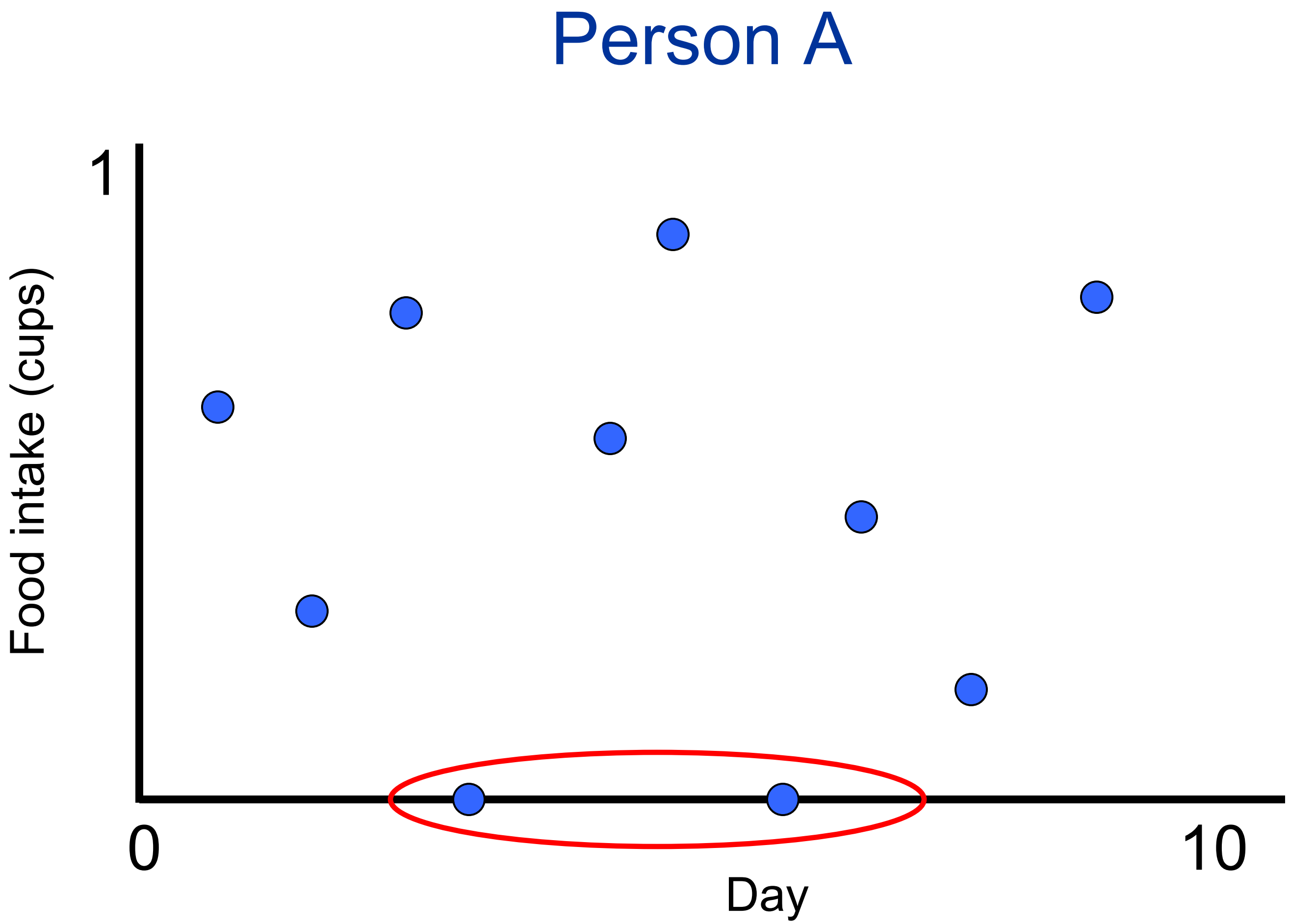
Usual intake for a given individual



Slide 19

This plot illustrates usual intake of a food over time for a hypothetical person over 10 days. The horizontal axis represents time in days, and the vertical axis represents food intake in cups. The dots represent intake on one day. The dashed line represents usual intake, about ½ cup.

True probability of consumption

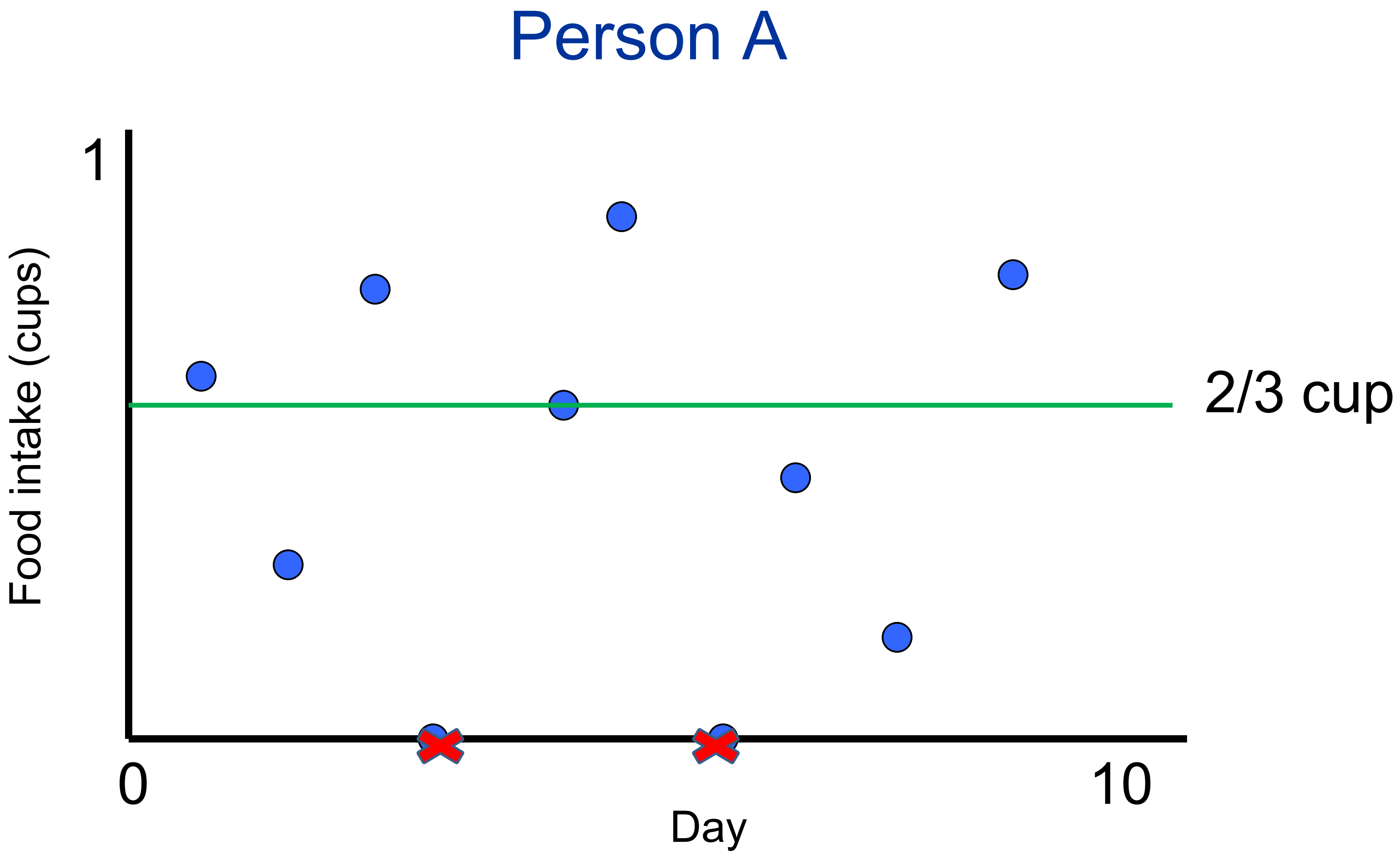


8/10 = 80%

Slide 20

You can see that person A did not consume any of the food on the fourth or seventh day, consuming the food on only 80 percent of the days, so the probability of consumption was 80 percent.

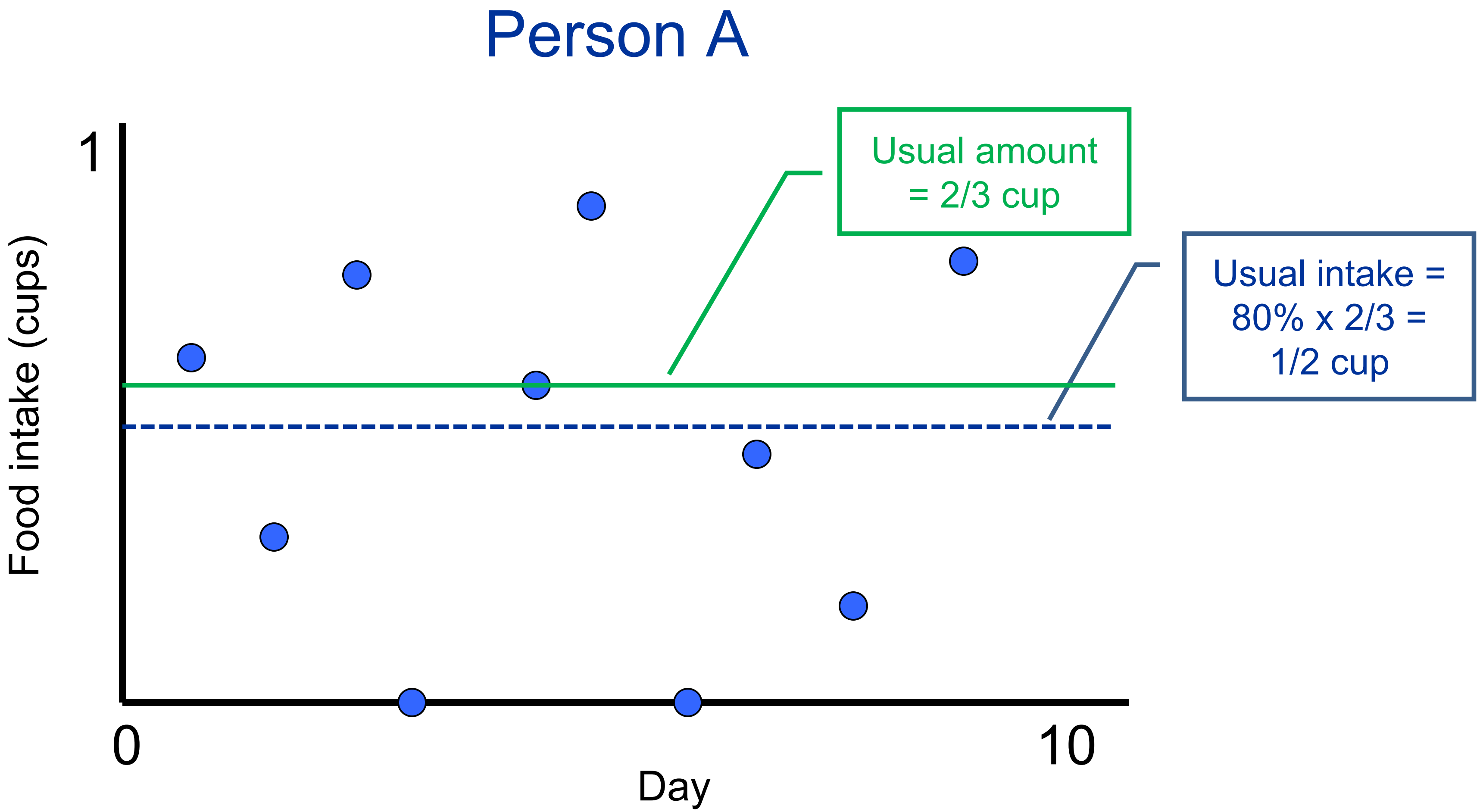
True consumption-day amount



Slide 21

If we ignore the two days where the food was not consumed, we can see by the green solid line that the average consumption-day amount is about ⅔ of a cup.

Probability x consumption-day amount



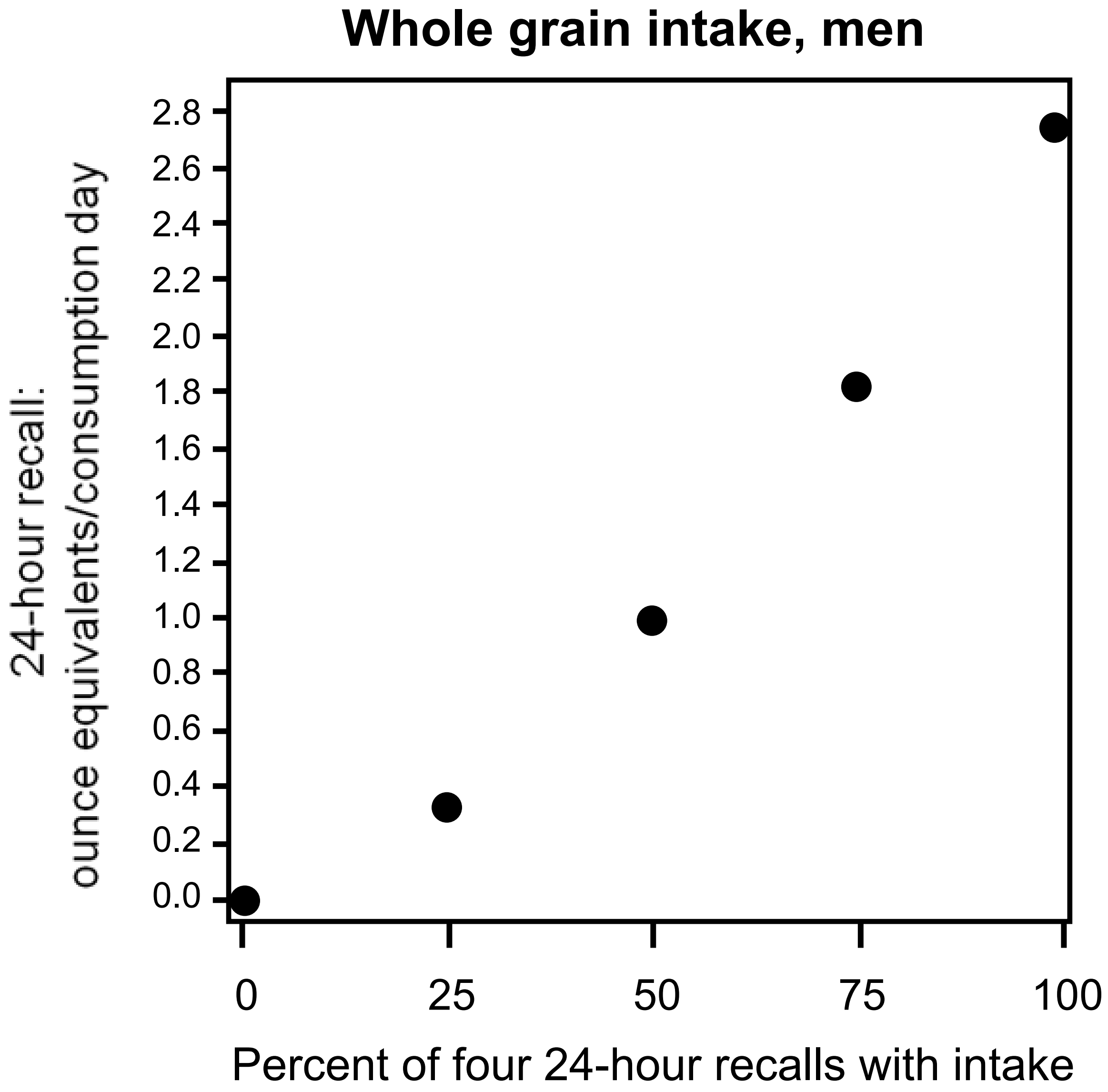
Usual amount = 2/3 cup

Usual intake = 80% x 2/3 = 1/2 cup

Slide 22

Finally, we calculate the usual intake by multiplying ⅔ cup by 80 percent, giving us the usual intake of ½ cup.

Correlation between probability and amount



Source: EATS

Slide 23

It is also important to note that we often see correlation between probability and amount. This is illustrated in this plot, which is of 24 hour recall whole grain intake data for men from the Eating at America's Table Study, which collected four 24 hour recalls. Along the vertical axis is 24 hour recall reported ounce equivalent consumption, and along the horizontal axis is the percent of the four 24 hour recalls that have reported intake. As you can see, there is a positive correlation between the proportion of days that whole grains are consumed and the amount eaten on the consumption occasion. This is not surprising; it just means that men who eat whole grains more frequently tend to eat more of them, perhaps because they like them. We see this for about 80 percent of foods. There are a few food groups that don't exhibit this correlation, such as dark green vegetables; people tend to consume the same serving size no matter often they eat it.

