



# Estimating usual intake distributions for multivariate dietary variables

Raymond J. Carroll, PhD  
Texas A&M University

## Slide 1

Hello and welcome to the ninth session in the Measurement Error Webinar Series. I'm Sue Krebs-Smith from the U.S. National Cancer Institute and I'll be moderating today's webinar.

Before we get started with today's presentation, please note that the webinar is being recorded so that we can make it available on our Web site. All phone lines have been muted and will remain that way throughout the webinar. There will be a question and answer session following the presentation; you can use the Chat feature to submit a question.

A reminder: You can find the slides for today's presentation on the Web site that has been set up for series participants. The URL is available in the Notes box at the top left of the screen. Other resources available include the glossary of key terms and notation, and the recordings of the preceding webinars.

Now I'd like to introduce our presenter for today, Dr. Raymond Carroll. Dr. Carroll is Distinguished Professor of Statistics at Texas A&M University. He is a member of the Faculties of Nutrition and of Toxicology and holds a courtesy appointment in the Department of Epidemiology and Biostatistics. He has been principal investigator of a major NCI grant for the development of statistical methodology since 1990. Dr. Carroll's work on statistical methodology has found application in a broad variety of fields, including nutritional epidemiology. Today Dr. Carroll will discuss the estimation of usual intake distributions for multivariate dietary variables.

# 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 2

As everyone knows, this series is dedicated to the memory of Arthur Schatzkin, who I knew for 20 years from my various visits to NCI as a real visionary in the field. It's a real loss that he's not with us.

# 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 3

And of course I want to acknowledge my collaborators. We've been working together. Larry Freedman and I wrote our first paper in nutritional epidemiology in 1991, and the whole group has been working closely together now for many years, and this webinar is a byproduct of that.

# West Texas / East Texas ☹️

**Palo Duro Canyon, the Grand Canyon of Texas**

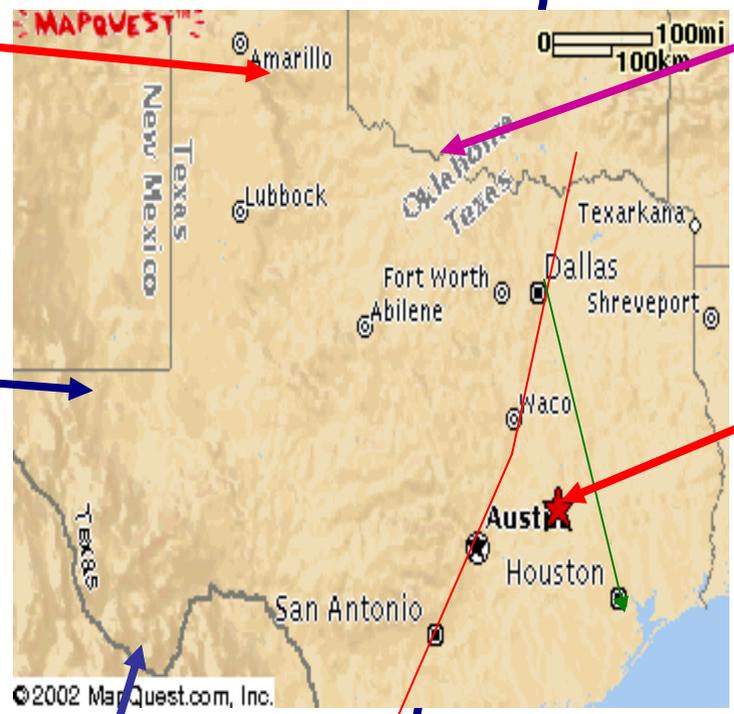


**Wichita Falls, Wichita Falls, that's my hometown**

**Guadalupe Mountains National Park**

**College Station, home of Texas A&M University**

**Big Bend National Park**



**I-35** **I-45**

## Slide 4

I'm a Texan and I thought, "Texans are always proud of being Texans." So I thought I'd just say a little bit about where I'm from. I'm from Texas A&M University, which is in College Station, strategically located between Austin and Houston; in other words, in the middle of nowhere. And I grew up in the northern central part of the state in a town called Wichita Falls. And those of you who haven't been to Texas may think it's totally flat, but there are three spectacular areas. One is the Big Bend National Park, which is just an unbelievable, high desert that is actually Alpine at the highest areas. The Guadalupe Mountains National Park has 2,800-meter mountains and is also quite spectacular. And then there's the Palo Duro Canyon, which is the Grand Canyon of Texas.

# Palo Duro Canyon of the Red River



## Slide 5

It's on the Red River of Texas, and this is a picture both of me and the Palo Duro Canyon, a fabulous hiking place just south of Amarillo.

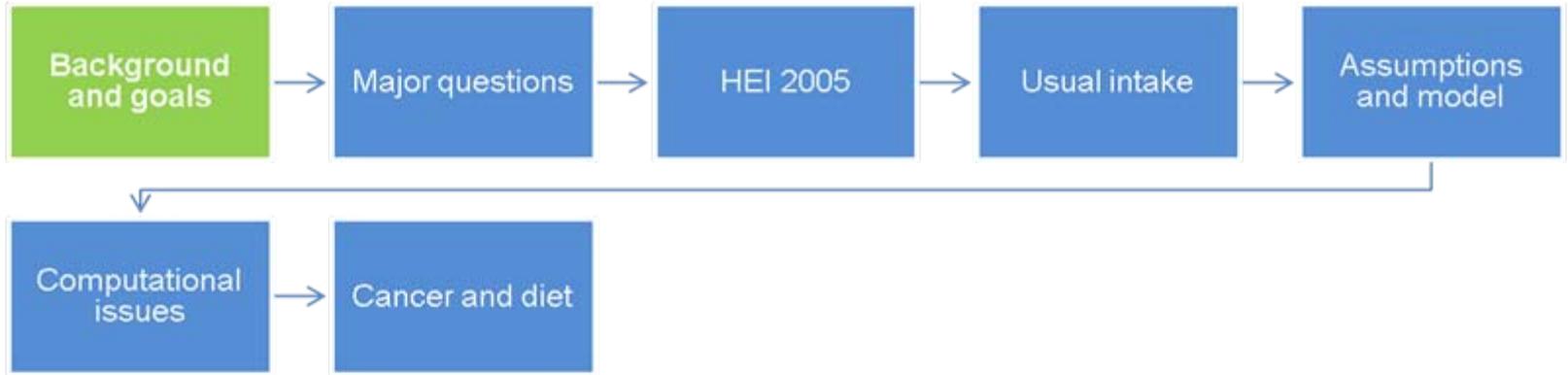
# Acknowledgments

- This work represents part of the Ph.D. dissertation of Saijuan Zhang at Texas A&M (*Annals of Applied Statistics*, 2011, volume 5, 1456-1487)



## Slide 6

And a lot of this work took place as part of the Ph.D. dissertation of Saijuan Zhang at Texas A&M, which appeared this year in *Annals of Applied Statistics*. She is now working at Merck in the pharmaceutical industry.



# BACKGROUND AND GOALS

## Slide 7

Okay, so the first thing we want to talk about is some background and some of our goals, but then we'll go on to the questions about what we're doing and our methodology and results.

# Past lectures

- Previous lectures have considered the relationship of dietary intakes and health effects
- Previous lectures have talked about distributions of dietary intakes

## Slide 8

So the previous lectures have considered the relationship of dietary intakes with health effects and they've talked about distributions of dietary intakes.

## Past lectures

- In those past cases, and in this talk, when understanding the relationship of health effects and dietary intakes, it is typical that 24HR is available only in a small sub-study
- Future talks will discuss web based instruments; it will become possible to use 24HR recall in large health effects studies

## Slide 9

In those cases and in this talk, when you're trying to understand the relationship of health effects and dietary variables or when you're trying to do nutritional surveillance, it's typical that a 24 hour recall is available. In the health effects thing, it would only be available in a small substudy or calibration study.

Future talks are going to discuss Web-based instruments where it may become possible to use the 24 hour recall in large health effects studies.

## Past lectures

- In these past lectures, the number of dietary components have been small
- In this talk, the number will be large
- The highly multivariate nature of diet analysis necessitates a different approach

## Slide 10

In the previous lectures, the numbers of dietary components have been reasonably small, and the difference between this lecture and the others is that in my instance the number will be reasonably large and will be a multivariate problem. And our thinking has always been that the highly multivariate nature of diet analysis needs a new approach from what you've seen before.

## The 24HR recall

- The 24HR is a good measure of intake on a single day, but as a measure of usual intake it does not account for day-day variability
- The sample mean 24HR value can be used as an estimate of the population mean usual intake
- The sample distribution of 24HR is not a good estimate of the population distribution of usual intake

## Slide 11

So part of our background is that the 24 hour recall is a good measure of intake on a single day but it's not a very good measure of long-term intake because it doesn't account for day-to-day variability. The sample mean of the 24 hour recall values across a sample can be used as an estimate of the population mean. But the sample distribution of the 24 hour recalls is not a good estimate of the population distribution. And I'll repeat some slides that point this out again.

# Context

- There will eventually be other instruments that capture dietary intake on a single day
- Nutritionists want to understand longer term average daily intake, not intake on 1 day
- We call this usual intake
- This is needed for both epidemiology and for surveillance

## Slide 12

So the context here is there are going to eventually be other instruments that capture dietary intake on a single day, and nutritionists are going to want to understand longer-term, average daily intake, not intake on one day, which we call usual intake. And we want to use this both in epidemiology and for nutritional surveillance.

# Context

- The data we use are from the NHANES 2001-2004 survey of children aged 2-8 in the U.S.
- The data set has 2,638 children with a 24HR
- There are 1,103 with two 24HR
- This is a real survey, and survey weights are incorporated into the analysis (details skipped)

### Slide 13

The data we're going to use are from the NHANES 2001-2004 survey of children age 2 to 8 in the U.S. The data set that we're working with has a little more than 2,600 children who have 24 hour recalls; 1,100 of them have two 24 hour recalls and the other 1,500 only have one 24 hour recall. It's a real survey, a real sample survey, and survey weights are incorporated into the analysis, but I'm not going to talk about them because they are just technical details that are of less interest.

## Context

- Dietary quality indices are an appealing way to summarize the multivariate nature of diet
- Later, we will define and discuss one such index, the Healthy Eating Index – 2005 (HEI-2005)

## Slide 14

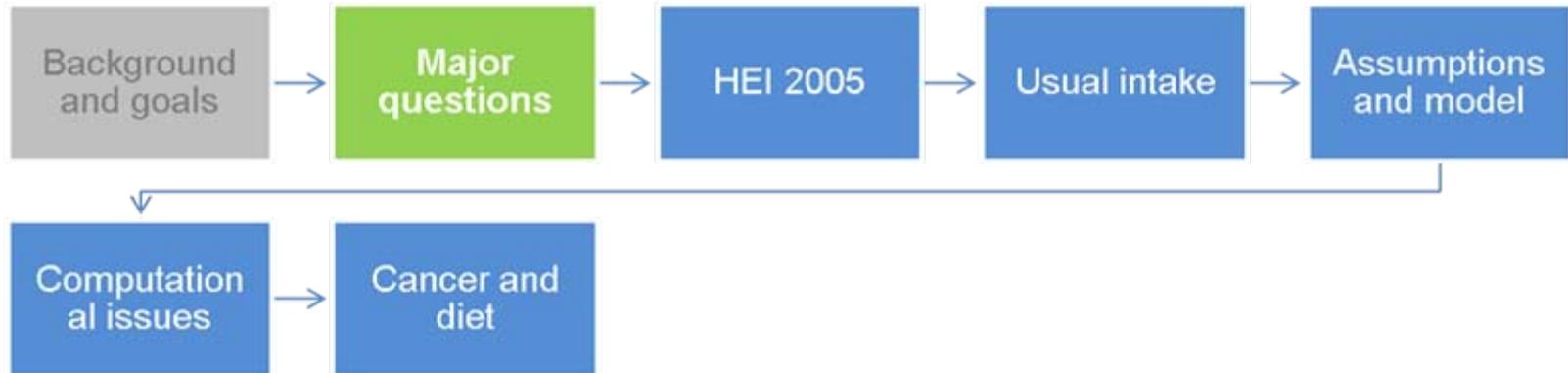
Okay, the context we're dealing with is about dietary quality indices, which is an appealing way to summarize the multivariate nature of diet. And I'm going to define and discuss just one such index, the one we've been working with, the Healthy Eating Index 2005, or HEI-2005. And I'll try to be careful because there are other HEIs floating around, so I'll emphasize the 2005 HEI.

