National Cancer Institute

measurementERRORwebinar series

Estimating usual intake distributions for multivariate dietary variables

Raymond J. Carroll, PhD
Texas A&M University

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

---

measurementERRORwebinar series

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

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

Presenters and Collaborators

Sharon Kirkpatrick
Series Organizer

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

Wake Forest School of Medicine · The Gertner Institute · STATISTICS TEXAS A&M UNIVERSITY · NATIONAL CANCER INSTITUTE · Office of Dietary Supplements National Institutes of Health · USDA

Introduction

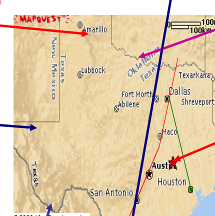
West Texas ☺ / East Texas ☹

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



I-45

I-35

Big Bend National Park

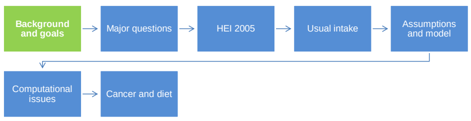
Introduction

Palo Duro Canyon of the Red River

Introduction

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)



BACKGROUND AND GOALS

Past lectures

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

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

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

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

Context

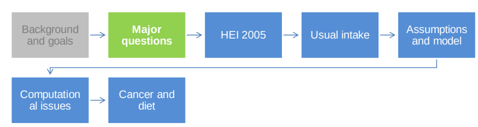
- The data we use are from the NHANES 2001-2004 survey of children aged 2-8 in the U.S.
- The data set has 2,638 children with a 24HR
- There are 1,103 with two 24HR
- This is a real survey, and survey weights are incorporated into the analysis (details skipped)

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)

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



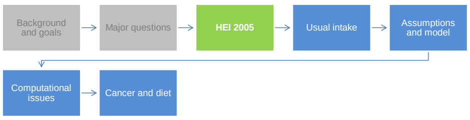
MAJOR QUESTIONS

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?

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



HEALTHY EATING INDEX 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

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	≥0.8 cup equiv. per 1,000 kcal	No Fruit
Whole Fruit (not juice)	5	≥0.4 cup equiv. per 1,000 kcal	No Whole Fruit
Total Vegetables	5	≥1.1 cup equiv. per 1,000 kcal	No Vegetables
Dark Green and Orange Vegetables and Legumes	5	≥0.4 cup equiv. per 1,000 kcal	No Dark Green or Orange Vegetables or Legumes
Total Grains	5	≥3.0 oz equiv. per 1,000 kcal	No Grains
Whole Grains	5	≥1.5 oz equiv. per 1,000 kcal	No Whole Grains
Milk	10	≥1.3 cup equiv. per 1,000 kcal	No Milk
Meat and Beans	10	≥2.5 oz equiv. per 1,000 kcal	No Meat or Beans
Oils	10	≥12 grams per 1,000 kcal	No Oil
Saturated Fat	10	≤7% of energy	≥15% of energy
Sodium	10	≤0.7 gram per 1,000 kcal	≥2.0 grams per 1,000 kcal
Calories from Solid Fats, Alcoholic beverages, and Added Sugars (SoFAAS)	20	≤20% of energy	≥50% of energy

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”

The Healthy Eating Index 2005

Maximum Score=5	Maximum Score=10	Maximum Score=20
• Whole fruit • Total fruit • Whole grains • Total grains • DOL • Total vegetables	• Milk • Meat and beans • Oils • Saturated fat • Sodium	• SoFAAS

Maximum Score=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

HEI 2005

The Healthy Eating Index 2005

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

measurementERRORwebinarseries

25 Introduction to measurement error in dietary intake data

HEI 2005

The Healthy Eating Index 2005

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

$$= 0 \quad \text{if} \geq 15$$
$$= 10 \quad \text{if} <= 7$$
$$= 8\text{-}8*(\text{density}\text{-}10)/5 \quad \text{if} > 10 \text{ but } < 15$$
$$= 10\text{-}2*(\text{density}\text{-}7)/3 \quad \text{if} > 7 \text{ but } < 10$$

measurementERRORwebinarseries

26 Introduction to measurement error in dietary intake data

HEI 2005

The Healthy Eating Index 2005

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

$$= 0 \quad \text{if} \geq 50$$
$$= 20 \quad \text{if} <= 20$$
$$= \text{linearly interpolated otherwise}$$

measurementERRORwebinarseries

27 Introduction to measurement error in dietary intake data

HEI 2005

The Healthy Eating Index 2005

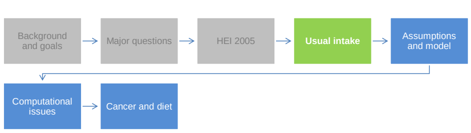
- Now we focus on long term HEI total score, not short term