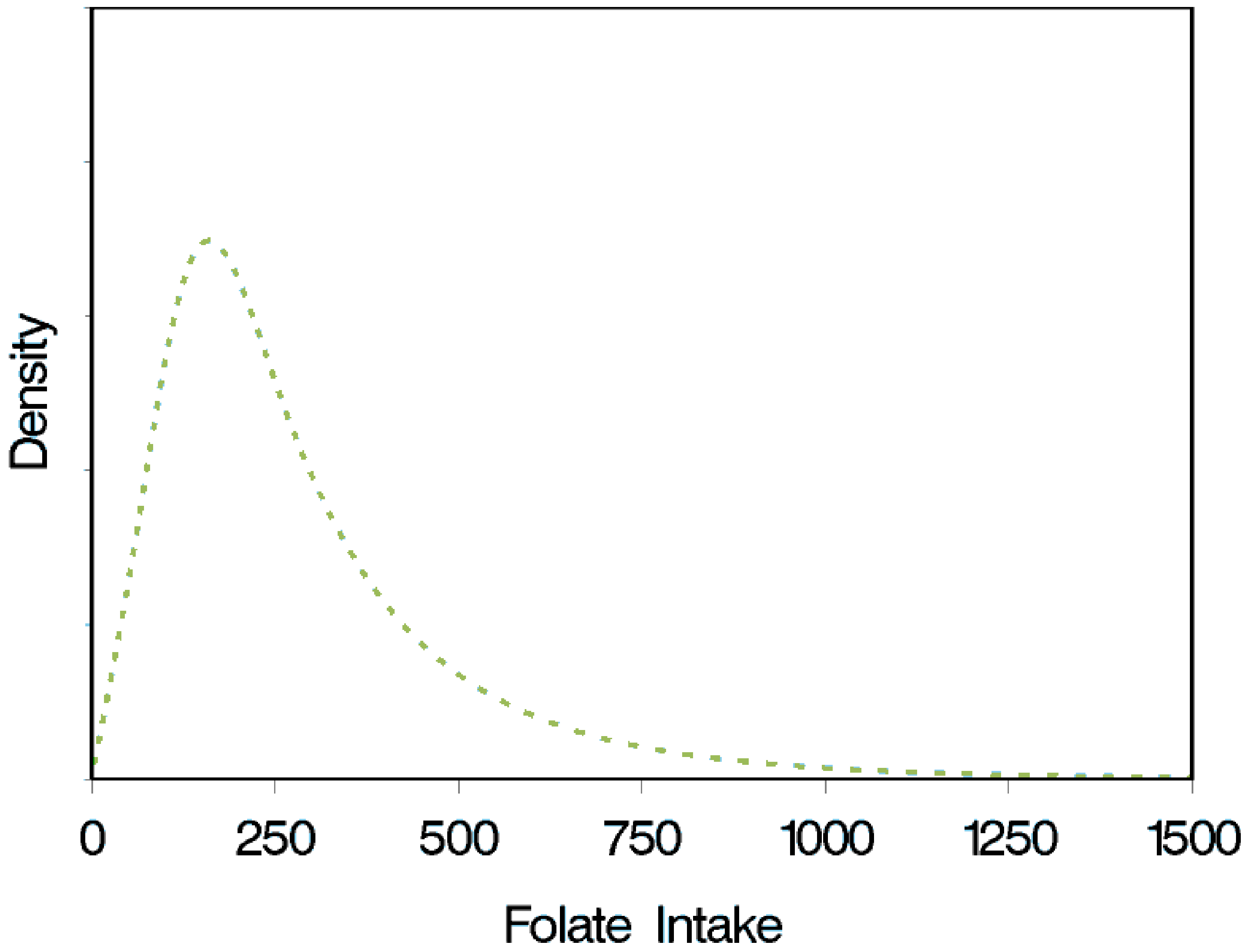
Key concepts

- Consumption patterns vary across dietary constituents
- Usual intake is comprised of probability to consume and consumption-day amount
- **Dietary intake data are often skewed**
- Current dietary assessment measures are prone to error

Slide 24

As was discussed in webinar 2, and as we saw for both total and whole grains, dietary intake data are often skewed.

Skewness



Slide 25

This is an example of folate intake, showing some values up to 1,500 micrograms per day. These right-skewed data are common in studies of dietary intake, even for episodically consumed foods and nutrients.

Key concepts

- Consumption patterns vary across dietary constituents
- Usual intake is comprised of probability to consume and consumption-day amount
- Dietary intake data are often skewed
- **Current dietary assessment measures are prone to error**

Slide 26

And, as was addressed in detail in webinar 1, current dietary assessment measures are prone to error. In the next few slides, I will briefly summarize some of the slides from webinar 1 that describe measurement error in FFQs and 24HRs.

Within-person error: Summary of Webinar 1

- Day-to-day variation
- Random error in reporting

Random

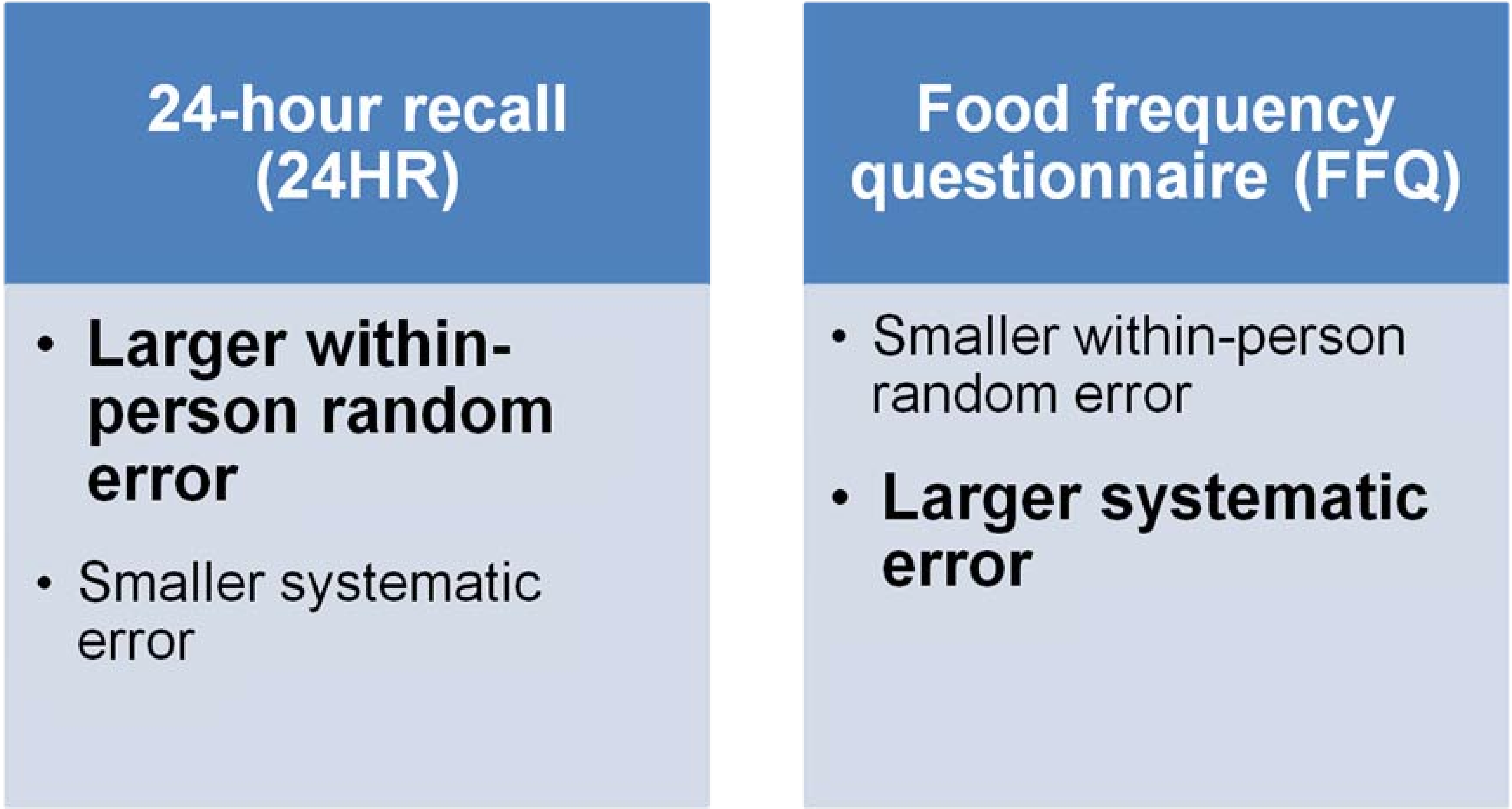
- Additive error
- Intake-related bias
- Person-specific bias

Systematic

Slide 27

There are roughly five sources of within-person error: day-to-day variation and random error in reporting, which are random and cannot be distinguished, and we simply refer to as random within-person error; and additive error; intake-related bias; and person-specific bias, which are systematic and lead to bias in estimation. Recall that the systematic error can shift both the mean and variance of the distribution of usual intake.

OPEN findings: Structure of measurement error



24-hour recall (24HR)

- **Larger within-person random error**
- Smaller systematic error

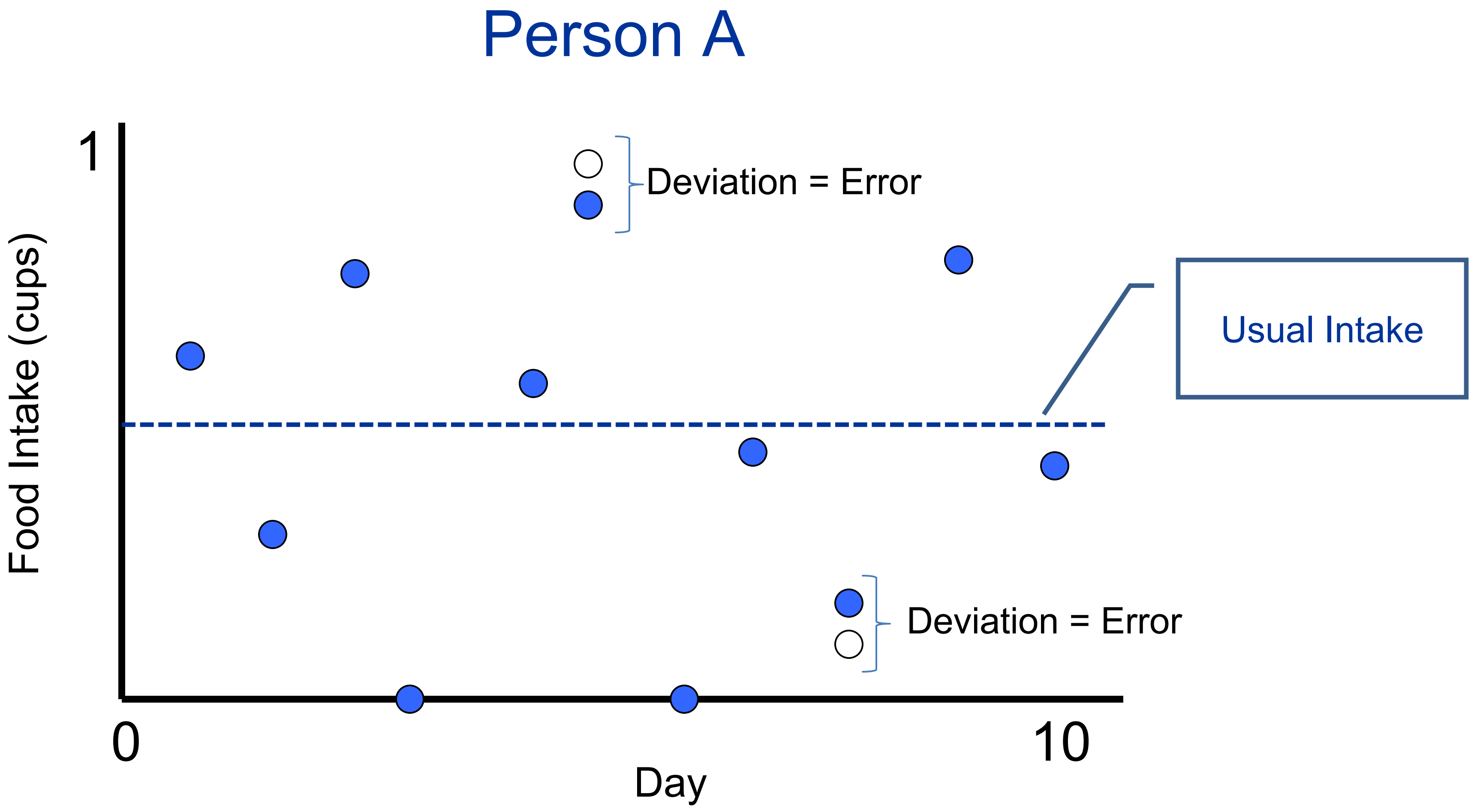
Food frequency questionnaire (FFQ)

- Smaller within-person random error
- **Larger systematic error**

Slide 28

The findings of the OPEN Study, which Sharon discussed in webinar 1, suggest that 24-hour recalls have larger within-person random error than FFQs but smaller systematic error. The random error in the 24-hour recall is driven by day-to-day variation in intake and other random errors that affect reporting from day to day. The error in the FFQ, on the other hand, is driven by inaccuracies associated with the task of recalling long-term intake as well as features of the instrument such as the finite food list and the relative lack of detail about foods consumed.

Reported intake



Slide 29

So, what instrument do we choose? Well, the 24 hour recall is the instrument most commonly used in dietary surveillance, and is the instrument I will refer to in the remainder of the webinar. I added two dots to this earlier plot, representing that often only two days of 24 hour recall are available, and that these can be reported with random error. This is an important assumption—that the 24 hour recall is prone to random but not systematic error. With only two days of 24HR, we are able to estimate usual intake by accounting for this random within-person error by using statistical methods.

Unbiasedness: 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 30

This assumption of unbiasedness is a working assumption for the 24HR. As Sharon discussed in webinar 1, the critical assumption of unbiasedness does not hold in practice for the two biomarkers studied in the OPEN study, energy and protein. However, in the OPEN study, the 24HR was subject to less systematic bias than the FFQ. Unfortunately, we don’t have other recovery biomarkers to know how well it works for other dietary components. Essentially, the 24HR is one of the best dietary assessment tools that we have, and we proceed under the working assumption that 24HRs are unbiased for usual intake.

Key concepts

Challenges

Statistical modeling

Example

Simulations

Conclusions

CHALLENGES TO ASSESSING USUAL INTAKE

Slide 31

So, now that we’ve reviewed some key concepts, let’s move on to challenges to assessing usual intake for episodically consumed foods and nutrients.

Challenges of modeling episodically consumed constituents

- Account for measurement error
- Account for skewness
- Model probability and amount
- Allow for correlation between probability and amount
- Incorporate covariates

Slide 32

In the previous few slides and in the earlier webinars, we have seen that dietary data are prone to measurement error, and tend to be skewed. Therefore, our models must accommodate these challenges. In addition, with episodically consumed dietary constituents, we have to model consumption patterns; that is, we need to model both parts of usual intake—probability of consumption and the consumption-day amount, as well as the correlation exhibited by these two components of usual intake. Finally, there is often interest in incorporating covariates, and there may be interest in incorporating different sets of covariates into the model for probability and amount. So, I'm going to briefly discuss each of these challenges in a bit more detail in this section, and then I'll get into the details of the statistical modeling in the next section.