# Context

- Our goal is to give realistic estimates of dietary intake distributions and dietary quality indices that account for the day-to-day variability
- We also want to estimate the real relationship between nutrition and health outcomes, while accounting for day-to-day variability
- We focus on the first problem, but also do an analysis on the latter

## Slide 15

The first goal here is to give realistic estimates of dietary intake distributions and dietary quality indices that account for the day-to-day variability. And then we also want to estimate the real relationship between nutrition and health outcomes while accounting for day-to-day variability in 24 hour recalls. So I'm going to focus mostly on the first problem, the surveillance problem, but we'll also do an analysis of the epidemiology problem at the end.



# MAJOR QUESTIONS

## Slide 16

So here are some questions that we want to address. There are basically two questions.

## Major questions about usual intake

- What is the distribution of the usual dietary pattern scores, such as the HEI-2005 or the Mediterranean index?
- What is the relationship of usual dietary pattern scores and health outcomes?

## Slide 17

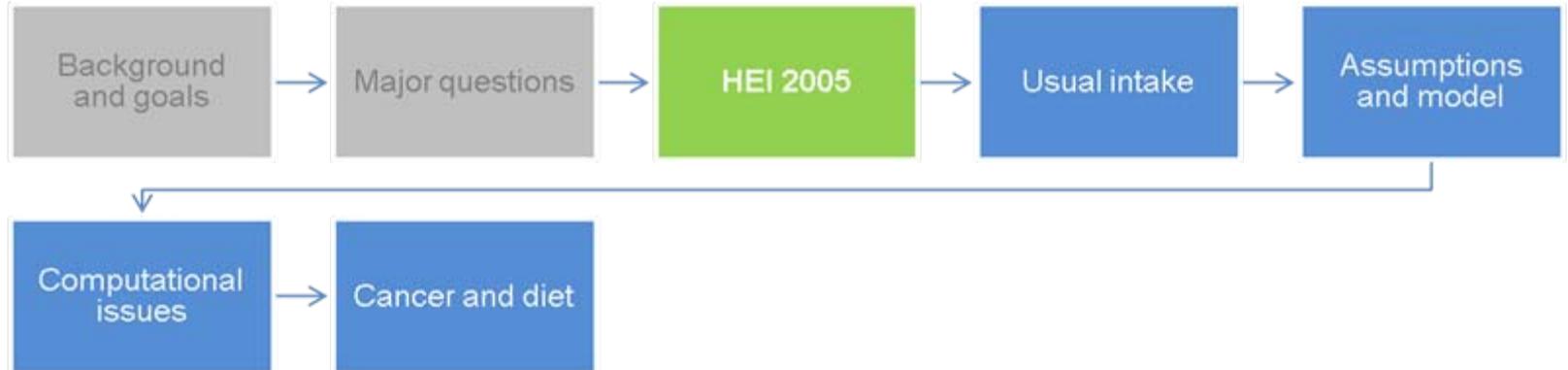
One is: What's the distribution of the usual dietary pattern scores, such as HEI-2005 or something like the Mediterranean index? Now, the data I have are only for HEI-2005 but the methodology applies to other indices. And, second, the relationship of the usual dietary pattern scores and health outcomes is of interest.

## Background Point

- Many researchers are developing new tools for dietary assessment
- One is the web-based ASA24, although this is just one example
- The future holds hope for being able to do multiple 24HR or other measures on an individual.
- However, large surveys such as NHANES will typically only have at most two 24HR

## Slide 18

I should point out that there is a very exciting thing, as many of you know, going on in developing new tools for dietary assessments. The Web-based ASA24 developed at NCI is one example but it's not the only example. And it's exciting to think that in the future we're going to be able to get multiple 24 hour recalls or multiple other measures on individuals, and that will help us in all sorts of problems in nutrition. But large surveys such as NHANES will typically have, at most, two 24 hour recalls. So even in the future, we'll still be dealing with this problem.



# HEALTHY EATING INDEX 2005

## Slide 19

All right, so let me tell you about the HEI-2005.

# The Healthy Eating Index 2005

- The HEI-2005 is a composite score based on the Dietary Guidelines for Americans (USDA and HHS)
- This is just one example of a multivariate dietary pattern score
- It is a multi-component dietary quality index involving energy-adjusted values of dietary components (you have heard about energy adjustment previously)
- It ranges from 0 to 100, with higher scores better

## Slide 20

It's a composite score based on the USDA and HHS Dietary Guidelines for Americans. It's, as I said, one example of a multivariate dietary pattern score. I'm going to go through a bit to show you how multivariate it really is. And it has multiple components, each of which is adjusted for energy. You heard about energy adjustment previously. So all of the components of diet that are measured in HEI-2005 will be adjusted for energy. And the final score, the total score, ranges from 0 to 100, and you want to be at 100 and you don't want to be at 0. I'm closer to 100 than 0, that's for sure.



# Healthy Eating Index—2005

**THE HEALTHY EATING INDEX (HEI)** is a measure of diet quality that assesses conformance to Federal dietary guidance. The original HEI was created by the U.S. Department of Agriculture (USDA) in 1995. Release of new Dietary Guidelines for Americans in 2005 motivated a revision of the HEI. The food group standards are based on the recommendations found in My Pyramid (see Britten *et al.*, *Journal of Nutrition Education and Behavior* 38(6S) S78-S92). The standards were created using a density approach, that is, they are expressed as a percent of calorie or per 1,000 calories. The components of the HEI-2005 and the scoring standards are shown below.

## Health Eating Index—2005 component and standards for scoring

Component	Maximum points	Standard for maximum score	Standard for minimum score of zero
Total Fruit (includes 100% juice)	5	$\geq 0.8$ cup equiv. per 1,000 kcal	No Fruit
Whole Fruit (not juice)	5	$\geq 0.4$ cup equiv. per 1,000 kcal	No Whole Fruit
Total Vegetables	5	$\geq 1.1$ cup equiv. per 1,000 kcal	No Vegetables
Dark Green and Orange Vegetables and Legumes	5	$\geq 0.4$ cup equiv. per 1,000 kcal	No Dark Green or Orange Vegetables or Legumes
Total Grains	5	$\geq 3.0$ oz equiv. per 1,000 kcal	No Grains
Whole Grains	5	$\geq 1.5$ oz equiv. per 1,000 kcal	No Whole Grains
Milk	10	$\geq 1.3$ cup equiv. per 1,000 kcal	No Milk
Meat and Beans	10	$\geq 2.5$ oz equiv. per 1,000 kcal	No Meat or Beans
Oils	10	$\geq 12$ grams per 1,000 kcal	No Oil
Saturated Fat	10	$\leq 7\%$ of energy	$\geq 15\%$ of energy
Sodium	10	$\leq 0.7$ gram per 1,000 kcal	$\geq 2.0$ grams per 1,000 kcal
Calories from Solid Fats, Alcoholic beverages, and Added Sugars (SoFAAS)	20	$\leq 20\%$ of energy	$\geq 50\%$ of energy

## Slide 21

This is a picture I got from the Web that the USDA put out, and it lists some of the variables. You can see total fruit, whole fruit, total veggies, DOLs—I'll combine all these—total grains, whole grains, etc. There are 12 components in the HEI, plus the adjustment for energy. And there's a formula here but I'll talk about it in bigger font.

# The Healthy Eating Index 2005

- Some abbreviations follow
- A component of HEI-2005 is dark green and orange vegetables and legumes, which we call **DOL**
- Another component is calories from solid fats, alcoholic beverages and added sugars, which we call **SoFAAS**
  - These might be thought of as “empty calories”

## Slide 22

I'm going to call DOLS dark green and orange vegetables and legumes. And the other thing that has an acronym is calories from solid fats, alcoholic beverages, and added sugars, which are the infamous SoFAAS, which might be thought of as empty calories.

# The Healthy Eating Index 2005

## Maximum Score=5

- Whole fruit
- Total fruit
- Whole grains
- Total grains
- DOL
- Total vegetables

## Maximum Score=10

- Milk
- Meat and beans
- Oils
- Saturated fat
- Sodium

## Maximum Score=20

- SoFAAS

## Maximum Score=100

## Slide 23

And the Healthy Eating Index, in its 12 components has various maximum scores, so the fruits, the grains, the DOLs, and the total vegetables have a maximum score of 5. Milk, meat and beans, oils, saturated fat, and sodium have a maximum score of 10. And [for] SoFAAS, you can get a maximum score of 20. And so the total—if you get a maximum score on everything, you get 100.

# The Healthy Eating Index 2005

- The scores assigned to each component are nonlinear functions because of truncations
- Total fruit for example is measured as

$$\text{Adjusted Total Fruit} = \frac{\text{Cups}}{\text{Energy}/1000}$$

- The score increases linearly up to 0.8 equivalents per 1,000 kilocalories with a maximum score of 5, and does not increase with intakes above 0.8 cup equivalents per 1,000 kilocalories

## Slide 24

And the formulas look like this. So total fruit is one of the components, and what we do is we adjust total fruit by kilocalories, so that's the adjusted total fruit. And then the score increases linearly up to 0.8 equivalents per 1,000 calories, or 1 kilocalorie, with a maximum score of 5. And it's not linear because what happens is it's nice and linear up to 0.8 cup equivalents. Then, after that it's flat; the score stays at 5. So it has a change point as it goes along. And all the components are like this in the HEI.

# The Healthy Eating Index 2005

- For saturated fat, energy adjusted intake is the percentage of energy from saturated fat

## Slide 25

For saturated fat, energy intake is the percentage of energy from saturated fat.

# The Healthy Eating Index 2005

- The HEI-2005 score for energy-adjusted saturated fat is:

$$= 0 \quad \text{if } \geq 15$$

$$= 10 \quad \text{if } \leq 7$$

$$= 8 - 8 * (\text{density} - 10) / 5 \quad \text{if } > 10 \text{ but } < 15$$

$$= 10 - 2 * (\text{density} - 7) / 3 \quad \text{if } > 7 \text{ but } < 10$$

## Slide 26

And it has a fairly complicated way of measuring things. If you're below 7 percent of your calories coming from saturated fat, you get the maximum score. If you're above 15 percent, you get 0, and in between there, there're these formulas, which are sort of hard to sort out without a graph. But that's how they all work. All of these components have this little bit of nonlinearity in them.

# The Healthy Eating Index 2005

- The HEI-2005 score for SoFAAS as a percentage of energy is:

= 0 if  $\geq 50$

= 20 if  $\leq 20$

= linearly interpolated otherwise

## Slide 27

This is the empty calories. If your percentage of energy from SoFAAS is more than 50 percent, you get a score of 0. If it's less than 20 percent, you get a score of 20. And it's linearly interpreted otherwise.

I could go through all the other nine, but you get the idea of what it's like.

