National Cancer Institute

measurementERRORwebinar series

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

Kevin W. Dodd, PhD
National Cancer Institute

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

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

measurementERRORwebinar series

---

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.

Objectives

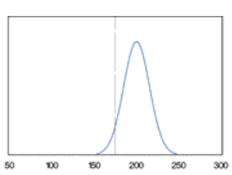
- Participants will have an understanding of:
  - Considerations in estimating usual intakes of nutrients and foods consumed nearly daily by nearly all persons
  - Assumptions made in current approaches to estimating usual intake distributions
  - Statistical modeling techniques and data requirements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

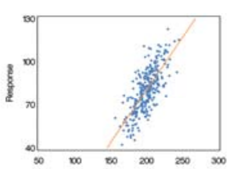
measurementERRORwebinar series

Objectives

Two main areas of interest

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





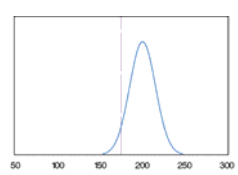
- Estimating diet-health relationships: regression coefficients

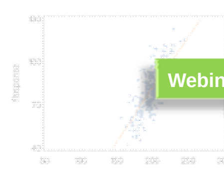
measurementERRORwebinar series

Objectives

Two main areas of interest

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





- Estimating diet-health relationships: regression coefficients

Webinars 6-8, 12

measurementERRORwebinar series

Objectives

Daily versus episodic consumption

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

measurementERRORwebinarseries

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

Objectives

Daily versus episodic consumption

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

Webinar 3

measurementERRORwebinarseries

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

Outline

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

measurementERRORwebinarseries

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

Basic assumptions → Building a model → Estimating distributions → The role of covariates

BASIC ASSUMPTIONS

measurementERRORwebinarseries

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

Basic assumptions

Focus is on usual intake

Usual intake = long-term average daily intake

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

measurementERRORwebinarseries

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

Basic assumptions

Challenge

Usual intakes are not directly observable

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

measurementERRORwebinarseries

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

Assessment strategies fall between two extremes

Usual intake = long-term average daily intake

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

Two classes of measurement error in instruments

- **Random:** Average of repeats = true value
  - Instrument is **accurate**, or **unbiased**
  - May not be **precise**
- **Systematic:** Average of repeats ≠ true value
  - Instrument is **inaccurate**, or **biased**
  - Systematic bias can occur in many ways

Potential sources of error in instruments: FFQ

– Cognitively challenging

– Limited food list/portion size choices

\+ No need for repeated application (high reproducibility)

Potential sources of error in instruments: 24HR

\+ Less cognitively challenging

\+ Open-ended format

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

Comparison of measurement error structures

24-hour recall (24HR)	Food frequency questionnaire (FFQ)
• Larger within-person random error • Smaller systematic error	• Smaller within-person random error • Larger systematic error

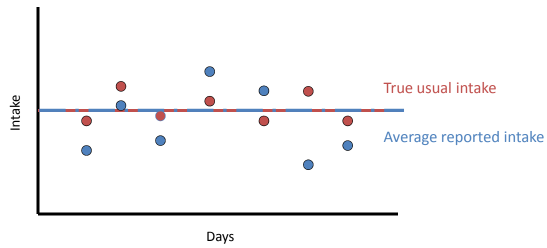
Rationale for using short-term instruments

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

Basic assumptions

Main assumption

24HR unbiased for individual-level usual intake



Basic assumptions

Working assumption

- Unbiasedness of 24HR is a **working assumption**
- Required to proceed with development of methods
- May be more or less justified depending on dietary component of interest

Basic assumptions → **Building a model** → Estimating distributions → The role of covariates

BUILDING A MODEL

Building a statistical model for 24HRs

Typical data scenario

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

Notation

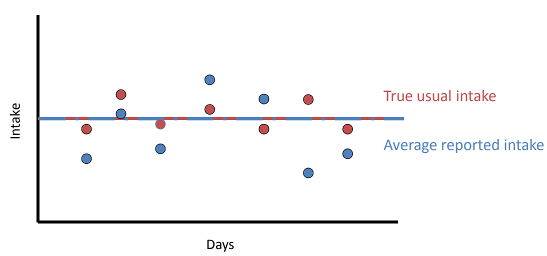
- Observations denoted by R_{ij} , usual intake by T_i
- Individuals indexed by subscript $i = 1,2,\ldots,N$
- Repeats indexed by subscript $j = 1,2,\ldots,J$