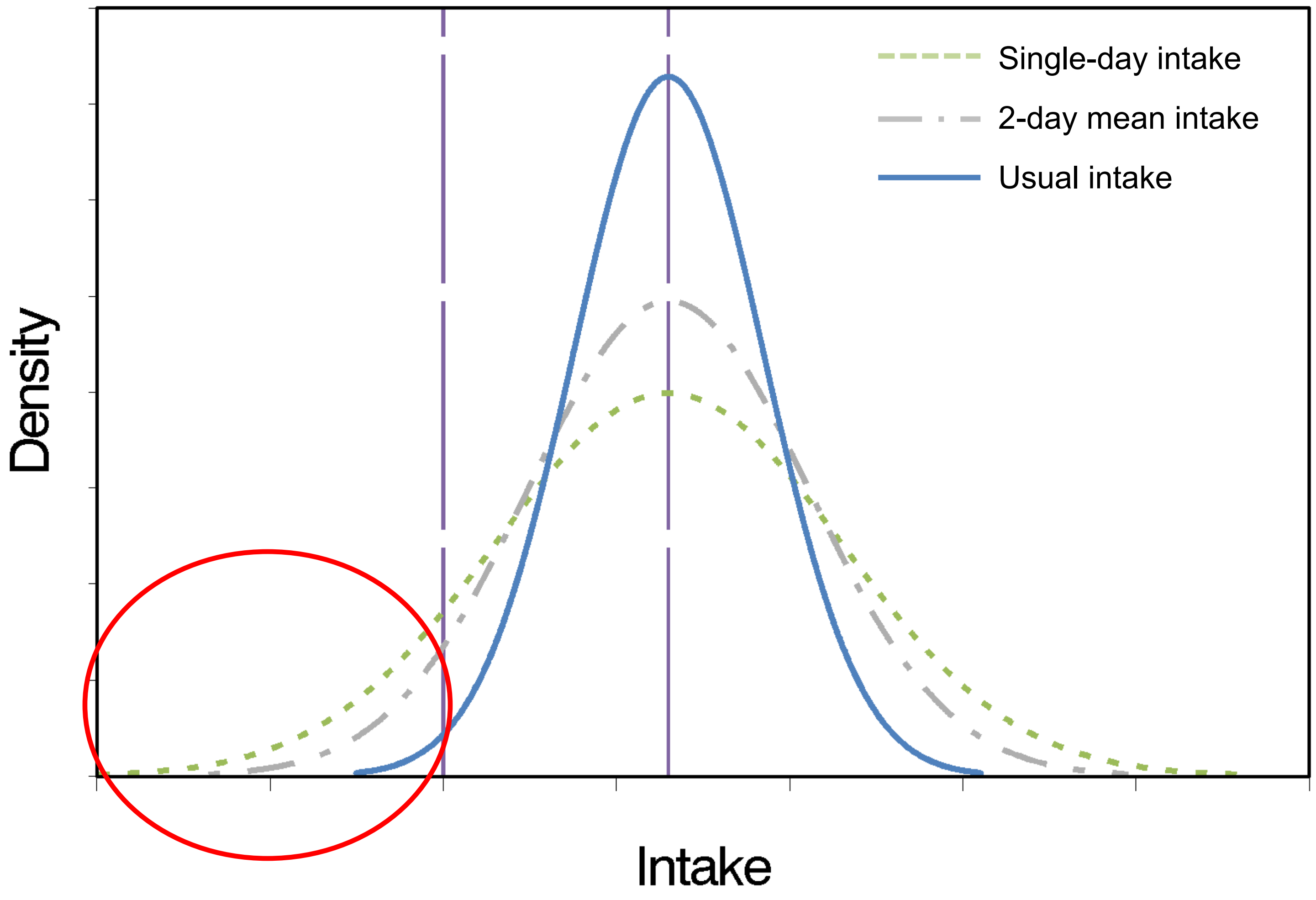
Challenges of modeling episodically consumed constituents

- **Account for measurement error**
- Account for skewness
- Model probability and amount
- Allow for correlation between probability and amount
- Incorporate covariates

Slide 33

[No notes.]

Effect of within-person variation



Slide 34

Let’s discuss the impact of random error on the estimation of intake distributions, where we typically rely on 24 hour recall data. Here, we are looking at a hypothetical distribution of intake. The widest distribution, shown by the greenish dashed line, shows intake based on a single day; the gray line shows intake based on two days; and the solid blue line shows the actual usual intake. The distribution of intake based on one day is wider and flatter compared with the usual intake distribution. The graph also shows that averaging over two days may help somewhat to alleviate the effects of within-person error, but certainly is not sufficient to account for that error. An important implication here is that if we do not account for measurement error, it will result in overestimates of inadequate or excess intake. This is illustrated by the vertical lavender line on the left side of the slide. If this represented a hypothetical threshold, we can see that both the one day of recall and the two-day mean overestimate the proportion of the population in the tail-animation-, which suggests that simple analysis methods based on averaging are not generally satisfactory.

Methods of correcting for random measurement error

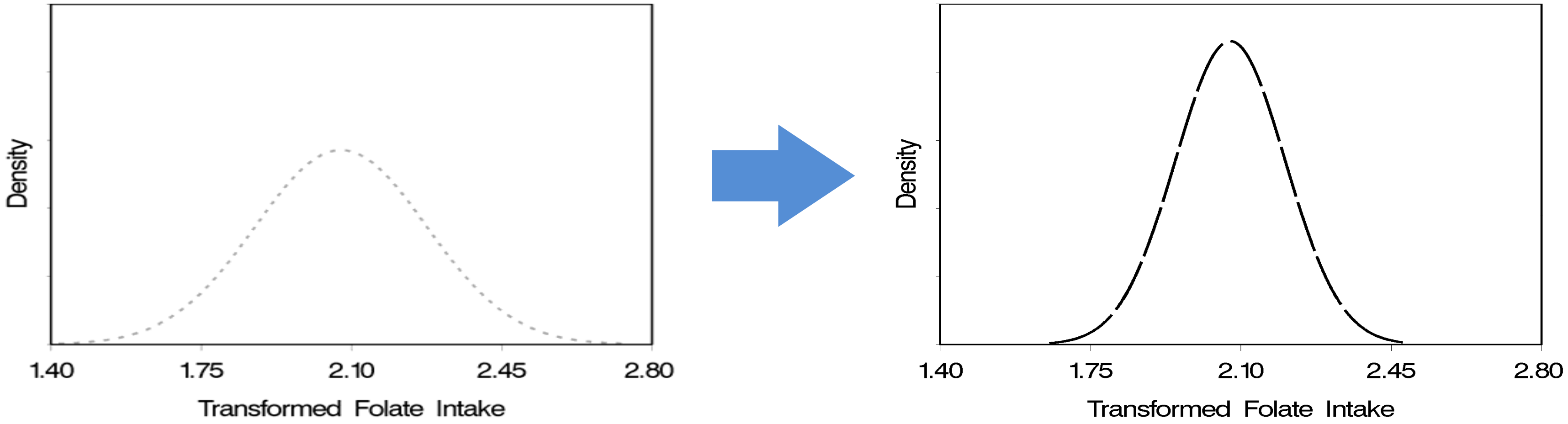
- Estimating usual intake distributions using short-term instruments – some existing methods:
  - Iowa State University Foods (ISUF) Method
  - National Cancer Institute (NCI) Method
  - EFCOVAL Consortium Multiple Source Method (MSM)
  - Statistical Program for Age-adjusted Dietary Assessment (SPADE)

Slide 35

A number of methods have been proposed to estimate usual intake distributions for episodically consumed foods using short-term instruments like the 24 hour recall, including those on this slide, which I have listed in chronological order.

General approach to estimating usual intake

- Separate between-person from within-person variation
  - Assume normality
- Estimate distribution with only between-person variation



Slide 36

The general approach used by these methods is to separate the within- and between-person variation and to remove the within-person variation using statistical modeling. This is done using a normality assumption because the normal distribution has a number of nice properties, one of which is that it is defined by two parameters, the mean and a constant variance, as was discussed by Dr. Kevin Dodd in webinar 2. And, as we've said, this modeling makes an important assumption—that the 24HR is only subject to random error. So, to explain this graphically, our goal is to go from the wider distribution on the left to the narrower one on the right, which does not exhibit within-person variation.

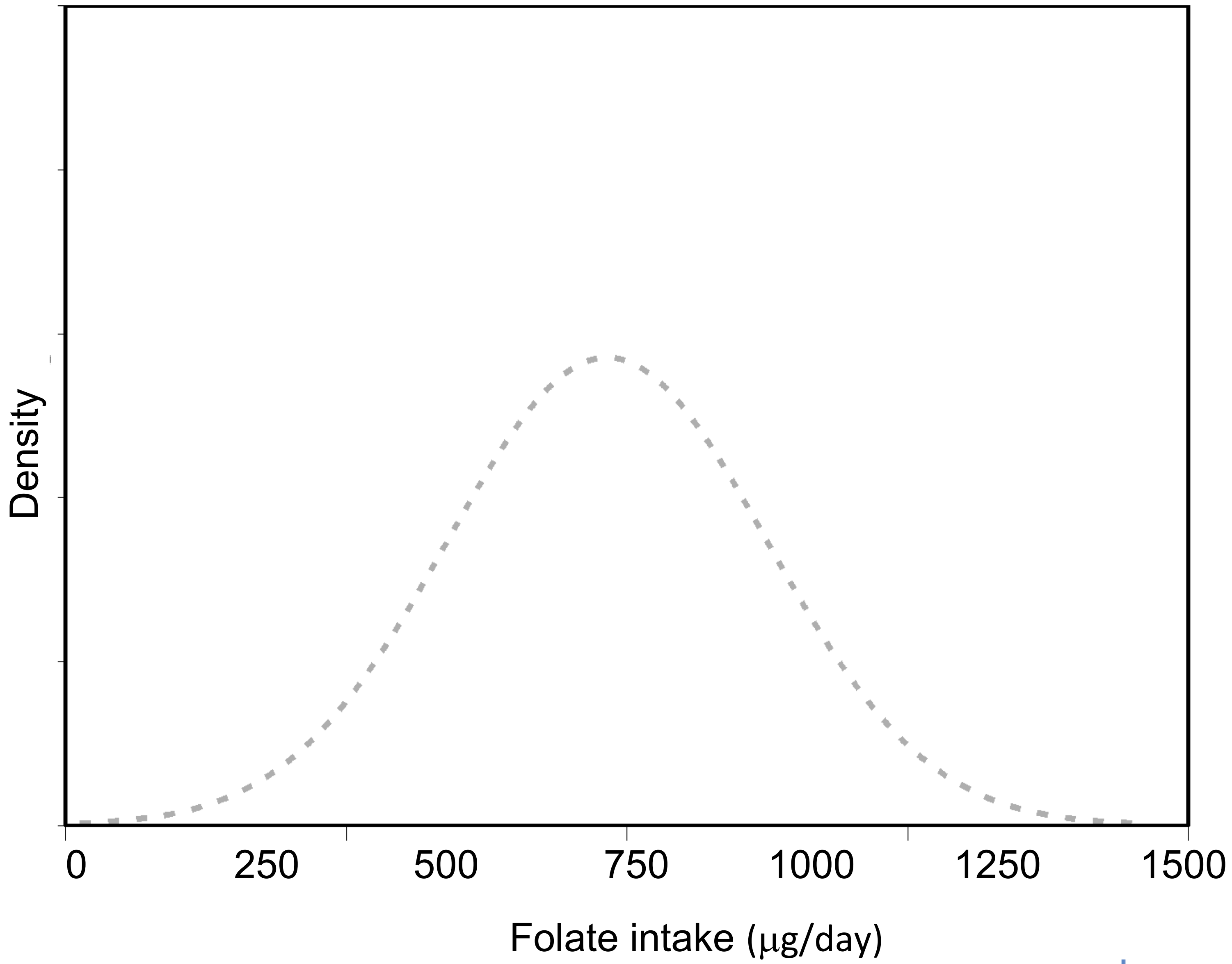
Challenges of modeling episodically consumed constituents

- Account for measurement error
- **Account for skewness**
- Model probability and amount
- Allow for correlation between probability and amount
- Incorporate covariates

Slide 37

[No notes.]

Skewness



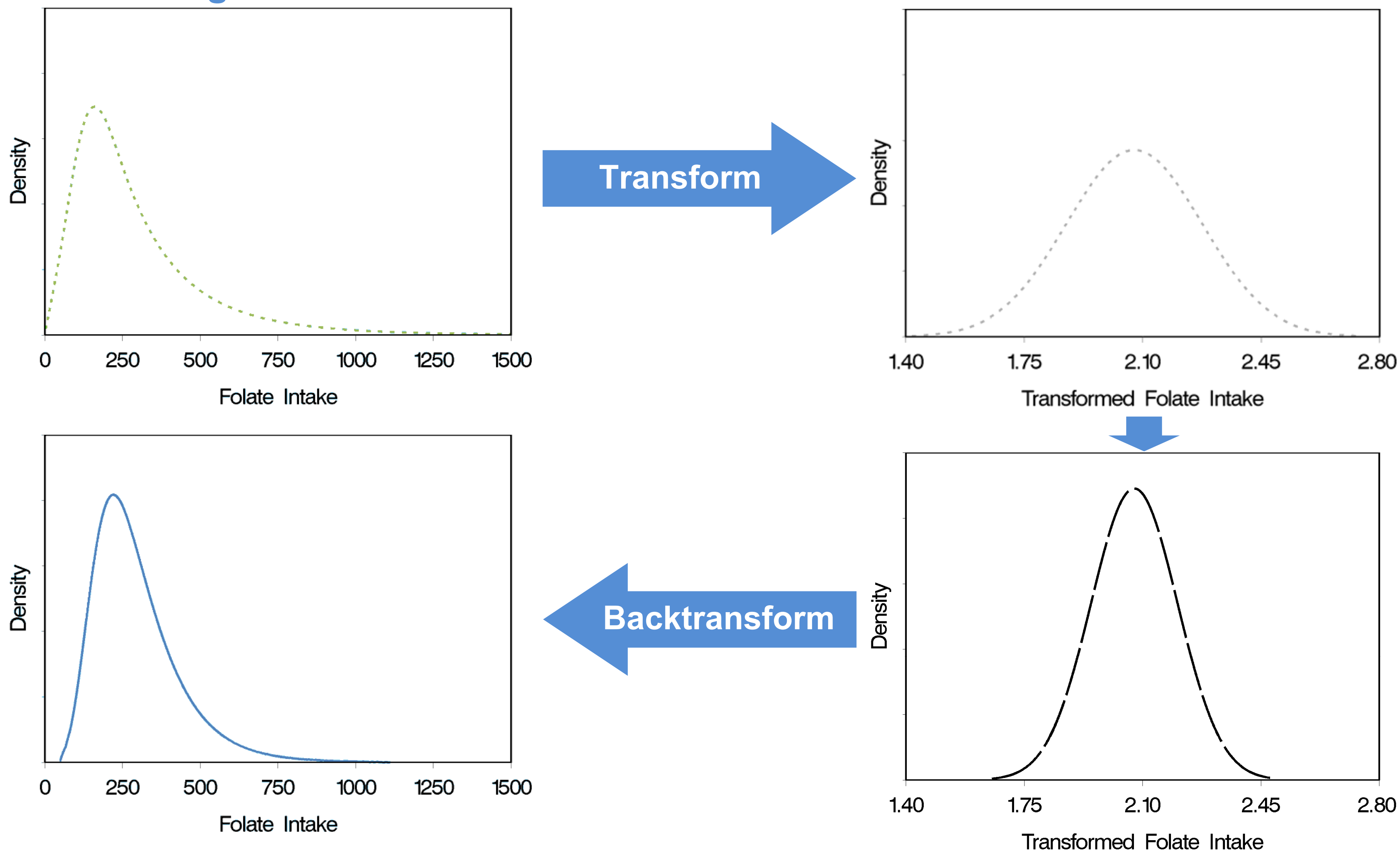
Slide 38

You may have noticed on the previous slide the nice bell curve shape of the distributions. Because when we separate between-person from within-person variation it's very helpful to assume that the data are normally distributed, we'd like our raw data to look like this as we've seen in the previous two slides-animation- but, unfortunately, they usually look like this.