# The Healthy Eating Index 2005

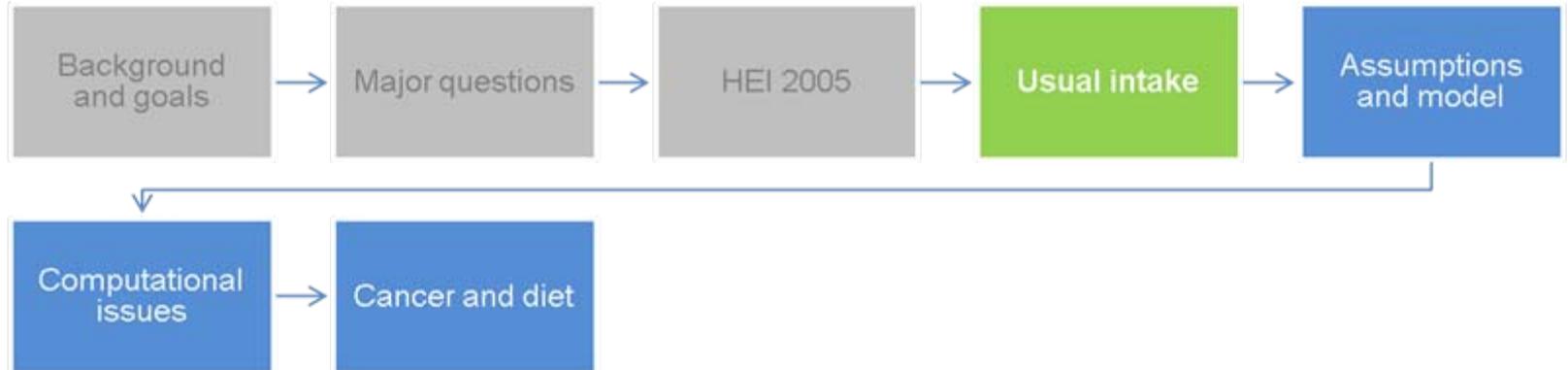
- Now we focus on long term HEI total score, not short term

$$\text{Adjusted Whole Fruit} = \frac{\text{Cups}}{\text{Energy}/1000}$$

- In this formula, “Cups” is long-term daily average number of cups consumed
- “Energy” is long-term daily average number of calories consumed
- These are called **Usual Intakes**

## Slide 28

And now what I want to do is talk about long-term HEI total score. I don't want to talk about short-term HEI total scores based on a single 24 hour recall. So the long-term adjusted whole fruit is going to be the long-term daily average number of cups of whole fruit consumed divided by the long-term daily average of kilocalories. And these are going to be my usual intakes, which will lead to the long-term HEI total score.



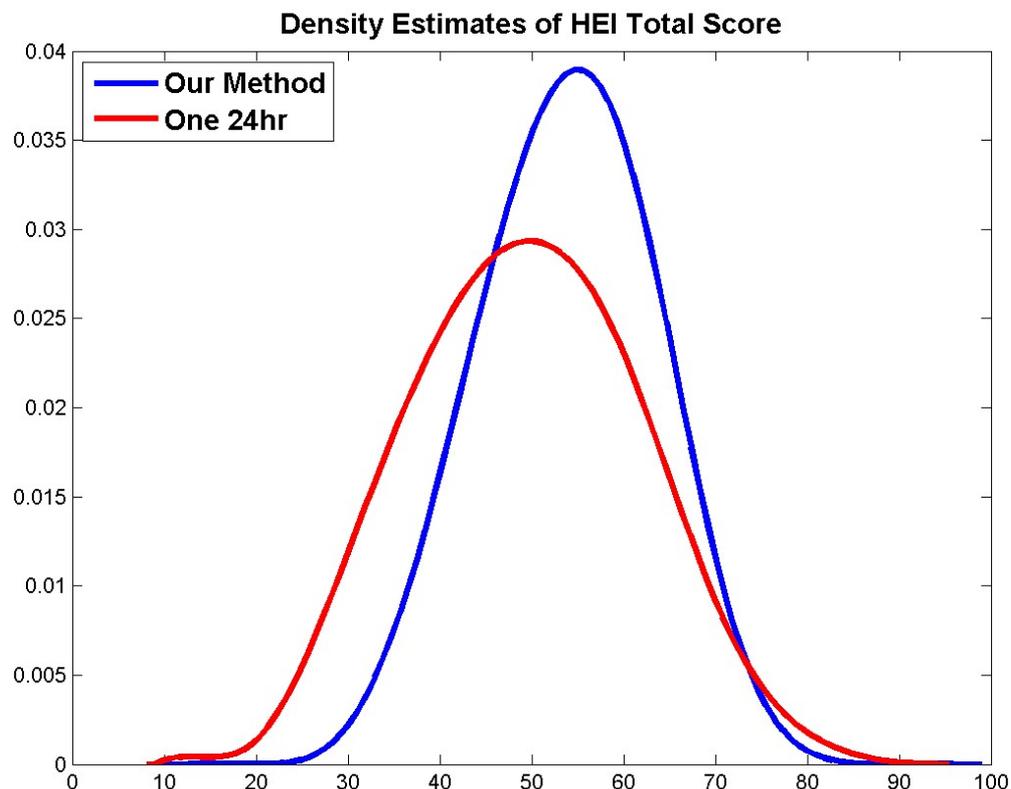
# RESULTS FOR THE DISTRIBUTIONS OF USUAL INTAKES

## Slide 29

So that's the HEI total score. Here are some results from that HEI total score.

## Results: HEI total score density

- This is the actual result; the 24HR overestimates the % with diet scores  $< 30$  and overestimates the % with diet scores  $> 80$



## Slide 30

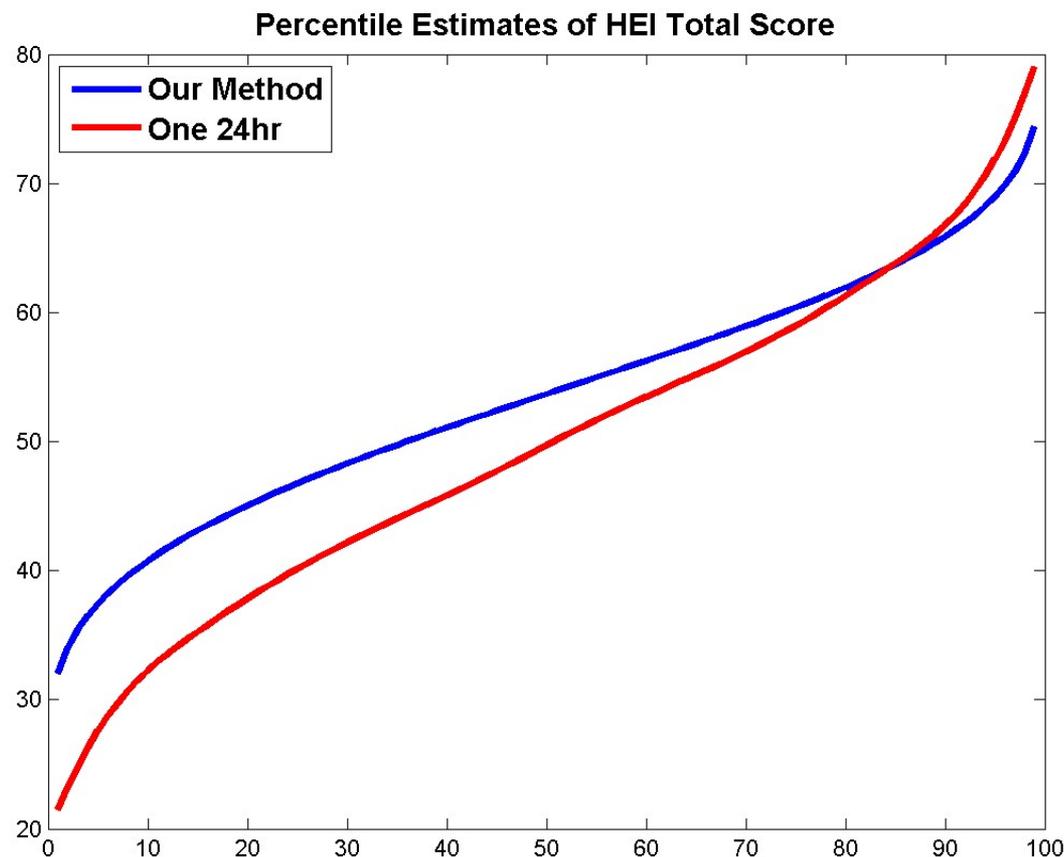
Instead of introducing the methods first and getting to the results, I thought I'd give you the results first and then the methods.

This is the actual result in children age 2 to 8 in the United States for the total score. And in red is the total score, the distribution of the total score. This is a histogram. These are smooth histograms. And it's the distribution; the red is that of the single 24 hour recall, and the blue is our method, which adjusts for the day-to-day variability in 24 hour recalls.

And you can see that what happens with a single 24 hour recall is that you're overestimating the percentage of diet scores less than 30 and you're also overestimating the percentage of diet scores which are greater than 80. So there are biases in both directions.

## Result: HEI total score percentiles

- Notice that 8% of children have an HEI-2005 total score < 40; the single 24HR says 25% do

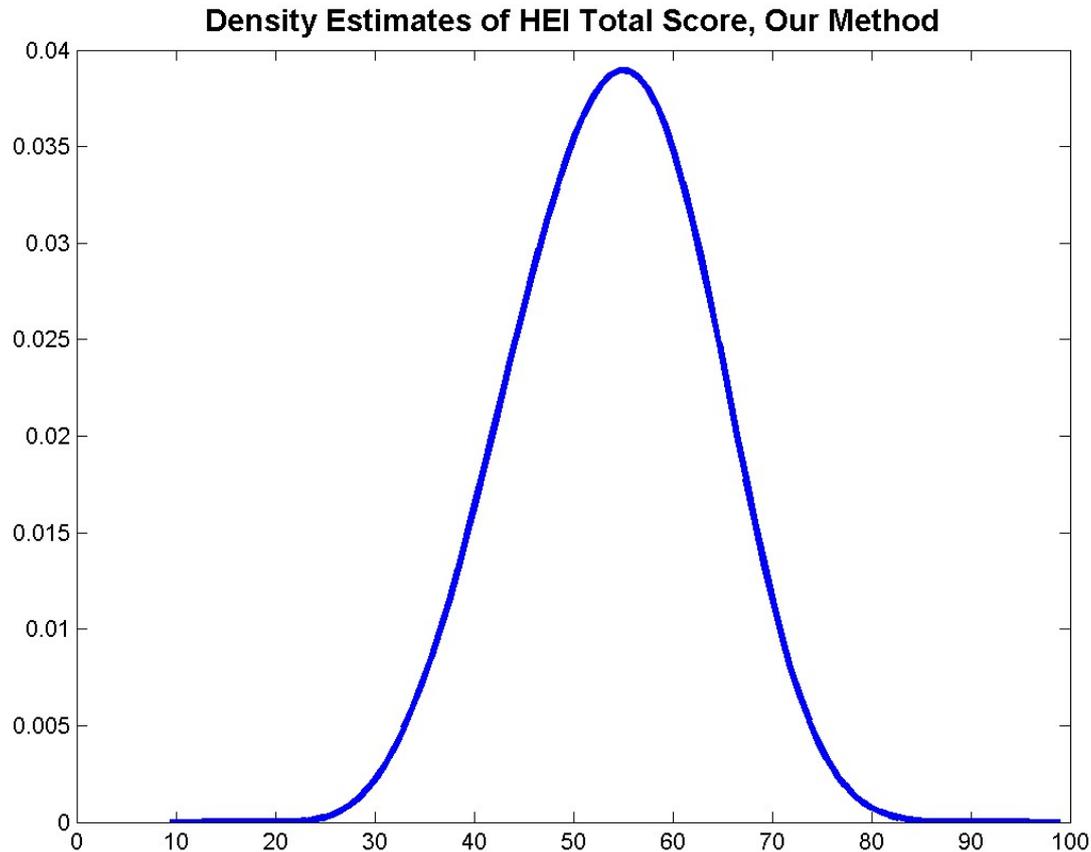


## Slide 31

Here is a picture of not a very good plot, really. If I'm looking for the 50<sup>th</sup> percentile, I will try to do this. There we are. I'll put it right here. That's the 50<sup>th</sup> percentile. I just go up and then read off. So the median for the 24 hour recall is about 50 and the median for our method in blue is also a little bit more than 50. I think the crucial thing to see in this graph is that we estimate that 8 percent, which is still a pretty large number, but 8 percent of children have an HEI-2005 total score, of less than 40, whereas the single 24 hour recall says 25 percent have a total score of less than 40. So it's quite a big discrepancy. And the other thing happens on the other end in high scores.