Building a statistical model for 24HRs

Implications of unbiasedness assumption

24HR unbiased for individual-level usual intake

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



Building a statistical model for 24HRs

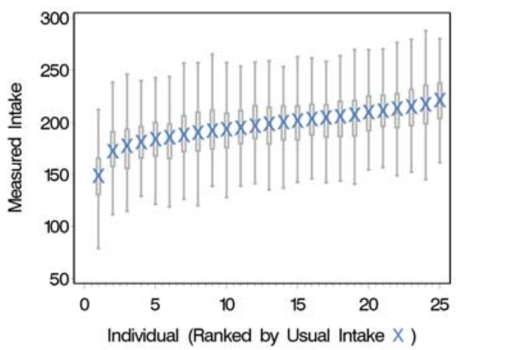
Implications of unbiasedness assumption

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

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

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

Within- and between-person variation

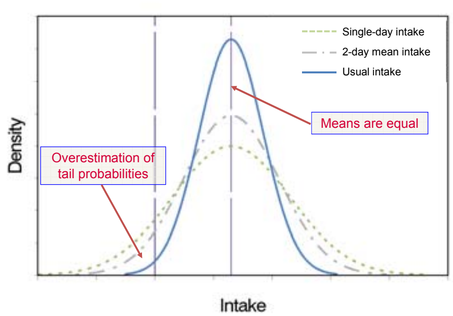


Limitations of unbiasedness assumption

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

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

Effect of within-person variation



Effect of within-person variation

- Population mean usual intake may be well estimated by simple averaging methods
- Percent of population with usual intake below/above cutoff values may be very biased – **modeling necessary**

What does "modeling" entail?

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

Foundation of the model

- Each recall is usual intake plus a deviation

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

Foundation of the model

- Each recall is usual intake plus a deviation

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

Within-person deviation

Foundation of the model

- Each recall is usual intake plus a deviation

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

Within-person deviation

- Within-person deviations cancel out across days

Foundation of the model

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

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

Foundation of the model

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

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

Between-person deviation

Foundation of the model

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

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

Between-person deviation

- Between-person deviations cancel out across the population

Foundation of the model

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

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

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

Building a statistical model for 24HRs

Common variance assumption

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

Building a statistical model for 24HRs

Distributional assumptions

- Statistically convenient to assume that both types of deviations follow a parametric probability distribution
- The normal distribution is a common choice
  - Naturally parameterized by mean and variance
  - Dependence between deviations can be completely modeled via correlation

Building a statistical model for 24HRs

Basic statistical model for 24HRs

Within-person deviations are:

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

Building a statistical model for 24HRs

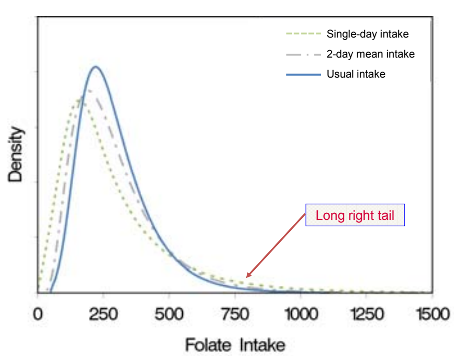
Basic statistical model for 24HRs

Between-person deviations are:

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

Building a statistical model for 24HRs

Complications of skewed distributions



Building a statistical model for 24HRs

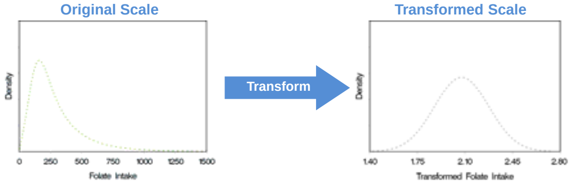
Common nonlinear transformations

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

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

Building a statistical model for 24HRs

Extended model for 24HRs



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

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

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

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

measurementERRORwebinarseries

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

Building a statistical model for 24HRs

A warning about means of transformed data

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

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

- Taylor series argument:

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

- Extra terms involve "higher-order moments"

measurementERRORwebinarseries

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

Building a statistical model for 24HRs

Summary

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

measurementERRORwebinarseries

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

Basic assumptions → Building a model → Estimating distributions → The role of covariates

ESTIMATING DISTRIBUTIONS

measurementERRORwebinarseries

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

Estimating distributions from the model

Data requirements

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

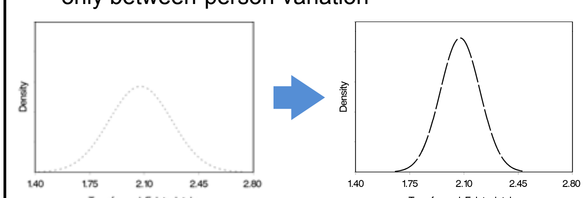
measurementERRORwebinarseries

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

Estimating distributions from the model

General approach – no transformations (yet!)