Accounting for nonlinear transformations

Original Scale

Transformed Scale

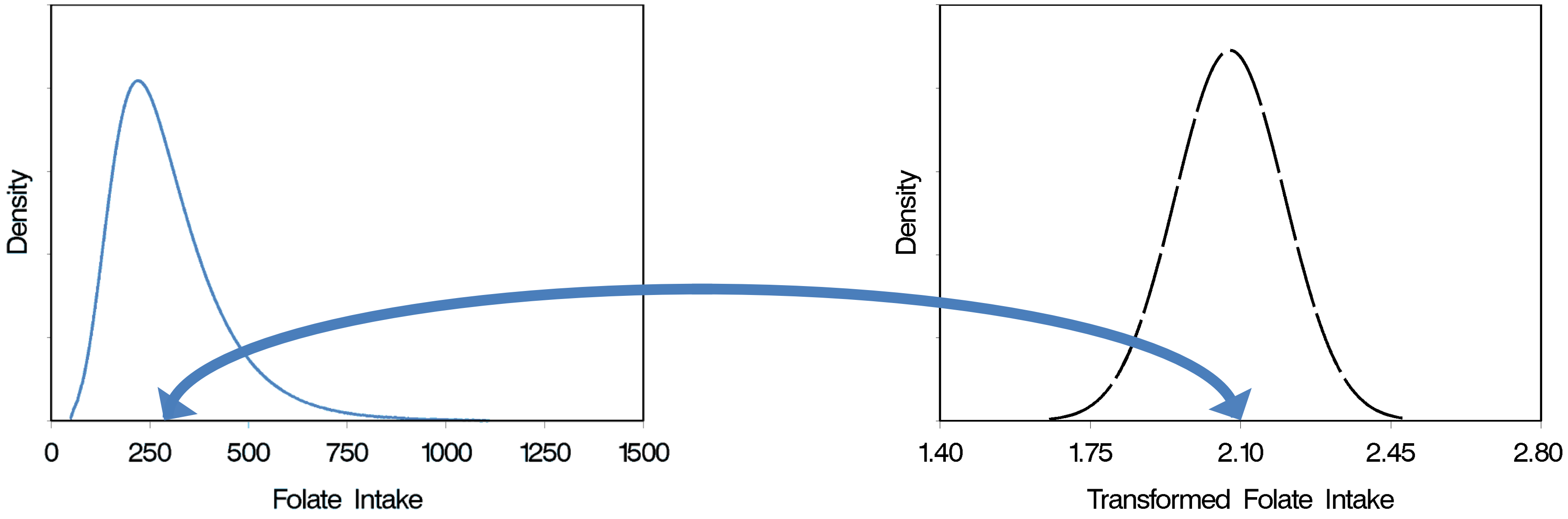


Slide 39

There is a way to deal with skewness, however. You may remember from a statistics class that when data are skewed to the right, we apply a power transformation to pull them in. This is exactly what we do in the modeling. We start with the 24HR data in the original scale and apply this type of transformation that is similar to a square root or log transformation to obtain a normal distribution. Now, the within-person variation still exists in the transformed scale, but it is possible to apply a statistical model to remove its effect and get the more narrow distribution that exhibits only between-person variation. Because we used a transformation but would like usual intake on the original scale, we have to make one last step to translate back to the original scale. All of the methods I mentioned use this general framework for transformation, although the types of transformations vary.

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 40

When we do these backtransformations, our goal is to estimate a quantile on the original scale that corresponds to one in the normal distribution. That is, we want the median and other quantiles on the transformed scale to be mapped back to the quantiles on the original scale. We also want the mean on the original scale to match the mean on the transformed scale.

Backtransformation

- Mean of transformed data ≠ transformation of mean on the original scale
- With nonlinear transformation is used, the estimated quantile is an integral that can be calculated/ approximated in several ways
  - Taylor series approximation
  - Numerical integration for known distribution
    - Quadrature formulas, e.g., Gauss-Hermite
    - Monte Carlo integration

Slide 41

However, as Kevin discussed in webinar 2, taking the mean of transformed data is not the same as transforming the mean when the transformation is nonlinear. So, we have to find some way of approximating the estimated quantile in the backtransformation step. As Kevin mentioned, the quantile is an integral that can be approximated in a couple of ways—using an approximation such as the Taylor series approximation, or by numerical integration using quadrature formulas, or Monte Carlo integration. The basic idea behind quadrature is to approximate the integral by breaking it into smaller pieces and to sum the value of the integrand over a series of points using specific weighting coefficients for each point. The trapezoidal rule is a simple case of this method with which you may be familiar. In contrast to quadrature, Monte Carlo integration uses random points to approximate the integral. Whatever method we use, it is an important step to ensure that the data are backtransformed to the original scale, so that the mean on the original scale is equivalent to the mean after the backtransformation.

Challenges of modeling episodically consumed constituents

- Account for measurement error
- Account for skewness
- **Model probability and amount**
- Allow for correlation between probability and amount
- Incorporate covariates

Slide 42

[No notes.]

Model probability and amount: Two-part model

- For episodically-consumed dietary constituents, we fit two statistical models:
  - Probability
    - Mixed model logistic regression
  - Amount
    - Mixed model linear regression

Slide 43

Previously, I mentioned that it is helpful to divide usual intake into two parts, probability of consumption, and the consumption-day amount. All of the methods that I mentioned estimate these two parts of usual intake, although they differ in the way in which this is done. In the NCI method, in order to model episodically consumed foods, we fit a two-part model, where the first part is a mixed-model logistic regression and the second part is a mixed-model linear regression on a transformed scale. I use the term here “mixed model” to describe a model that has both fixed effects, as in usual regression, and random, or person-specific effects. The “mix” of the fixed and random effects gives it the name “mixed model.”

Two-part model: Person-specific effects

- Also known as random effects
- Latent
- Constant for an individual
- Captures how an individual’s value deviates from the average after adjusting for covariates, if appropriate
- Both the probability and amount models incorporate person-specific effects

Slide 44

These person-specific or random effects are latent; that is, we can’t directly measure them. They are constant for an individual, and they capture how an individual’s value deviates from the average, after adjusting for covariates if we have them in the model. We can think of them as a person’s tendency to consume a certain food or nutrient, and the variability of the person-specific effect, therefore, captures the variability in the population. We incorporate person-specific effects in both parts of the model, although each part of the model has a different effect.

Challenges of modeling episodically consumed constituents

- Account for measurement error
- Account for skewness
- Model probability and amount
- **Allow for correlation between probability and amount**
- Incorporate covariates

Slide 45

[No notes.]

Modeling correlation

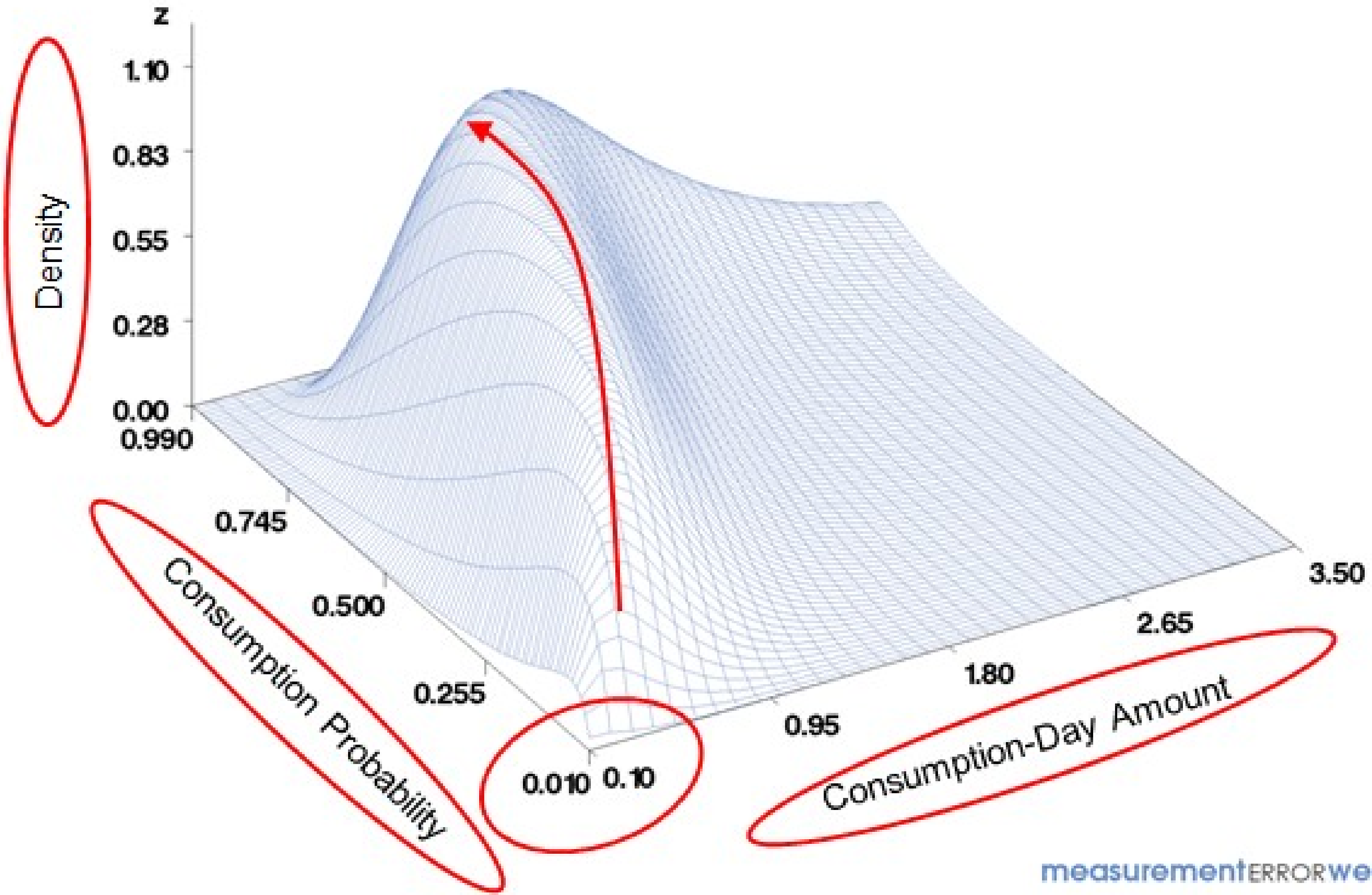
- Model probability and amount simultaneously
- Correlation between person-specific effects
  - Probability of consumption and consumption day amount
- Covariates

Slide 46

As you may recall, most foods exhibit correlation between the probability of consumption and the consumption-day amount. Therefore, it is important that a statistical method for estimating usual intake be able to accommodate this correlation. In the NCI method, this is done by allowing the person-specific effects to be correlated. With correlated effects, we must model probability and amount simultaneously. In addition, the two parts of the model may share the same covariates, although the covariates may also be different, if desired.

Estimating distributions

Joint Distribution of Probability and Amount



measurementERRORwebinar series

Slide 47

This 3-dimensional plot shows the joint distribution of probability and amount with positive correlation that we often see. Along the axis to the right side of the graph, we have the consumption-day amount; on the axis to the left of the slide, we see the consumption probability; the vertical axis represents the distribution density. Notice that the lowest points of the graph are at the point closest to the bottom, so that as probability increases, consumption-day amount also increases, so the lower probability is associated with lower amounts and the higher probability is associated with the higher amounts. This is what we are modeling with the correlated two-part model.

Challenges of modeling episodically consumed constituents

- Account for measurement error
- Account for skewness
- Model probability and amount
- Allow for correlation between probability and amount
- **Incorporate covariates**

Slide 48

[No notes.]

Types of covariates

- **Individual-specific**
  - Affect true intake on all days
  - e.g., gender/age/race-ethnicity
- **Time-dependent**
  - Affect true intake on specific days
  - e.g., season/weekday
- **Nuisance**
  - Affect reporting error
  - e.g., interview sequence/mode effects

Slide 49

In addition, there is often interest in incorporating covariates into statistical modeling. For example, we may want to correct for individual-specific effects that affect true intake, such as gender, age, or race and ethnicity. One reason we may want to include these types of covariates is to make different estimates of the distribution of intake for different subpopulations in an efficient manner.

We may also be interested in adjusting for time-dependent covariates that affect true intake on certain days such as season or weekend vs weekday. Finally, we may want to adjust for nuisance variables that are related to reporting error. These can include interview sequence effects, with the first interview usually reporting higher intake, and mode effects, such as in-person vs. telephone administration.

Key concepts → Challenges → Statistical modeling → Example → Simulations → Conclusions

STATISTICAL MODELING: NCI METHOD

Slide 50

I’ve already briefly discussed the NCI method. Now, I will discuss it in more detail in this section.

NCI Method: Overview

Two-part model: Episodically-consumed constituents

- Part 1: Probability
  - Mixed model logistic regression
    - Can incorporate covariates

Slide 51

This is an overview of the model used in the NCI method, which was developed by me and other members of the Surveillance Measurement Error Working Group.

Part 1 is a mixed-model logistic regression to model probability that may incorporate covariates.

NCI Method: Overview

Two-part model: Episodically-consumed constituents

- Part 2: Amount
  - Mixed model linear regression
    - Transformed scale – accounts for skewness
    - Can incorporate covariates
    - Separates between-/within-person random error

Slide 52

Part 2 is a mixed-model linear regression that models amount on a transformed scale to account for skewness. This model can incorporate covariates, and also separates within- and between-person random error to account for measurement error.

NCI Method: Overview

Two-part model: Episodically-consumed constituents

- Link
  - Person-specific effects are correlated
  - May share covariates

Slide 53

Finally, the two parts of the model are linked by allowing the person-specific random effects to be correlated. This models the correlation we often see between probability and amount.

The two parts of the model may also be linked by sharing covariates.

Definitions

- Let T_{ij} be true intake for a person *i* on day *j*
  - Let p_i be true probability to consume
  - $p_i = \Pr(T_{ij} > 0 \mid i)$
  - Let A_i be the true average consumption-day amount
  - $A_i = E[T_{ij} \mid i, T_{ij} > 0]$
- True usual intake $T_i = E[T_{ij} \mid i] = p_i A_i$

Slide 54

I’m going to describe the model in statistical notation briefly before turning to an example. First, we let Tij be the true intake for a person i on day j. We also let pi be the true probability to consume; that is, it is the probability that the true intake is greater than zero, given the person, i. Next, we let Ai be the true average consumption-day amount; that is, the average true intake, given that the individual consumed on that day. Finally, we can see that true usual intake is the product of the probability and amount.

Assumptions

Let R_{ij} be intake reported on the 24HR for a person *i* on day *j*

i. A food is reported on 24HR if and only if consumed

   – Therefore, probability of consumption on recall is the same as the probability of true consumption

$$\Pr(R_{ij} > 0|i) = \Pr(T_{ij} > 0|i) = p_i$$

ii. 24HR is unbiased for usual amount consumed on a consumption day

$$E[R_{ij} \mid i; R_{ij} > 0] = A_i$$

$\Rightarrow$ 24HR is unbiased for true usual intake

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

Slide 55

Unfortunately, we don’t have truth, but we have the 24 hour recall reported intake, Rij, instead.

We make some important assumptions in the model. First, we assume that if someone ate a food that he or she reported it on the 24 hour recall, and if a food is reported on the 24 hour recall, then it was consumed. This means that the probability, pi, is the same as the probability that the 24 hour recall is greater than zero. Second, we assume that the 24 hour recall is unbiased for the consumption-day amount. This is another way of saying that the 24 hour recall only has random error. Although there may be some noise with the 24 hour recall estimates, we assume that, on average, they estimate the true consumption-day amount.

Putting these two assumptions together, we assume that the 24 hour recall is unbiased for the true usual intake, T sub i.

Part I: Probability to consume

- Mixed model logistic regression

$$\Pr(R_{ij} > 0 \mid \mathbf{X}_{1i}, u_{1i}) = h(\beta_{10} + \boldsymbol{\beta}'_{\mathbf{X}1}\mathbf{X}_{1i} + u_{1i})$$

- Where *h*( ) is the logistic function,
- $\mathbf{X}_{1i}$ is a vector of covariates, and
- u_{1i} is a person-specific random effect
  - Allows a person's value to differ from that defined by covariates
  - $u_{1i} \sim N(0, \sigma^2{}_{u1})$

Slide 56

As I mentioned previously, the probability to consume is modeled using a mixed-model logistic regression. Here, we use the h to indicate the logistic function.

Because it's a mixed model, it has both fixed effects corresponding to the covariates, X1i, and random effects, called u1i. So, u1i is a person-specific random effect that allows a particular person's value to differ from that defined by the covariates. We assume the person-specific random effects are normally distributed.

Assumptions revisited

Let R_{ij} be intake reported on the 24HR for a person *i* on day *j*

i. A food is reported on 24HR if and only if consumed

   – Therefore, probability of consumption on recall is the same as the probability of true consumption

$$P(R_{ij} > 0|i) = P(T_{ij} > 0|i)$$

ii. 24HR is unbiased for usual amount consumed on a consumption day

$$\mathrm{E}[R_{ij} \mid i;\, R_{ij} > 0] = A_i$$

$\Rightarrow$ 24HR is unbiased for true usual intake

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

iii. **On transformed scale the reported amount has additive and independent measurement error**

Slide 57

Recall the assumptions that we made about the 24 hour recall earlier. We have to add one more assumption to this list that is related to the amount part of the model. Specifically, we assume that, on the transformed scale, the amount has additive and independent measurement error.

Part II: Amount on consumption days

- Mixed model linear regression on *g(·)* Scale

$$\boxed{g(R_{ij},\gamma \mid R_{ij} > 0; \mathbf{X}_{2i}, u_{2i}) = \beta_{20} + \boldsymbol{\beta}'_{\mathbf{X}2}\mathbf{X}_{2i} + u_{2i} + e_{ij}}$$

- where *g( )* is the Box-Cox transformation,
- $\mathbf{X}_{2\mathrm{i}}$ is a vector of covariates,
- $u_{2i} \sim N(0, \sigma^2_{u2})$ is a person-specific random effect,
- $e_{ij} \sim N(0, \sigma^2_{e})$ is within-person random error

Slide 58

In particular, on the transformed scale, we model the consumption-day amount as a mixed model with fixed effects corresponding to the covariates, X2; the random effect, u2i; and within-person random error, e. Both the random effects and the within-person errors are assumed to be normally distributed, additive, and independent.