## Result: HEI total score density

- Essentially no children in the U.S. have a total HEI score of greater than 80



## Slide 32

And literally now this is just our estimate of the distribution of the total scores, the histogram of the total scores. And essentially no children in the U.S., we estimate, have a total HEI-2005 score greater than 80.

## A vignette

- Recently, the White House Task Force on Obesity was considering a goal that all children would have a HEI-2005 usual intake total score  $> 80$
- The 99th percentile = 79.4

### Slide 33

And that's sort of interesting because a year or year and a half ago the White House Task Force on Obesity was considering a goal that all children would have an HEI-2005 usual intake total score of greater than 80. And we estimate that the 99<sup>th</sup> percentile is actually less than 80; it's 79.4 is our estimate.

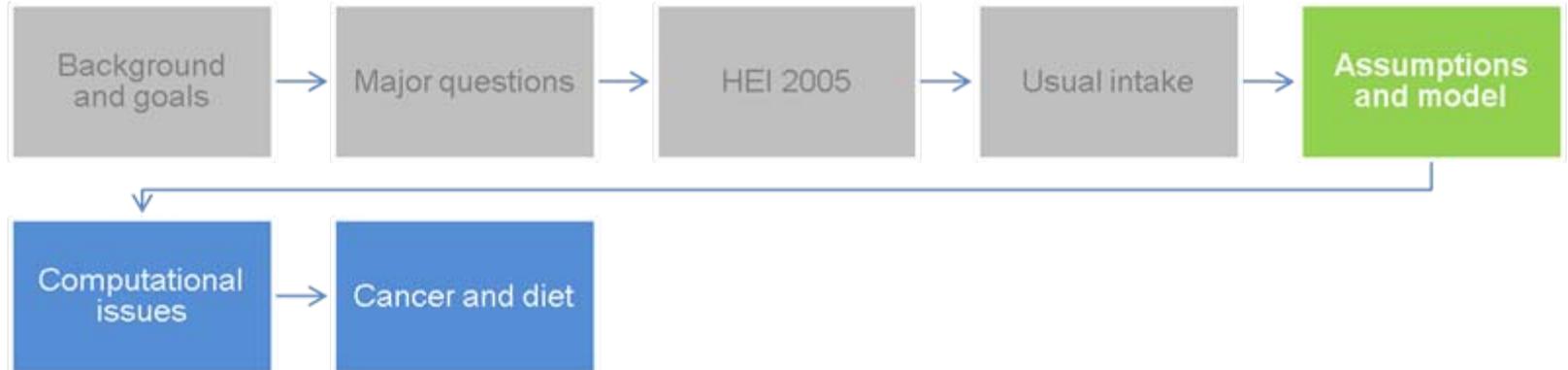
## A vignette

- Given our results and other information, the Task Force changed its goal to have children move to a mean of 80

### Slide 34

So when they heard about our results and got other information, the Task Force changed its goal to have children move to a mean of 80.

Let's go back. It's still ambitious because we estimate that the mean is 53, a little bit more than 53. So to meet this goal is ambitious and hopefully it can be done in the near future.



# ASSUMPTIONS AND MODEL

## Slide 35

Okay, so those are the results. The basic point of the results is that using a single 24 hour recall can result in some very large biases when you're looking at the distributions of intakes and HEI total scores.

## Modeling assumption

- Assumption: 24HR's are unbiased measures of usual intake on a given day
- This fixes discussion and states that 24HR's pretty accurately reflect a single day's intake
- The next few slides are a repeat of what you have seen previously, but still important

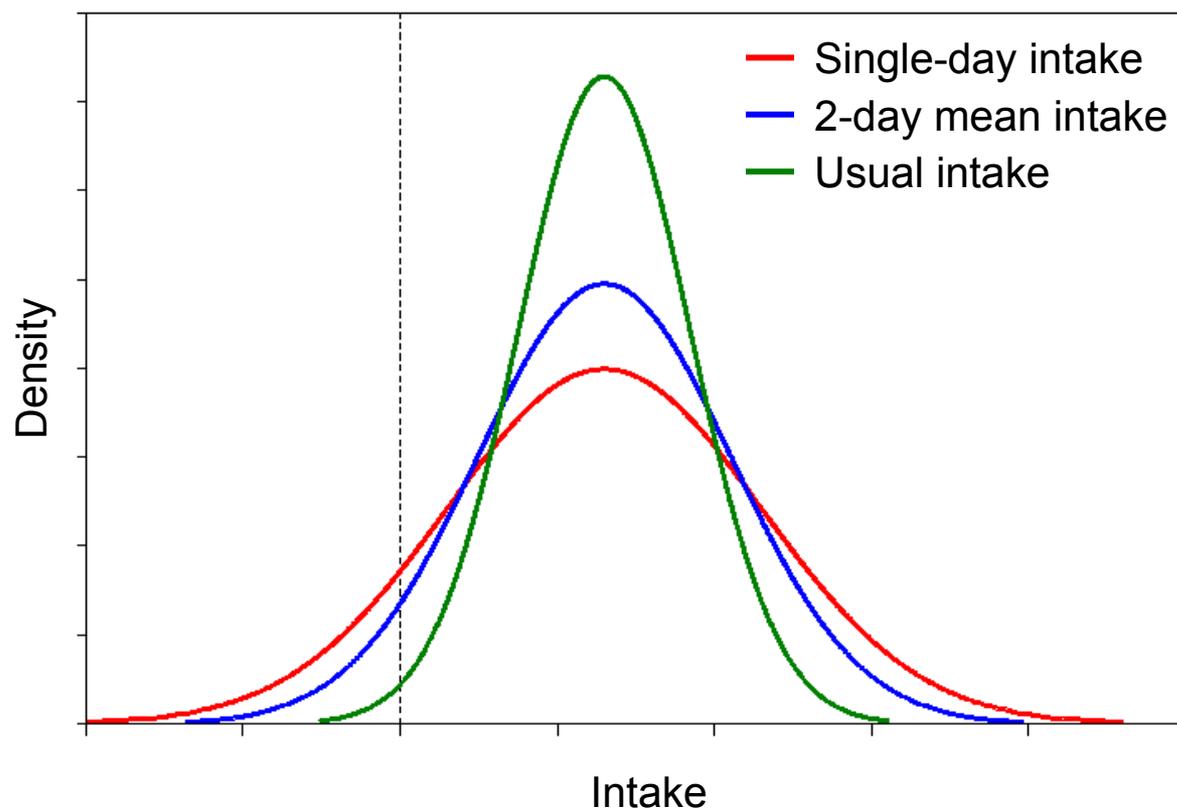
## Slide 36

Our assumption that we've made is that 24 hour recalls are unbiased measures of usual intake on a given day. And it fixes the discussion and makes it statistically feasible, and it also basically states that 24 hour recalls pretty accurately reflect a single day's intake. But what we want to do is take into account the day-to-day variability.

And I'm going to repeat a few slides that you've seen previously, in at least Janet Tooze's talk, but they are still important and they're still useful to remember what's going on here.

# Nutrient data do NOT look like this

- This classical picture points out though that day-to-day variability makes the 24HR recall more variable than usual intake

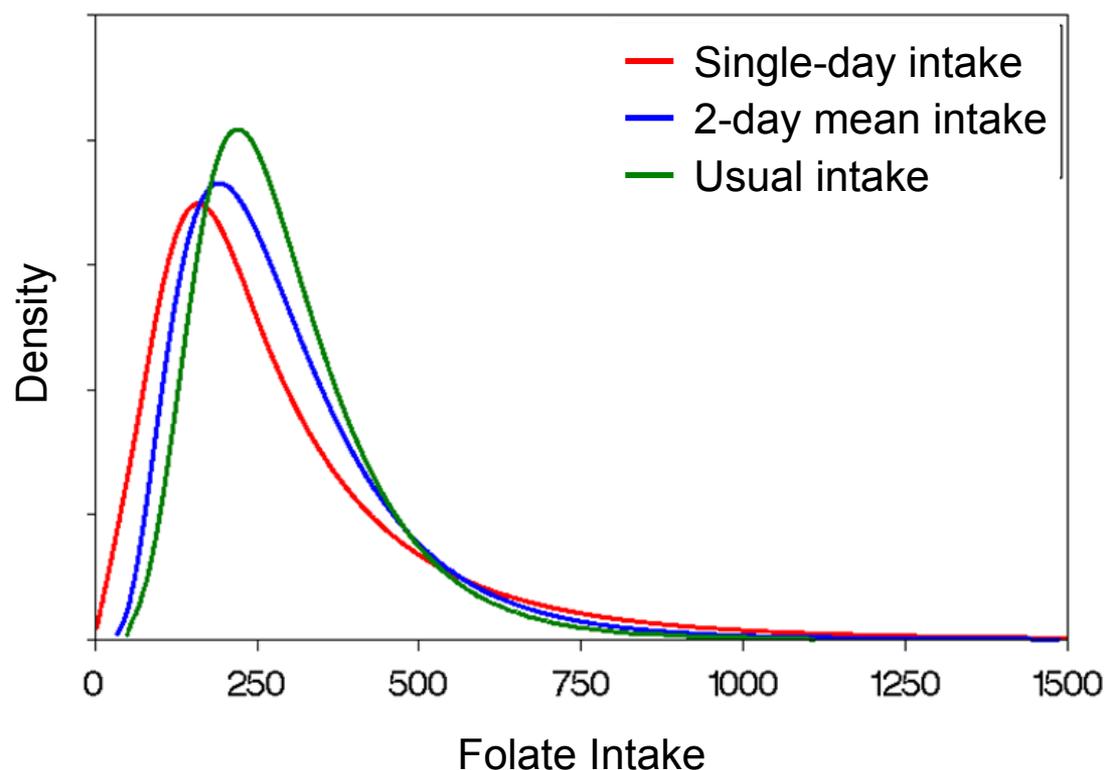


### Slide 37

So the first thing is that nutrient data don't look like this. There's a large literature on measurement error and measurement error modeling and corrections, and they are often based on the idea that the data are beautifully bell shaped; in particular, such as the nutrient intake data. And that's not the case in general, but this is the beautiful picture where in green is the usual intake, in red is a single day's intake. And the point of this picture is that while the two methods have the same mean, the single day's 24 hour recall is much more variable than usual intake.

# Folate: right-skewed distributions

- Note here that a single 24HR is shifted left compared to usual intake, although the means are the same due to some unusually high days of intake



### Slide 38

This is a picture you've seen from the Eating at America's Table Study for folate intake. The green, again, is usual intake; red is a single day's intake. And you'll notice once again that the single day's intake has shifted to the left, so it's much more right skewed than usual intake. It has more values that are low and it has more values that are high, just as we've seen in the HEI-2005 total scores.