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

- 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**

measurementERRORwebinarseries

28 Introduction to measurement error in dietary intake data



RESULTS FOR THE DISTRIBUTIONS OF USUAL INTAKES

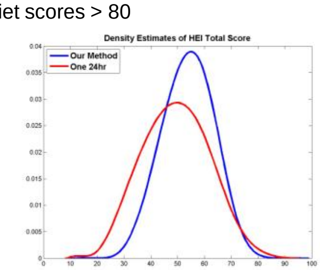
measurementERRORwebinarseries

29 Introduction to measurement error in dietary intake data

Usual intake

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

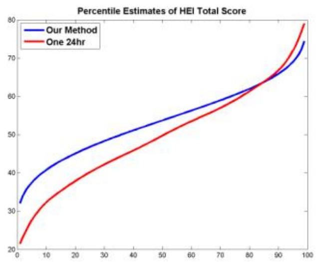


measurementERRORwebinarseries

30 Introduction to measurement error in dietary intake data

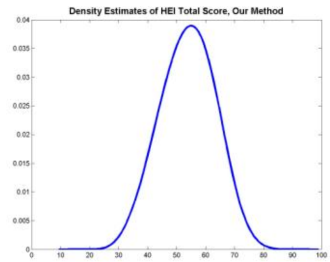
Result: HEI total score percentiles

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



Result: HEI total score density

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

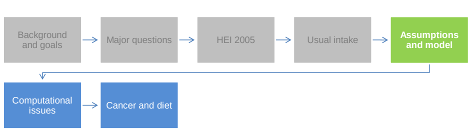


A vignette

- 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

A vignette

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



ASSUMPTIONS AND MODEL

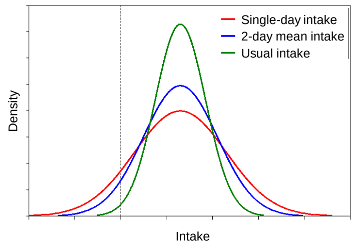
Modeling assumption

- 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

Assumptions and model

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



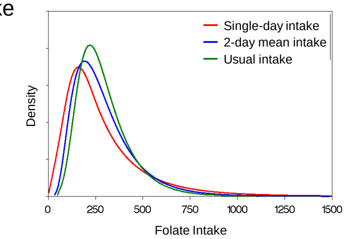
measurementERRORwebinarseries

37 Introduction to measurement error in dietary intake data

Assumptions and model

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



Source: EATS

measurementERRORwebinarseries

38 Introduction to measurement error in dietary intake data

Assumptions and model

Transformations

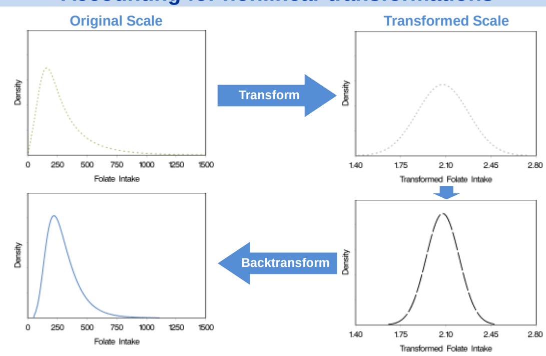
- To deal with the skewness, it is typical to transform the data so that day-to-day variation has a nice Gaussian-like distribution
- One analyzes in this transformed scale, and then back-transforms to the original nutrient scale
- Here is an illustration