Link

- Person-specific effects have bivariate normal distribution

$$(u_{1i}, u_{2i})' \sim BVN(\mathbf{0}, \mathbf{\Sigma})$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{u1}^{2} & \rho\sigma_{u1}\sigma_{u2} \\ \rho\sigma_{u1}\sigma_{u2} & \sigma_{u2}^{2} \end{bmatrix}$$

Slide 59

Finally, we assume the two random effects, u1i and u2i, have a bivariate normal distribution; that is, they are correlated, indicated here by the parameter, rho.
-animation-

Fitting the model

- Implemented in SAS macro MIXTRAN that calls PROC NLMIXED

- Download at http://riskfactor.cancer.gov/diet/usualintakes/macros.html

Slide 60

The model is fit by maximizing the likelihood using a quasi-Newton optimization of a log likelihood approximated by adaptive Gaussian quadrature. As I mentioned earlier, quadrature is a method of numerically estimating an integral. This is implemented in a SAS macro that is called PROC NLMIXED. The macro may be downloaded at the Web site address given here.

The MIXTRAN macro works by fitting a fixed-effects model and two uncorrelated logistic and linear models for probability and amount to obtain starting estimates of the model parameters. It then fits the two models simultaneously with the shared correlation parameter. SAS is required to run this macro.

Estimating the distribution

- Use Monte Carlo approach to generate bivariate distribution of random effects using estimated model parameters
  - Approximates integral using a numeric approach
- Combine with empirical distribution of fixed effects
- Backtransform estimate and multiply by estimated probability
  - Taylor series
  - 9-point quadrature method - recommended
- Estimate percentiles

Slide 61

I’ve described the NCI method for fitting the model, but fitting the model is really just the first step of the NCI method. To estimate the distribution, we use a Monte Carlo approach to generate the bivariate distribution of random effects using the parameters that were estimated in the modeling.

As I mentioned earlier, Monte Carlo is a numeric approach for approximating integrals like quadrature is. It uses random points to estimate the distribution, however, so you can think of it being a way of simulating the distribution based on the model parameters. Most of the time, we use 100 realizations of the random effects for each person in the data set. We then combine the estimated random effects with the empirical distribution of fixed effects that come from the data. Finally, we backtransform the estimates and multiply by the estimated probability to get an estimate of usual intake on the transformed scale.

Originally in the NCI method, we had used a Taylor series approximation to backtransform the data, as Kevin described in webinar 2. This method works very well in many cases. However, when there is a large amount of within-person variation compared with between-person variation, this method does not work as well as the 9-point quadrature method that the ISU method uses. For this reason, we now recommend the quadrature method for general use.

Finally, percentiles may be estimated from sample quantiles from the representative sample of the 100 times N backtransformed values, and the sample fraction that falls below a given cutoff comprises the estimate of the proportion of the population with usual intake above or below that cutoff.

Estimating the distribution

- Implemented in SAS macro DISTRIB
- Currently uses Taylor series approximation
- 9-point approximation to be added
- Download at http://riskfactor.cancer.gov/diet/usualintakes/macros.html

Slide 62

The Monte Carlo distribution method is implemented in the SAS macro DISTRIB, which may be found at the same Web site as the MIXTRAN macro. Currently, the DISTRIB macro uses the Taylor series approximation; we are in the process of updating the macros and incorporating the 9-point approximation.

Key concepts
Challenges
Statistical modeling
Example
Simulations
Conclusions

EXAMPLE

Slide 63

Now, let’s turn to an example.

Eating at America's Table Study (EATS)

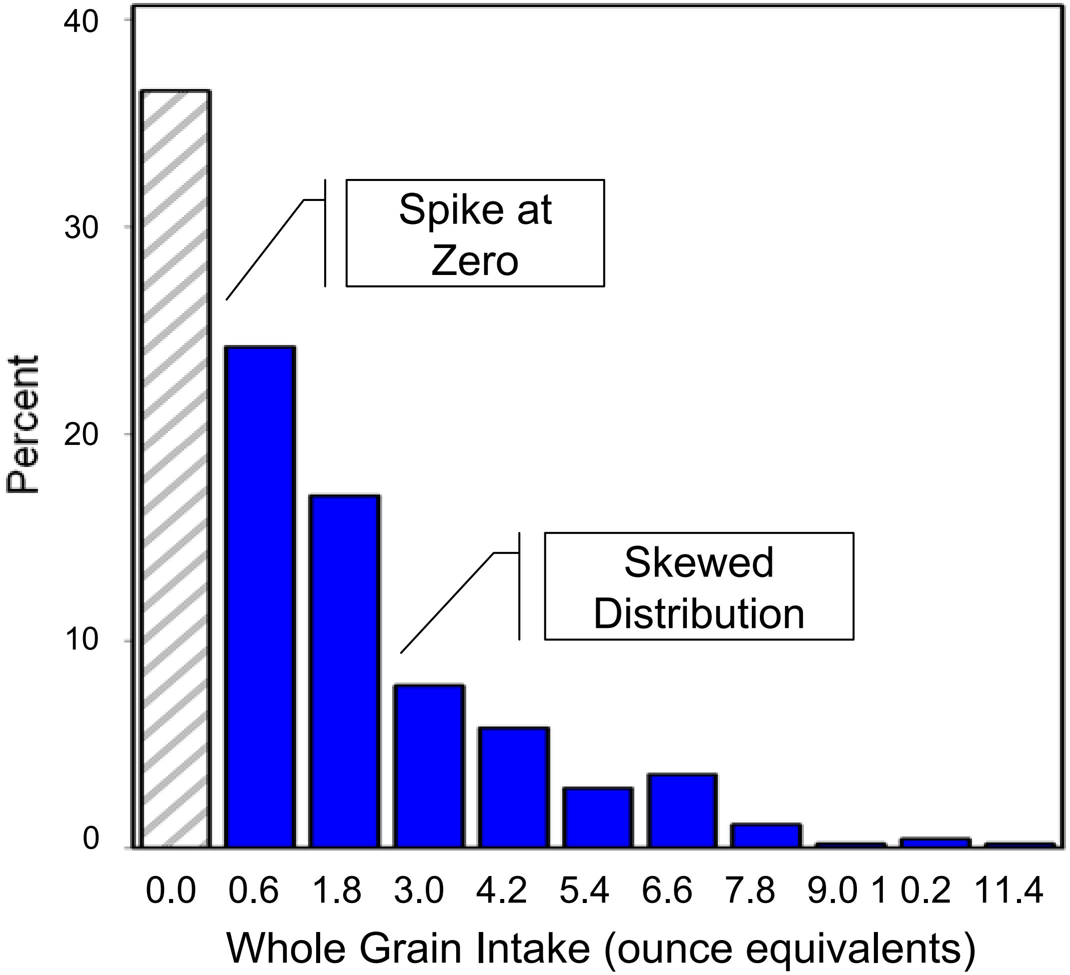
- Men and women, 20-70 years
- Nationally representative sampling of 12,615 telephone numbers
  - Approximately 1600 recruited
- Four 24HRs, one in each season
- After one year: FFQ about past year
- 965 respondents completed four 24HRs and FFQ

Slide 64

This example is from the Eating at America's Table study, or EATS. This study was conducted on men and women ages 20 to 70 years old. Approximately 1,600 subjects were recruited from a nationally representative sample; 965 respondents completed four 24 hour recalls, one in each season, and an FFQ, NCI's Diet History Questionnaire, about the previous year.

EATS example: Whole grains

- 36% of men have no consumption on a given day

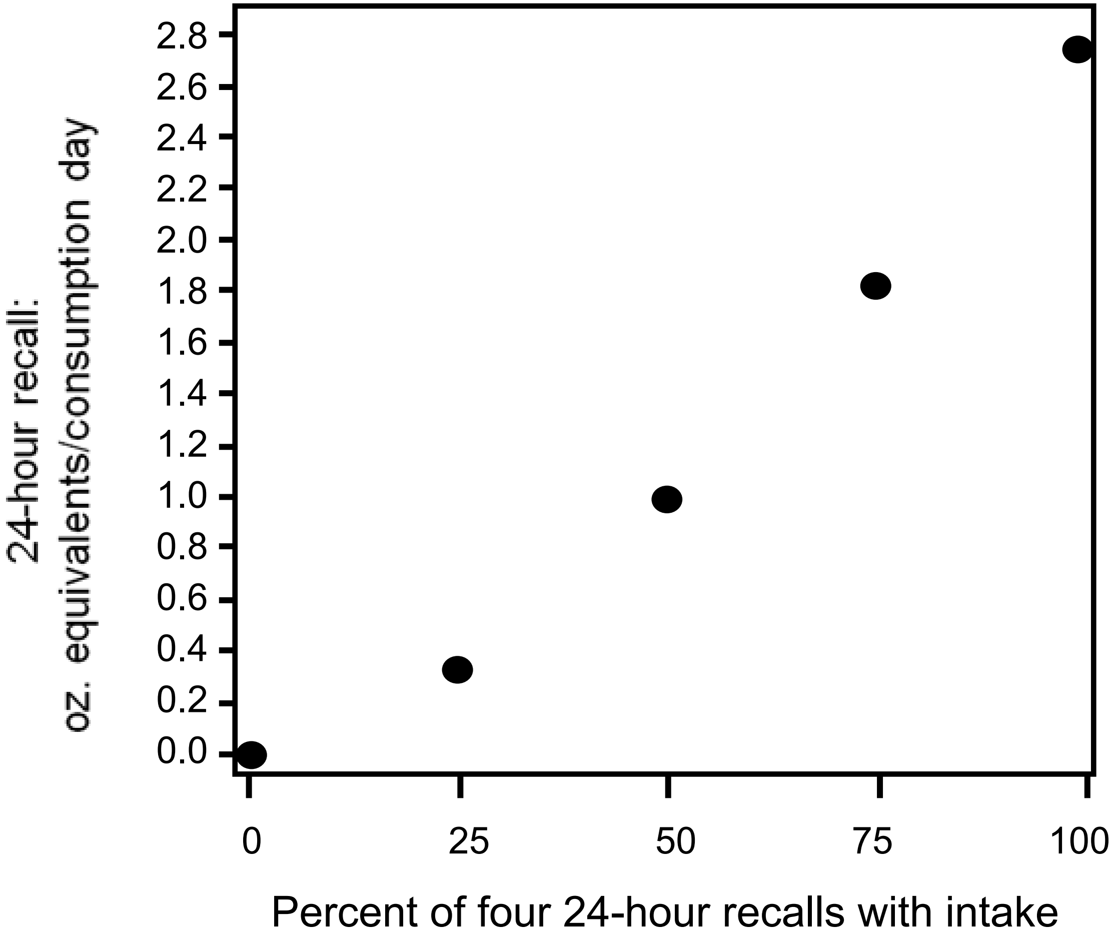


Slide 65

I am going to give an example of estimating whole grain intake for men in EATS. You've seen this plot previously. Recall that over a third of the men in EATS did not consume any whole grains on the first 24 hour recall. On the consumption days, the data were skewed.

EATS example: Whole grains

- Correlation between probability and amount is around 0.3



Source: EATS

Slide 66

Furthermore, the correlation between the probability and amount for whole grains was about 0.3, indicating that the men who consumed whole grains more frequently consumed more of them.

MIXTRAN macro: Call

- %include "C:\NHANES\Macros\mixtran_macro_v1.1.sas";
- %MIXTRAN(data=men, response=r_g_whl_tot, foodtype=gwhlm, subject=nid, repeat=intaken, covars_prob=, covars_amt=, outlib=webinar, modeltype=corr, titles=1, printlevel=2);
- Parameter estimates and predicted values are saved in datasets:
  - outlib._param_modeltype_foodtype_vcontrol
    - webinar._param_corr_gwhlm_
  - outlib._pred_modeltype_foodtype_vcontrol
    - webinar._pred_corr_gwhlm_

Slide 67

So, I wanted to show you how I fit these data in SAS.

The first line here shows how I load the macro into SAS. In particular, I have stored the MIXTRAN macro in a folder called “C:\NHANES\Macros.” You would just want to make sure you update this to wherever you have the macro stored on your computer. This %include statement will read all of the macro code into SAS so that it is ready to execute the macro.

Next, I’m going to go through each statement I put into the macro call, indicated by %mixtran. A SAS macro is a useful technique for rerunning a block of code when you want only to change a few variables. In these statements, I am defining the macro variables, so that the macro will replace the generic macro variable name such as “data” with the real name of the data set I’m using, which I’ve called “men”-animation-. Next, the response variable is defined as “r_g_whl_tot,”
-animation- which is the variable that describes the total intake of whole grains on the 24 hour recall for a particular day.

I’ll stop here and note that it’s important that the data be arranged in one row per day, and sorted by id and day. In this case, I have four recall days, so I have four observations for each man.-animation-

The foodtype variable is a little different from the previous definitions because it doesn’t correspond to a variable in the data set. Instead, it is used to label the data sets. I’ve decided to call my foodtype “gwhlm,” but I could have chosen another name. -animation-

In the next statement, I’ve defined the variable that identifies the subject or participant, “nid.” -animation-

Then, I define the variable that identifies the day of the 24 hour recall, from 1 to 4. It is called “intaken.” -animation-

If I wanted to incorporate covariates into the probability part of the model, I could put them in here, with spaces in between the variables. However, I’ve fit the distribution here with no covariates, so I just leave it blank. I’ve also left the amount blank.

I could have chosen to include variables like age or race here; weekend requires special macro variables. -animation-

The “outlib” variable defines the name of the library where the output data sets will be stored. In this case, I’ve already defined a permanent SAS directory called “webinar” using a libname statement. If I wanted to create temporary data sets, I could use the word “work” here. -animation-

Because this is an episodically consumed food, *modeltype=corr* is specified. This fits the two-part model with correlated random effects. There are also options to fit an uncorrelated model or an amount-only model for daily consumed constituents. -animation-

The macro variable "titles" saves one line for a title supplied by the user. Other numbers could be specified. The "printlevel" is 2, which prints the output from the NLMIXED runs and the summary.

The MIXTRAN macro saves parameter estimates and predicted values in two data sets. They have the general form given here, and the specific file produced from my call is given below. The term "vcontrol" is an optional macro parameter that may be specified in the model when you have difficulty getting the model to converge and want to put in starting values from a previous run, and it is appended at the end of the file name.

If you would like more information on the macros, a user guide is included on the Web site.

[This page intentionally blank.]

MIXTRAN macro: Output

- Correlated model with printlevel=2 produces:
  - 3 sets of NLMIXED output
  - Summary of the Uncorrelated model runs
    - Parameters
    - AIC and -2 log likelihood
  - Summary of the Correlated model runs
    - Parameters
    - AIC and -2 log likelihood with comparison to uncorrelated model

Slide 68

With printlevel=2, the MIXTRAN macro will produce three sets of NLMIXED output for the correlated model. It also prints out a summary of the uncorrelated and correlated model runs, including the parameter estimates and the AIC and -2 log likelihood for the models. For the correlated model, it compares the -2 log likelihood to the uncorrelated model and calculates the p-value associated with the comparison. If the test is significant, it indicates that there is correlation between probability and amount, and that the model with correlated person-specific effects should be used.

MIXTRAN macro: Uncorrelated

```
                                    Men
                    Results from Fitting Uncorrelated Model
                        Response Variable: r_g_whl_tot

                           Convergence Status:
  Probability Model -- NOTE: GCONV convergence criterion satisfied.
  Amount Model -- NOTE: GCONV convergence criterion satisfied.

Parameter                Name                       Estimate     Std Err Prob>|t|

P01_INTERCEPT            Intercept--bi                0.7311     0.0871   0.0000
P_LOGSDU1                Reparam Var(u1)--bi          0.2571     0.0857   0.0029
A01_INTERCEPT            Intercept--in                0.5538     0.0504   0.0000
A_LAMBDA                 lambda--in                   0.3134     0.0195   0.0000
A_LOGSDE                 Resid, Reparam--in           0.1408     0.0269   0.0000
A_LOGSDU2                Reparam Var(u2)--in         -0.4863     0.0893   0.0000


   Name                                Value           Sum

   AIC--bi                             2228.41               .
   AIC--amount                         4252.40         6480.81
   -2 Log Likelihood--bi               2224.41               .
   -2 Log Likelihood--amount           4244.40         6468.81
```

measurementERRORwebinar series

Slide 69

Here is the summary table output from the uncorrelated model. This is equivalent to modeling the probability of consumption and the consumption-day amount separately, but only works if there is no correlation between these two parts of the model. -animation-

It's important to make sure that you check the message that the model has converged. -animation-

The parameters that are marked with P correspond to the probability part of the model, -animation- and the parameters marked with A correspond to the amount part of the model. -animation- "Lambda" is the Box-Cox parameter.