# Transformations

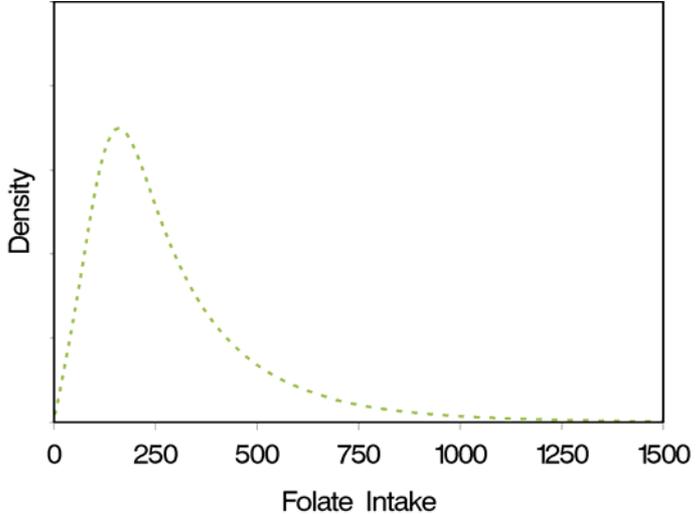
- To deal with the skewness, it is typical to transform the data so that day-to-day variation has a nice Gaussian-like distribution
- One analyzes in this transformed scale, and then back-transforms to the original nutrient scale
- Here is an illustration

### Slide 39

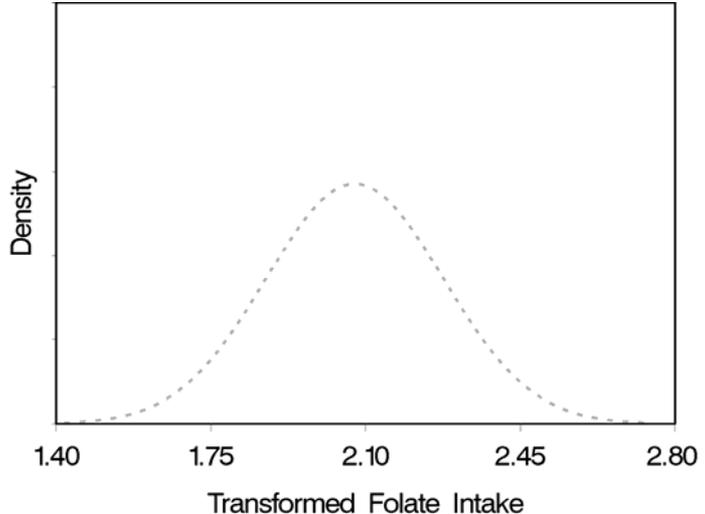
And so to deal with the skewness, we try and go back to that beautiful picture I had, the nice bell-shaped curve. Everybody loves nice bell-shaped curves, Gaussian-like distributions. And so what we do, and it's typical to do this, is we transform the scale of the nutrient, do a measurement error correction, and then back-transform to the original nutrient scale.

# Accounting for nonlinear transformations

Original Scale

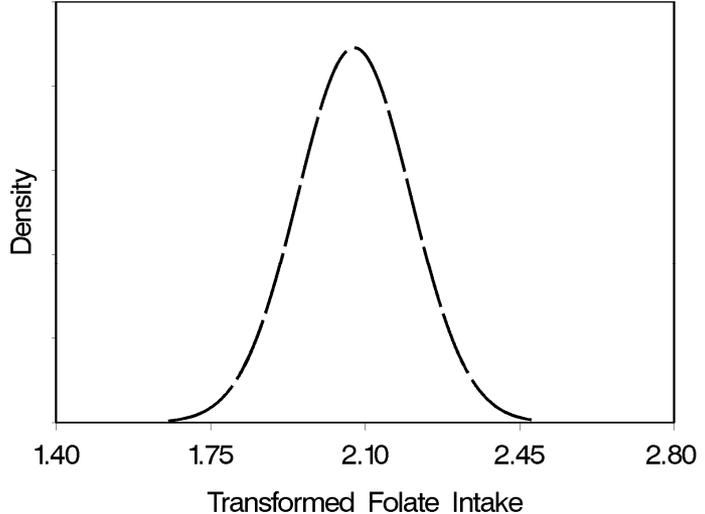
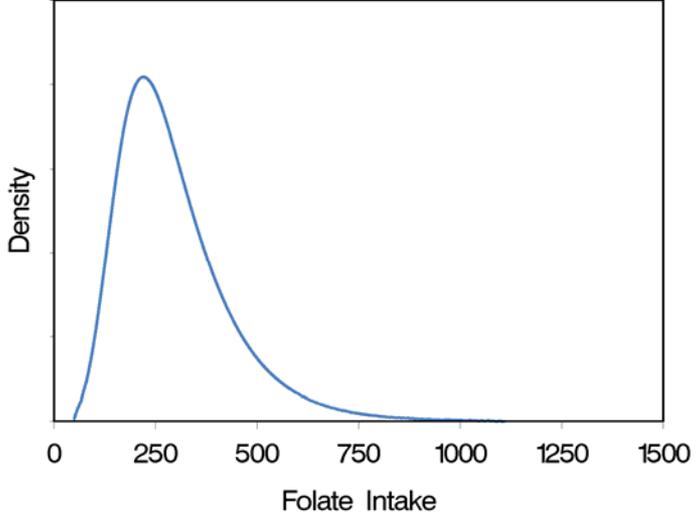


Transformed Scale



Folate Intake

Transformed Folate Intake



## Slide 40

So this is an illustration; you've seen this before. But this is what we do. We look at that folate intake distribution. We say, "Oh, my goodness; that's ugly. It's skewed." We then transform it so it looks like it's bell shaped, but that's a single day's intake so it's too variable. We then do a measurement analysis in this transformed scale, get still a bell-shaped curve, but a lower variability. And then we back-transform to the original scale so we're talking about distributions of folate intake, not distributions of transformed folate intakes. And we're going to do this on all 12 of the components of the HEI-2005.

## Episodically consumed foods

- The HEI-2005 has 6 components that are episodically consumed
- Among children aged 2-8 in the U.S., here are the percentages of reported non-consumption on a 24HR

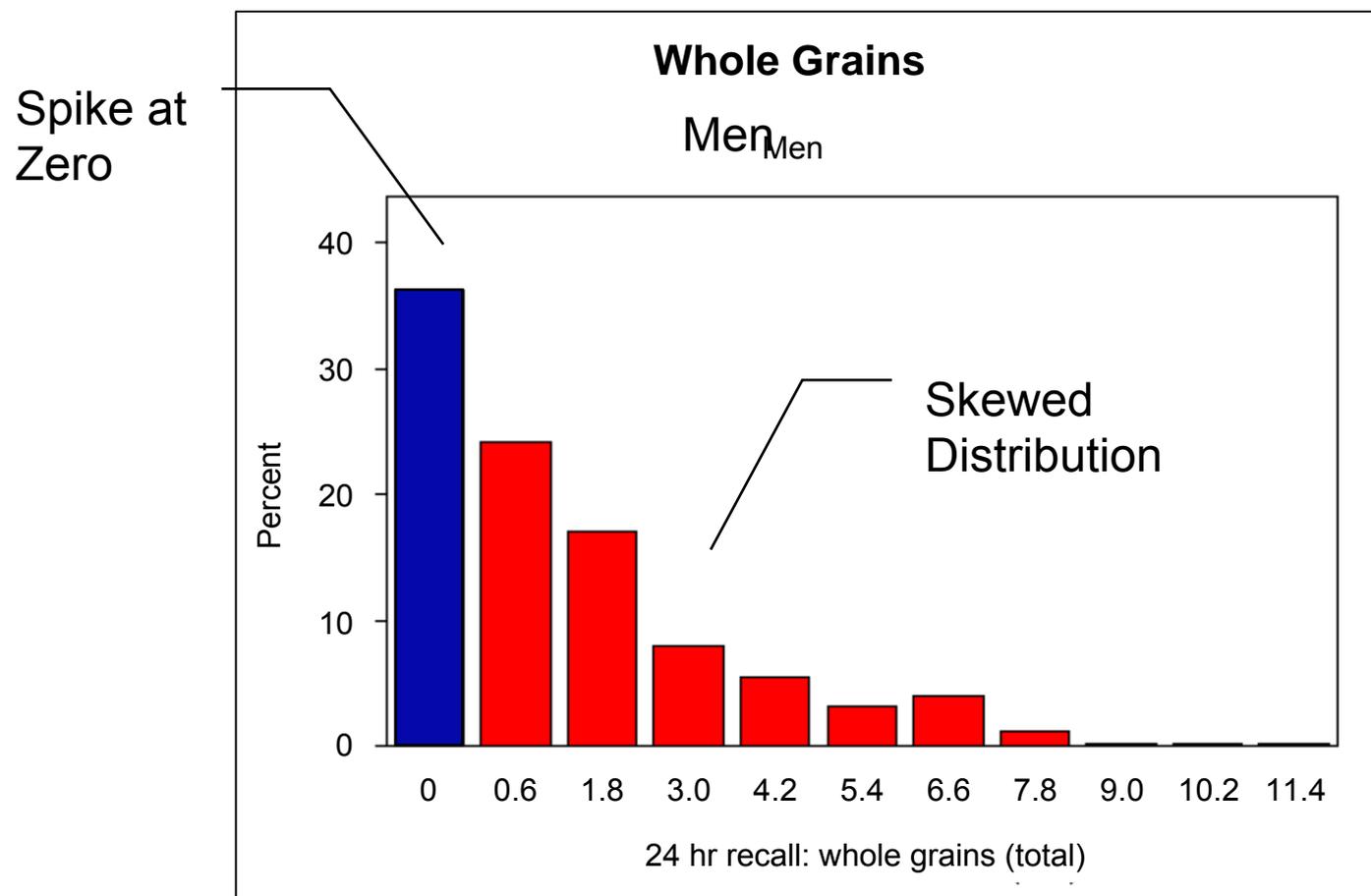
Total fruit	17%
Whole fruit	40%
Whole grains	42%
Total veggies	3%
DOL	50%
Milk	12%

## Slide 41

The thing that makes HEI-2005 challenging is that it has episodically consumed components and it has lots of them. So in this picture, you can see here this is the percentage of 24 hour recalls in the NHANES study for children where they report no intake of these six quantities. So 40 percent of the 24 hour recalls in the data set report no intake of whole fruit; 42 percent, no intake of whole grains; and 50 percent, no intake of DOLs; and total fruit and milk and total veggies are lower. But this is a very challenging problem. It's a really hard problem because there are six of these things and, in general, there can be even more that are episodically consumed.

# Challenges to estimation – foods

- Observed food intakes are often zero



## Slide 42

So what you get—another picture that you've seen—this is a whole grains picture for men from EATS. What you get is a spike at zero where there is reported no whole grain intake, and then on consumption days, a very right-skewed distribution. So it's the combination of the blue and the red that makes it a fun statistical problem. We want to take into account the fact that these zeros are overrepresented for usual intake.

## Model for a single food

- For a single food, as in a previous lecture, we have developed a flexible modeling framework, which we call the NCI Method
- For SAS programs based on NLMIXED, see <http://riskfactor.cancer.gov/diet/usualintakes/>
- For the HEI-2005 analysis, we are building upon our earlier work that was done at NCI and other sites; the previous lecture discussed one episodically consumed component plus energy

### Slide 43

And for a single food, we developed a flexible modeling framework, which we call the NCI method. Even though I'm at Texas A&M and was a collaborator on this, most of the work was done at NCI, and this is a reasonable thing to call it.

It's based on the SAS program called NLMIXED, nonlinear mixed models, and as you know, you can get it from the Web.

So in the HEI-2005 analysis, what we're doing, really, is trying to go from a single episodically consumed dietary component, which the NCI method can handle, to multiple dietary components, which the NCI method cannot handle.

## Need for Multivariate Model

- It is possible to get estimates of the distribution of each energy-adjusted dietary component and each HEI-2005 dietary score component, **SEPARATELY**
- This approach allows estimating the mean of the HEI total score in a population
- It does not allow estimation of **percentiles** of the HEI-2005 total score
- Percentiles require a multivariate model

## Slide 44

It's important to understand that we actually need a multivariate model. I can run the NCI method on every dietary component in the HEI-2005 separately; it works great and we've done this. And that allows you to estimate the mean of the HEI total score in a population. But by running it separately and then adding up things, it doesn't get you the distribution of the total scores. The only thing you can get by running the NCI method 12+ times is the mean and not the percentiles. And so the percentiles need the multivariate model.