measurementERRORwebinarseries

39 Introduction to measurement error in dietary intake data

Assumptions and model

Accounting for nonlinear transformations



measurementERRORwebinarseries

40 Introduction to measurement error in dietary intake data

Assumptions and model

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%

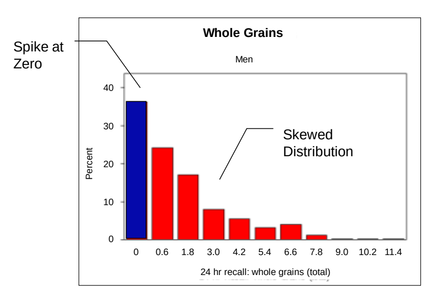
measurementERRORwebinarseries

41 Introduction to measurement error in dietary intake data

Assumptions and model

Challenges to estimation – foods

- Observed food intakes are often zero



Source: EATS

measurementERRORwebinarseries

42 Introduction to measurement error in dietary intake data

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

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

So, what's the big deal?

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

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

So, what's the big deal?

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

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

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

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, ε will denote day-to-day variability

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

$$W_{Fijk} > 0$$

Amount

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

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

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

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

Covariance Matrices

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

Covariance Matrices

- Our latent variable model for HEI-2005 has 19 components whose variances and correlations need to be modeled Σ
- 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

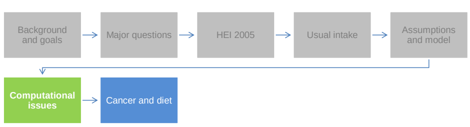
Energy

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

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

- We assume that

$$(u_{i1},\dots,u_{i5}) = \text{Normal}(0,\Sigma_u)$$
$$(\varepsilon_{ij1},\dots,\varepsilon_{ij5}) = \text{Normal}(0,\Sigma_\varepsilon)$$



COMPUTATIONAL ISSUES

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

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

Computation and more

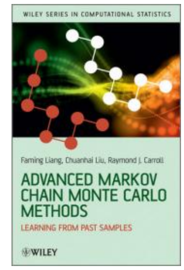
- The computational issue is that the components of the day-to-day variability, the epsilons, are all correlated
- So too are the components of the individual usual intake, the u's
- Maximum likelihood requires integration (area under the curve of a function)

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

Computation

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



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

Computation

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

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

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.

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,…,B, generate

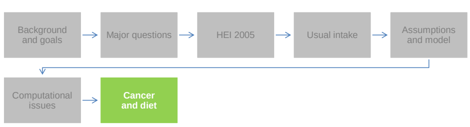
$$\tilde{u}_{ib} = \mathrm{Normal}(0, \Sigma_u)$$

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

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

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



RESULTS FOR CANCER AND DIET RELATIONSHIPS

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

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

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

Cancer and diet

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

measurementERRORwebinar series

73 Introduction to measurement error in dietary intake data

Cancer and diet

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

measurementERRORwebinar series

74 Introduction to measurement error in dietary intake data

Cancer and diet

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^{th} to the 90^{th} 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

measurementERRORwebinar series

75 Introduction to measurement error in dietary intake data

Cancer and diet

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

measurementERRORwebinar series

76 Introduction to measurement error in dietary intake data

Cancer and diet

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

measurementERRORwebinar series

77 Introduction to measurement error in dietary intake data

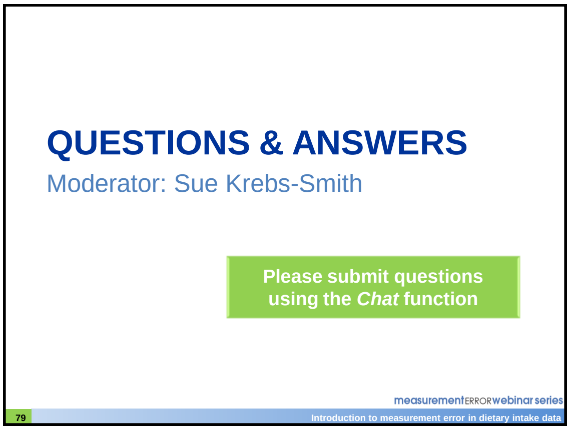
Cancer and diet

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

measurementERRORwebinar series

78 Introduction to measurement error in dietary intake data


QUESTIONS & ANSWERS
Moderator: Sue Krebs-Smith
Please submit questions using the Chat function
measurementERRORwebinar series
79
Introduction to measurement error in dietary intake data


National Cancer Institute
measurementERRORwebinar series
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
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health