In the next table, -animation- the AIC for each part of the model is given, along with the sum -animation- and the corresponding values for -2 log likelihood.

MIXTRAN macro: Correlated

```
                                  Men
                  Results from Fitting Correlated Model
                     Response Variable: r_g_whl_tot

 Convergence Status:
   NOTE: GCONV convergence criterion satisfied.

 Parameter                Name                      Estimate    Std Err Prob>|t|

 P01_INTERCEPT            Intercept--bi               0.7249    0.0862  <.0001
 P_LOGSDU1                Reparam Var(u1)--bi         0.2436    0.0853  0.0045
 A01_INTERCEPT            Intercept--in               0.4168    0.0558  <.0001
 A_LAMBDA                 lambda--in                  0.3108    0.0194  <.0001
 A_LOGSDE                 Resid, Reparam--in          0.1349    0.0267  <.0001
 A_LOGSDU2                Reparam Var(u2)--in        -0.4058    0.0847  <.0001
 Z_U                      Z-trans of Correlation      0.9356    0.2187  <.0001

Name                Value     Diff in -2ll   p-value

   AIC                  6446.72            .            .
   -2 Log Likelihood    6432.72     36.09        0.0000
```

Slide 70

The correlated model summary output is similar to the previous page. I’m just going to point out the additional test here for the comparison to the uncorrelated model, which, in this case, is highly significant, indicating that there is positive correlation between the probability of whole grain consumption and the amount consumed on the consumption day.

DISTRIB macro

- %include "C:\NHANES\Macros\distrib_macro_v1.1.sas";
- %DISTRIB (seed=0, nsim_mc=100, modeltype=corr, pred=webinar._pred_gwhlm, param=webinar._param_gwhlm, outlib=webinar, cutpoints=.1 .25 .33 .5 .66 .75 1 1.5 2 2.5 3 3.5 4, ncutpnt=13, subject=nid, titles=1, food=gwhlm);
- Outputs one SAS data set that contains descriptive statistics for usual intake:
  - outlib.descript_food_freq_var
    - webinar.descript_gwhlm_

Slide 71

The DISTRIB macro is used to estimate the distribution of usual intake, after estimating the model. It uses the input from the MIXTRAN macro. As before, I've used the %include statement to load the DISTRIB macro into SAS. Next, I call the macro with the %DISTRIB statement.-animation-

So, remember how I stated that the Monte Carlo procedure uses random numbers? All this statement does is set the seed in SAS to generate the random numbers—0 selects the date and time as the random seed. -animation-

The "ncsim_mc" macro variable sets the number of Monte Carlo simulations per person in the data set. In this case I've used 100. -animation-

As before, the modeltype is "corr." -animation-

I'm now inputting the variables from the previous run—first predicted values -animation-

Then parameters. -animation-

As before, the library I'm saving to is called "webinar." -animation-

This "cutpoints" statement allows me to calculate the estimated percent of the population below these cutpoints, here, from 0.1 to 4 ounce equivalents of whole grains. -animation-

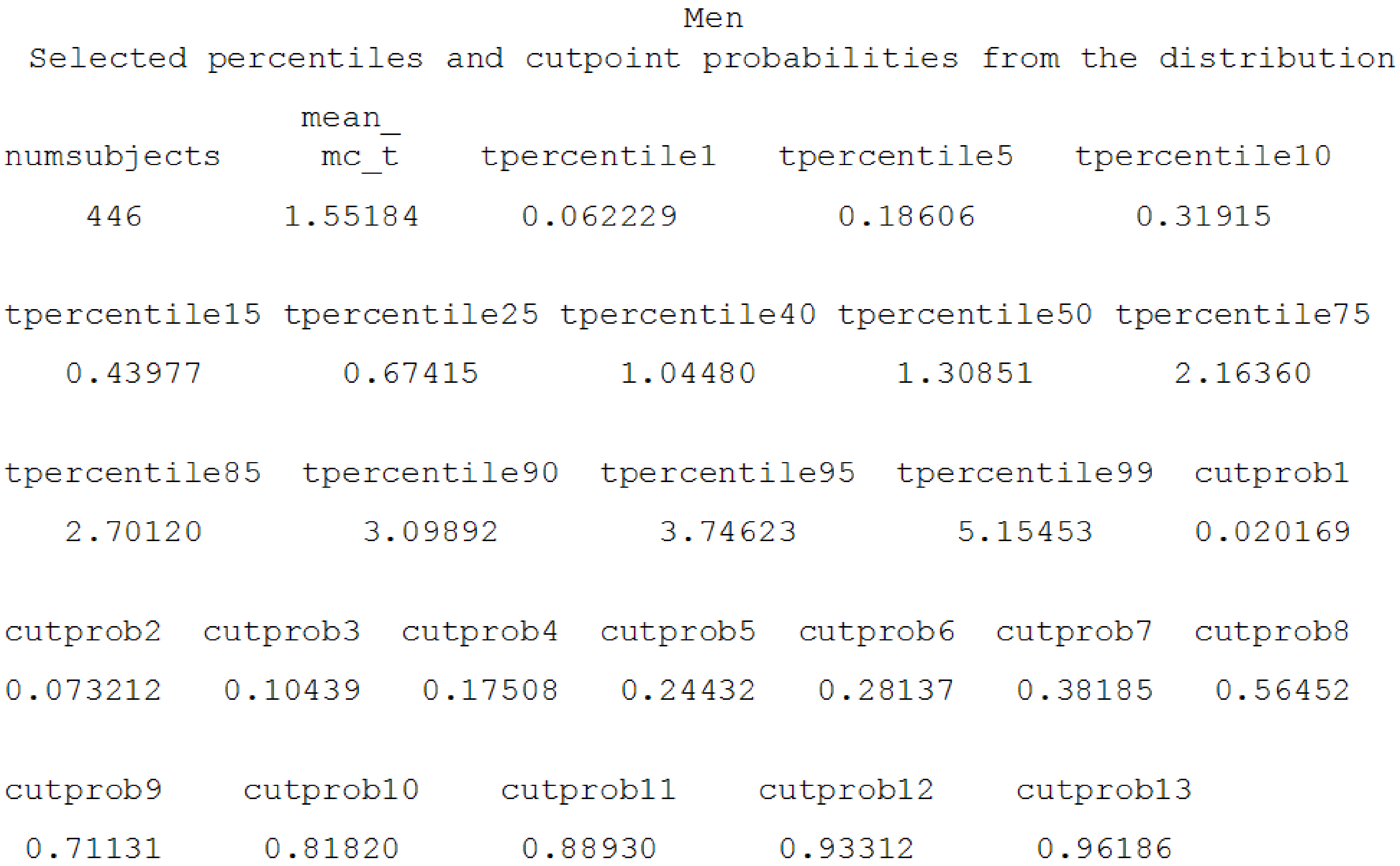
In the next macro variable, I simply put in how many cutpoints there are; in this case there are 13. -animation-

Again, the subject variable is "nid." -animation-

I save a line for a title. -animation-

And I've decided to call my food (or dietary constituent) "gwhlm." This is used to name the output data set given below.

DISTRIB macro: Output

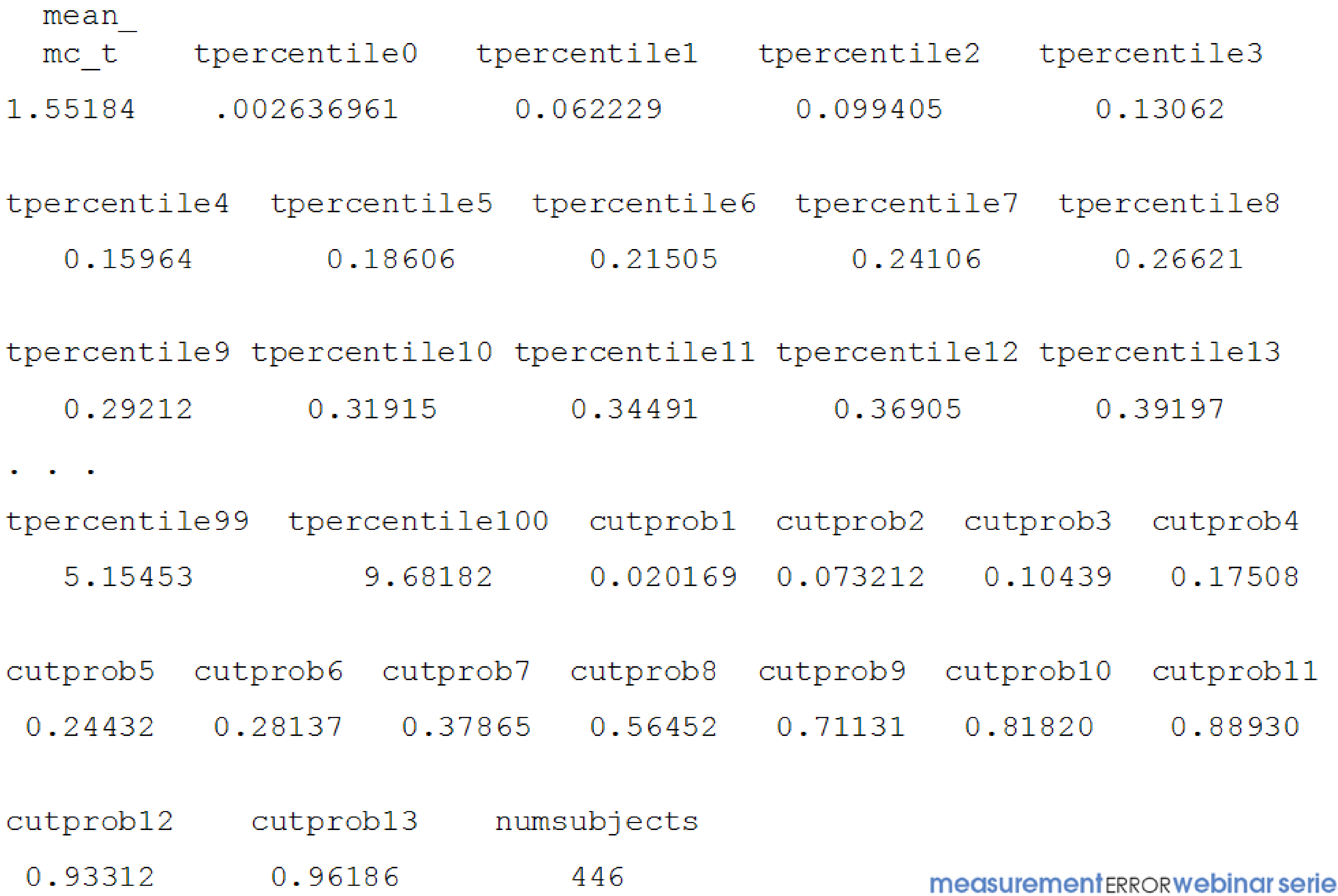


```
                                  Men
  Selected percentiles and cutpoint probabilities from the distribution

                mean_
numsubjects      mc_t     tpercentile1    tpercentile5    tpercentile10

    446        1.55184     0.062229        0.18606         0.31915


tpercentile15 tpercentile25 tpercentile40 tpercentile50 tpercentile75

   0.43977       0.67415       1.04480       1.30851       2.16360


tpercentile85  tpercentile90  tpercentile95  tpercentile99  cutprob1

   2.70120        3.09892        3.74623        5.15453     0.020169


cutprob2  cutprob3  cutprob4  cutprob5  cutprob6  cutprob7  cutprob8

0.073212   0.10439   0.17508   0.24432   0.28137   0.38185   0.56452


cutprob9    cutprob10    cutprob11    cutprob12    cutprob13

 0.71131     0.81820      0.88930      0.93312      0.96186
```

Slide 72

This is the output that prints from the DISTRIB macro. It prints selected percentiles and the N and mean from the estimated distribution, as well as all 13 cutpoint probabilities.

DISTRIB macro: descript_gwhlm_ dataset



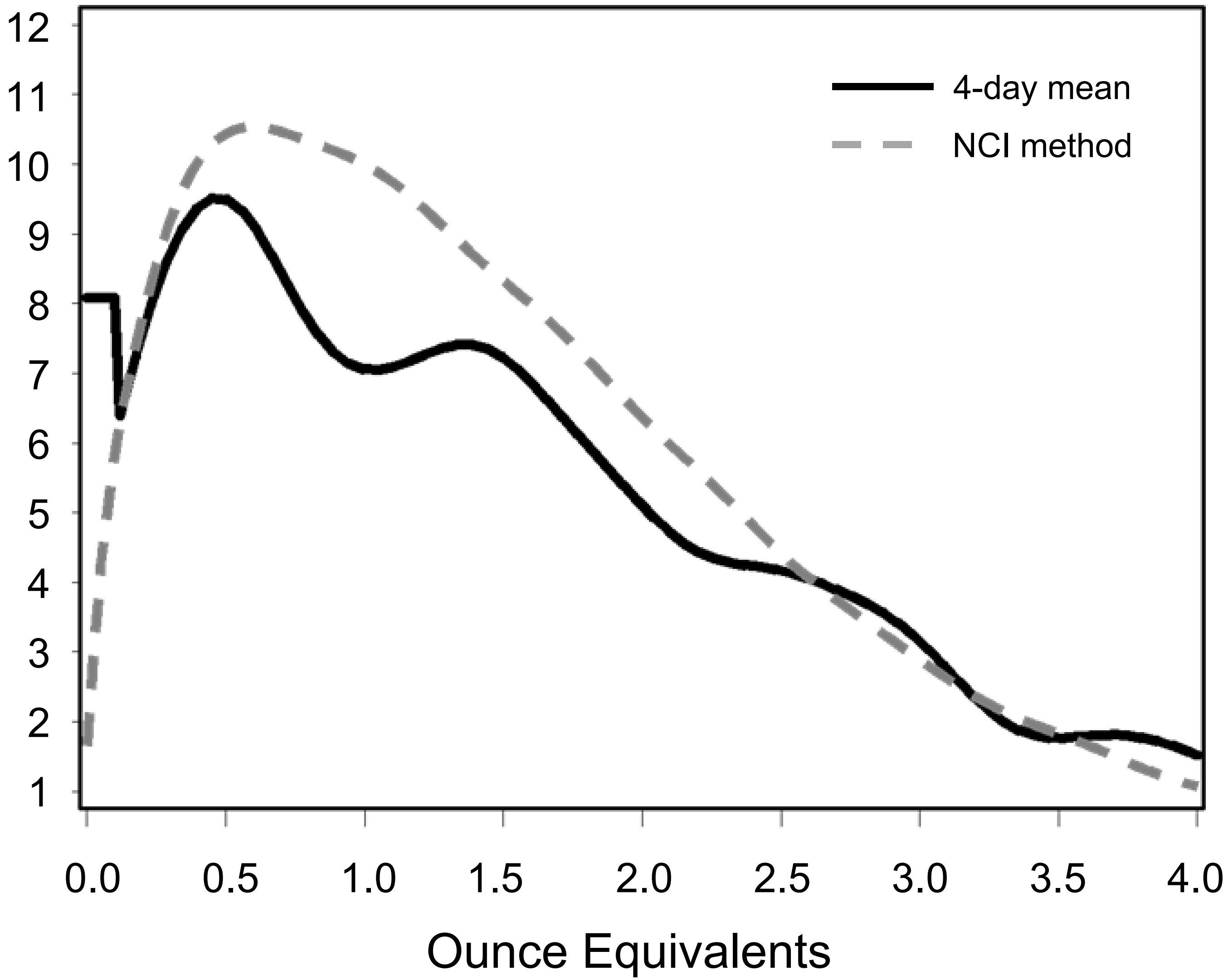
```
  mean_
  mc_t     tpercentile0    tpercentile1    tpercentile2    tpercentile3

1.55184     .002636961       0.062229        0.099405         0.13062


tpercentile4  tpercentile5  tpercentile6  tpercentile7  tpercentile8

   0.15964       0.18606       0.21505       0.24106       0.26621


tpercentile9 tpercentile10 tpercentile11 tpercentile12 tpercentile13

   0.29212      0.31915      0.34491      0.36905      0.39197

. . .

tpercentile99  tpercentile100  cutprob1  cutprob2  cutprob3  cutprob4

   5.15453          9.68182    0.020169  0.073212   0.10439   0.17508


cutprob5  cutprob6  cutprob7  cutprob8  cutprob9  cutprob10  cutprob11

 0.24432   0.28137   0.37865   0.56452   0.71131   0.81820    0.88930


cutprob12    cutprob13    numsubjects

 0.93312      0.96186         446
```

Slide 73

And here is a partial printout of the descript_gwhlm_ data set that is output by the macro. It gives the mean, N, and all of the percentiles from 0 to 100, along with the 13 cutpoint probabilities.