## So, what's the big deal?

- HEI is complex, because it has 6 episodically-consumed foods, 6 daily-consumed foods and nutrients, and energy

## Slide 45

And I just think it's a really cool problem for statisticians to work on, and it's a really cool problem because the HEI-2005 is an interesting way of measuring dietary consumption.

## So, what's the big deal?

- The bottom line is that when we turn to things like the HEI-2005, we have three problems
- Problem #1: The dimensionality of the integration is too great for PROC NLMIXED to run as a computer program, because of the many dimensions of diet quality
- So, we're stuck: without a new approach, software does not exist to analyze the HEI-2005

## Slide 46

So we have to do other things now. That's the bottom line. There is no current way to analyze the HEI-2005 to do distributions of usual total scores. The NLMIXED cannot do these things. It's technically infeasible for them to use it. And so we need a new approach. I'll talk more about why it's technically infeasible.

## So, what's the big deal?

- Problem #2: Figure out a model that can allow analysis of HEI-2005
- Problem #3: Compute!

## Slide 47

So what we have to do is figure out a model that allows analysis of the HEI-2005 total score, and we have to figure out how to compute those things.

I'm going to spend a little bit of time on the model and less time on the computation, which is highly technical.

# A multivariate model

- Here,  $i$  will denote person
- Also,  $j$  will denote replicate of the 24HR
- Finally,  $k$  will denote an index
- There are 6 episodically consumed dietary components
- There are 6 daily consumed components
- There is also energy
- I will illustrate in the case of 2 foods and energy

## Slide 48

We use the same notation as from other talks:  $i$  is going to be a person and  $j$  is going to be a replicate of the 24 hour recall;  $k$  is going to be an index, depending on which dietary component you have. There are six episodically consumed dietary components, six daily consumed, plus there's energy. So  $k$  could, in principle, run to be a pretty big number. What I'm going to do, because we'd need a 90 inch-screen to do the whole thing, is I'm going to just do two foods and energy and show you what the model looks like.

## A multivariate model

- I will just do 2 foods plus energy here, and briefly mention what happens with many foods, nutrients and energy
- We have to formulate the consumption model to allow day-to-day energy to be correlated with day-to-day consumption
- We use a choice-based probit model for this task

## Slide 49

What we have to do is formulate consumption to allow day-to-day energy to be correlated with day-to-day consumption, but we need to know: What do we mean by consumption? What do we mean by a distribution on consumption days?

## Multivariate model

- Generically,  $X$  will denote covariates
  - Demographics
  - Food frequency questionnaire if available
- Generically,  $u$  denotes how people with the same covariates differ from one another in their long term intake
- Finally,  $\varepsilon$  will denote day-to-day variability

## Slide 50

So first of all, I'm going to say that, generically,  $X$  is going to denote covariates—demographics or the food frequency questionnaire if it's available. And I'm going to make little  $u$  be the person-specific biases. These are a really crucial concept that you've seen before; namely, that even if I and another person share the same demographics, we're going to have different long-term intakes because we're different. And so this acknowledges that people can't—their dietary intakes can't be just described by demographics. And then, finally, epsilons are going to be the day-to-day variability that floats around in here.

# Consumption?

- For  $k=1,3$ , define a latent variable

$$W_{Fijk} = X_i^T \beta_k + u_{ik} + \varepsilon_{ijk}$$

- Consumption of the food for person  $i$  on day  $j$  is distributionally equivalent to a probit model defined through

$$W_{Fijk} > 0$$

## Slide 51

So there are three components. The first thing is consumption. Well, consumption is a funny concept, but we use a technique that comes from econometrics, and that's called choice-based sampling. And what it says is it says that, well, this is a latent variable. I have the X data here; that's the demographics, the effect of demographics. The  $u$  is the effect of the individual person, and the epsilon is the effect of day-to-day variability.

And I will never see this  $W$ , but I'm going to say that consumption of the food is equivalent to this latent variable, this unseen variable, being positive. And this is a really good way to model consumption.

# Amount

- When a food is consumed, it is positive, so we use transformations
- The Box-Cox transformation is denoted by:

$$g_{\text{tr}}(x, \gamma) = \frac{x^{\gamma} - 1}{\gamma}$$

## Slide 52

And when the food is consumed, it's positive; we've seen it's badly skewed. And so we're going to transform the data. We use the Box-Cox transformation, which is the logs, the square roots, the cube roots, etc., which are really very effective methods to transform data to normality.

# Amount

- For  $k = 2, 4$ , we have a second latent variable, involving consumption of the food

$$g_{tr}(W_{Fijk}, \gamma_k) = X_i^T \beta_k + u_{ik} + \varepsilon_{ijk}$$

- We get to observe this latent variable only if there is consumption, i.e., only if

$$W_{Fi,j,k-1} > 0$$

## Slide 53

Then, after we transform the data, we're going to say that on consumption days the transformed data also follow the same kind of model. There is the effect of the demographics, the effect of individual people, and the effect of day-to-day variability. Sometimes, this  $W$  is observed; on a consumption day, it's observed. But it's not observed on a nonconsumption day, so it's latent on a nonconsumption day but we get to see it on a consumption day.

# Covariance Matrices

- A covariance matrix is denoted with the symbol  $\Sigma$
- It describes the variances of each latent variable and their correlations

## Slide 54

And then, finally—well, finally is a little early here. I'm going to denote a covariance matrix with the symbol sigma. We have to have at least one Greek symbol in here other than a beta, and the sigma denotes the variances and the correlations of each of the variables.

# Covariance Matrices

- Our latent variable model for HEI-2005 has 19 components whose variances and correlations need to be modeled  $\Sigma$
- There are 2 for each episodically consumed component, 1 for each daily-consumed component, and 1 for energy
- So, the covariance matrix is of dimension 19

## Slide 55

And, finally, now I'm going to add in energy. Our latent variable turns out in HEI-2005 to have 19 components. So it's a 19-dimensional covariance matrix. There are two variables for episodically consumed components: one for daily consumed components and one for energy. So we have a pretty large dimensional covariance matrix.

# Energy

- Energy (k=5) is always positive, so we observe

$$g_{\text{tr}}(Y_{\text{Eijk}}, \gamma_k) = X_i^T \beta_j + u_{ij} + \varepsilon_{ijk}$$

- We assume that

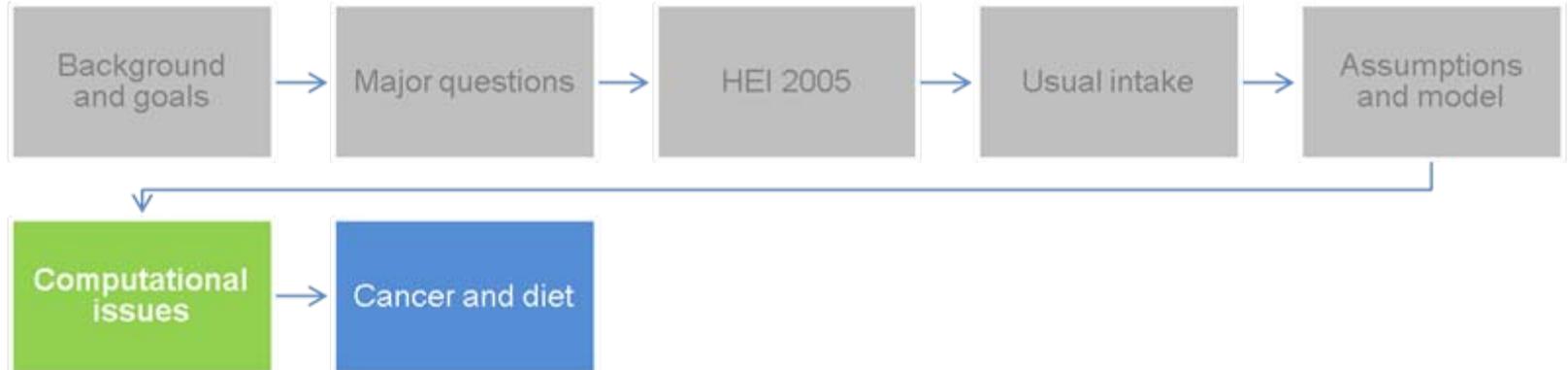
$$(u_{i1}, \dots, u_{i5}) = \text{Normal}(0, \Sigma_u)$$

$$(\varepsilon_{ij1}, \dots, \varepsilon_{ij5}) = \text{Normal}(0, \Sigma_\varepsilon)$$

## Slide 56

And energy, of course, is transformed. It's skewed, and we transform it in the same way with demographics, individual effects, and day-to-day variability. And our model poses that on the transformed scale, the individual effects are normally distributed, and on the transformed scale the day-to-day variability is normally distributed.

And our big model is just a generalization of this to add all 12 components of the HEI-2005.



# COMPUTATIONAL ISSUES

## Slide 57

So having defined a model—and 19 dimensions is a lot of dimensions to work with, but it's feasible to do this and we have some nice results—I want to show you a little bit about computational issues and discuss them.

# Computation and more

- There are many **technical issues** related to fitting the resulting model
- These are of great interest to biostatisticians, but may be less interesting (or boring) for everyone else
- The full details can be found in a paper in the *Annals of Applied Statistics*. See the (8 page!) appendix

## Slide 58

There are a lot of technical issues related to fitting this model for the HEI-2005. And I'm a technical statistician and so I've found it extremely interesting to work with, and Saijuan Zhang got a Ph.D. out of it, but it really would be boring for you, just mind-numbing. In fact, the details of the computations take eight published pages in the *Annals of Applied Statistics*, and I guess I could amuse my friends by showing them, but I won't show them to you.

# Computation and more

- Our model is an example of a nonlinear mixed effects (or random effects) model
- The key point is that because of the 19 components in the model, standard software such as SAS NLMIXED will not run and give answers in my lifetime

## Slide 59

But what it is in the framework of statistics is an example of a high-dimensional, highly nonlinear, mixed effects model with excess zeros caused by the episodic consumption of the six dietary components. It ends up with 19 components in the model, and SAS NLMIXED, which is a very good program if there are only 1 or 2 dietary components, simply will not run. We've tried it, and it doesn't run; it just doesn't give answers. And I don't think it's a matter of having a big enough computer. It's a matter of just the problem is too large so it just doesn't run.

# Computation and more

- The computational issue is that the components of the day-to-day variability, the epsilons, are all correlated
- So too are the components of the individual usual intake, the  $u$ 's
- Maximum likelihood requires integration (area under the curve of a function)

## Slide 60

And the computational issue is that the day-to-day variabilities are correlated, so my consumption of calories and my consumption of SoFAAS are going to be correlated. If I have a lot of SoFAAS, I'm going to have a lot of calories, most likely. And it's also true that how individuals differ from one another—those components, the  $u$ 's, are also going to be correlated, because if I tend to want to have whole fruits, I'm likely to tend to want to have non-whole fruits.

And the standard statistical technique called maximum likelihood requires integration, areas under the curve of a function, but we have 19 dimensions. And it's hard enough doing 1-dimensional integration, but doing 19-dimensional integration is, in this context, not really feasible.