- Separate within- from between-person variation
- Estimate usual intake distribution that exhibits only between-person variation



measurementERRORwebinarseries

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

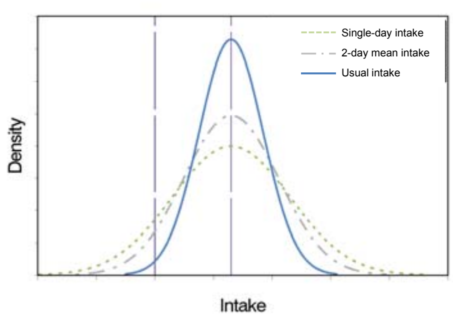
Two general approaches

- Model-Assisted (M-A) – rescales observed individual mean distribution
- Model-Based (M-B) – estimates distributions from theoretically-derived quantities

Two general approaches

- Model-Assisted (M-A) – rescales observed individual mean distribution
- Model-Based (M-B) – estimates distributions from theoretically-derived quantities

Rationale for the Model-Assisted approach



Rationale for the Model-Assisted approach

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

- For a sample of single 24HRs:

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

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

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

Implementing the Model-Assisted approach

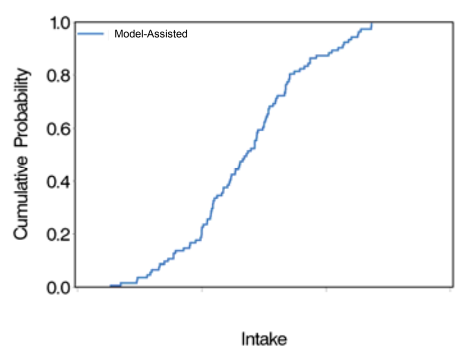
$$R_{ij} = \mu + u_i + \varepsilon_{ij}, \quad \mathrm{Var}(u_i) = \sigma_u^2, \quad \mathrm{Var}(\varepsilon_{ij}) = \sigma_\varepsilon^2$$

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

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

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

Features of the Model-Assisted approach



Interpretation of scaled means

- The scaled means r_i are not intended to be estimates of individual usual intake
- The distribution of scaled means has the same mean and variance as the distribution of usual intakes in the population
  - Distributions coincide for normal distributions
  - Agreement only approximate otherwise

Features of the Model-Assisted approach

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

Two general approaches

- Model-Assisted (M-A) – rescales observed individual mean distribution
- Model-Based (M-B) – estimates distributions from theoretically-derived quantities

Rationale for the Model-Based approach

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

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

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

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

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

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

Rationale for the Model-Based approach

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

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

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

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

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

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

Quantile from the distribution of u_i

Implementation using Monte Carlo simulation

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

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

- Create simulated usual intake (pseudo-value)

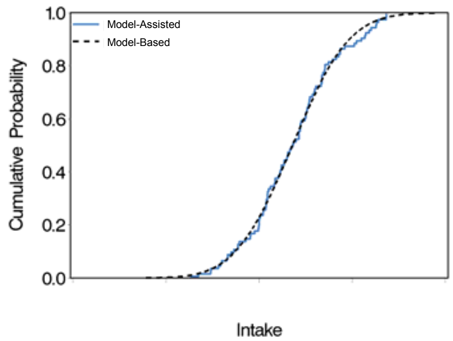
$$r_k = \hat{\mu} + u_k$$

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

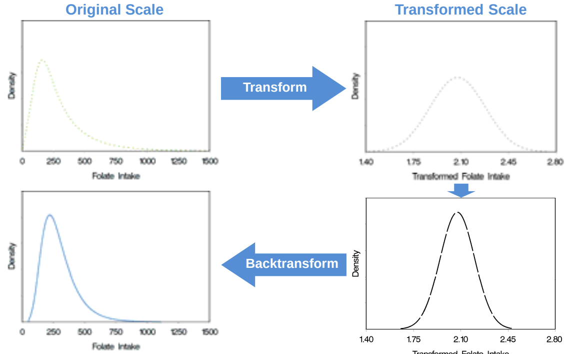
Features of the Model-Based approach

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

Features of the Model-Based approach

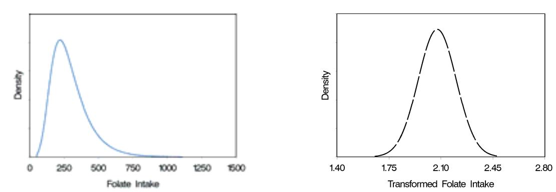


Accounting for nonlinear transformations



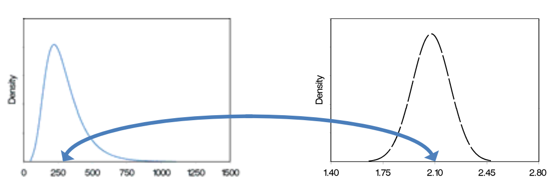
Estimating quantiles when transformations are used

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



Estimating quantiles when transformations are used

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



Estimating quantiles when transformations are used

- With no transformation used:

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

- Estimated quantile is a linear function

Estimating quantiles when transformations are used

- With nonlinear transformation *g* used:

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

- Estimated quantile is an integral
- Can be calculated/approximated several ways

Integration provides the "backtransformation"

- Taylor series approximation (Dodd, 2006):

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

- Exact calculation for normal ε (Hoffmann, 2002)
- Numerical integration for known ε distribution
  - Quadrature formulas, e.g., Gauss-Hermite
  - Monte Carlo integration

Estimation approaches when transformations used

- Both Model-Assisted and Model-Based approaches can be extended
- If transformation *g* achieves the desired distribution of ε terms, Taylor series approximation may be poor
  - Alternatives use all moments, not just two

Evolution of estimation methods

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

* NRC method incorporates transformations under alternative assumptions

Software availability for estimation methods

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

Summary

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

Basic assumptions → Building a model → Estimating distributions → **The role of covariates**

THE ROLE OF COVARIATES

The role of covariates

The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics

The role of covariates

Dietary Reference Intakes: Estimated Average Requirements

Food and Nutrition Board, Institute of Medicine, National Academies

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

The role of covariates

The need for subpopulation estimates

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

The role of covariates

One answer is to stratify sampled data

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

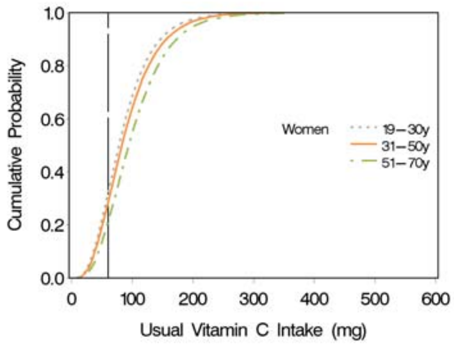
The role of covariates

The need for subpopulation estimates

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

The role of covariates

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

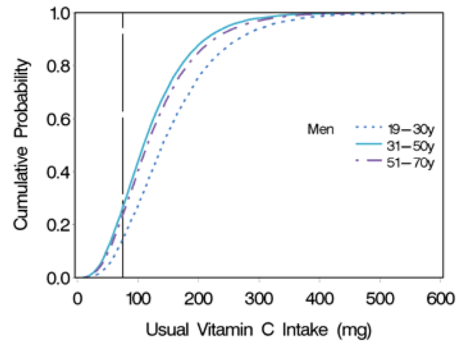


measurementERRORwebinarseries

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

The role of covariates

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



measurementERRORwebinarseries

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

The role of covariates

Limitations of stratification approach

- When multiple factors thought to influence diet are considered,
  - Subsample sizes decrease dramatically
  - Analysis burden increases
- Allowing covariates in the statistical models can overcome this limitation

measurementERRORwebinarseries

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

The role of covariates

A mixed model formulation

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

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

measurementERRORwebinarseries

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

The role of covariates

A mixed model formulation including covariates

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

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

measurementERRORwebinarseries

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

The role of covariates

Types of covariates

- **Individual-level**: affects true intake on all days, e.g., gender, age, smoker/nonsmoker status
- **Time-dependent**: affects true intake on specific days, e.g., season, weekday
- **Nuisance**: affects reporting error, e.g., interview sequence, mode (telephone vs. in-person)

measurementERRORwebinarseries

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

Potential benefits of incorporating covariates

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

Potential benefits of incorporating covariates

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

Potential benefits of incorporating covariates

- Overall bias due to nuisance effects can be corrected
- In epidemiologic applications, less **unstructured** variation is better

Webinar 10

Estimating distributions with covariates in the model

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

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

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

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

Estimating distributions with covariates in the model

- Model-Assisted approach retains observed joint distribution of individual-level covariates
  - Some covariate combinations may be rare
  - M-B: draw $\mathbf{X}_0$ at random from observed joint distribution to mimic this behavior
- Model-Based approach also offers a choice to perform **direct standardization**
  - Draw $\mathbf{X}_0$ from a **standard population**

Estimating distributions with covariates in the model

- Model-Assisted and Model-Based similar unless
  - Important covariate(s) are omitted, and/or
  - Exact normality does not hold
- Discrepancy between Model-Assisted and Model-Based distributions useful as a diagnostic

Direct standardization for time-dependent covariates

- Overall usual intake is weighted average of time-dependent usual intake
- Weights come from the standard population, e.g., for weekend/weekday effects:

Standard Population for Weekdays/Weekend Days

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

Explicit adjustments for nuisance effects

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

Types of covariates allowed in available methods

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

* fractional polynomial option for age

Summary

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

QUESTIONS & ANSWERS

Moderator: Sharon Kirkpatrick

Please submit questions using the *Chat* function

National Cancer Institute

measurementERRORwebinarseries

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

Estimating usual intake distributions for foods and nutrients consumed episodically

Janet Tooze
Wake Forest University

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