Example

Whole grains (men): Distribution



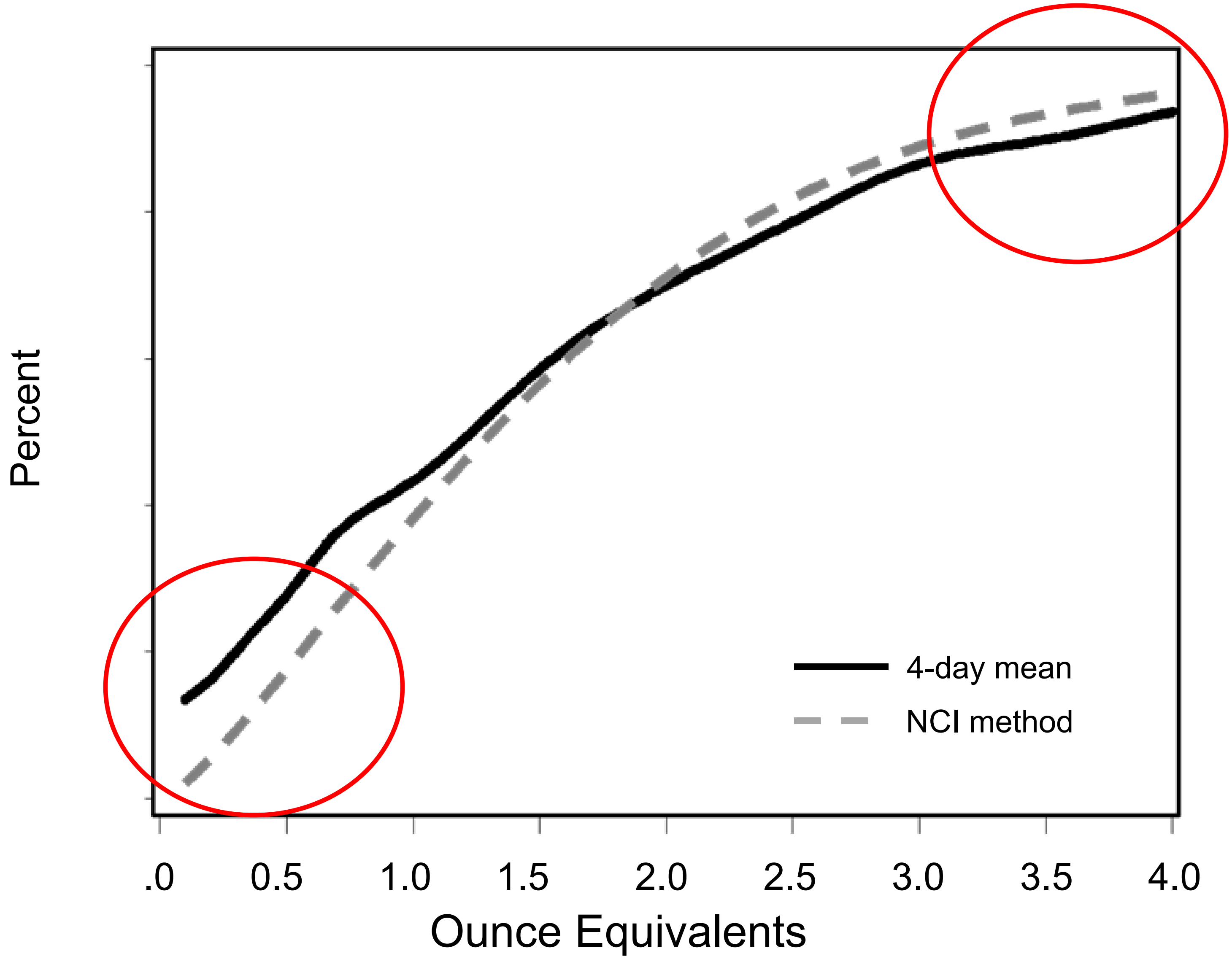
Source: EATS

Slide 74

And these are smoothed histograms showing the approximate distribution of whole grains for men in EATS. Ounce equivalents of intake are shown on the horizontal axis, and the percent of participants in the group is given on the vertical axis; these are kernel smoothed estimates. You can see here a big spike, indicating that 8 percent of the men did not consume whole grains on any of the four days in EATS. In contrast, we see a smaller percentage in the tail for the NCI method.

Example

Whole grains (men): Cumulative distribution



Source: EATS

Slide 75

However, the smoothed histogram is a little difficult to interpret due to the bumpiness from the data, and the semi-continuous nature of the four-day mean data, so I am also showing the cumulative distribution function for men.

From this we can estimate the proportion of men consuming more than a certain number of ounce equivalents per day. You can see here, -animation- that the four-day mean estimates a higher proportion of men have intake under about 1 ounce equivalent in the left side of the plot. The four-day mean also diverges from the NCI method in the other tail of the distribution, -animation- indicating that a higher percentage of men are below recommended intakes compared with the four-day mean.

Example

Whole grains: % above cutpoints

Ounce Equivalents	Gender	% Above (4-day mean)	% Above (NCI Method)
1/3	Men	78.7%	89.7%
	Women	74.2%	84.5%
1	Men	56.7%	62.1%
	Women	39.5%	42.6%
3	Men	13.5%	11.1%
	Women	3.9%	1.1%

Source: EATS; Tooze et al, 2006

Slide 76

We can see what these estimated percentages are in this table, for both men and women. For men, the four-day mean method estimates that about 21 percent of men are consuming less than ⅓ ounce equivalents of whole grains, but the NCI method estimates this number is about 10 percent. In the upper tail, the four-day mean estimates that about 14 percent of men eat more 3 ounce equivalents per day, but the NCI method estimates that this number is only 11 percent.

SIMULATIONS

Slide 77

[No notes.]

Simulations: Whole grains

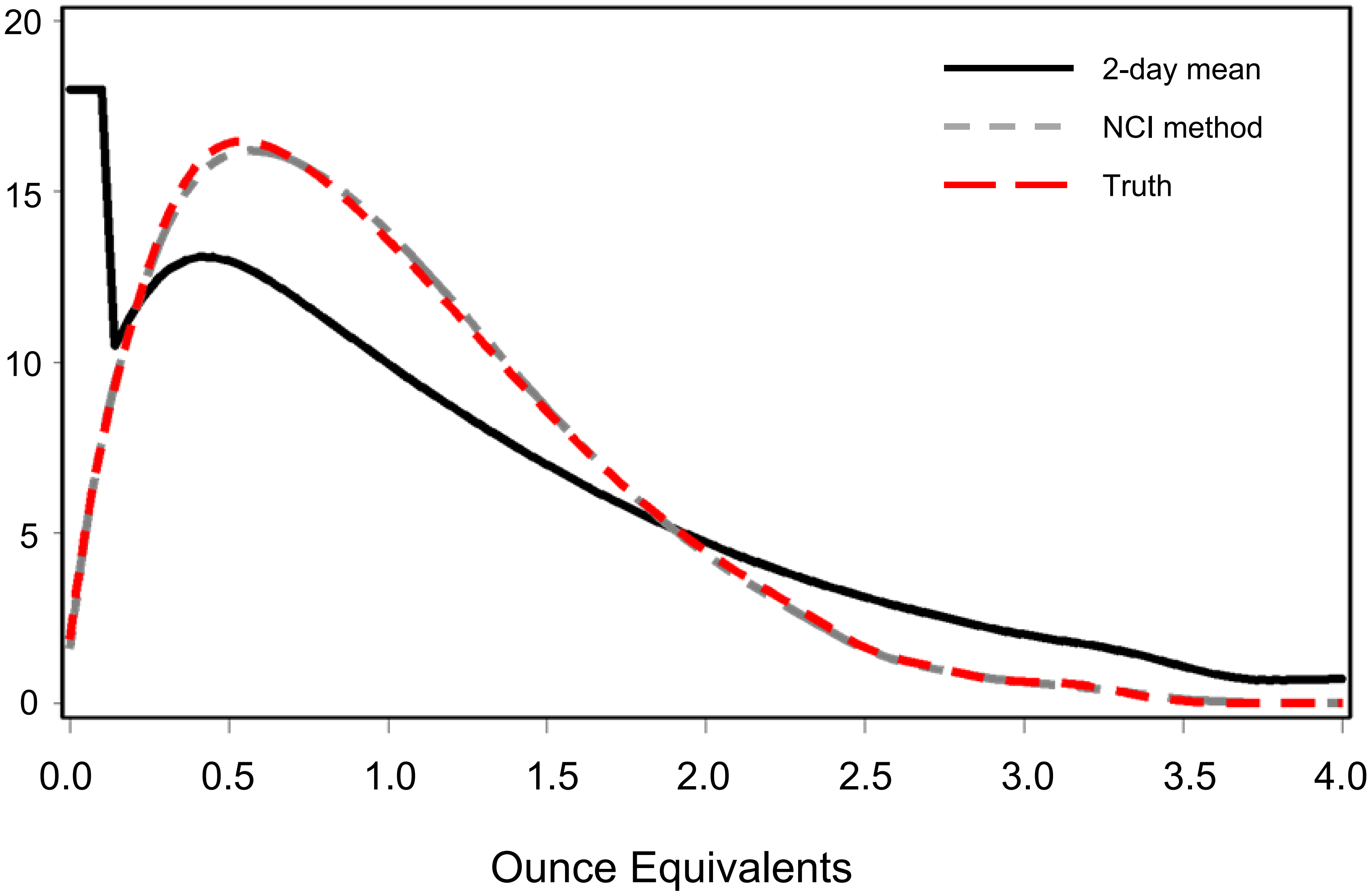
- Data were simulated based on EATS
  - Women
  - Probability and amount are correlated (r=0.23)
- 300 data sets of 2000 individuals
  - Simulate 365 days per person
  - Truth defined as the mean of 365 days
  - Fit model 300 times using only 2 days and take average
- Compare truth to the NCI method and 2-day mean

Slide 78

So, it is interesting to compare the NCI method with four-day mean for the EATS data. However, we were interested in seeing how our method would compare with truth. In order to do this, we simulated data that were very similar to the EATS data. They were simulated to have the same 24HR consumption pattern as was seen for women in EATS. In EATS, probability and amount were correlated.

We generated 300 data sets of 2,000 individuals each, with 365 simulated recalls for each person. The mean of the 365 days was considered to be "truth." We then fit the model on two of the recall days 300 times and took the average and compared it to the truth from all 600,000 data points, and to using a two-day mean.

Simulations: Distribution



Slide 79

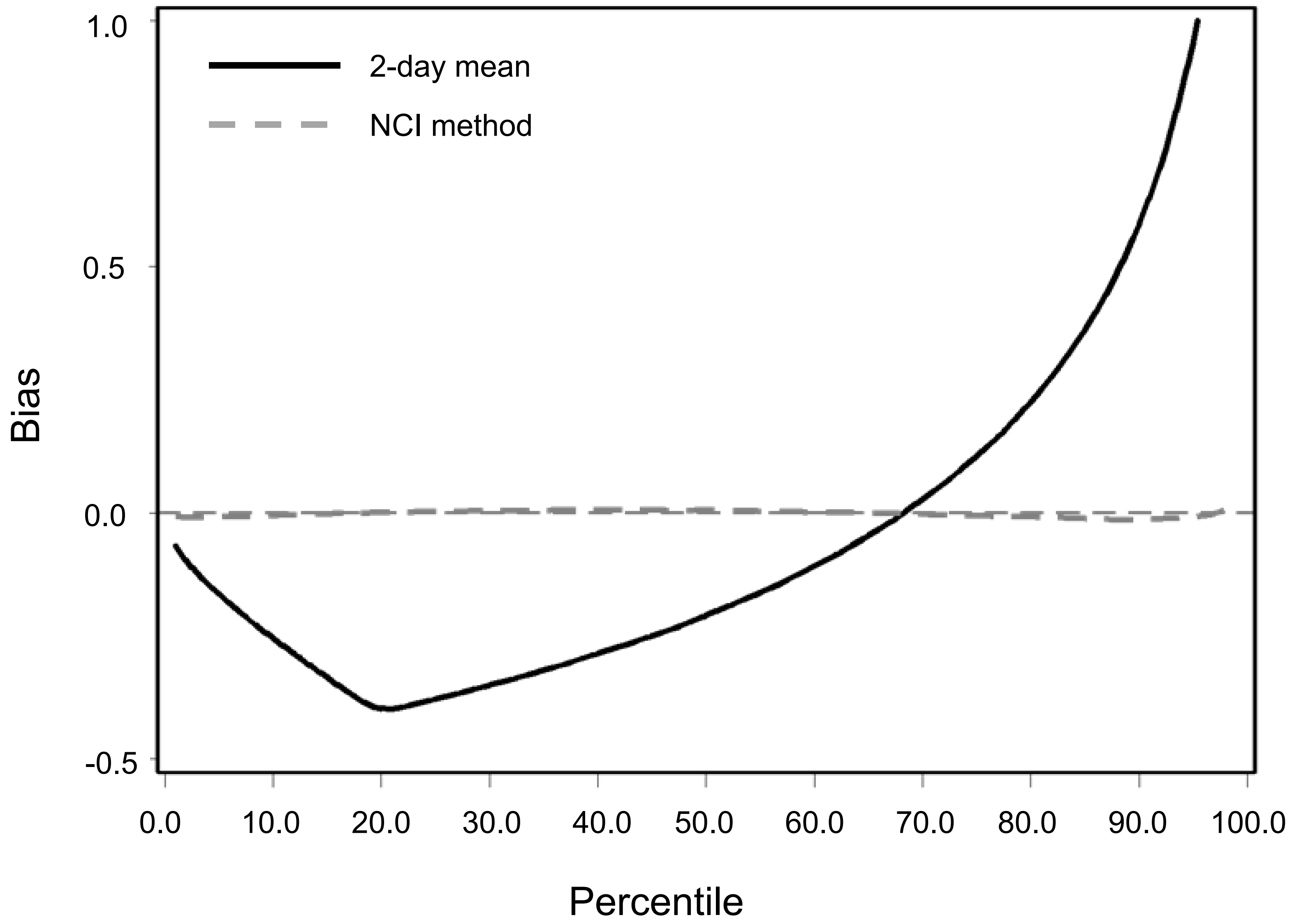
Here are the estimated distribution curves from smoothed histograms.

In these curves:

- The red dashed line represents truth.
- The gray dashed line, which is very similar to the red line, is the NCI method.
- And the two-day mean is shown in black.

You can see the NCI method is very close to truth and the two-day mean has a big spike at zero and a long tail.