# Computation

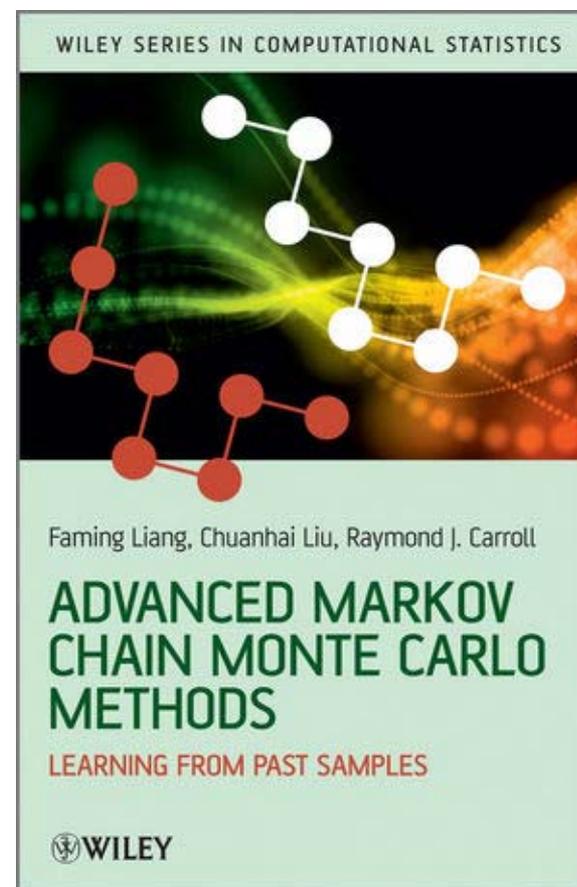
- We have developed a fully parametric model, with transformations and lots of latent variables
- All parameters are free to roam or be on a fixed interval
- Even so, standard software will not work because of the integration

## Slide 61

And while I love our model—we've worked really hard to make it computationally feasible, with parameters that are free to roam about as little happy people, or that are on a fixed interval—standard software doesn't work.

# Computation

- Statisticians have made huge gains in computing integrals using Monte-Carlo techniques
- There is a vast literature, including my book!



## Slide 62

So what did we do? Well, there has been a revolution in statistics and statistical computing in the last 20 years using Monte Carlo techniques. Statisticians, computer scientists, bioinformaticians—they're all moving to Monte Carlo techniques for computation of integrals. There's an enormous literature now on how to do this, and I can't resist advertising a book that my genius colleague, Faming Liang, wrote with myself and Chuanhai Liu, who is at Purdue. So we know a lot about Monte Carlo computation. We know how to use Monte Carlo to do integration.

# Computation

- The most commonly used computational method to do the integration is called **Markov Chain Monte Carlo**
- It generally uses what are called **Gibbs sampling** and **Metropolis-Hastings steps**
- It is an iterative numerical procedure; in this particular case, we had to write our own program to do the computation

## Slide 63

And the most commonly used method is something called Markov Chain Monte Carlo. I've resisted the temptation to say MCMC, which is the term in statistics that people use. And those of you who are statisticians or know a friendly statistician may have heard of Gibbs sampling and Metropolis-Hastings steps.

# Computation

- We used Markov Chain Monte Carlo to do the integration
- We got standard errors using Balanced Repeated Replication from the survey sample literature

## Slide 64

We were able to use Markov Chain Monte Carlo to do the integration. Mostly, the steps were very easy but we had to write our own program to do the computation. So there is no standard software that can be used in this particular problem.

Because we were doing a survey sample, we used balanced repeated replication from the survey sample literature to get the standard errors. And we used Markov Chain Monte Carlo to get the estimates of all the parameters.

# Computation

- If one consumes whole fruit one also consumes total fruit, so we separate out whole fruit and fruit juice
- Same for total grains and whole grains
- Same for total vegetables and DOL

## Slide 65

You have to be careful in this. So if you consume a whole fruit, you consume total fruit. So we separated them out and treated them as two different dietary components. We did the same thing for whole grains and total grains, and we did the same thing for total veggies and DOLs.

# Computation

- After fitting the model, we get total fruit by adding together whole fruit and fruit juices
- Having obtained model estimates, we used Monte Carlo in a non-clever way to get the distributions of energy-adjusted usual intakes, joint HEI-2005 scores, total HEI-2005 scores, etc.

## Slide 66

And then, after that, we get total fruit by adding together whole fruit and fruit juices. So we have the parameters estimated. We fit our model. It describes the usual intake of these usual energy-adjusted intakes of these 12 dietary components in the HEI, and we're able to then use more Monte Carlo to get the distributions of the total scores and the joint distributions of each of the component scores.

# Computation

- The measurement error corrected usual HEI-2005 score can be represented as

$$T(X_i, \tilde{\beta}, \Sigma_u, \Sigma_\varepsilon, \tilde{u}_i)$$

- For  $b = 1, \dots, B$ , generate

$$\tilde{u}_{ib} = \text{Normal}(0, \Sigma_u)$$

- Estimate the distribution of the total score by the (weighted) empirical distribution of

$$T(X_i, \tilde{\beta}, \Sigma_u, \Sigma_\varepsilon, \tilde{u}_{ib})$$

## Slide 67

I know this is a really ugly thing to write down, but what I have and what the HEI-2005 score really consists of is a very complicated function,  $T$ , which involves all sorts of cool things: covariates; parameters in the regression associated with the covariates; the individual—how people differ from one another; the covariates matrix—the covariates matrix of day-to-day variability; and the person-specific bias, which we can't see. We can't actually observe that but we can use the computer to generate the person-specific biases. And then what we do is we make up—it's a statistically allowable thing to do—we put in these things we've generated by computer and then we get the empirical distribution with survey-related empirical distribution of these generated total scores. And that's how we get the distributions that you've seen in our previous slides.

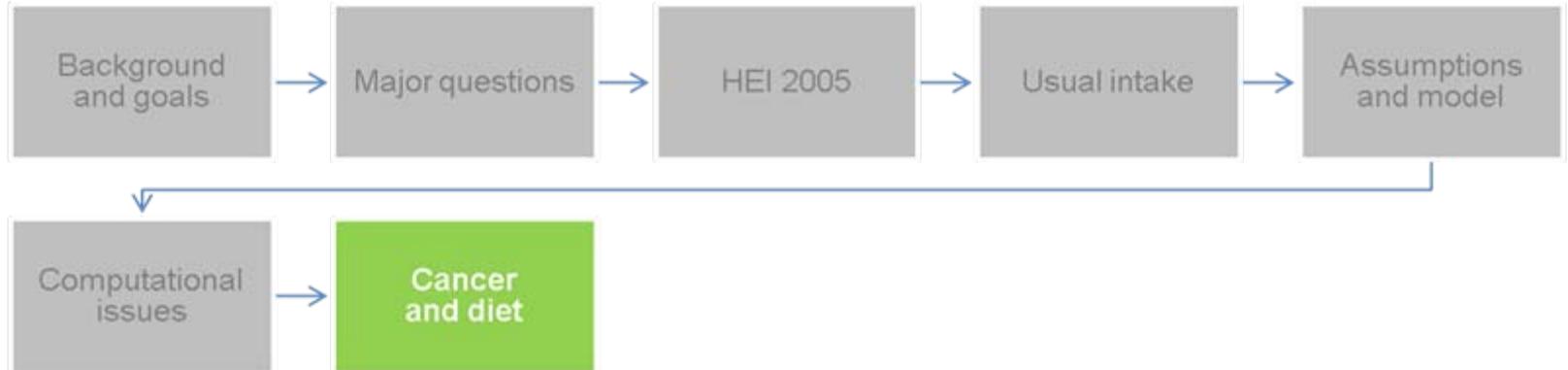
## Result verification

- For each food and nutrient, the previous lecture showed that it is possible to use standard nonlinear mixed effects software to get the distribution of adjusted usual intake and HEI-2005 component score, **one at a time, not jointly**
- Our results are in very close agreement with these results
- However, as mentioned previously, we can also do the multivariate case and estimate the distribution of the HEI-2005 total score

## Slide 68

It's a very complicated method; it is a complicated method. It's very easy to program but it looks like a complicated method. But we're very concerned about whether we programmed it correctly. Saijuan and I wrote our own separate codes. Dennis Buckman, who works with INS, wrote his code, and then we got the same answers and we also verified this by looking at the NCI model, the NCI method. And we took each component score separately, one at a time, and our answers have been in very close agreement with these results.

So we're really confident about our numbers and, of course, you have to do a multivariate model if you're interested in the distribution of the total score.



# RESULTS FOR CANCER AND DIET RELATIONSHIPS

## Slide 69

And then, finally, I want to give a few brief results about epidemiology.

## Relationship with health effects

- We have applied the model to the analysis of the NIH-AARP Diet and Health Study
- The outcome was colorectal cancer, separately for men and women
- The general goal is to study association of dietary patterns, assessed using dietary quality indices adjusted for measurement error, and a health outcome

## Slide 70

We've applied the model to the analysis of the NIH-AARP Diet & Health Study. And our outcome was colorectal cancer. And we've done this analysis separately for men and women. And what we want to do is study the association of dietary patterns using dietary quality indices such as HEI-2005, adjusted for measurement error and day-to-day variability, and the health outcome—namely, colorectal cancer.

So this is our goal. The key point is we want to get the dietary quality indices effect after adjusting for measurement error.

## Relationship with health effects

- We did a survival analysis using person years
- Variables in the model include age, ethnicity, education, BMI, smoking status, physical energy, energy and hormone replacement therapy (for women)
- The HEI total score was also in the model, in a loglinear continuous risk model

## Slide 71

So we did a survival analysis using person-years, and there were a number of variables in the model: age categories, ethnicity, education, BMI, smoking status, energy, and hormone replacement therapy for women. There wasn't any physical activity energy in the model.

And the HEI total score was also in the model, as was energy, as I said before. We fit this as a log linear continuous risk model.

## Relationship with health effects

- The first analysis done was using the FFQ for the HEI-2005 total score as well as energy
- The second was a measurement error corrected analysis, based on regression calibration
- The same covariates were used to fit the HEI-2005 total score model in a calibration sub-study

## Slide 72

Because it's a continuous model, we're able to use regression calibration. So the analysis was done first using the food frequency questionnaire and computing the HEI-2005 total score and energy using the food frequency questionnaire. And then the second analysis was our measurement error corrected approach based on our big model and using regression calibration.

The same covariates were used to fit the HEI-2005 total score in the calibration study. So the only 24 hour recalls we have are in a calibration study of about 900 people for each gender.

## Relationship with health effects

- We then used Monte-Carlo to implement the regression calibration and compute the expectation of energy and HEI-2005 total score given the observed covariates
- Bootstrapping was performed to estimate standard errors for the regression calibration analysis

### Slide 73

We then used Monte Carlo—that's a theme here—to do the regression calibration for both energy and the total score. And we used bootstrapping to estimate standard errors in the regression calibration analysis.

## Relationship with health effects

- What we expect to find is that the analysis based on the FFQ will have relative risks closer to 1.00 than will the measurement error corrected analysis
- There are two error-prone elements (HEI-2005 and energy) and 37 other covariates, so simple characterization of the effects of measurement error are not really possible

## Slide 74

The results are going to be pretty much what we expect. What we've seen before is that one of the effects of measurement error is typically to get relative risks or baseline hazards which are closer to 1 than a measurement error corrected analysis. That is, the effect of using an error-prone instrument is to lead to a diluted effect in appearance, and that is that you underestimate the relative risks associated with a bad diet.

So here we have a pretty good model. We did two error-prone components: the total score and energy. We have 37 other covariates for men and 38 for women. And so this is a pretty complicated problem where it's hard to actually tell exactly what's going to happen, so we had to actually do the calculations and do the theory here.