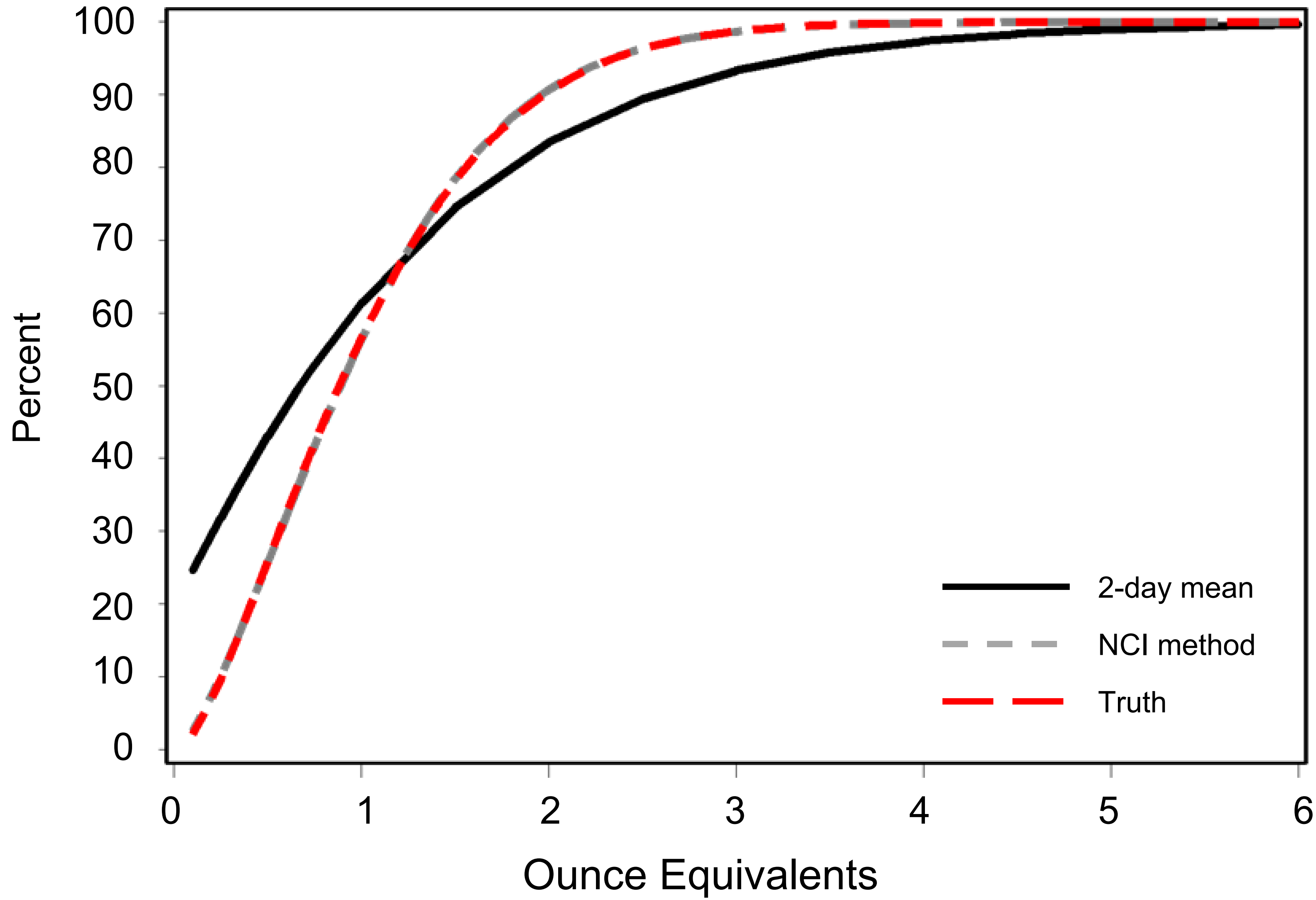
Simulations: Bias



Slide 80

This plot shows that the NCI method, as shown by the gray dashed line, consistently had low bias, especially compared with the two-day mean.

Simulations: Cumulative distribution



Slide 81

This plot shows the cumulative distribution function, with the dashed red line again representing truth. The NCI method and the two-day mean lines cross around the mean, showing that if you just want to estimate mean intake, all of the methods are good, but you see large differences in the tails

Simulations: Percent consuming below cutoffs

Ounce Equivalents	% Above (Truth)	% Above (2-day mean)	% Above (NCI Method)
1/3	86.1	65.9	87.3
1/2	75.8	57.9	76.2
1	44.5	38.4	43.3
2	9.2	16.1	8.3
3	1.2	6.1	1.1
4	0.0	2.7	0.1

Slide 82

Here are the percentages above the cutoffs for the simulations. In this case, we see that the NCI method is closer to truth than the two-day mean, especially in the tails. For example, the table shows that for the simulated data, only 1.2 percent of the simulated population consumed more than 3 ounce equivalents of whole grains per day, but the two-day mean estimated the value to be 6.1 percent. The NCI method was close to truth, with a value of 1.1 percent.

Summary of simulation studies

- The NCI Method is less biased than the 2-day mean
- For estimating the mean of the distribution both methods do well
- In the tails of the distribution
  - NCI Method is close to truth
  - Simple 2-day mean overestimates the proportion of the population in the tails

Slide 83

So, to summarize the simulation results for whole grains: The NCI method is less biased for estimating usual intake than the two-day mean.

When estimating the percent above or below a certain number of servings:

- At the mean
  - Both methods do well.
- But above or below the mean
  - The NCI method was very close to truth.
  - Simple mean tends to overestimate the proportion who consume in the tails.

Key concepts

Challenges

Statistical modeling

Example

Simulations

Conclusions

CONCLUSIONS

Slide 84

[No notes.]

Conclusions

- The two-part model is appropriate for the estimation of the usual intake for episodically consumed foods
- The NCI Method meets the following challenges:
  - Accounts for measurement error
  - Accounts for skewness
  - Models probability and amount
  - Allows for correlation between probability and amount
  - Incorporates covariates

Slide 85

So, in conclusion, the two-part model I've presented here is appropriate for the estimation of the usual intake for episodically consumed foods and other dietary constituents. The NCI method, in particular, accounts for the challenges of episodically consumed constituents, including measurement error, skewness, separation of probability and amount, as well as their correlation, and is able to incorporate covariates into modeling.

QUESTIONS & ANSWERS

Moderator: Sharon Kirkpatrick

Please submit questions using the *Chat* function

Slide 86

[No notes.]

Measurement Error Webinar 3 Q&A

Question: **How do you decide if a dietary component is episodically consumed or not?**

That's a good question. There's no hard-and-fast rule but, generally, what we see is that if you see more than 5 to 10 percent zero intake, it's beneficial to use a two-part model. *(J. Tooze)*

And what about dietary components that are episodically consumed by some subgroups but not others? The example given is milk that might be episodically consumed among non-Hispanic blacks but not non-Hispanic whites. If you want a population estimate, how do you handle that?

Well, there's a couple of ways you can do it. For example, we might use age and race/ethnicity as variables in the modeling, as covariates for both the probability and the amount. And we can estimate different estimates for the subpopulations defined by these variables. Now, I didn't really have time to go into it today in detail, but you have to be really careful that you're appropriately modeling the variance components when you do this. There are ways of checking this statistically. So, it would be possible to do it within the modeling framework; it would also be possible to just do a stratified analysis. If you're really concerned about making estimates for one subgroup, you might want to consider doing a stratified analysis for episodically consumed components for that subgroup and fit, perhaps, just an amount model for a subgroup that was not episodically consumed. *(J. Tooze)*

Does the NCI method provide confidence intervals for the percentiles of the usual intake distribution?

You can get standard errors of the estimates by using bootstrap methods. We have a paper that illustrates this when you have nutrients or daily consumed dietary constituents. We haven't actually implemented it in the macros yet, however. (Note, when using complex surveys, you can use balanced repeated replication to obtain standard errors rather than bootstrap.) *(J. Tooze)*

How would you apply these statistical methods to three-day food records?

Three-day food records are generally treated as an eating occasion and as a short-term instrument itself, so you would need to have repeats of the three-day food record, probably a week or so apart, on at least a subset of the population in order to be able to separate out the between- and within-person error. And then you could proceed in a similar manner to what I described for 24-hour recalls. *(J. Tooze)*

Would you recommend using the NCI method when you have data from two or fewer 24-hour recalls?

You can use the NCI method if you have data on two 24-hour recalls on at least a subset of the population. You can't use it if you only have one recall. You really need to have two on at least some people to be able to figure out what proportion of the variability is within subjects and which is between subjects, and if you just have one recall, the sources of error are put together. *(J. Tooze)*

This question is a little bit related. In practice, if you have only two 24-hour recalls per person, are there problems getting the model to converge?

Well, no, not really. There can be problems getting the model to converge. I won't lie about that. Modeling is really more of an art than a science, but it's probably more limited by the number of people than the number of recalls, generally, when we have convergence problems. Of course, having more recalls is helpful; having more people, of course, is helpful. And then it can also depend on the proportion of people who have zeros and who are included in estimating the probability, and it can also depend on having enough people to make sure that at least some of them consume the food or nutrient on the two days in order to be able to fit the model.

I'll just comment that with convergence problems in general, I've generally found the best thing to do is to try to update the starting values that are used in the nonlinear mixed-effects model and PROC NLMIXED. And we do have that function that I briefly touched on to be able to do that in the macro. *(J. Tooze)*

This is a follow-up question to an earlier one about food records. Is there any benefit to collecting a three-day food record as opposed to two or more recalls?

I don't know that I'm the one to answer that question. You really would have to think about the different sources of error that you might see. Of course, you would have a somewhat longer period than the 24-hour recall, which could give you some benefit. However, some studies have shown that food records can be more reactive, that when people think about what they're eating, they might say to themselves, "Oh, my, I ate that many Hershey's kisses?!" And then they may modify what they eat. So there can be some concern that food records might not reflect usual intake as well as the 24-hour recall, especially if people don't know when the 24-hour recall is coming. But I'm a statistician; I'm not a nutritionist. And so I won't comment on those aspects much more.

Statistically, I don't think there's really much difference in the modeling, the only kind of caveat to that is it's possible to have some autocorrelation, perhaps, when, let's say, you eat a food on one day and then you eat leftovers the next day. And currently, we don't model autocorrelation or serial correlation in this modeling approach. *(J. Tooze)*

Another related question: Do you know if there has been a study comparing distributions based on recalls to those based on three-day or multiple-day diet records?

No, I'm not aware of any study that's done that. *(J. Tooze)*

Has the NCI method been validated against known biomarkers for energy expenditure or protein?

No, we haven't done directly for energy or for protein intake. Larry Freedman does have a paper that incorporates biomarkers and estimating distributions of intake, but we haven't written any specific manuscripts to do that with the NCI method explicitly. *(J. Tooze)*

Changing topics now: What do you do if the density of the two-day means shows two peaks?

Well, I guess I didn't really get into this, either, but we do make the assumption in this that we don't have two peaks, that we just have the one spike at zero and that the non-zero data are able to be transformed to normality, and so if you can't do that, if it's not possible to transform your data to normality, then the method is not going to work perfectly in the way it's supposed to. I haven't really done any simulations to be able to tell you how much off it may be. But that certainly is a good point, that you should look at your data. I always suggest to people that they should look at their data graphically, and that's part of the reason I showed those histograms, because I think that's always an important first step, to look at your data and make sure that it can be appropriately transformed.
(J. Tooze)

Thinking of zero intakes, what if reporting of zero intake is biased; for example, if overweight individuals overreport zero intake of high-calorie food?

Well, that would be a systematic error, and so that would be a kind of a person-specific bias and the general impact of person-specific bias is to not impact the mean of the distribution, but to impact the variability. And so it could lead to biased estimates of the distribution of usual intake. You know, lots of people have tried to look at this question, including me, trying to figure out if there are certain factors that may impact underreporting or trying to predict underreporting, and it turns out it's very difficult to do. I did some analyses of the Observing Protein and Energy Nutrition Study (OPEN Study) data and found that I could only explain about 10 percent of the error in the FFQ and 25 percent of the error in the recalls by looking at things like BMI, fear of negative evaluation, social desirability, dieting … things like that. And so at this point, there are not really good ways of correcting for those factors, and so that's why we just make this working assumption that the 24-hour recall is only subject to random error. *(J. Tooze)*

Following up on an earlier question, what proportion of the sample do you need to have a second recall for to properly estimate episodic food intake?

It really can vary a lot by the food. As I said, you want to make sure that you have a sufficient number of people who would have the food intake on both days, and so you want to have at least 50 to 100 people who would have the food intake on both days. So you need to take into account what the probability of consumption on any given day is, and then use that to calculate how many people you would need to have the repeat days. And, of course, you also then would need to think about whether you wanted to make estimates for certain subpopulations and things like that. So it depends a lot on the food or nutrient of interest. If you wanted to model something like soy, which is very episodically consumed, you would need a lot more people than if you wanted to look at something like fruit consumption or even whole grains that are not quite as episodically consumed.

Sticking with the topic of episodic consumption, there is a question about how to handle vegetarians, assuming that they consume a number of components episodically or have zero consumption for particular components.

Well, I guess it would depend on whether your entire population were vegetarians or you somehow wanted to model them within your population. You could certainly use an indicator as to whether or not someone was a vegetarian as a covariate in your modeling and make separate estimates. And this kind of relates back to the earlier question about different groups having different types of patterns. You would want to really think about it and be careful in the modeling. You can try doing it as a subgroup analysis within the model, or you could try to stratify for that group. *(J. Tooze)*

Can you speak to the use of an FFQ in estimation of intake distributions?

Yes, that's a good question. I'm glad that someone asked that. Those of you who have heard the earlier talks on the NCI method and may have read the papers may have noticed that, originally—and I think this is still in many people's minds—the NCI method was developed as a way of augmenting 24-hour recall data with the FFQ. And the reason that I didn't emphasize the FFQ as a covariate today—and you notice I didn't include it as a covariate in my example—is that it turns out that when you're really interested in the distribution of usual intake, the FFQ is not that helpful. It's just because we're interested in describing the distribution of intake and so it's not going to make that much difference whether the FFQ is included as a covariate or not for surveillance. Now, in later webinars, when we start to talk about diet-health relationships, the presenters will talk about how the FFQ can be incorporated, and it turns out that with predicting individual intake, it can be really helpful to include the FFQ as a covariate in the modeling. And so I've really deemphasized the FFQ because it's not always available on everyone who has a 24-hour recall. For example, in NHANES, it's not available on everyone. So you can really limit your sample if you require an FFQ and you may also limit people by literacy or income or some things that might be related to the FFQ nonresponse, and it's just not that helpful. I actually have estimates for the whole grain that show using the FFQ versus not using the FFQ, and they're almost identical. And so you saw in the example that the whole grains was very close to truth in our simulations, and adding the FFQ into that didn't change that relationship at all. It barely made a difference. *(J. Tooze)*

How would you apply the NCI method to multiply imputed data?

That's a good question and something I don't have a good answer for because it's something that's actually come up with some extensions of the NCI method we've been working on; multiple imputation has been difficult. So it's not something that I've tried yet, so I don't really want to speak to that. *(J. Tooze)*

This question, I think, is in follow-up to the walk-through that you did of the macros and the output. The question is about how you would present the findings, so when using the NCI method for episodic

component, is there a recommendation in terms of statistics that you would report in a journal article?

Well, I think it just depends again on what you're trying to show. A lot of times what we've done, and what I did today, was I reported the percent that are below cutoff. And I think the reason for that is just because that's what's often recommended in, let's say, dietary guidelines or some other types of dietary recommendations or dietary goals, and so we often use those. But sometimes we're just interested in reporting the median, perhaps, with a standard error of the estimate, or the 25th or 75th percentile, to show an estimate of spread. So I think it just depends on what your goal is in using this method. *(J. Tooze)*

Can the methodology be adapted for use in an office or institutional menu, for example, to analyze two days from a cycle of menus?

I'm not sure that I understand. Can you tell me a little bit more about what institutional menus are? *(J. Tooze)*

I think they're talking about maybe like a nursing home or a school feeding program where they don't actually have intake data but they have menus and it might be only for two menu days. *(S. Kirkpatrick)*

I don't know. I think you would use it similarly, but we've actually just developed these methods for 24-hour recall data, which is a little bit different, I guess, than menu data. It's an interesting thought to try to apply it to something like that. Just off the top of my head, it seems like it would be possible to do it with two days. I would probably, I guess, select two or maybe a few more days, at random, from the month if that was what you wanted to estimate, but nonconsecutive days also. *(J. Tooze)*

How important is the correlation between the random effects, and does this differ by dietary component?

It turns out to be really important and it can shift the estimate of the distribution if you don't account for the correlation and it exists. In the EATS study, we looked at correlation between probability of consumption and amount consumed for the My Pyramid food groups. There were somewhere around 27, maybe 25 to 30, of these food groups, and we found that for about 80 percent of them that they exhibited the positive correlation between probability and amount, and we found that if we didn't account for it, then it would cause a problem in the estimation. So most foods have it; not all do. I mentioned in the talk dark green vegetables don't tend to; people tend to consume about the same amount regardless of how often they eat it. And sometimes things like yogurt, let's say, which might come in standard-sized containers generally, like 6 ounce containers, it's going to be less likely to have that type of pattern as well. But we do see it in most foods, and actually it's also very common in other types of zero inflated data as well, not just dietary data but physical activity and things like that also. *(J. Tooze)*

This goes back to the programs or the methods that you mentioned earlier. The question is whether these other methods have software that could be used to analyze episodically consumed components.

Which other methods? *(J. Tooze)*

In one of your slides, you talked about the ISU method and a couple of others. *(S. Kirkpatrick)*

Oh, yes, okay. The ISU method that I think Kevin mentioned in the last webinar does have software. It has a software called C SIDE, which I have used to model episodically consumed foods. The only caveat is that it should not be used if there is correlation between probability and amount. It will print out a warning and say, "Don't use this method," because it assumes the two parts are independent. And I believe the Multi Source Method also has software available. *(J. Tooze)*

[This page intentionally blank.]

National Cancer Institute

measurementERRORwebinar series

Next Session

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

Estimating usual dietary intake distributions: Accounting for complex survey design in modeling usual intake

Kevin W. Dodd, PhD
National Cancer Institute

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institutes
of Health

Slide 87

That brings today's webinar to a close. Please join us next week for Webinar 4, when Dr. Kevin Dodd will discuss accounting for complex survey design in modeling usual intake when estimating usual dietary intake distributions.