## Relationship with health effects

- We have applied the model to the analysis of the NIH-AARP Diet and Health Study
- Using the HEI-2005 total score from the FFQ, the relative risk for going from the 10<sup>th</sup> to the 90<sup>th</sup> percentile for women is estimated as 0.80
- After measurement error correction, it is 0.62
- Note the attenuation in the FFQ that we expected

## Slide 75

And here's what we get. If you use the FFQ, the relative risk for going from the 10<sup>th</sup> to the 90<sup>th</sup> percentile for women is estimated as 0.8. And after measurement error correction—this is based on the total score—we get an analysis that says the relative risk is .062. So this is what we expect in measurement error—that the impact of measurement error is to dilute the effect, sending it closer to the null value of 1. And this is a pretty good attenuation here due to the measurement error and biases in the food frequency questionnaire.

## Relationship with health effects

- The 95% confidence interval on the relative risk ignoring measurement error for women is 0.68 – 0.98, with a p-value = 0.04
- For usual intake, the CI is 0.45 – 0.93, with a p-value = 0.02
- The fact that the p-value is smaller for the measurement error analysis has to do with the complex data structure

## Slide 76

The T values also change. And this is an unusual thing that happens because of the large dimensionality of the problem. The 95 percent confidence interval for the relative risk, ignoring measurement error, for women is from 0.68 to 0.98 with a p value of 0.04. For our usual intake analysis, the confidence interval is longer but it's shifted way down and its p value is even smaller, 0.02. That's not what you'd typically see—that the p value goes down—but it does happen, especially here in this highly multivariate context.

So again, we have attenuation, and in this context we actually have the effect of measurement error to raise the p value almost to the null level.

# Summary

- 24HR recalls have great day-to-day variability
- Adjusting for this variability to estimate the distributions of usual intakes of multiple episodically consumed foods and nutrients has been unsolved and is extremely challenging
- We have provided the first solution to the problem

## Slide 77

So I'm going to summarize here what we've done: 24 hour recalls have large amounts of day-to-day variability. And our problem was a problem of adjusting for this day-to-day variability to try and get usual intakes when we have multiple episodically consumed foods and multiple nutrients. And that's a problem that's been unsolved and it's extremely challenging. But we were able to provide what is right now the first solution to the problem, and we're applying this to other quality indices as we're speaking.

# Summary

- The methods allow us to understand dietary patterns, estimate distributions, consider risk models, etc.
- The NCI has a working version of the model fitting in SAS, which is under development

## Slide 78

The method that we developed allows us to understand dietary patterns, estimate distributions, consider risk models, etc. And we're very happy with that. The original programs that I wrote and that Saijuan Zhang wrote were written in MATLAB, which is not critically user-friendly language. It's good for scientific computing, but not for field use. And so NCI is developing a SAS version of our model that will be a generalization of the NCI method in these more complicated problems.

# QUESTIONS & ANSWERS

Moderator: Sue Krebs-Smith

Please submit questions  
using the *Chat* function

## Slide 79

So thanks very much for your attention, and I guess Sue will now take over.

Thank you Dr. Carroll. We'll now move on to the question and answer period of the webinar. *(S. Krebs-Smith)*

## Measurement Error Webinar 8 Q&A

**Question:** One thing that you mentioned was the assumption that the 24 hour recalls are accurate measures of dietary intake and that they're unbiased. But although that's an assumption, we know that that's not really true. So given that there are some systematic biases, do we know how the results would be affected with this bias?

Well, we don't really know what the biases are for anything seen in the HEI-2005 other than energy, where there is doubly labeled water to understand the bias. We found in dealing with protein and with energy that the effect of bias is there but it's modest compared to the effect of the day-to-day variability, which is really quite large (*R. Carroll*)

**When you were talking about episodically consumed foods, you mentioned the HEI-2005 that you said had six components that were episodically consumed. One of those was total veggies, with 3 percent that didn't report any consumption of total veggies. How do you determine episodic foods? Three percent sounds like a low number for determining episodic consumption.**

It is a low number. How I did it was I referred to my friendly nutritionist, who suggested what to do. But I also redid the analysis by substituting in for those 3 percent people the one-half of the minimum amount of vegetable consumption across the consumption days, and the analysis doesn't change very much. The thing you don't want to do is go really down to an extreme. I think anything below 3 percent is probably a bad idea to treat as episodically consumed. If it's like half a percent, any statistical method, especially with these 24 hour recalls, is going to have a hard time estimating the consumption part of the model, the probability of consumption, because the probability of consumption would be so high. (*R. Carroll*)

**Fruit seems like an aggregate of the consumption of different fruits: apples, oranges, and so on. Do you think it's better to apply this method before or after aggregating to get total fruit consumption?**

Oh, it would be better to aggregate. From a computational point of view and a numerical stability point of view, it's better to aggregate because if you were to run this analysis on apples separately and bananas or whatever separately, you'd start to raise the dimensionality of the problem very, very, very rapidly. And I think it's going to be really hard at some point to get very reliable answers. It's possible to run the model but

unless you're really interested in the different components for the HEI analysis, it's certainly better to aggregate. *(R. Carroll)*

**You have talked about the analysis with HEI-2005, but you also mentioned that this technique that you have would be workable for other dietary scores such as the Mediterranean dietary score. Have you applied your method to the Mediterranean dietary score? And what would be the results from that, if so?**

No, I haven't applied it to the Mediterranean diet score. Part of NCI's programming is to build a methodology that will enable any sort of dietary pattern score to be done. So our current programs were geared to various HEI measures on a one-off kind of thing, and we know how to do it completely generally, and this is why we have a professional programmer working on putting together a SAS program. *(R. Carroll)*

**One characteristic of the HEI-2005 is it adjusts all of the individual components for energy intake but, of course, absolute energy intake is not figured into the score. Can you comment on this a bit?**

Well, that question has been raised as I've given talks, and from the perspective of, for example, risk measurement, we do include energy in the analysis. But the total score is just defined, and the HEI-2005 total score was defined, and we went with the definitions to try and understand the statistical properties of that particular measure. But it isn't a problem to work with different measures that don't deal always with energy-adjusted dietary components. *(R. Carroll)*

From a nutritional point of view, I'll just add to that that the HEI-2005, without including absolute energy intake, then reflects, really, the quality of the mix of foods. And you can look at any set of foods, because you have this density-based standard. So we've found that [to be] a useful feature of the HEI-2005 for those reasons. *(S. Krebs-Smith)*

**The distributions of total scores are remarkably normally distributed. Would you expect that?**

No, this was just with the kids and that's just what happens with those data. We've worked with the NHANES for men over the age of 20, and they are actually quite left skewed; they're not even right skewed. They are left skewed, and very, very noticeably left skewed. So it's going to depend on the population as to whether things are normal, right skewed, or left skewed. *(R. Carroll)*

**Why do you think the 24 hour recalls are shifted to the left?**

It's because there are lots of zeros in them, and so what happens is if you apply—for most of the dietary components that are episodic, more is better, more results in a higher score. And because the scores are highly nonlinear, whenever you get a zero, you get a zero in the score, and it just moves the thing over. *(R. Carroll)*

**It seems your analysis doesn't include the possibility of never consumers. There may be some components where, for some individuals, they truly do not consume the component at all. What would happen if you needed to include that possibility in the model?**

In this program, we don't think that happens. We've done some analyses and don't think that for the HEI-2005 that's going to be much of an issue. But we are working on—I have a postdoc and a graduate student working with me and with people at NCI on developing methods to handle episodically consumed foods with never consumers in this complicated context, and we're programming as we speak. *(R. Carroll)*

**Related to the model itself, when you were describing the different terms that you would use in the model, you said that k would denote an index. Can you describe what you meant by an index, because it didn't seem like you meant the HEI index but some...**

That's fine. I'm sorry; k describes, for example, if I have, let's say whole fruits, I have two things going on with whole fruits. One is whether that's consumed or not, and secondly, I have going on what is consumed on consumption days. So I said k=1 for whether there is consumption and k=2 for amounts on consumption days. And then I would do this for each of the six components, so I'd have another k, so 3 for, say, fruit juices consumption, and 4 for the amount of fruit juices on consumption days, and I work that up. So it's really not a very good—"index" is probably the wrong phrase. *(R. Carroll)*

**So when you were talking about those 19 dimensions, those were 19 k's in your model?**

Yeah, 19 k's, yes indeed. *(R. Carroll)*

**The NLMIXED—can you talk a little bit more about the issue of integration being too great and why that's such an issue, why that wouldn't work?**

So, what happens is that NLMIXED uses a method called maximum likelihood, and so it tries to integrate out—in other words, get rid of those

u's, the person-specific biases, because they are not observed. And so the maximum likelihood can only work on things that are actually observable. And so it tries to integrate out these person-specific biases, the 19 dimensional u's, and that, simply, NLMIXED just physically can't handle that high a dimension. So any of these high-dimensional problems that involve latent variables such as the person-specific effects are going to have to go with some version of the Monte Carlo computation. When you get past about dimension 3, there's really no place else to go but Monte Carlo. *(R. Carroll)*

**Regarding the run time for the computations and how long it takes to run the model for the total HEI score distribution, I wonder if you could talk about that in relation to the models that you developed for this, and then compare that, maybe, to what it's like to run a single food with the NCI method?**

Right, well, the single food adjusted for energy, the NLMIXED, if I just look at that problem alone, we have a paper where we showed that the single food—this is not the NCI method, although it's a generalization of it. The Monte Carlo computation is much, much faster. It's on the order of three minutes in MATLAB on a good machine. For the HEI, it's a couple of hours for all components to run. I'm very, very conservative about these things because I don't mind seven hours or three hours or four hours if I'm going to put some numbers in a paper. There are ways of cutting it down to about 25 minutes, but I prefer to—in the Monte Carlo computation you have to do a lot of repetitive things and I like to do lots and lots of repetitive things to guarantee the stability of the final answers. But we have no trouble doing bootstrapping or balance repeated replication on these sorts of sample sizes. It's not very long. *(R. Carroll)*

**Do you expect the timing would be similar with the SAS programs that NCI is developing?**

Yes, it should be pretty similar. I haven't taken into account any of the really fancy features of MATLAB for parallel computing and things. I've tried to keep it very vanilla so that the SAS program timings would be roughly the same. And some of the initial analyses that Dennis Buckman has done suggest that the timing is going to be about the same. And the method is very, very numerically stable, which is not true with NLMIXED with episodically consumed food plus energies. They sometimes have lots of problems with instability. *(R. Carroll)*

**You use MCMC, which is usually used in Bayesian statistics. Are you doing a Bayesian analysis?**

No, I'm not doing a Bayesian analysis. Bayesian analyses are, especially in surveys, almost impossible. I find it's a very controversial field, how to make confidence intervals, u being Bayesian, in sample surveys. The Markov Chain Monte Carlo, which is used in Bayesian statistics, is also a perfectly acceptable statistical method for non-Bayesians. There is a large theory that says it's effective. If you could compute those integrals and do maximum likelihood, then you would get the same answers using these MCMC things. So it's just a way of getting a maximum likelihood estimate that's actually feasible to compute. *(R. Carroll)*

**In the example that you gave from the NHANES analysis, you had mentioned that there were a certain number of children with a single 24 hour recall and another set with two 24 hour recalls. Did you include the children that had only one recall in your analysis?**

Oh yes, for sure; they contain valuable information and actually lower the standard errors by 25 percent, which is nontrivial. For sure, we used all the data. More 24 hour recalls are good. So having one on 1,500 people — remember, 24 recalls are, by assumption, unbiased for the mean. So if nothing else, they can be used to really increase the precision of the estimate of the mean. *(R. Carroll)*

**Never throw away information, right?**

Never throw away information, yeah. *(R. Carroll)*

[This page intentionally blank.]

Next Session

Tuesday, November 22, 2011  
10:00-11:30 EST

**Combining self-report dietary  
assessment instruments to reduce  
the effects of measurement error**

Douglas Midthune  
National Cancer Institute

## Slide 80

Thank you very much, Dr. Carroll, and thanks to our audience for joining today's webinar. Please join us next week for webinar 10, when Douglas Midthune will discuss combining instruments as a strategy to reduce the effects of measurement error.